Econometrics

Rómulo A. Chumacero[1]

October, 2025

[1] Department of Economics. University of Chile. rchumace@econ.uchile.cl

Contents

Preface

This book is intended for graduate and advanced undergraduate students, as well as practitioners with a working knowledge of statistics, econometrics, calculus, and algebra.

It distills more than thirty years of my teaching experience into a coherent set of lecture notes, refined through hundreds of classroom hours and enriched by feedback from students across disciplines. The material is organized into three parts, which together can support two to three full courses. The first part lays the foundation, core concepts essential for all subsequent work and typically covered in an introductory semester. The second part focuses on time series econometrics, and the third addresses advanced and specialized topics that deepen or complement earlier material.

To truly master econometrics, one must learn to visualize and internalize complex ideas. In my view, this process involves three essential stages: first, understanding the intuition behind what is being done; second, grasping the operative mechanics of how it is done; and third, bringing abstract concepts to life by implementing them from scratch, at least once. This process transforms passive understanding into active command.

Like a good wine, command of econometrics improves with time and repetition. The first time we encounter new material, we may, if fortunate, grasp its big picture. But the subtleties, refinements, and sheer elegance of econometric methods only emerge through repeated engagement and, in my experience, through the attempt to explain these ideas to others who are unfamiliar with them. That is when the flavors mature, the intricacies surface, and the full power of the tools becomes transparent and usable.

In an era of rapidly advancing AI, the value of simply providing code has diminished. What remains essential, and distinctly human, is understanding: knowing what we are doing, why we are doing it, and how. Econometrics is not just a toolbox; it is a way of thinking.

Few things, in my experience, are as pedagogically powerful, or as transformative, as watching a theoretical idea come to life in code. That is where real understanding begins.

My personal preference is Gauss (by Aptech), which I find especially powerful and flexible. Still, recognizing the accessibility and popularity of other platforms, I provide companion code in R, EViews, and Python.

Ultimately, the comparative advantage of econometricians lies in bringing empirical content to economic theory. The topics covered here reflect those I have found most relevant—whether teaching, publishing, or applying econometrics in real-world settings.

I have been fortunate to teach hundreds of dedicated students, many of whom have shaped and improved the content of this book. Several of my lectures, covering much of the material presented here, are available on my YouTube channel: `https://www.youtube.com/rchumacero`. I invite you to visit and subscribe—this may be my best shot at turning 1.6K subscribers into a cult following in the niche world of econometrics. Please note that the lectures are in Spanish. You can also find additional material, research papers (mostly in English), and other resources relevant to this book on my website: `https://rchumace.econ.uchile.cl`.

Many excellent textbooks already exist. This one does not aim to replace them. Rather, it integrates the best features of my favorites, offering a coherent and intuitive framework for understanding, teaching, and mastering econometrics. To further support this goal, each chapter concludes with a brief summary of selected references. These are not meant to be exhaustive surveys, but guided tours, designed to complement and deepen the reader's understanding. The references include: (i) key sources from which the chapter draws or that closely align with its content; (ii) introductory material for readers who may need to build intuition or review prerequisites; and (iii) additional or advanced references that extend the material, formalize it further, or offer alternative perspectives. These curated suggestions aim to help readers navigate the broader literature and enrich their econometric journey.

Part I

The Foundations

Chapter 1

Ordinary Least Squares

1.1 Introduction

Ragnar Frisch (one of the founders of the Econometric Society) is credited with coining the term econometrics. Econometrics aims to give empirical content to economic relationships by uniting three key ingredients: economic theory, economic data, and statistical methods. Neither theory without measurement nor measurement without theory is sufficient for explaining economic phenomena. It is their union that is essential to understanding economic relationships.

Social scientists generally must accept the conditions under which their subjects act and the responses occur. As economic data come almost exclusively from non-experimental sources, researchers cannot specify or choose the level of a stimulus and record the outcome. They can only observe the natural experiments that take place. In this sense, economics—like meteorology—is an observational science.

For example, many economists have studied the influence of monetary policy on macroeconomic conditions, yet the effects of actions by central banks continue to be widely debated. Some of these controversies might be resolved if a central bank could experiment with monetary policy over repeated trials under identical conditions, allowing the isolation of policy effects more precisely. However, no one can turn back the clock to try various policies under essentially the same circumstances. Each time a central bank contemplates an action, it faces a new set of conditions. The actors and technologies have changed. The social, economic, and political environments are different. To learn from one episode in economic history and apply it effectively to another, one must understand both the similarities and the differences between the past, present, and future.

In this context, Ordinary Least Squares (OLS) provides a natural starting point for empirical analysis. It is the most widely used estimation method in applied economics and often serves as the benchmark against which other methods are judged. Despite its simplicity, OLS embodies a set of assumptions that are frequently misunderstood or taken for granted. Understanding when and why these assumptions matter is essential for responsible empirical work.

This chapter develops the finite-sample and asymptotic properties of OLS. Section 1.2 introduces the correlation coefficient and highlights the dangers of interpreting statistical association as causal. Section 1.3 presents the simple linear regression model with one explanatory variable. The OLS estimator is derived from first principles, and its connection to correlation is discussed. This sets the stage for the general regression framework introduced in Section 1.4, where regression is formalized through the lens of conditional expectation and the regression error. Section 1.5 derives the multivariate OLS estimator and examines its key properties. Section 1.6 extends the model to accommodate linear restrictions on the coefficients. Finally, Section 1.7 turns to inference, emphasizing the distinction between finite-sample and asymptotic results, and warning that t- and F-distributions only apply exactly under normality.

Throughout, the emphasis will be on both rigor and intuition. We aim not only to derive results, but to understand what they mean — when they apply, when they do not, and what to do in each case. Later chapters will build on this foundation, introducing alternative estimators designed to address violations of the assumptions behind OLS.

1.2 Correlation and Causation

In empirical work, one of the first questions we ask is whether two variables are related. For example: Do countries with larger governments grow more slowly? Do education levels affect wages? Does monetary policy influence inflation? These questions all concern association, and a natural place to begin is to ask whether two variables move together in a systematic way.

1.2.1 Population Covariance and Correlation

Let (x, y) be two random variables defined on a probability space. The covariance between them is defined as:

$$Cov\left(y, x\right) = \mathcal{E}\left(yx\right) - \mathcal{E}\left(y\right)\mathcal{E}\left(x\right).$$

If x and y tend to be simultaneously above or below their means, the covariance is positive. If one tends to be above its mean when the other is below, the covariance is negative. If the two are statistically independent, the covariance is zero — although the converse is not true unless linearity is assumed.

Covariance is a valuable concept, but it depends on the scale of measurement. For example, if x is measured in kilograms and then converted to grams, the covariance changes proportionally. This makes it hard to compare covariances across variables with different units.

To overcome this, we define the correlation coefficient, which standardizes the covariance by the standard deviations of x and y:

$$\rho = \frac{Cov\left(y, x\right)}{\sqrt{V\left(y\right)}\sqrt{V\left(x\right)}}.$$

This yields a dimensionless number in the interval $[-1, 1]$, where $\rho = 1$ indicates perfect positive linear association, $\rho = -1$ indicates perfect negative linear association, and $\rho = 0$ indicates no linear association.

1.2.2 Sample Estimation

In practice, we do not observe the population distribution of (x, y). We observe a sample $\{y_t, x_t\}_{t=1}^{T}$, and compute the sample covariance:

$$s_{xy} = T^{-1}\sum_{t=1}^{T}\widetilde{x}_t\widetilde{y}_t,$$

where data expressed in deviations from the (sample) mean is defined as:

$$\widetilde{z}_t = z_t - \overline{z}, \quad \overline{z} = T^{-1}\sum_{t=1}^{T} z_t, \quad z = y, x.$$

We then define the sample correlation coefficient as:

$$r = T^{-1} \sum_{t=1}^{T} \frac{\widetilde{x}_t}{s_x} \frac{\widetilde{y}_t}{s_y} = \frac{s_{xy}}{s_x s_y}, \text{ where } s_z = \sqrt{T^{-1} \sum_{t=1}^{T} \widetilde{z}_t^2}, \quad z = y, x.$$

The sample correlation r, known as Pearson's correlation coefficient, serves as an estimator of the population correlation ρ, and inherits its basic interpretation: it measures the strength and direction of linear association in the data.[1]

1.2.3 Interpretation and Limitations

Correlation is a useful summary statistic, but it has important limitations. It captures association, not direction. It does not distinguish cause from effect. And it tells us nothing about the mechanisms linking two variables.

Consider, for example, the observed correlation between investment and economic growth. One might be tempted to conclude that higher investment leads to faster growth — a plausible story. But the reverse could also hold: fast-growing economies may attract more investment. Alternatively, both investment and growth could be driven by a common third factor, such as institutional quality or the rule of law. In countries with stable legal frameworks and enforceable property rights, both variables may perform better. Without modeling these mechanisms explicitly, the correlation coefficient alone cannot adjudicate between these explanations.

This is the classic post hoc, ergo propter hoc fallacy — the idea that because two things move together, one must cause the other. But as we will see, correlation can arise:

- From genuine causal mechanisms
- From reverse causation
- From common shocks or omitted variables
- From mechanical or definitional relationships

[1] Other measures of association exist, such as Spearman's rank correlation and Kendall's tau, which are based on ranks rather than magnitudes and are better suited to detecting monotonic (not necessarily linear) relationships.

- Or even by pure coincidence

The last case — spurious correlation — is especially deceptive. Yule (1926) famously illustrated this danger by documenting a strong historical correlation between mortality rates and the proportion of Church of England marriages in England. The implication — that civil marriage is lethal — was offered tongue-in-cheek, but the statistical lesson remains valid.

Modern datasets provide similarly absurd examples. These associations are real in the data, but economically meaningless. They arise from shared trends, unmodeled time dependence, or simple coincidence.[2] The problem is not in the numbers — it's in the lack of structure behind them.

In short, correlation is not causation. It cannot substitute for theory, nor justify economic conclusions on its own. If our goal is to understand economic relationships and evaluate counterfactuals, we need a model that defines directionality, accounts for residual variation, makes assumptions explicit, and tests them.

The next section introduces the simplest such framework: the simple linear regression model, in which one variable is modeled as a linear function of another plus a residual. This is our first step toward moving from descriptive statistics to econometric explanation.

1.3 The Simple Linear Regression Model

To move beyond descriptive statistics, we need a model — one that defines what we aim to explain, what we condition on, and what we treat as unobserved. The simple linear regression model offers such a framework while remaining analytically transparent and conceptually rich.

We begin by postulating a linear relationship between two variables, x_t and y_t, of the form:

$$y_t = \beta_1 + \beta_2 x_t + u_t, \quad t = 1, .., T.$$

This model separates the variation in y_t into two components:

- A systematic part, $\beta_1 + \beta_2 x_t$, which depends linearly on x_t.

[2] For a humorous collection of such absurd statistical relationships — including correlations between Nicolas Cage films and swimming pool drownings — see Vigen (2015). See also his website at `www.tylervigen.com`.

- An unsystematic part, u_t, which captures everything not explained by x_t.

We interpret x_t as the explanatory (or independent) variable, and y_t as the dependent variable. The coefficient β_1 is the intercept — the expected value of y when $x = 0$ — and β_2 is the slope, which reflects the marginal change in y associated with a one-unit increase in x, conditional on u_t being unrelated to x_t.

1.3.1 Why Estimate? And Why OLS?

In practice, we do not observe the parameters β_1 and β_2; we must estimate them using data. Suppose we have T observations of the pair $\{y_t, x_t\}$. Each pair gives us an equation in the unknowns β_1 and β_2, but the system is overdetermined: we have more equations than unknowns.

Geometrically, this means that there is no line that passes exactly through all the points (unless all the observations lie perfectly on a line — a degenerate case). Thus, we must choose a line that approximates the data in some sense.

But how? There are infinitely many ways to draw a line through a cloud of points. Some lines minimize the maximum deviation. Others minimize the sum of absolute deviations. Still others may enforce symmetry, or smoothness, or robust behavior under outliers. The Ordinary Least Squares (OLS) method chooses the line that minimizes the sum of squared vertical deviations — that is, it solves:

$$\min \sum_{t=1}^{T} \left(y_t - \beta_1 - \beta_2 x_t\right)^2 .$$

This choice is popular because it is analytically tractable, it delivers closed-form solutions, and under certain assumptions, it yields desirable statistical properties (discussed below).

But it is not unique in principle. OLS is one choice among many — a decision rule, not a law of nature. There are situations where OLS performs poorly: for instance, when there are outliers, or when minimizing squared error does not align with the economic objective. In later chapters, we will encounter alternative estimators that address these concerns.

1.3.2 Deriving the OLS Estimator

Let us now derive the OLS estimators formally.

We aim to choose the values of β_1 and β_2 that minimize the sum of squared residuals (SSR) of the sample:

$$S_T\left(\beta_1, \beta_2\right) = \sum_{t=1}^{T}\left(y_t - \beta_1 - \beta_2 x_t\right)^2.$$

The OLS estimator is obtained by deriving the objective function with respect to β_1 and β_2 and finding the values that equate these conditions to 0:

$$\begin{aligned} \left.\frac{\partial S_T\left(\beta\right)}{\partial \beta_1}\right|_{\widehat{\beta}} &= -2\sum\left(y_t - \widehat{\beta}_1 - \widehat{\beta}_2 x_t\right) = 0 \\ \left.\frac{\partial S_T\left(\beta\right)}{\partial \beta_2}\right|_{\widehat{\beta}} &= -2\sum x_t\left(y_t - \widehat{\beta}_1 - \widehat{\beta}_2 x_t\right) = 0. \end{aligned}$$

As there are two equations and 2 unknowns, the solutions are:

$$\begin{aligned} \widehat{\beta}_1 &= \overline{y} - \widehat{\beta}_2 \overline{x} \\ \widehat{\beta}_2 &= \frac{s_{xy}}{s_x^2} = r\frac{s_y}{s_x} = \frac{\sum_{t=1}^{T}\widetilde{x}_t\widetilde{y}_t}{\sum_{t=1}^{T}\widetilde{x}_t^2}. \end{aligned}$$

These formulas express the slope as a ratio of the sample covariance between x and y to the sample variance of x, and the intercept as the value needed to ensure the line passes through the point $\left(\overline{x}, \overline{y}\right)$.

Thus, $\left(\widehat{\beta}_1, \widehat{\beta}_2\right)$ satisfy the following properties:

- $\widehat{\beta}_1, \widehat{\beta}_2$ minimize SSR for the sample considered.
- They ensure that the lines passes through the mean point $(\overline{x}, \overline{y})$.
- The sum of the estimated residuals ($\widehat{u}_t \equiv y_t - \widehat{\beta}_1 - \widehat{\beta}_2 x_t$) is 0.
- The estimated residuals $\widehat{u}_t$ are uncorrelated (in the sample) with x_t.

These are not assumptions; they are algebraic consequences of the least squares minimization.

The expressions above define the OLS estimators — functions of the sample data used to infer unknown population parameters. Since they depend on the data, these are random variables: they vary from sample to sample. That is what makes them estimators.

When applied to a specific sample — that is, when we plug in observed values — these estimators yield estimates. For example, in a given dataset we might obtain $\widehat{\beta}_1 = 2, \widehat{\beta}_2 = 1.13$. These numbers are estimates: realizations of the corresponding estimators.

Understanding the difference between estimators and estimates is essential. Only estimators have sampling distributions, bias, and variance. Statistical inference concerns properties of estimators — not the particular numbers they produce in a single sample.

1.3.3 From Correlation to Conditional Expectation

OLS goes beyond correlation by modeling how y changes with x. It does this by treating y as a response variable and x as a conditioning variable. Implicitly, the model asserts:

$$\mathcal{E}\left(\left.y_t\right| x_t\right) = \beta_1 + \beta_2 x_t,$$

with the assumption that $\mathcal{E}\left(\left.u_t\right| x_t\right) = 0$. This is a substantive statement: it means the errors u_t are uncorrelated with x_t and carry no information that could improve prediction once x_t is known.

This shift — from symmetric association to asymmetric explanation — is the essence of regression. It marks the departure from descriptive statistics toward structural modeling.

In the next section, we introduce the general linear regression model, extending the ideas developed here to multiple explanatory variables using matrix notation.

1.4 The General Linear Regression Model

An econometrician has the observational data $\{w_1, w_2, ..., w_T\}$, where each w_t is a vector of data. Partition $w_t = (y_t, x_t)$ where $y_t \in \mathbb{R}$, $x_t \in \mathbb{R}^k$. Let the joint density of the variables be given by $f(y_t, x_t, \theta)$, where θ is a vector of unknown parameters.

In econometrics we are often interested in the conditional distribution of one set of random variables given another set of random variables (e.g., the conditional distribution of consumption given income, or the conditional distribution of wages given individual characteristics). Recalling that the joint density can be written as the product of the conditional density and the marginal density, we have:

$$f(y_t, x_t, \theta) = f(y_t \,|x_t, \theta_1)f(x_t, \theta_2),$$

where $f(x_t, \theta_2) = \int_{-\infty}^{\infty} f(y_t, x_t, \theta)dy$ is the marginal density of x.

Regression analysis can be defined as statistical inferences on θ_1. For this purpose we can ignore $f(x_t, \theta_2)$, provided there is no relationship between θ_1 and θ_2.[3] In this framework, y is called the 'dependent' or 'endogenous' variable and the vector x is called the vector of 'independent' or 'exogenous' variables.

In regression analysis we usually want to estimate only the first and second moments of the conditional distribution, rather than the whole parameter vector θ_1 (in certain cases the first two moments characterize θ_1 completely). Thus we can define the conditional mean $m\,(x_t, \theta_3)$ and conditional variance $g\,(x_t, \theta_4)$ as

$$\begin{aligned} m\,(x_t, \theta_3) &= \mathcal{E}\,(y_t \,|x_t, \theta_3\,) = \int_{-\infty}^{\infty} yf(y|\, x_t, \theta_1)dy \\ g\,(x_t, \theta_4) &= \int_{-\infty}^{\infty} y^2 f(y|\, x_t, \theta_1)dy - [m\,(x_t, \theta_3)]^2\,. \end{aligned}$$

The conditional mean and variance are random variables, as they are functions of the random vector x_t. If we define u_t as the difference between y_t and its conditional mean,

$$u_t = y_t - m\,(x_t, \theta_3)\,,$$

we obtain:

$$y_t = m\,(x_t, \theta_3) + u_t. \tag{1.1}$$

Other than (y_t, x_t) having a joint density, no assumptions have been made to develop (1.1).

Proposition 1 *Properties of* u_t :

1. $\mathcal{E}\,(u_t \,|x_t\,) = 0,$
2. $\mathcal{E}\,(u_t) = 0,$
3. $\mathcal{E}\,[h(x_t)u_t] = 0$ *for any function* $h\,(\cdot)\,,$
4. $\mathcal{E}\,(x_t u_t) = 0.$

[3] In this case we say that x is "weakly exogenous" for θ_1.

Proof. 1. By definition of u_t and the linearity of conditional expectations,[4]

$$\begin{aligned}\mathcal{E}\left(u_t\left|x_t\right.\right) &= \mathcal{E}\left[y_t - m\left(x_t\right)\left|x_t\right.\right] \\ &= \mathcal{E}\left[y_t\left|x_t\right.\right] - \mathcal{E}\left[m(x_t)\left|x_t\right.\right] \\ &= m\left(x_t\right) - m\left(x_t\right) = 0.\end{aligned}$$

2. By the law of iterated expectations and the first result,[5]

$$\mathcal{E}\left(u_t\right) = \mathcal{E}\left[\mathcal{E}\left(u_t\left|x_t\right.\right)\right] = \mathcal{E}\left(0\right) = 0.$$

3. By essentially the same argument,

$$\begin{aligned}\mathcal{E}\left[h(x_t)u_t\right] &= \mathcal{E}\left[\mathcal{E}\left[h(x_t)u_t\left|x_t\right.\right]\right] \\ &= \mathcal{E}\left[h(x_t)\mathcal{E}\left[u_t\left|x_t\right.\right]\right] \\ &= \mathcal{E}\left[h(x_t)\cdot 0\right] = 0.\end{aligned}$$

4. Follows from the third result, setting $h(x_t) = x_t$. ■

Equation (1.1) plus the first result of Proposition 1 are often stated jointly as the regression framework:

$$\begin{aligned}y_t &= m\left(x_t, \theta_3\right) + u_t \\ \mathcal{E}\left(u_t\left|x_t\right.\right) &= 0.\end{aligned}$$

This is a framework, not a model, because no restrictions have been placed on the joint distribution of the data. These equations hold true by definition.

Given that the moments $m\left(\cdot\right)$ and $g\left(\cdot\right)$ can take any shape (usually nonlinear), a regression model imposes further restrictions on the joint distribution and on u (the regression error). If we assume that $m\left(\cdot\right)$ is linear we obtain what is known as the linear regression model:

$$m\left(x_t, \theta_3\right) = x_t'\beta,$$

where β is a k-element vector. Finally, let

$$\underset{T\times 1}{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_T \end{bmatrix}, \quad \underset{T\times k}{X} = \begin{bmatrix} x_1' \\ \vdots \\ x_T' \end{bmatrix} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,k} \\ \vdots & \ddots & \vdots \\ x_{T,1} & \cdots & x_{T,k} \end{bmatrix}, \quad \underset{T\times 1}{u} = \begin{bmatrix} u_1 \\ \vdots \\ u_T \end{bmatrix}.$$

[4] The linearity of conditional expectations states that $\mathcal{E}\left[g\left(x\right)y\left|x\right.\right] = g\left(x\right)\mathcal{E}\left[y\left|x\right.\right]$.

[5] The law of iterated expectations states that $\mathcal{E}\left[\mathcal{E}\left[y\left|x,z\right.\right]\left|x\right.\right] = \mathcal{E}\left[y\left|x\right.\right]$.

Definition 2 *The Linear Regression Model (LRM) is:*

1. $y_t = x_t'\beta + u_t$ *or* $Y = X\beta + u,$

2. $\mathcal{E}\left(u_t\left|x_t\right.\right) = 0,$

3. *rank*$(X) = k$ *or* $\det\left(X'X\right) \neq 0,$

4. $\mathcal{E}\left(u_t u_s\right) = 0\ \forall t \neq s.$

The most important assumption of the model is the linearity of the conditional expectation. Furthermore, this framework considers that x provides no information for forecasting u and that X is of full rank. Finally, it is assumed that u_t is uncorrelated with u_s.[6]

Definition 3 *The Homoskedastic Linear Regression Model (HLRM) is the LRM plus*

5. $\mathcal{E}\left(u_t^2\left|x_t\right.\right) = \sigma^2$ *or* $\mathcal{E}\left(uu'\left|X\right.\right) = \sigma^2 I_T.$

This model adds the auxiliary assumption that $g\left(\cdot\right)$ is conditionally homoskedastic.

Definition 4 *The Normal Linear Regression Model (NLRM) is the LRM plus*

6. $u_t \sim \mathcal{N}\left(0, \sigma^2\right).$

Posing and additional assumption, this model has the advantage that exact distributional results are available for the OLS estimators and tests statistics. It is not very popular in current econometric practice and, as we will see, is not necessary to derive most of the results that follow.

1.5 OLS Estimation

This section defines the OLS estimator of β and shows that it is the best linear unbiased estimator.[7] The estimation of the error variance is also discussed.

[6] Ocassionally, we will make the assumption of serial independence of $\{u_t\}$ which is stronger than no correlation, although both concepts are equivalent when u is normal.

[7] The method of least squares was first published by Adrien-Marie Legendre in 1805, but Carl Friedrich Gauss claimed prior use dating back to 1795. Gauss later formalized the method in his 1809 work on celestial mechanics. The priority dispute lingered, but both contributions helped establish least squares as the dominant tool for astronomical and statistical estimation.

1.5.1 Definition of the OLS Estimators of β and σ^2

Define the sum of squares of the residuals (SSR) function as:

$$\begin{aligned} S_T(\beta) &= (Y - X\beta)'(Y - X\beta) \\ &= Y'Y - 2Y'X\beta + \beta'X'X\beta. \end{aligned}$$

The OLS estimator $(\widehat{\beta})$ minimizes $S_T(\beta)$. The First Order Necessary Conditions (FONC) for minimization are:

$$\left.\frac{\partial S_T(\beta)}{\partial \beta}\right|_{\widehat{\beta}} = -2X'Y + 2X'X\widehat{\beta} = 0,$$

which yield the normal equations $X'Y = X'X\widehat{\beta}$.

Proposition 5 $\arg\min_{\beta} S_T(\beta) = \widehat{\beta} = (X'X)^{-1}(X'Y)$.

Proof. Using the normal equations we obtain $\widehat{\beta} = (X'X)^{-1}(X'Y)$. To verify that $\widehat{\beta}$ is indeed a minimum we evaluate the Second Order Sufficient Conditions (SOSC)

$$\left.\frac{\partial^2 S_T(\beta)}{\partial \beta \partial \beta'}\right|_{\widehat{\beta}} = 2X'X,$$

which show that $\widehat{\beta}$ is a minimum, as $X'X$ is a positive definite matrix. ■

Three important implications are derived from this theorem: First, $\widehat{\beta}$ is a linear function of Y. Second, even if X is a non stochastic matrix, $\widehat{\beta}$ is a random variable as it depends on Y which is itself a random variable. Finally, in order to obtain the OLS estimator we require $X'X$ to be of full rank.

Given $\widehat{\beta}$, we define

$$\widehat{u} = Y - X\widehat{\beta}, \tag{1.2}$$

and call it the least squares residuals. Using $\widehat{u}$, we can estimate σ^2 by

$$\widehat{\sigma}^2 = T^{-1}\widehat{u}'\widehat{u}.$$

Using (1.2), we can write

$$Y = X\widehat{\beta} + \widehat{u} = PY + MY,$$

where $P = X\left(X'X\right)^{-1}X'$ and $M = I - P$. Given that $\widehat{u}$ is orthogonal to X (that is, $\widehat{u}'X = 0$), OLS can be regarded as decomposing Y into two orthogonal components: a component that can be written as a linear combination of the column vector of X and a component that is orthogonal to X. Alternatively, we can call PY the projection of Y onto the space spanned by the column vectors of X and MY the projection of Y onto the space orthogonal to X. These properties are illustrated in Figure 1.1.[8]

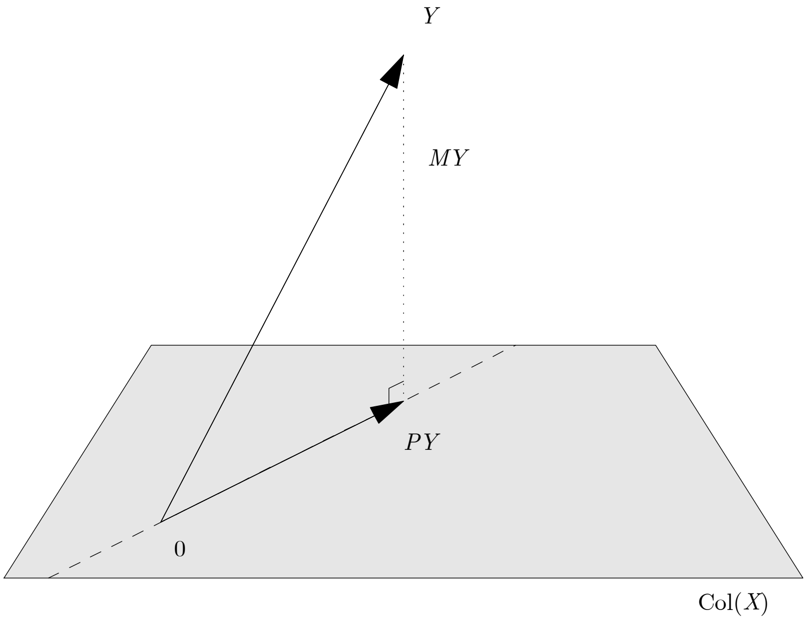

Figure 1.1: Orthogonal Decomposition of Y

Proposition 6 *Let X be an $T \times k$ matrix of rank k. A matrix of the form $P = X\left(X'X\right)^{-1}X'$ is called a projection matrix and has the following properties:*

i) $P = P' = P^2$ (Hence P is symmetric and idempotent),

ii) rank$(P) = k$,

iii) the characteristic roots (eigenvalues) of P consist of k ones and T-k zeros,

[8] The column space of X is denoted by Col(X).

iv) if $Z = Xc$ for some vector c, then $PZ = Z$ (hence the word projection),

v) $M = I - P$ is called the annihilator matrix and is also symmetric and idempotent with rank T-k, the eigenvalues consist of T-k ones and k zeros, and if $Z = Xc$, then MZ$= 0$,

vi) P can be written as $G'G$, where $GG' = I$, or as $v_1v_1' + v_2v_2' + ... + v_rv_r'$ where v_i is a vector and $k = rank(P)$.

Proof. Left as an exercise. ■

1.5.2 Gaussian Quasi-Maximum Likelihood Estimator

Now we relate a traditional motivation for the OLS estimator. The *NLRM* is $y_t = x_t'\beta + u_t$ with $u_t \sim \mathcal{N}(0, \sigma^2)$.

The density function for a single observation is

$$f\left(y_t \left| x_t, \beta, \sigma^2\right.\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_t - x_t'\beta)^2}{2\sigma^2}}$$

and the log-likelihood for the full sample is

$$\begin{aligned} \ell_T\left(\beta, \sigma^2; Y \left| X\right.\right) &= \ln\left[\prod_{t=1}^{T} f\left(y_t \left| x_t, \beta, \sigma^2\right.\right)\right] \\ &= \sum_{t=1}^{T} \ln f\left(y_t \left| x_t, \beta, \sigma^2\right.\right) \\ &= -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{t=1}^{T}\left(y_t - x_t'\beta\right)^2 \\ &= -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln\left(\sigma^2\right) - \frac{1}{2\sigma^2}S_T(\beta). \end{aligned}$$

Proposition 7 *In the NLRM, $\widehat{\beta}_{MLE} = \widehat{\beta}_{OLS}$.*

Proof. The FONC for the maximization of $\ell_T(\beta, \sigma^2)$ are:

$$\begin{aligned} \left.\frac{\partial \ell_T(\beta, \sigma^2)}{\partial \beta}\right|_{\widehat{\beta}, \widehat{\sigma}^2} &= \frac{1}{\widehat{\sigma}^2}\left(X'Y - X'X\widehat{\beta}\right) = 0 \\ \left.\frac{\partial \ell_T(\beta, \sigma^2)}{\partial \sigma^2}\right|_{\widehat{\beta}, \widehat{\sigma}^2} &= -\frac{T}{2\widehat{\sigma}^2} + \frac{\left(Y - X\widehat{\beta}\right)'\left(Y - X\widehat{\beta}\right)}{2\widehat{\sigma}^4} = 0. \end{aligned}$$

Thus, $\widehat{\beta}_{MLE} = (X'X)^{-1}(X'Y)$ and $\widehat{\sigma}^2 = T^{-1}\widehat{u}'\widehat{u}$.[9] ■

This result is obvious since $\ell_T(\beta, \sigma^2)$ is a function of β only through $S_T(\beta)$. Thus, MLE maximizes $\ell_T(\beta, \sigma^2)$ by minimizing $S_T(\beta)$. Due to this equivalence, the OLS estimator $\widehat{\beta}$ is frequently referred to as the "Gaussian MLE", the "Gaussian Quasi-MLE", or the "Gaussian Pseudo-MLE".[10]

1.5.3 The Mean and Variance of $\widehat{\beta}$ and $\widehat{\sigma}^2$

Proposition 8 *In the LRM,* $\mathcal{E}\left[\left(\widehat{\beta} - \beta\right)|X\right] = 0$ *and* $\mathcal{E}\left(\widehat{\beta}\right) = \beta$.

Proof. From previous results,

$$\begin{aligned} \widehat{\beta} &= (X'X)^{-1}X'Y = (X'X)^{-1}X'(X\beta + u) \\ &= \beta + (X'X)^{-1}X'u. \end{aligned}$$

Then

$$\begin{aligned} \mathcal{E}\left[\left(\widehat{\beta} - \beta\right)|X\right] &= \mathcal{E}\left[(X'X)^{-1}X'u\,|X\right] \\ &= (X'X)^{-1}X'\mathcal{E}(u\,|X) \\ &= 0. \end{aligned}$$

Applying the law of iterated expectations, $\mathcal{E}\left(\widehat{\beta}\right) = \mathcal{E}\left[\mathcal{E}\left(\widehat{\beta}\,|X\right)\right] = \beta$. ■

Thus, $\widehat{\beta}$ is unbiased for β. Indeed it is conditionally unbiased (conditional on X), which is a stronger result.

Proposition 9 *In the HLRM,* $\mathcal{V}\left(\widehat{\beta}\,|X\right) = \sigma^2(X'X)^{-1}$ *and* $\mathcal{V}\left(\widehat{\beta}\right) = \sigma^2\mathcal{E}\left[(X'X)^{-1}\right]$.

Proof. Since $\widehat{\beta} - \beta = (X'X)^{-1}X'u$,

$$\begin{aligned} \mathcal{V}\left(\widehat{\beta}\,|X\right) &= \mathcal{E}\left[\left(\widehat{\beta} - \beta\right)\left(\widehat{\beta} - \beta\right)'|X\right] \\ &= \mathcal{E}\left[(X'X)^{-1}X'uu'X(X'X)^{-1}|X\right] \\ &= (X'X)^{-1}X'\mathcal{E}[uu'|X]X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}. \end{aligned}$$

[9] Verify that the SOSC are satisfied.

[10] The term "quasi" ("pseudo") is used for misspecified models. In this case, the normality assumption was used to construct the likelihood and the estimator, but may be believed not to be true.

Thus, $\mathcal{V}\left(\widehat{\beta}\right)=\mathcal{E}\left[\mathcal{V}\left(\widehat{\beta}\,|X\right)\right]+\mathcal{V}\left[\mathcal{E}\left(\widehat{\beta}\,|X\right)\right]=\sigma^2\mathcal{E}\left[\left(X'X\right)^{-1}\right].$ ■

This result is derived from the assumptions that u is uncorrelated and homoskedastic. The variance-covariance matrix of $\widehat{\beta}$ measures the precision with which the relationship between Y and X is estimated. Some of its features are: First, and most obvious, the variance of $\widehat{\beta}$ grows proportionally with σ^2 (the volatility of the unpredictable component). Second, although less obvious, as the sample size increases, the variance-covariance matrix of $\widehat{\beta}$ should decrease (we will provide formal arguments in this regard when we analyze the asymptotic properties of OLS). Finally, it also depends on the volatility of the regressors; as it increases, the precision with which we measure β will be enhanced. Thus, we generally "prefer" a sample of X that is more volatile, given that it would better help us to uncover its association with Y.

Proposition 10 *In the LRM, $\widehat{\sigma}^2$ is biased.*

Proof. We know that $\widehat{u}=MY$. It is trivial to verify that $\widehat{u}=Mu$. Then, $\widehat{\sigma}^2=T^{-1}\widehat{u}'\widehat{u}=T^{-1}u'Mu$. This implies that

$$\begin{aligned}\mathcal{E}\left(\widehat{\sigma}^2\,|X\right) &= T^{-1}\mathcal{E}\left[u'Mu\,|X\right]\\ &= T^{-1}\mathrm{tr}\,\mathcal{E}\left[u'Mu\,|X\right]\\ &= T^{-1}\mathcal{E}\left[\mathrm{tr}\left(u'Mu\right)|X\right]\\ &= T^{-1}\mathcal{E}\left[\mathrm{tr}\left(Muu'\right)|X\right]\\ &= T^{-1}\sigma^2\mathrm{tr}\left(M\right)\\ &= \sigma^2\left(T-k\right)T^{-1}.\end{aligned}$$

Applying the law of iterated expectations we obtain $\mathcal{E}\left(\widehat{\sigma}^2\right)=\sigma^2\left(T-k\right)T^{-1}$. ■

To derive this result we used the facts that σ^2 is a scalar (tr denotes the trace of a matrix), that the expectation is a lineal operator (thus tr and $\mathcal{E}$ are interchangeable), that tr(AB)=tr(BA), and that M is symmetric in which case tr(M) $=\sum_{i=1}^{T}\lambda_i$, where λ_i denotes the i-th eigenvalue of M (here we used the results of Proposition 6).

Proposition 10 shows that $\widehat{\sigma}^2$ is a biased estimator of σ^2. A trivial modification yields an unbiased estimator for σ^2; $\widetilde{\sigma}^2=\left(T-k\right)^{-1}\widehat{u}'\widehat{u}$.

Proposition 11 *In the NLRM,* $\mathcal{V}\left(\widehat{\sigma}^2\right) = T^{-2}2\left(T-k\right)\sigma^4.$

Proof. Left as an exercise. ■

From the results derived so far, three facts are worth mentioning: First, with the exception of Proposition 11 none of the results derived required the assumption of normality of the error term. Second, while $\widehat{\sigma}^2$ is biased, it coincides with the maximum likelihood estimator of the variance under the assumption of normality of u and, as we will show later, it is consistent. Finally, both the variance-covariance matrix of $\widehat{\beta}$ and $\widehat{\sigma}^2$ depend on σ^2 which is unknown, thus in practice we use estimators of the variance-covariance matrix of the OLS estimators by replacing σ^2 with $\widehat{\sigma}^2$ or $\widetilde{\sigma}^2$. For example, the estimator of the variance-covariance matrix of $\widehat{\beta}$ is $\widehat{\mathcal{V}}\left(\widehat{\beta}\right) = \widetilde{\sigma}^2\left(X'X\right)^{-1}.$

1.5.4 $\widehat{\beta}$ is BLUE

Definition 12 *Let* $\widehat{\theta}$ *and* θ^* *be estimators of a vector parameter* θ*. Let* A *and* B *be their respective mean squared error matrices; that is* $A = \mathcal{E}\left(\widehat{\theta}-\theta\right)\left(\widehat{\theta}-\theta\right)'$ *and* $B = \mathcal{E}\left(\theta^*-\theta\right)\left(\theta^*-\theta\right)'$*. We say that* $\widehat{\theta}$ *is better (or more efficient) than* θ^* *if* $c'\left(B-A\right)c \geq 0$ *for every vector* c *and every parameter value and* $c'\left(B-A\right)c > 0$ *for at least one value of* c *and at least one value of the parameter.*[11]

Once we made precise what we mean by *better*, we are ready to present one of the most famous theorems in econometrics;

Theorem 13 (Gauss-Markov) *The Best Linear Unbiased Estimator (BLUE) is* $\widehat{\beta}$*.*

Proof. Let $A = \left(X'X\right)^{-1}X'$, then $\widehat{\beta} = AY$. Consider any other linear estimator b. Without loss of generality, let $b = \left(A+C\right)Y$. Then,

$$\mathcal{E}\left(b\left|X\right.\right) = \left(X'X\right)^{-1}X'X\beta + CX\beta = \left(I+CX\right)\beta.$$

[11] This definition can also be stated as $B \geq A$ for every parameter value and $B \neq A$ for at least one parameter value (in this context, $B \geq A$ means that $B-A$ is positive semi-definite and $B > A$ means that $B-A$ is positive definite).

For b to be unbiased we require $CX = 0$ to hold, in which case

$$\mathcal{V}\left(b\left|X\right.\right) = \mathcal{E}\left[\left(A+C\right)uu'\left(A+C\right)'\right].$$

As $\left(A+C\right)\left(A+C\right)' = \left(X'X\right)^{-1} + CC'$, we obtain

$$\mathcal{V}\left(b\left|X\right.\right) = \mathcal{V}\left(\widehat{\beta}\left|X\right.\right) + \sigma^2 CC'.$$

Then, $\mathcal{V}\left(b\left|X\right.\right) \geq \mathcal{V}\left(\widehat{\beta}\left|X\right.\right)$, as CC' is a positive semi-definite matrix. ■

Despite its popularity, the Gauss-Markov theorem is not very powerful. It restricts our quest of alternative candidates to those that are both linear and unbiased estimators. There may be a "nonlinear" or biased estimator that can do better in the metric of Definition 12. Furthermore, OLS ceases to be BLUE when the assumption of homoskedasticity is relaxed. If both homoskedasticity and normality are present, we can rely on a stronger theorem which we will discuss later (the Cramer-Rao lower bound).

1.5.5 Analysis of Variance (ANOVA)

By definition,

$$Y = \widehat{Y} + \widehat{u}.$$

Subtracting $\overline{Y}$ (the sample mean of Y) from both sides we have,

$$Y - \overline{Y} = \left(\widehat{Y} - \overline{Y}\right) + \widehat{u}.$$

Thus

$$\left(Y - \overline{Y}\right)'\left(Y - \overline{Y}\right) = \left(\widehat{Y} - \overline{Y}\right)'\left(\widehat{Y} - \overline{Y}\right) + 2\left(\widehat{Y} - \overline{Y}\right)'\widehat{u} + \widehat{u}'\widehat{u},$$

but $\widehat{Y}'\widehat{u} = Y'PMY = 0$ and $\overline{Y}'\widehat{u} = \overline{Y}\imath'\widehat{u} = 0$ when the model contains an intercept (more generally, if $\imath$ lies in the space spanned by X).[12] Thus

$$\left(Y - \overline{Y}\right)'\left(Y - \overline{Y}\right) = \left(\widehat{Y} - \overline{Y}\right)'\left(\widehat{Y} - \overline{Y}\right) + \widehat{u}'\widehat{u}.$$

[12] We define $\imath$ as a row vector of ones.

This is called the analysis of variance formula, often written as

$$TSS = ESS + SSR,$$

where TSS, ESS, and SSR stand for "Total sum of squares", "Equation sum of squares" and "Sum of squares of the residuals", respectively. The equation R^2 (also known as the centered coefficient of determination) is defined as

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS} = 1 - \frac{Y'MY}{Y'LY},$$

where $L = I_T - T^{-1}\imath\imath'$. Therefore, provided that the regressors include a constant, $0 \leq R^2 \leq 1$. If the regressors do not include a constant, R^2 can be negative because, without the benefit of an intercept, the regression could do worse (tracking the dependent variable) than the sample mean.

The equation measures the percentage of the variance of Y that is accounted for in the variation of the predicted value $\widehat{Y}$. R^2 is typically reported in applied work and is frequently referenced as "measure" or "goodness" of fit. This label is inappropriate, as R^2 does not measure the adequacy or "fit" of a model.[13]

It is not even clear if R^2 has an unambiguous interpretation in terms of forecast performance. To see this, note that the "explanatory" power of the models $y_t = x_t\beta + u_t$ and $y_t - x_t = x_t\gamma + u_t$ with $\gamma = \beta - 1$ are the same. The models are mathematically identical and yield the same implications and forecasts. Yet their reported R^2 will differ greatly. For illustration, suppose that $\beta \simeq 1$. Then the R^2 from the second model will (nearly) equal zero, while the R^2 from the first model can be arbitrarily close to one. An econometrician reporting the near-unit R^2 from the first model might claim "success", while an econometrician reporting the $R^2 \simeq 0$ from the second model might be accused of a poor fit. This difference in reporting is quite unfortunate, since the two models and implications are mathematically identical. The bottom line is that R^2 is not a measure of fit and should not be interpreted as such.

This phenomenon is illustrated in Figure 1.2, which compares the two models using simulated data. The discrepancy in R^2 values is stark, despite the fact that both models fit the data equally well in substantive terms. The

[13] More unfortunate is the claim that the R^2 measures the percentage of the variance of y that is "explained" by the model. An econometric model, by itself, doesn't explain anything. Only the combination of a good econometric model and sound economic theory can, in principle, "explain" a phenomenon.

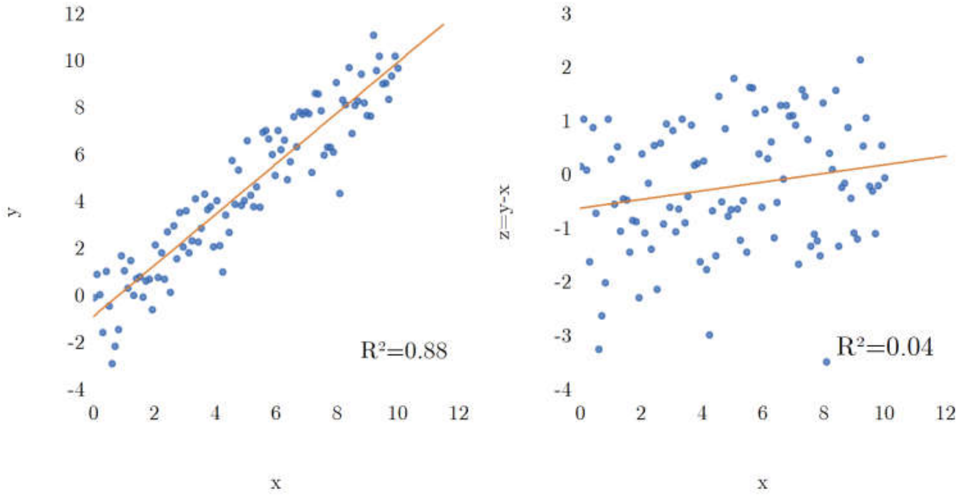

Figure 1.2: The R^2 Fallacy: Why Good Models Can Look Bad

bottom line is that R^2 is not a measure of fit and should not be interpreted as such.

Another interesting fact about R^2 is that it necessarily increases as regressors are added to the model. As by definition the OLS estimate minimizes the SSR, by adding additional regressors, the SSR cannot increase; it either can stay the same, or (more likely) decrease. But the TSS is unaffected by adding regressors, so that R^2 either stays constant or increases. To counteract this effect, Theil proposed an adjustment, typically called $\overline{R}^2$ (or "adjusted" R^2) which penalizes model dimensionality and is defined as:

$$\overline{R}^2 = 1 - \frac{SSR}{TSS}\frac{T}{T-k} = 1 - \frac{\widetilde{\sigma}^2}{\widehat{\sigma}_y^2}.$$

While often reported in applied work, this statistic is not used that much today as better model evaluation criteria have been developed, which we will discuss on Chapter 2.

1.5.6 OLS Estimator of a Subset of β

Sometimes we may not be interested in obtaining estimates of the whole parameter vector, but only of a subset of β. Partition

$$X = \begin{bmatrix} X_1 & X_2 \end{bmatrix}$$

and

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}.$$

Then $X'X\widehat{\beta} = X'Y$ can be written as:

$$X_1'X_1\widehat{\beta}_1 + X_1'X_2\widehat{\beta}_2 = X_1'Y \qquad (1.3a)$$
$$X_2'X_1\widehat{\beta}_1 + X_2'X_2\widehat{\beta}_2 = X_2'Y. \qquad (1.3b)$$

Solving for $\widehat{\beta}_2$ and reinserting in (1.3a) we obtain

$$\widehat{\beta}_1 = \left(X_1'M_2X_1\right)^{-1} X_1'M_2Y$$

and

$$\widehat{\beta}_2 = \left(X_2'M_1X_2\right)^{-1} X_2'M_1Y,$$

where $M_i = I - P_i = I - X_i\left(X_i'X_i\right)^{-1} X_i'$ (for $i = 1, 2$).

These results can also be derived using the following theorem:

Theorem 14 (Frisch-Waugh-Lovell) *$\widehat{\beta}_2$ and $\widehat{u}$ can be computed using the following algorithm:*

1. *Regress Y on X_1, obtain residual $\widetilde{Y}$,*
2. *Regress X_2 on X_1, obtain residual $\widetilde{X}_2$,*
3. *Regress $\widetilde{Y}$ on $\widetilde{X}_2$, obtain $\widehat{\beta}_2$ and residuals $\widehat{u}$.*

Proof. Left as an exercise. ■

In some contexts, the Frisch-Waugh-Lovell (FWL) theorem can be used to speed computation, but in most cases there is little computational advantage of using it.[14] There are, however, two common applications of the FWL theorem, one of which is usually presented in introductory econometrics courses:

[14] A few decades ago, a crucial limitation for conducting OLS estimation was the computational cost of inverting even moderately sized matrices and the FWL was invoked routinely.

the demeaning formula for regression; the other deals with ill-conditioned problems.

The first application can be constructed as follows: Partition $X = \begin{bmatrix} X_1 & X_2 \end{bmatrix}$ where $X_1 = \imath$ is a vector of ones, and X_2 is the matrix of observed regressors. In this case,

$$\begin{aligned} \widetilde{X}_2 &= M_1 X_2 = X_2 - \imath \left(\imath'\imath\right)^{-1} \imath' X_2 \\ &= X_2 - \overline{X}_2 \end{aligned}$$

and

$$\begin{aligned} \widetilde{Y} &= M_1 Y = Y - \imath \left(\imath'\imath\right)^{-1} \imath' Y \\ &= Y - \overline{Y}. \end{aligned}$$

which are ‘demeaned’.

The FWL theorem says that $\widehat{\beta}_2$ is the OLS estimate from a regression of $\widetilde{Y}$ on $\widetilde{X}_2$, or $y_t - \overline{Y}$ on $x_{2t} - \overline{X}_2$:

$$\widehat{\beta}_2 = \left(\sum_{t=1}^{T} \left(x_{2t} - \overline{X}_2\right) \left(x_{2t} - \overline{X}_2\right)' \right)^{-1} \left(\sum_{t=1}^{T} \left(x_{2t} - \overline{X}_2\right) \left(y_t - \overline{Y}\right)' \right).$$

Thus, the OLS estimator for the slope coefficients is a regression with demeaned data.

The other application is more useful. In our analysis we assumed that X is full rank ($X'X$ is invertible). Suppose for a moment that X_1 is full rank but that X_2 is not. In that case β_2 cannot be estimated, but β_1 still can be estimated as follows:

$$\widehat{\beta}_1 = \left(X_1' M_2^* X_1\right)^{-1} X_1' M_2^* Y,$$

where M_2^* is formed using X_2^* that has columns equal to the maximal number of linearly independent columns of X_2.

1.6 Constrained Least Squares (CLS)

In this section we shall consider the estimation of β and σ^2 when there are certain linear constraints on the elements of β. We shall assume that the constraints are of the form:

$$Q'\beta = c, \tag{1.4}$$

where Q is a $k \times q$ matrix of known constants and c is a q-vector of known constants. We shall also assume that $q < k$ and $\text{rank}(Q) = q$.

1.6.1 Derivation of the CLS Estimator

The CLS estimator of β, denoted by $\overline{\beta}$, is defined to be the value of β that minimizes the SSR subject to the constraint (1.4). The Lagrange expression for the CLS minimization problem is

$$\mathcal{L}(\beta,\gamma) = (Y - X\beta)'(Y - X\beta) + 2\gamma'(Q'\beta - c),$$

where γ is a q-vector of Lagrange multipliers corresponding to the q constraints. The FONC are

$$\begin{aligned} \left.\frac{\partial \mathcal{L}}{\partial \beta}\right|_{\overline{\beta},\overline{\gamma}} &= -2X'Y + 2X'X\overline{\beta} + 2Q\overline{\gamma} = 0 \\ \left.\frac{\partial \mathcal{L}}{\partial \gamma}\right|_{\overline{\beta},\overline{\gamma}} &= Q'\overline{\beta} - c = 0. \end{aligned}$$

The solution for $\overline{\beta}$ is

$$\overline{\beta} = \widehat{\beta} - (X'X)^{-1} Q \left[Q'(X'X)^{-1} Q\right]^{-1} \left(Q'\widehat{\beta} - c\right). \tag{1.5}$$

The corresponding estimator of σ^2 can be defined as

$$\overline{\sigma}^2 = T^{-1}\left(Y - X\overline{\beta}\right)'\left(Y - X\overline{\beta}\right).$$

1.6.2 CLS as BLUE

It can be shown that (1.5) can be expressed as

$$\overline{\beta} = \beta + R(R'X'XR)^{-1} R'X'u,$$

where R is a $k \times (k - q)$ matrix such that the matrix (Q, R) is nonsingular and $R'Q = 0$.[15] Therefore $\overline{\beta}$ is unbiased and its variance-covariance matrix is given by

$$\mathcal{V}\left(\overline{\beta}\right) = \sigma^2 R(R'X'XR)^{-1} R'.$$

Now define the class of linear estimators $\beta^* = D'Y - d$ where D' is a $k \times T$ matrix and d is a k-vector. This class is broader than the class of linear estimators considered in the unconstrained case because of the additive

[15] Such a matrix can always be found and is not unique, and any matrix that satisfies these conditions will do.

constants d. We did not include d previously because in the unconstrained model the unbiasedness condition would ensure $d = 0$. Here, the unbiasedness condition $\mathcal{E}\left(D'Y - d\right) = \beta$ implies $D'X = I + GQ'$ and $d = Gc$ for some arbitrary $k \times q$ matrix G. We have $\mathcal{V}\left(\beta^*\right) = \sigma^2 D'D$ and CLS is BLUE because of the identity

$$\begin{aligned} & D'D - R\left(R'X'XR\right)^{-1}R' \\ = & \left[D' - R\left(R'X'XR\right)^{-1}R'X'\right]\left[D' - R\left(R'X'XR\right)^{-1}R'X'\right]', \end{aligned}$$

where we have used $D'X = I + GQ'$ and $R'Q = 0$.

1.7 Inference with Linear Constraints

In this section we shall regard the linear constraints (1.4) as a testable hypothesis, calling it the null hypothesis. For now we will assume that the normal linear regression model holds and derive the most frequently used tests in the OLS context.[16]

1.7.1 The t Test

The t test is an ideal test to use when we have a single constraint, that is, $q = 1$. As we assumed that u is normally distributed, so is $\widehat{\beta}$; thus under the null hypothesis we have

$$Q'\widehat{\beta} \overset{a}{\backsim} \mathcal{N}\left[c, \sigma^2 Q'\left(X'X\right)^{-1}Q\right].$$

With $q = 1$, Q' is a row vector and c is a scalar. Therefore

$$\frac{Q'\widehat{\beta} - c}{\left[\sigma^2 Q'\left(X'X\right)^{-1}Q\right]^{1/2}} \sim \mathcal{N}\left(0,1\right). \tag{1.6}$$

This is the test statistic that we would use if σ were known. As

$$\frac{\widehat{u}'\widehat{u}}{\sigma^2} \sim \chi^2_{T-k}, \tag{1.7}$$

[16] We will discuss the case of inference in the presence of nonlinear constraints and departures from normality of u later. For those impatient, none of the results derived here change when these assumptions are relaxed (at least asymptotically).

it can be shown that (1.6) and (1.7) are independent, hence:

$$t_T = \frac{Q'\widehat{\beta} - c}{\left[\widetilde{\sigma}^2 Q' \left(X'X\right)^{-1} Q\right]^{1/2}} \sim S_{T-k},$$

which is Student's t with $T-k$ degrees of freedom. Only now we have invoked the assumption of normality of u and, as shown later, it is not necessary for (1.6) to hold (in large samples).

That is, if u is not normally distributed, the exact finite-sample result no longer holds. The t distribution is no longer a valid reference distribution. In such case, one must rely instead on asymptotic theory: under weak regularity conditions (discussed later), the t-statistic converges in distribution to a standard normal. This asymptotic approximation is typically used in empirical work when normality cannot be assumed.

If we were interested in testing a single hypothesis of the form:

$$\mathrm{H}_0 : \beta_1 = 0,$$

we would define $Q = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}'$ and $c = 0$, in which case we would obtain the familiar t test

$$t_T = \frac{\widehat{\beta}_1}{\sqrt{\widehat{\mathcal{V}}_{1,1}}},$$

where $\widehat{\mathcal{V}}_{1,1}$ is the 1,1 component of the estimator of the variance-covariance matrix of $\widehat{\beta}$.

With these tools we can construct confidence intervals C_T for β_i. As C_T is a function of the data, it is random. Its objective is to cover β_i with high probability. The coverage probability is $\Pr\left(\beta \in C_T\right)$. We say that C_T has $(1-\alpha)\,\%$ coverage for β if $\Pr\left(\beta \in C_T\right) \to (1-\alpha)$. We construct a confidence interval as follows:

$$\Pr\left[\widehat{\beta}_i - z_{\alpha/2}\sqrt{\widehat{\mathcal{V}}_{i,i}} < \beta_i < \widehat{\beta}_i + z_{\alpha/2}\sqrt{\widehat{\mathcal{V}}_{i,i}}\right] = 1-\alpha,$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the distribution being considered (asymptotically, the normal distribution; in small samples, the Student's t distribution). The most common choice for α is 0.05. If $|t_T| < z_{\alpha/2}$, we cannot reject the null hypothesis at an $\alpha\%$ significance level; otherwise the null hypothesis is rejected.

An alternative approach to reporting results, is to report a p-value. The p-value for the above statistic is constructed as follows. Define the tail probability, or p-value function

$$p_T = p\left(t_T\right) = \Pr\left(|Z| \geq |t_T|\right) = 2\left(1 - \Phi\left(|t_T|\right)\right).$$

If the p-value p_T is small (close to zero) then the evidence against H_0 is strong. In a sense, p-values and hypothesis tests are equivalent since $p_T \leq \alpha$ if and only if $|t_T| \geq z_{\alpha/2}$. The p-value is more general, however, in that the reader is allowed to pick the level of significance α.[17]

A confidence interval for σ can be constructed as follows

$$\Pr\left[\frac{(T-k)\,\widetilde{\sigma}^2}{\chi^2_{T-k,1-\alpha/2}} < \sigma^2 < \frac{(T-k)\,\widetilde{\sigma}^2}{\chi^2_{T-k,\alpha/2}}\right] = 1 - \alpha. \tag{1.8}$$

1.7.2 The F Test

When $q > 1$ we cannot apply the t test described above and use instead a simple transformation of what is known as the Likelihood Ratio Test (which we will discuss at length later). Under the null hypothesis, it can be shown that

$$\frac{S_T\left(\overline{\beta}\right) - S_T\left(\widehat{\beta}\right)}{\sigma^2} \sim \chi^2_q.$$

As in the previous case, when σ^2 is not known, a finite sample correction can be made by replacing σ^2 with $\widetilde{\sigma}^2$, in which case we have

$$\frac{S_T\left(\overline{\beta}\right) - S_T\left(\widehat{\beta}\right)}{\widetilde{\sigma}^2} = \frac{T-k}{q}\frac{\left(Q'\widehat{\beta} - c\right)'\left[Q'\left(X'X\right)^{-1}Q\right]^{-1}\left(Q'\widehat{\beta} - c\right)}{\widehat{u}'\widehat{u}} \tag{1.9}$$

which has a $F_{q,T-k}$ distribution. Once again, as in the case of t tests, we reject the null hypothesis when the value computed exceeds the critical value.

Again, if u is not normally distributed, the exact finite-sample result no longer holds. The F distribution is no longer valid a reference distribution. In such case, one must rely instead on asymptotic theory, where the F-statistic converges to a chi-squared distribution (divided by its degrees of freedom).

Moreover, when sample sizes are moderate and the distribution of errors is unknown or suspected to be non-normal, one may wish to go beyond

[17] GAUSS tip: to compute $p\left(t\right)$ use $2 * \texttt{cdfnc}\left(t\right)$.

asymptotics. A promising approach in such settings is to apply bootstrap methods, which provide refined finite-sample approximations to the distribution of test statistics. Bootstrap inference does not require normality and often improves upon the accuracy of asymptotic inference. These techniques will be discussed in detail in Chapter 2.

1.7.3 Tests for Structural Breaks

Suppose we have a two-regimes regression

$$\begin{aligned} Y_1 &= X_1\beta_1 + u_1 \\ Y_2 &= X_2\beta_2 + u_2, \end{aligned}$$

where the vectors and matrices have T_1 and T_2 rows respectively ($T = T_1 + T_2$). Suppose further that

$$\mathcal{E}\begin{bmatrix} u_1 \\ u_2 \end{bmatrix}\begin{bmatrix} u_1' & u_2' \end{bmatrix} = \begin{bmatrix} \sigma_1^2 I_{T_1} & 0 \\ 0 & \sigma_2^2 I_{T_2} \end{bmatrix}.$$

We want to test the null hypothesis $\mathrm{H}_0 : \beta_1 = \beta_2$. First, we will derive an F test assuming homoskedasticity among regimes and later we will relax this assumption. To apply the test we define

$$Y = \underline{X}\beta + u,$$

where:

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}, \quad \underline{X} = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}, \quad \text{and} \quad u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}.$$

Applying (1.9) we obtain:

$$\frac{T_1 + T_2 - 2k}{k}\frac{\left(\widehat{\beta}_1 - \widehat{\beta}_2\right)'\left[(X_1'X_1)^{-1} + (X_2'X_2)^{-1}\right]^{-1}\left(\widehat{\beta}_1 - \widehat{\beta}_2\right)}{Y'\left[I - \underline{X}\left(\underline{X}'\underline{X}\right)^{-1}\underline{X}'\right]Y} \sim F_{k, T_1+T_2-2k}, \tag{1.10}$$

where $\widehat{\beta}_1 = \left(X_1'X_1\right)^{-1}X_1'Y_1$ and $\widehat{\beta}_2 = \left(X_2'X_2\right)^{-1}X_2'Y_2$.

Alternative, the same result can be derived as follows: Define the sum of squares of the residuals under the alternative of structural change,

$$S_T\left(\widehat{\beta}\right) = Y'\left[I - \underline{X}\left(\underline{X}'\underline{X}\right)^{-1}\underline{X}'\right]Y$$

and the sum of squares of the residuals under the null hypothesis

$$S_T\left(\overline{\beta}\right) = Y'\left[I - X\left(X'X\right)^{-1}X'\right]Y.$$

It is easy to show that

$$\frac{T_1 + T_2 - 2k}{k}\frac{S_T\left(\overline{\beta}\right) - S_T\left(\widehat{\beta}\right)}{S_T\left(\widehat{\beta}\right)} \sim F_{k,T_1+T_2-2k}. \tag{1.11}$$

In this case an unbiased estimate of σ^2 is

$$\widetilde{\sigma}^2 = \frac{S_T\left(\overline{\beta}\right)}{T_1 + T_2 - 2k}.$$

Before we remove the assumption that $\sigma_1 = \sigma_2$ we will first derive a test of the equality of the variances. Under the null hypothesis (same variances across regimes) we have

$$\frac{\widehat{u}_i'\widehat{u}_i}{\sigma^2} \sim \chi^2_{T_i-k} \quad \text{for } i = 1, 2.$$

Because these chi-square variables are independent, we have

$$\frac{T_2 - k}{T_1 - k}\frac{\widehat{u}_1'\widehat{u}_1}{\widehat{u}_2'\widehat{u}_2} \sim F_{T_1-k,T_2-k}.$$

Unlike the previous tests, a two-tailed test should be used here, because a large or small value of the test is a reason to reject the null hypothesis.

If we remove the assumption of equal variances among regimes and focus on the hypothesis of equality of the regression parameters, the tests are more involved. We will concentrate on the case in which $k = 1$, where a t test is applicable. It can be shown (though this is not trivial) that

$$t_T = \frac{\widehat{\beta}_1 - \widehat{\beta}_2}{\sqrt{\frac{\widetilde{\sigma}_1^2}{X_1'X_1} + \frac{\widetilde{\sigma}_2^2}{X_2'X_2}}} \sim S_v,$$

where

$$v = \frac{\left[\frac{\widetilde{\sigma}_1^2}{X_1'X_1} + \frac{\widetilde{\sigma}_2^2}{X_2'X_2}\right]^2}{\frac{\widetilde{\sigma}_1^4}{(T_1-1)\left(X_1'X_1\right)^2} + \frac{\widetilde{\sigma}_2^4}{(T_2-1)\left(X_2'X_2\right)^2}}.$$

A cleaner way to perform this type of tests is through the use of direct Likelihood Ratio Tests (which we will discuss in depth later).

Even though structural change (or Chow) tests are popular, modern econometric practice is skeptic with respect to the way in which they are described above, particularly because in these cases the econometrician sets in an *ad-hoc* manner the point at which to split the sample. Recent theoretical and empirical applications are working on treating the period of possible break as an endogenous latent variable.

1.8 Further Reading

This chapter's material is developed in several standard texts. Among graduate-level references, Amemiya (1985), Hayashi (2000), and Ruud (2000) are particularly rigorous. Hansen (2022) offers a clear and conceptually well-structured alternative, while Greene (2012) and Baltagi (1999) provide comprehensive treatments with broad applied coverage. Mittelhammer, Judge, and Miller (2000) frame the regression model within a general statistical decision-theoretic approach.

For more introductory treatments, Johnston and DiNardo (1997) and Pindyck and Rubinfeld (1997) offer accessible presentations with an emphasis on applications. Hill, Griffiths, and Lim (2011) is a well-structured and widely used entry point. Ramanathan (1993) connects statistical methods and econometric applications with a focus on implementation. Stock and Watson (2015) provides a clean and modern introduction to regression, balancing intuition with formal development.

On the mathematical side, Dhrymes (2000) develops the essential linear algebra and probability tools required for formal econometrics. Harville (1997) is a definitive reference on matrix algebra, written from a statistician's point of view. Mittelhammer (1996) introduces mathematical statistics with applications to economic modeling and inference.

Yule (1926) is a foundational paper on spurious correlation, notable for both its historical importance and enduring relevance. Vigen (2015) complements this lesson with real data and humorous pairings that illustrate how easily one can find statistically strong but substantively meaningless associations.

1.9 Workout Problems

1. Prove that independence implies no correlation but that the contrary is not necessarily true. Give an example of variables that are uncorrelated but not independent.

2. Let y, x be scalar dichotomous random variables with zero means. Define $u = y - \mathrm{Cov}(y, x)\left[\mathcal{V}(x)\right]^{-1}$. Prove that $\mathcal{E}(u\,|x) = 0$. Are u and x independent?

3. Let y be a scalar random variable and x a vector random variable. Prove that $\mathcal{E}\left[y - \mathcal{E}(y\,|x)\right]^2 \leq \mathcal{E}\left[y - w(x)\right]^2$ for any function w.

4. Prove that if $\mathcal{V}(u_t) = \sigma^2$, $\mathcal{V}(\widehat{u}_t) = (1 - h_t)\,\sigma^2$. Find an expression for h_t.

5. Prove Proposition 6.

6. Prove Proposition 11.

7. In Theorem 13 we used the fact that $(A + C)(A + C)' = (X'X)^{-1} + CC'$. Prove this.

8. Prove that when a constant is included, $R^2 = 1 - (Y'MY/Y'LY)$, with L being as defined in section 1.5.5.

9. Derive the variance-covariance matrix of $\widehat{\beta}_2$ defined in section 1.5.6.

10. Prove Theorem 14.

11. Prove that (1.5) is the CLS estimator.

12. Prove that the CLS estimator can be expressed as $\overline{\beta} = \beta + R\left(R'X'XR\right)^{-1} R'X'u$ and obtain $\mathcal{V}\left(\overline{\beta}\right)$.

13. Show that $(\widehat{u}'\widehat{u})\,\sigma^{-2} \sim \chi^2_{T-k}$.

14. Demonstrate (1.8).

15. Derive equations (1.9), (1.10), and (1.11).

16. Prove that to test the null $\mathrm{H}_0 : \beta_i = 0$ for all i except the constant, the F test is equivalent to $(T-k)\,R^2/\left[\left(1-R^2\right)(k-1)\right]$.

Chapter 2

Simulation and Resampling

2.1 Introduction

Most of the methods for estimation and hypothesis testing used in econometrics have statistical properties that are known only asymptotically. This is true for nonlinear models of all types, for linear simultaneous equations models, and even for OLS once we dispense the strong assumption of fixed regressors or that the error terms are normal and independent and identically distributed (i.i.d. for short). In practice, "exact" finite sample theory can rarely be used to interpret estimates or test statistics. Often we want to know whether the theoretical properties of our models are reasonably good approximations for the problems at hand.

The purpose of this document is to introduce some procedures that can be used to conduct numerical simulation exercises and is organized as follows: Section 2.2 presents a brief review of the basic methods that may be used to generate numbers that behave as outcomes from some stochastic process. Section 2.3 discusses the fundamental concepts underlying Monte Carlo experiments. Finally, Section 2.4 describes a data-based simulation method known as the bootstrap, which is a simulation procedure that does not require explicit assumptions about the true probability model and may provide more refined approximations to the sampling distribution than first-order asymptotic results.

2.2 Pseudo-Random Number Generation

Computer programs designed to simulate outcomes of a stochastic process are based on numerical rules or algorithms. Given some starting value, the algorithm generates a sequence of numbers $\{u_i\}_{i=1}^N$ that behave as if drawn from a particular distribution $\mathcal{F}$. If we know the starting value and the algorithm used to generate a set of numbers, we can actually replicate the sequence of outcomes. Thus, computer-generated outcomes are actually deterministic and not truly random. For this reason, numerically generated outcomes are said to be pseudo-random numbers.

Computing packages (such as GAUSS) provide commands and procedures that generate pseudo-random draws from many common parametric families. It is important to understand the steps involved in generating them because one may encounter nonstandard simulation problems for which there are no preprogrammed numerical tools. In addition, some pitfalls in the application of pseudo-random number generators can be avoided by understanding their structure.

2.2.1 Generating $\mathcal{U}(0,1)$ Pseudo-Random Numbers

Although probability models of economic behavior are rarely based on uniform distributions, the $\mathcal{U}(0,1)$ distribution serves as an important basis for simulating the outcomes of other nonuniform random variables. To demonstrate the basic mechanics of generating $\mathcal{U}(0,1)$ pseudo-random numbers, we focus on a particular algorithm known as the linear congruential generator.

In simple terms, the linear congruential rule generates a set of N values $\{u_i\}_{i=1}^N$ in the $(0,1)$ interval by forming the ratio of two integers I_i/n and reporting the fractional reminder u_i for $i = 1, \cdots, N$. The integer in the denominator (n) is a fixed value known as the modulus. To form a sequence of pseudo-random numbers, the integer in the numerator (I_i) changes for each number of the sequence. The numerator sequence begins from a starting value I_0, which is known as the seed. Subsequent integers are generated by the linear progression $I_i = \alpha + \beta I_{i-1}$ for $i = 1, \cdots, N$ and fixed integers α and β. Thus, a different sequence of numbers can be generated by changing the seed (I_0) or one of the other constants $(\alpha, \beta$ or $n)$.

Owing to the finite number of integers that may be represented on a modern 32-bit computer, the linear congruential algorithm will eventually repeat the sequence for adequately large N. The period of an algorithm is an integer s such that $u_i = u_{i+s}$ for each i. This characteristic is a limitation of the computer hardware and not the software, and thus virtually all numerical pseudo-random number generators have a finite period. Well-written algorithms will have a period that is long enough to avoid the problem of repeating sequences in typical applications. However, in order to avoid repetition of the generated numbers, a new linear congruential rule should be used if the number of draws exceeds n.[1]

2.2.2 Generating Continuous Pseudo-Random Numbers

Given a reliable generator of $\mathcal{U}(0,1)$ pseudo-random variables, there are several ways in which one can generate pseudo-random variates that appear to be drawn from any desired distribution. We will briefly discuss two general techniques: The transformation method and the rejection method, as well as special methods applicable to certain cases of interest.

Transformation Method

This method (also known as the probability integral method) is based on the fact that the range of a cumulative distribution function (c.d.f.) is the 0-1 interval. Thus, if w is distributed according to a strictly increasing c.d.f. $\mathcal{F}(w)$, $u = \mathcal{F}(w)$ must be distributed as $\mathcal{U}(0,1)$. For any u, we can invert the c.d.f. so as to obtain $w = \mathcal{F}^{-1}(u)$. To obtain a sequence of w_i's distributed according to $\mathcal{F}(w)$, we simply generate a sequence of u_i's distributed as $\mathcal{U}(0,1)$ and subject each of them to the transformation $\mathcal{F}^{-1}(u)$. This is illustrated in Figure 2.1 which shows that any value of u on the vertical axis, is mapped uniquely via $\mathcal{F}^{-1}(u)$ into a corresponding value w on the horizontal axis.

This method works well when $\mathcal{F}^{-1}(u)$ is inexpensive to compute. One of such cases is that of the exponential distribution, for which the probability

[1] GAUSS tip: The GAUSS procedure `rndu(d,e)` generates a `d`×`e` matrix of pseudo-random outcomes from the $\mathcal{U}(0,1)$ distribution and is based on the linear congruential rule. The parameters α, β and n can be set with the `rndcon`(), `rndmult`(), and `rndmod`() procedures. The initial seed is generated by the system clock when GAUSS is invoked. Thereafter, the seed is updated each time `rndu` is called. The seed can also be set by the user with the `rndseed`() procedure or the `rndus`() procedure. Under the default options, GAUSS uses the largest possible modulus for a 32-bit computer ($n = 2^{31} - 1 = 2,147,483,647$), and the period s equals n.

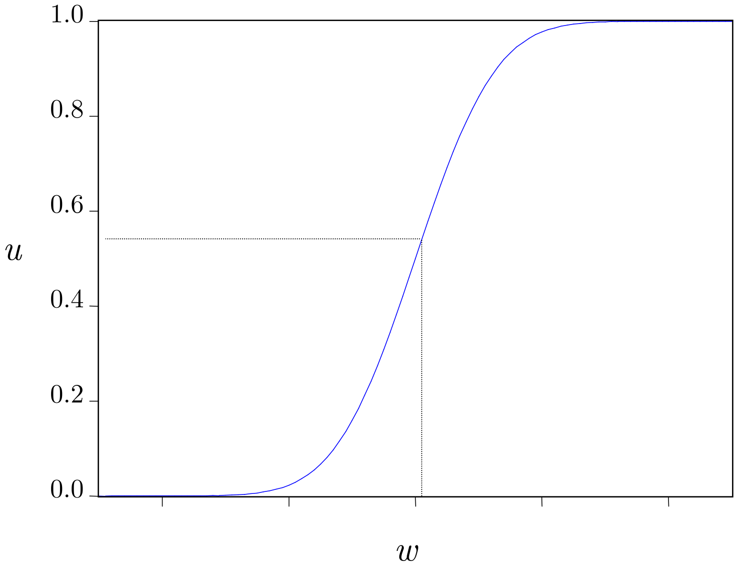

Figure 2.1: Transformation Method

density function (p.d.f.) is

$$f\left(w\right)=\theta e^{-\theta w}$$

and the corresponding c.d.f. is

$$\begin{aligned}\mathcal{F}\left(w\right) &= \int_0^w f\left(s\right)ds\\ &= \int_0^w \theta e^{-\theta s}ds\\ &= -e^{-\theta s}\Big|_0^w\\ &= 1-e^{-\theta w}.\end{aligned}$$

Setting u equal to $\mathcal{F}\left(w\right)$ and solving, we find

$$w=\mathcal{F}^{-1}\left(u\right)=-\frac{1}{\theta}\ln\left(1-u\right).$$

The transformation method can also be used to generate pseudo-random normal variates, but it requires a certain amount of computation because there is no closed form expression for the standard normal c.d.f. $\Phi(\cdot)$ or its inverse $\Phi^{-1}(\cdot)$.[2] An alternative technique that is widely used is the Box-Muller method. It uses the fact that if u_1 and u_2 are independent random variates from $\mathcal{U}(0,1)$, then the variates

$$\begin{aligned} w_1 &= (-2\ln(u_1))^{1/2}\cos(2\pi u_2) \\ w_2 &= (-2\ln(u_1))^{1/2}\sin(2\pi u_2) \end{aligned}$$

are independent random variates from $\mathcal{N}(0,1)$. The major problem with this technique is that it relies heavily on the independence of u_1 and u_2. If the random number generator that is used to generate them is not a good one, they may exhibit some dependence, and the resulting variates w_1 and w_2 may be neither normal nor independently distributed.

Rejection Method

Unlike the transformation method, the rejection method can be used when the p.d.f. $f(w)$ is known.

In its simplest version, the rejection method requires that the domain of $f(w)$ be a finite interval on the real line, say the interval $[a,b]$. One starts by obtaining two random variates from $\mathcal{U}(0,1)$, say u_1 and u_2. The first of these is transformed into v_1, a random variate from $\mathcal{U}(a,b)$, while the second is transformed into v_2, a random variate from $\mathcal{U}(0,h)$, where h is a number at least as large as the maximum of $f(w)$.

Once v_1 and v_2 have been obtained, v_2 is compared with $f(v_1)$. If v_2 exceeds $f(v_1)$, the proposed random variate v_1 is rejected and another pair (v_1, v_2) is drawn. If v_2 is less than or equal to $f(v_1)$, v_1 is accepted and w is set equal to it. This method is illustrated in Figure 2.2 where the point B yields a value of w and the point A is rejected.

It is easy to see why the rejection method works. Although we pick v_1 initially from $\mathcal{U}(a,b)$, we accept only if $v_2 < f(v_1)$, and the probability that this will happen is proportional to $f(v_1)$. This version of the rejection method

[2] GAUSS tip: The command `rndn(a,b)` generates an $\mathtt{a} \times \mathtt{b}$ matrix of pseudo-random $\mathcal{N}(0,1)$ numbers. Although GAUSS does not use the transformation method to generate them, it is easy to devise an algorithm to do so. Let u be a pseudo-random realization of $\mathcal{U}(0,1)$ (generated for example using the command `rndu`). The command `cdfni` computes the inverse of a standard normal (of course, this inverse is numerically evaluated by GAUSS).

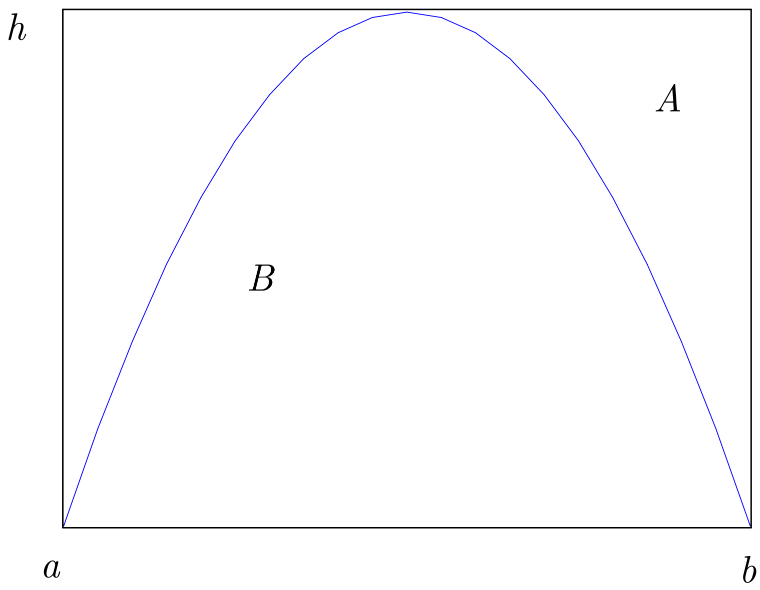

Figure 2.2: Rejection Method

is somewhat inefficient, since we have to generate, on average, $2h\,(b-a)$ random variates for every w that we actually obtain. If the density $f\,(w)$ has a tall spike, h will be large. If it has long tails, $b-a$ will be large. In either of these cases, $2h\,(b-a)$ will be large, and the method may be quite inefficient.

In a more general version of this method, that no longer requires a closed domain for w, the constant h is replaced by a function $h\,(v_1)$, with v_1 then drawn from a density proportional to $h\,(v_1)$. Then v_2 is chosen to be $\mathcal{U}\,(0, h\,(v_1))$ provided that $h\,(v_1) > f\,(v_1)$ for every v_1 on the domain of w. If it is easy to draw pseudo-random realizations of $h\,(\cdot)$, and the area under it is not too much larger that the area under $f\,(\cdot)$, this variant will work efficiently. Note that $h\,(\cdot)$ is not itself a density, since $h\,(v)$ must be larger than $f\,(v)$ for all v and hence must integrate to more than unity; in practice, it may be convenient to make $h\,(\cdot)$ proportional to some well-known density.[3]

[3] GAUSS uses a variant of this method to generate pseudo-random realizations of $\mathcal{N}\,(0, 1)$.

Another rejection method to generate pseudo-random outcomes of a standard normal (known as the Marsaglia-Bray method) is as follows: Generate v_1 and v_2 from $\mathcal{U}(-1,1)$. Define $d = v_1^2 + v_2^2$; if $d \geq 1$, reject v_1 and another pair (v_1, v_2) is drawn. If $d < 1$, define $w_i = v_i\sqrt{-2\ln(d)/d}$ $(i = 1, 2)$. The values w_1 and w_2 represent a pair of i.i.d. $\mathcal{N}(0,1)$ outcomes.

2.2.3 Generating Discrete Pseudo-Random Numbers

To simulate pseudo-random draws for a discrete random variable with distribution function $\mathcal{F}$ supported on a countable set, we can extend the idea underlying the transformation method for a c.d.f. to a step function. Suppose we have a discrete random variable X supported on the ordered set $\{x_1, x_2, \cdots\}$ with associated probabilities $p = \{p_1, p_2, \cdots\}$ such that $\sum_{i=1}^{\infty} p_i = 1$. The right continuous distribution function $\mathcal{F}$ for X is given by the step function

$$\mathcal{F}(x) = \sum_{i=1}^{\infty} p_i I(x_i \leq x),$$

where $I(x_i \leq x)$ is an indicator function that equals 1 if $x_i \leq x$ and 0 otherwise. As in the continuous case, we can apply the "inverse step function" to $\mathcal{U}(0,1)$ outcomes to generate pseudo-random outcomes w with the distribution $\mathcal{F}$.

More formally, let y be an outcome from the $\mathcal{U}(0,1)$ distribution. We can generate $w = x_1$ if $y \leq \mathcal{F}(x_1)$, and in general

$$w = x_i \quad \text{if} \quad \mathcal{F}(x_{i-1}) < y \leq \mathcal{F}(x_i) \quad \text{for } i \geq 2.$$

2.2.4 Performance of Pseudo-Random Number Generators

Good commercial pseudo-random number generators generally produce sequences that behave as if drawn from the target distribution and that, for practical purposes, do not repeat themselves on common applications. However, owing to the deterministic nature of the pseudo-random number generation, some algorithms may not provide sequences that mimic independent or identically distributed outcomes. If the results of a simulation exercise are especially sensitive to the properties of the pseudo-random numbers, or if there is doubt about the quality of an algorithm, it would be wise to evaluate the performance of the software.

To check the properties of a particular algorithm, researchers have devised a large number of diagnostic techniques, and many of them are applied from the literature of nonparametric statistics. For example, one of the most commonly used diagnostic tools is the Pearson's χ^2 test

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - c_i)^2}{c_i}.$$

The sample space for the target distribution is divided in k bins, o_i is the fraction of observed outcomes in the ith bin, and c_i is the fraction of expected outcomes in the ith bin under the target distribution. For example, to evaluate a $\mathcal{U}(0,1)$ pseudo-random number generator, we could divide the unit interval in k=10 bins so that $c_i = 1/10$ for a sample of N pseudo-random numbers and $i = 1, \cdots, 10$. As $N \to \infty$, the statistic is asymptotically χ^2_{k-1} under the null hypothesis that the sample was generated by the target distribution. Alternative measures to evaluate the performance of generators include the Kolmogorov-Smirnov and the Shapiro-Wilks (for normality) tests.

For the Kolmogorov-Smirnov test, define $O_i = \sum_{j=1}^{i} o_j$ and $C_i = \sum_{j=1}^{i} c_j$, then

$$KS = \max\left\{\max_i |O_i - C_i|, \max_i |O_{i-1} - C_i|\right\},$$

where O_0 is set equal to 0. This test statistic evaluates if at any point along the two cumulative distribution functions, the greatest distance is equal or greater than a critical value.[4]

2.3 Monte Carlo Experiments

Monte Carlo experiments are often used to analyze the finite sample properties of estimators or test statistics. The method is defined as a set of procedures in which quantities of interest are approximated by generating many pseudo-random realizations of some stochastic process and averaging them in some way. The method is computer intensive and is routinely used in several disciplines.[5]

[4] The critical values are reported in any statistics book. Sheskin (2002) presents this and other tests.

[5] The term "Monte Carlo" was coined in 1949 and was originally devised to evaluate integral functions that do not have tractable solutions. If it had been used at a later date, it might have been called the "Las Vegas Method".

In econometrics, a Monte Carlo experiment is typically comprised of several elements that the experimenter must specify:

a) A model and a set of estimators or tests associated with the model. The objective of the experiment is to assess their small sample properties.

b) A Data Generating Process (DGP). The DGP is usually a special case of the model and must specify the "true" values of the parameters, the laws of motion of the exogenous and endogenous variables, and the distributions of all the random variables involved.

c) Each experiment consists of a number of replications or samples (denoted by J). Each replication involves generating artificial samples of the data according to the DGP and calculating the estimates or test statistics of interest. Typically J must be large.

d) After J replications are performed, we have an equal number of estimators which are subjected to statistical analysis.

e) Many times, several experiments are performed by changing the sample size of each experiment, values of key parameters, etc. Results can be reported using response surfaces which we describe below.

The most important thing to recognize is that results from Monte Carlo experiments are necessarily random. At the very least, this means that results must be reported in a way which allows readers to appreciate the extent of experimental randomness. Moreover, it is essential to perform enough replications so that the results are sufficiently accurate for the purpose at hand.

Monte Carlo experiments are often used to obtain critical values for test statistics whose asymptotic distribution may or may not be known. An example may help us to clarify the steps that are involved in a Monte Carlo experiment. Assume that we want to evaluate the finite sample properties of the estimator of α for the following model:

$$y_i = \alpha + u_i; \qquad u_i \sim \mathcal{U}\left(-2, 2\right),$$

for a sample size $N = 30$. Assume further that we set $\alpha = 0.5$.

We generate an artificial sample j of errors $\left\{\widetilde{u}_i^j\right\}_{i=1}^N$ with which be obtain an artificial sample for y: $\widetilde{y}_i^j = 0.5 + \widetilde{u}_i^j$. Once we generated each sample j, we obtain an estimate of α that we denote by $\widetilde{\alpha}^j$ $(j = 1, \cdots, J)$. With these results we can obtain an approximation for the bias of the estimator:

$$\text{Bias} \simeq \frac{1}{J} \sum_{j=1}^{J} \left(\widetilde{\alpha}^j - 0.5 \right).$$

If $\widetilde{\alpha}^j = \frac{1}{N} \sum_{i=1}^{N} \widetilde{y}_i^j$, we can also compute an estimator of its variance $\widehat{\mathcal{V}} \left(\widetilde{\alpha}^j \right) = \frac{1}{N} \sum_{i=1}^{N} \left(\widetilde{y}_i^j - \widetilde{\alpha}^j \right)^2$ and the t statistic

$$\widetilde{t^j} = \frac{\widetilde{\alpha}^j - 0.5}{\sqrt{\widehat{\mathcal{V}} \left(\widetilde{\alpha}^j \right)}}.$$

If we wanted to obtain the empirical critical value for a test of $(1 - \delta)\,\%$ coverage, we simply sort $\left| \widetilde{t^j} \right|$ and pick the value corresponding to the $(1 - \delta)\, J$ observation of this new vector.[6] This is a very simple way to obtain critical values and may be useful when it is suspected that the asymptotic distribution is not a good approximation for the sample size or DGP at hand.[7]

2.3.1 Variance Reduction

Obtaining sufficiently accurate results from a Monte Carlo experiment can often require that a great many replications be computed. This is not always feasible. In some cases, the number of replications that is needed can be substantially reduced by using techniques for reducing the variance of experimental results. One of them uses antithetic variates.

The idea of antithetic variates is to calculate two different estimates of the quantity of interest in such a way that the two estimates are negatively correlated. Their average will then be substantially more accurate than either of them individually. Suppose that we wish to estimate some quantity θ, and that from a single Monte Carlo experiment we can obtain two different unbiased estimators of θ, say $\widehat{\theta}$ and $\widetilde{\theta}$. If these are antithetic variates, the pooled estimator

$$\overline{\theta} = \frac{1}{2} \left(\widehat{\theta} + \widetilde{\theta} \right),$$

[6] This may not be the best way to compute confidence intervals. See Section 2.4.6 below.

[7] GAUSS tip: Let `z` be an $a \times b$ matrix, the command `sortc(z,d)` generates a matrix that sorts `z` (in ascending order), taking column d as the basis.

has variance

$$\mathcal{V}\left(\overline{\theta}\right)=\frac{1}{4}\left(\mathcal{V}\left(\widehat{\theta}\right)+\mathcal{V}\left(\widetilde{\theta}\right)+2\mathrm{Cov}\left(\widehat{\theta},\widetilde{\theta}\right)\right),$$

where $\mathcal{V}\left(\widehat{\theta}\right)$ and $\mathcal{V}\left(\widetilde{\theta}\right)$ denote the variances of $\widehat{\theta}$ and $\widetilde{\theta}$ respectively. If $\mathrm{Cov}\left(\widehat{\theta},\widetilde{\theta}\right)$ is negative, $\mathcal{V}\left(\overline{\theta}\right)$ will be smaller than $\frac{1}{4}\left(\mathcal{V}\left(\widehat{\theta}\right)+\mathcal{V}\left(\widetilde{\theta}\right)\right)$, which is the variance that we would have obtained using the same number of replications to estimate θ from two independent experiments. Thus, the extent to which we can gain by using antithetic variates, depends on how strong the negative correlation between $\widehat{\theta}$ and $\widetilde{\theta}$ is.

One might ask why $\widehat{\theta}$ and $\widetilde{\theta}$ should receive equal weight in computing $\overline{\theta}$. Let us therefore consider the weighted estimator

$$\ddot{\theta}=w\widehat{\theta}+(1-w)\,\widetilde{\theta}.$$

Differentiating the variance of $\ddot{\theta}$ with respect to w and setting the result to zero, we find that

$$w=\frac{\mathcal{V}\left(\widehat{\theta}\right)-\mathrm{Cov}\left(\widehat{\theta},\widetilde{\theta}\right)}{\mathcal{V}\left(\widehat{\theta}\right)+\mathcal{V}\left(\widetilde{\theta}\right)-2\mathrm{Cov}\left(\widehat{\theta},\widetilde{\theta}\right)},$$

which is satisfied by setting $w=\frac{1}{2}$ whenever $\mathcal{V}\left(\widehat{\theta}\right)=\mathcal{V}\left(\widetilde{\theta}\right)$. In most cases, the variances of the two estimators will be equal, in which case it will be optimal to give the two of them equal weights.

In several applications (such as the regression model when u has a symmetric distribution) we can use each set of generated samples twice, with the sign reversed in the second time. Suppose for example that we wish to estimate the mean (α) of an i.i.d. process (y) with symmetric distribution. For each set of pseudo-random realizations $\left\{\widehat{y}_i^j\right\}_{i=1}^N$, we can generate another set $\left\{\widetilde{y}_i^j\right\}_{i=1}^N$, where $\widetilde{y}_i^j=-\widehat{y}_i^j$. Once we obtain an estimator of α with each artificial sample we can construct the estimator

$$\overline{\alpha}=\frac{1}{2J}\sum_{j=1}^{J}\left(\widehat{\alpha}_j+\widetilde{\alpha}_j\right),$$

in which case the variance is fully reduced.[8]

Perfect negative correlation of the antithetic variables will not occur in general. When it does, the problem is usually so simple that there is no need to perform Monte Carlo experiments. Less than perfect negative correlation often does occur, however, and it means that in certain cases the use of antithetic variates can greatly reduce the number of replications that are needed to estimate the first moments of an estimator.

2.3.2 Response Surfaces

One of the most difficult aspects of any Monte Carlo experiment is presenting the results in a fashion that makes them easy to comprehend. An approach that is often useful is to estimate response surfaces. This is simply a regression model in which each observation corresponds to one experiment, the dependent variable is some quantity that was estimated in the experiments, and the independent variables are functions of the various parameter values, which characterize the experiment.

This approach is particularly useful if a response surface adequately explains the experimental results, given that it may be easier to understand the behavior of the estimator of interest from parameters of a response surface than from several tables full of numbers. Furthermore, response surfaces may greatly reduce the problem of specificity of the experiment. What this means is that each experiment may arise from a different data generating process, in which case response surfaces may show how sensitive are the finite sample properties of the estimator of interest to several features of the experiment such as sample size, true values of the parameters, method of estimation, etc.

2.3.3 Examples

As discussed above, one of the main purposes for conducting Monte Carlo experiments is to evaluate whether or not asymptotic distribution theory is a useful method for approximating the finite sample properties of estimators and test statistics. As shown below, sometimes these approximations can be quite poor.

[8] This will happen everytime we consider linear functions of y, given that $\text{Cov}\left(\widehat{\theta}, \widetilde{\theta}\right) = -\mathcal{V}\left(\widehat{\theta}\right)$.

Consider the linear Gaussian regression

$$\begin{aligned} y_t &= \alpha + \beta x_{1,t} + \gamma x_{2,t} + u_t \\ \begin{bmatrix} x_{1,t} \\ x_{2,t} \end{bmatrix} &\sim \mathcal{N}\left(0, I_2\right) \\ u_t &\sim \mathcal{N}\left(0, \sigma^2\right). \end{aligned}$$

We set $\sigma = 3, \alpha = 0, \beta = 1, \gamma = 0.5$, and $T = 300$.

Consider two exercises: The first, in which the goal is to estimate β and test the hypothesis $\mathrm{H}_0 : \beta = 1$. The second has as parameter of interest the ratio of the regression slopes

$$\theta = \frac{\beta}{\gamma},$$

and, as in the other case, the goal is to estimate θ and test the hypothesis $\mathrm{H}_0 : \theta = 2$.

We estimate the parameters of the model by OLS, obtain $\widehat{\alpha}, \widehat{\beta}, \widehat{\gamma}$, and $\widehat{\theta} = \widehat{\beta}/\widehat{\gamma}$.

The asymptotic approximation to the distribution of $\widehat{\beta}$ is $\mathcal{N}\left(\beta, A\mathcal{V}\left(\widehat{\beta}\right)\right)$, where

$$\begin{aligned} A\mathcal{V}\left(\widehat{\beta}\right) &= T^{-1}\mathcal{E}u_t^2\mathcal{E}\left(x_{1,t}^2\right)^{-1} \\ &= \frac{1}{300} \cdot 3^2 \cdot 1 \\ &= \frac{3}{100} = 0.03. \end{aligned}$$

and $A\mathcal{V}\left(\cdot\right)$ denotes the asymptotic variance.

On the other hand, if we use a first-order asymptotic approximation and calculate the standard error for $\widehat{\theta}$ using the "delta method" we have

$$\widehat{\mathcal{V}}\left(\widehat{\theta}\right) = \widetilde{\sigma}^2 \widehat{H}'(X'X)^{-1}\,\widehat{H}, \qquad \widehat{H} = \begin{bmatrix} 0 \\ 1/\widehat{\gamma} \\ -\widehat{\beta}/\widehat{\gamma}^2 \end{bmatrix}.$$

As in the previous case, the asymptotic approximation to the distribution of $\widehat{\theta}$ is $\mathcal{N}\left(\theta, A\mathcal{V}\left(\widehat{\theta}\right)\right)$, where

$$
\begin{aligned}
A\mathcal{V}\left(\widehat{\theta}\right) &= T^{-1}\left[H'(\mathcal{E}x_i x_i')^{-1}H\right]\mathcal{E}u_t^2 \\
&= \frac{1}{300}\cdot\left[\begin{pmatrix} 0 & 2 & -4 \end{pmatrix}\begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} 0 \\ 2 \\ -4 \end{pmatrix}\right]\cdot 3^2 \\
&= \frac{1}{300}\cdot 20\cdot 3^2 = \frac{3}{5} = 0.6.
\end{aligned}
$$

Thus, the asymptotic approximations are $\widehat{\beta}\overset{a}{\sim}\mathcal{N}(1,0.03)$ and $\widehat{\theta}\overset{a}{\sim}\mathcal{N}(2,0.6)$. We can calculate the exact distribution of both estimators by simulation methods. 20,000 independent samples were generated from the above process, and on each sample, $\widehat{\beta}$ and $\widehat{\theta}$ were calculated.

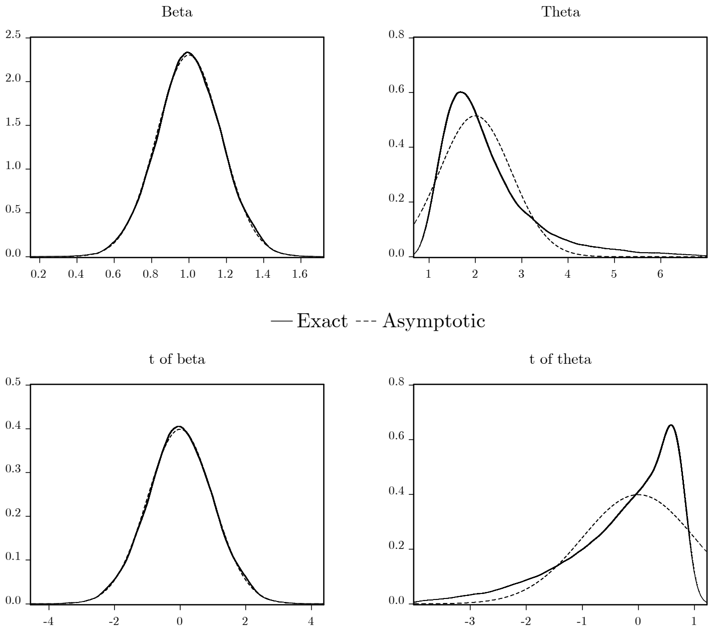

Figure 2.3: Exact and Asymptotic Distributions

The densities of both estimators were estimated non-parametrically (using a kernel density estimator). The exact densities (continuous lines), along with the asymptotic distributions (dashed lines) are displayed in the top panels of Figure 2.3. While the exact distribution mimics the asymptotic approximation for β, the divergence between them is quite dramatic for θ given that the exact density is skewed and thick-tailed. Numerically, we can calculate $\mathcal{E}\left(\widehat{\beta}\right) = 1.000, \mathcal{V}\left(\widehat{\beta}\right) = 0.030, \mathcal{E}\left(\widehat{\theta}\right) = 2.239$, and $\mathcal{V}\left(\widehat{\theta}\right) = 0.980$; which in the latter case are quite different from the values implied by the asymptotic approximation.

Inference based on t-ratios can be constructed for each artificial sample as $t_1 = \left(\widehat{\beta} - 1\right) / \left(\widehat{\mathcal{V}}\left(\widehat{\beta}\right)\right)^{1/2}$ and $t_2 = \left(\widehat{\theta} - 2\right) / \left(\widehat{\mathcal{V}}\left(\widehat{\theta}\right)\right)^{1/2}$. The asymptotic approximation for both tests is $\mathcal{N}(0,1)$. Using the simulated samples, we estimated the exact densities for t_1 and t_2, which are displayed along the asymptotic distributions in the bottom panels of Figure 2.3. Once again, we find that while the exact and asymptotic distributions of t_1 match, the divergence between them for the case of t_2 is dramatic. The exact distribution is highly skewed and non-normal.

The accuracy of hypothesis testing and confidence interval coverage depends on tail probabilities. In the case of β we calculate the exact probabilities $\Pr[t_1 > 1.645] = 0.051$, $\Pr[t_1 < -1.645] = 0.051$ and $\Pr[|t_1| > 1.96] = 0.053$, meaning that both one-tailed and two-tailed tests, and confidence intervals based on asymptotic approximations are accurate. On the other hand, these probabilities are $\Pr[t_2 > 1.645] = 0.000$, $\Pr[t_2 < -1.645] = 0.115$ and $\Pr[|t_2| > 1.96] = 0.084$ for the case of θ; implying that both one-tailed and two-tailed tests, and confidence intervals based on asymptotic approximations have significant Type I errors, given that the nominal Type I error for each of these tests is 0.05.[9]

The latter example shows that asymptotic approximations may be quite poor, even in very simple regression models with reasonably large samples.

2.3.4 Monte Carlo Standard Errors (MCSE)

In any Monte Carlo experiment, the statistics computed from the simulated samples are themselves subject to simulation noise, because they are based on a finite number of replications. The Monte Carlo Standard Error (MCSE)

[9] To compute the exact probability $\Pr[z > c]$ we simple obtain $\frac{1}{J}\sum_{i=1}^{J} I\left(z^j > c\right)$, where $I(\cdot)$ is an indicator function. GAUSS tip: Let `z` be a vector and `c` a constant, the exact probability can be computed using the command `meanc(z.>c)`.

quantifies the variability of these simulation-based estimates.

Suppose you compute an average from J replications:

$$\overline{\theta}_J = \frac{1}{J}\sum_{j=1}^{J}\widehat{\theta}_j,$$

where $\widehat{\theta}_j$ is the statistic of interest (e.g., an estimator, t-ratio, test size) in replication j. The MCSE of $\overline{\theta}_J$ is the standard error of the mean of the simulated quantity:

$$MCSE\left(\overline{\theta}_J\right) = \sqrt{\frac{1}{J\left(J-1\right)}\sum_{j=1}^{J}\left(\widehat{\theta}_j - \overline{\theta}_J\right)^2}.$$

This quantity reflects the precision of the simulation-based estimate. A small MCSE indicates that the simulation results are stable across replications. In contrast, a large MCSE suggests that the number of replications is insufficient, and that the reported averages may be unreliable.

To illustrate, consider estimating the empirical size of a nominal 5% test. Suppose the Monte Carlo experiment yields a rejection frequency of 0.048 based on 10,000 replications, with an MCSE of 0.002. This means that, due to simulation variability alone, the rejection frequency could plausibly differ from 0.048 by about ±0.002. In this case, the estimated size is close to the nominal level, and the precision is satisfactory.

In practice, the number of replications J should be chosen so that the MCSE is sufficiently small relative to the scale of the quantity being estimated. While 1,000 replications may suffice for many pedagogical exercises, more demanding applications may require 10,000 or more.

2.4 Bootstrap Resampling

The bootstrap was introduced by Efron (1979) as a computational alternative to analytic approximations of sampling distributions. While asymptotic theory provides elegant results under regularity conditions, these approximations may perform poorly in small samples or in models with nonstandard features, such as heteroskedasticity or nonlinearity. The bootstrap often provides more accurate inference by exploiting the actual structure of the data, reducing bias and better approximating the shape of the sampling distribution. This comes at the cost of computation, but not of stronger assumptions.

The bootstrap approach views the observed sample as a population. The distribution function for this population is the Empirical Distribution Function (EDF) of the sample, and the parameter estimates based on the observed sample are treated as the actual model parameters. Conceptually, we can examine properties of the estimators or test statistics in repeated samples drawn from a tangible data-sampling process that mimics the actual DGP. Although simulation results based on the bootstrap do not represent the exact finite sample properties of estimators and test statistics under the actual DGP, the bootstrap provides an approximation that improves as the size of the observed sample increases.

In recent years, the bootstrap procedure has gained acceptance in applied research for several reasons. First, it avoids most of the strong distributional assumptions required in traditional Monte Carlo simulation exercises. As such, researchers can simulate the DGP without having to fully specify a parametric probability model. Second, like Monte Carlo methods, the bootstrap procedure may be used to solve intractable estimation and inference problems by computation rather than reliance on asymptotic approximations, which may be very complicated in nonstandard problems. Third, bootstrap approximations are often equivalent to first-order asymptotic results in large samples, and may dominate them in many cases.

2.4.1 Notation

Consider a sample of T observations of Y with distribution function $\mathcal{F}(Y, \theta)$. Suppose we are interested in the sampling properties of a statistic Z_T, which may be an estimator, test statistic, or some other random variable of interest that is based on the outcomes of the DGP. To indicate the relationship between $\mathcal{F}(Y, \theta)$ and the statistic, we denote the distribution of Z_T as $\mathcal{G}_T(\mathcal{F})$. Even if we assume we know $\mathcal{F}$, as in the case of a fully parametric model, we may only be able to derive the distribution $\mathcal{G}_T$ in a limited number of cases. In general, we have to use the sample to tell us about $\mathcal{F}$ in order to recover information about $\mathcal{G}_T$.

2.4.2 The Empirical Distribution Function

The bootstrap procedure is a data-based approximation to $\mathcal{G}_T$ which replaces $\mathcal{F}$ with the EDF of the observed sample. Note that $\mathcal{F}(y) = \Pr(y_t \leq y) = \mathcal{E}[I(y_t \leq y)]$, where $I(\cdot)$ is the indicator function, so $\mathcal{F}(y)$ can be expressed as a population moment. A natural estimate is therefore the corresponding sample moment:

$$\mathcal{F}_T(y) = \frac{1}{T}\sum_{t=1}^{T} I(y_t \leq y).$$

$\mathcal{F}_T(y)$ is called the EDF and is a nonparametric estimate of $\mathcal{F}(y)$. Note that while $\mathcal{F}(y)$ may be either discrete or continuous, $\mathcal{F}_T(y)$ is by construction a (discontinuous) step function.

The EDF is a consistent estimator of the c.d.f.. To see this, note that for any y, $I(y_t \leq y)$ is an i.i.d. random variable with expectation $\mathcal{F}(y)$. Thus by the Weak Law of Large Numbers,[10] $\mathcal{F}_T(y) \xrightarrow{p} \mathcal{F}(y)$. It is also true that the convergence is uniform in the argument y (by the Glivenko-Cantelli theorem):

$$\sup_{y\in\mathbf{R}} |\mathcal{F}_T(y) - \mathcal{F}(y)| \xrightarrow{p} 0.$$

It also converges at rate $\sqrt{T}$. By the Central Limit Theorem,

$$\sqrt{T}[\mathcal{F}_T(y) - \mathcal{F}(y)] \xrightarrow{D} \mathcal{N}(0, \mathcal{F}(y)(1 - \mathcal{F}(y))).$$

To see the effects of sample size on the EDF, Figure 2.4 shows the EDF and c.d.f. for random samples of size $T = 25, 100, 250,$ and 1000. The pseudo-random draws are from the $\mathcal{N}(0,1)$ distribution. For $T = 25$, the EDF is only a crude approximation to the c.d.f., but the approximation appears to improve for large T. As the sample size increases, the EDF step function gets uniformly close to the true c.d.f..

The EDF is a valid distribution function and has a discrete probability distribution which puts probability mass $1/T$ at each point y_t, $t = 1, \cdots, T$.

2.4.3 Computation

Since the EDF $\mathcal{F}_T$ is multinomial (with T support points), in principle the distribution of Z_T could be calculated by direct methods. Since there are

[10] Chapter 8 provides a formal treatment of asymptotic theory.

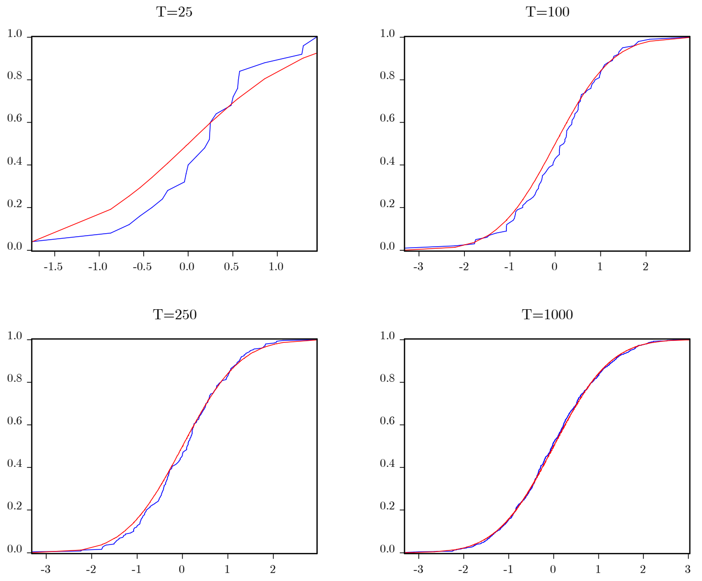

Figure 2.4: EDF of $\mathcal{N}(0,1)$ for Different Sample Sizes

2^T possible samples, however, such a calculation is computationally unfeasible unless T is very small. A popular alternative is to use a simulation to approximate the distribution. The algorithm is as follows:

- Pick the number of samples (J) to generate.
- For sample j, draw T vectors $\widetilde{y}_t^j$ randomly from $\{y_t\}_{t=1}^T$ (with replacement), and form the data $\left\{\widetilde{y}_t^j\right\}_{t=1}^T$.
- Calculate $\widetilde{Z}_T^j$ for each sample $(j = 1, \cdots, J)$.
- Given that each of the J draws of $\widetilde{Z}_T^j$ is i.i.d., the distribution function $\widehat{\mathcal{G}}_T$ may be estimated from the EDF of the $\widetilde{Z}_T^j$:

$$\widehat{\mathcal{G}}_T(v) = \frac{1}{J}\sum_{j=1}^{J} I\left(\widetilde{Z}_T^j \leq v\right).$$

The actual calculation of $\widehat{\mathcal{G}}_T(v)$ is typically unnecessary.

2.4.4 Bootstrap Estimation of Bias

The bias of $\widehat{\theta}$ is

$$\tau_T = \mathcal{E}\left(\widehat{\theta} - \theta\right).$$

Let $Z_T(\theta) = \widehat{\theta} - \theta$. Then

$$\begin{aligned} \tau_T &= \mathcal{E}\left(Z_T(\theta_0)\right) \\ &= \int x d\mathcal{G}_T(x, \mathcal{F}). \end{aligned}$$

The bootstrap counterpart calculated by simulation is

$$\begin{aligned} \widehat{\tau}_T &= \frac{1}{J}\sum_{j=1}^{J} \widetilde{Z}_T^j \\ &= \frac{1}{J}\sum_{j=1}^{J} \left(\widetilde{\theta}^j - \widehat{\theta}\right) \\ &= \overline{\overline{\theta}} - \widehat{\theta}. \end{aligned}$$

If $\widehat{\theta}$ is biased, it might be desirable to construct a bias-corrected estimator (one with reduced bias). Ideally, this would be

$$\theta^* = \widehat{\theta} - \tau_T,$$

but as τ_T is unknown (as the true value of the parameter is unknown), the bootstrap bias-corrected estimator is

$$\begin{aligned} \theta^* &= \widehat{\theta} - \widehat{\tau}_T \\ &= \widehat{\theta} - \left(\overline{\overline{\theta}} - \widehat{\theta}\right) \\ &= 2\widehat{\theta} - \overline{\overline{\theta}}. \end{aligned}$$

Note that the bias-corrected estimator is not $\overline{\overline{\theta}}$. Intuitively, the bootstrap makes the following experiment: Suppose that $\widehat{\theta}$ is the true value. Then, what is the average value of $\widehat{\theta}$ calculated for each sample? The answer is $\overline{\overline{\theta}}$. If this is lower than $\widehat{\theta}$, this suggests that the estimator is downward-biased, so the bias-corrected estimator of θ should be larger than $\widehat{\theta}$, and the best guess is the difference between $\widehat{\theta}$ and $\overline{\overline{\theta}}$. Similarly if $\overline{\overline{\theta}}$ is higher than $\widehat{\theta}$, then the estimator is upward-biased and the bias-corrected estimator should be lower than $\widehat{\theta}$.

Many estimation methods commonly used in econometrics—such as maximum likelihood (MLE), generalized method of moments (GMM), and Efficient Method of Moments—rely heavily on asymptotic approximations to justify inference. In subsequent chapters, we will examine these methods in detail. Bootstrap-based bias correction and standard error estimation will be revisited in those contexts, especially when analytical derivations are difficult or unreliable.

2.4.5 Bootstrap Estimation of Variance

Let $Z_T = \widehat{\theta}$. The variance of $\widehat{\theta}$ is

$$\mathcal{V}_T = \mathcal{E}\left(Z_T - \mathcal{E}\left(Z_T\right)\right)^2 .$$

The simulation estimate is

$$\widetilde{\mathcal{V}}_T = \frac{1}{J}\sum_{j=1}^{J}\left(\widetilde{\theta}^j - \overline{\overline{\theta}}\right)^2 .$$

A standard error for $\widehat{\theta}$ is the square root of the bootstrap estimate of the variance. In general, Monte Carlo approximations to bootstrap estimators of moments can be computed with relatively small simulation samples J, thus making them inexpensive to compute.

While this standard error may be calculated and reported, it is not clear if it is useful. The primary use of asymptotic standard errors is to construct confidence intervals, which are based on the asymptotic normal approximation to the t-ratio. However, the use of the bootstrap presumes that such asymptotic approximations might be poor, in which case the normal approximation is suspect. It appears superior to calculate bootstrap confidence intervals as discussed below.

2.4.6 Confidence Intervals

Efron's Percentile Interval

For a distribution function $\mathcal{G}_T(\mathcal{F})$, let $q_T(\alpha, \mathcal{F})$ denote its quantile function. This is the function which solves

$$\mathcal{G}_T\left(q_T(\alpha, \mathcal{F}), \mathcal{F}\right) = \alpha.$$

Let $q_T(\alpha)$ be the quantile function of the true sampling distribution, and $\widetilde{q}_T(\alpha) = q_T(\alpha, \mathcal{F}_T)$ denote the quantile function of the bootstrap distribution. Note that this function will change depending on the underlying statistic Z_T whose distribution is $\mathcal{G}_T$.[11]

Let $Z_T = \widehat{\theta}$ be an estimate of a parameter of interest. In $(1-\alpha)\,\%$ of samples, $\widehat{\theta}$ lies in the region $[q_T(\alpha/2), q_T(1-\alpha/2)]$. This motivates a confidence interval proposed by Efron:[12]

$$C_1 = \left[\widetilde{q}_T(\alpha/2), \widetilde{q}_T(1-\alpha/2)\right].$$

Computationally, this interval can be estimated from a bootstrap simulation by sorting the bootstrap estimates $\widetilde{\theta}$. Let the sorted numbers (in ascending order) be denoted by $\left\{\widetilde{\theta}^j\right\}_{j=1}^J$. Then $\widetilde{q}_\alpha^*$ is the $J\alpha$'th number in this ordered sequence. The estimate of a 95% Efron percentile interval is $[\widetilde{q}_T(0.025), \widetilde{q}_T(0.975)]$.

The interval C_1 is often used in empirical practice because it is easy to compute, simple to motivate, and it has the feature of being translation invariant. That is, if we define $\varphi = h(\theta)$ as the parameter of interest for a monotonic function h, the percentile method applied to this problem will produce the confidence interval $[h(\widetilde{q}_T(\alpha/2)), h(\widetilde{q}_T(1-\alpha/2))]$.

However, C_1 is in a deep sense very poorly motivated. It is useful if we introduce an alternative definition for C_1. Let $Z_T = \widehat{\theta} - \theta$ and let $q_T(\alpha)$ be the quantile function of its distribution. Then C_1 can alternatively be written as

$$C_1 = \left[\widehat{\theta} + \widetilde{q}_T(\alpha/2), \widehat{\theta} + \widetilde{q}_T(1-\alpha/2)\right].$$

[11] When $\mathcal{G}_T(\mathcal{F})$ is discrete, $q_T(\alpha, \mathcal{F})$ may not be unique, but we will ignore such complications.

[12] This is often called the percentile confidence interval.

This is a bootstrap estimate of the "ideal" confidence interval

$$C_1^0 = \left[\widehat{\theta} + q_T(\alpha/2), \widehat{\theta} + q_T(1-\alpha/2)\right].$$

The latter has coverage probability

$$\begin{aligned}\Pr\left(\theta \in C_1^0\right) &= \Pr\left[\widehat{\theta} + q_T(\alpha/2) \le \theta \le \widehat{\theta} + q_T(1-\alpha/2)\right] \\ &= \Pr\left[-q_T(1-\alpha/2) \le \widehat{\theta} - \theta \le -q_T(\alpha/2)\right] \\ &= \mathcal{G}_T(-q_T(\alpha/2), \mathcal{F}) - \mathcal{G}_T(-q_T(1-\alpha/2), \mathcal{F}).\end{aligned}$$

which generally is not $1-\alpha$! There is an important exception. If $\widehat{\theta} - \theta$ has a symmetric distribution, then $\mathcal{G}_T(-x, \mathcal{F}) = 1 - \mathcal{G}_T(x, \mathcal{F})$, so

$$\begin{aligned}\Pr\left(\theta \in C_1^0\right) &= \mathcal{G}_T(-q_T(\alpha/2), \mathcal{F}) - \mathcal{G}_T(-q_T(1-\alpha/2), \mathcal{F}) \\ &= (1 - \mathcal{G}_T(q_T(\alpha/2), \mathcal{F})) - (1 - \mathcal{G}_T(q_T(1-\alpha/2), \mathcal{F})) \\ &= \left(1 - \frac{\alpha}{2}\right) - \left(1 - \left(1 - \frac{\alpha}{2}\right)\right) = 1 - \alpha,\end{aligned}$$

and this idealized confidence interval is accurate. Therefore, C_1^0 and C_1 are designed for the case that $\widehat{\theta}$ has a symmetric distribution about θ. When that is not the case, C_1 may perform quite poorly. However, by the translation invariance argument presented above, it also follows that if there exists some monotonic function $h(\cdot)$ such that $h\left(\widehat{\theta}\right)$ is symmetrically distributed about $h(\theta)$, then the idealized percentile bootstrap method will be accurate.

Based on these arguments, many argue that the percentile interval should not be used unless the sampling distribution is close to unbiased and symmetric.

Hall's Percentile Interval

The problems with the percentile method outlined above can be circumvented by an alternative method. Let $Z_T = \widehat{\theta} - \theta$, then

$$1 - \alpha = \Pr\left[\widehat{\theta} - q_T(1-\alpha/2) \le \theta \le \widehat{\theta} - q_T(\alpha/2)\right],$$

so a bootstrap $(1-\alpha)\%$ confidence interval for θ would be

$$C_2 = \left[\widehat{\theta} - \widetilde{q}_T(1-\alpha/2), \widehat{\theta} - \widetilde{q}_T(\alpha/2)\right].$$

Notice that generally this is very different from the Efron interval C_1. They coincide in the special case that $\mathcal{G}_T$ is symmetric about $\widehat{\theta}$, but otherwise they differ.

Computationally, this interval can be estimated from a bootstrap simulation by sorting the bootstrap statistic $\widetilde{Z}_T^j = \left(\widetilde{\theta}^j - \widehat{\theta}\right)$, which is centered at the sample estimate $\widehat{\theta}$. These are sorted to yield the quantile estimates $\widetilde{q}_T(0.025)$ and $\widetilde{q}_T(0.975)$. The 95% confidence interval is then $\left[\widehat{\theta} - \widetilde{q}_T(0.975), \widehat{\theta} - \widetilde{q}_T(0.025)\right]$. This confidence interval is discussed in most theoretical treatments of the bootstrap, but is not widely used in practice.

2.4.7 Inference

One-Sided Hypothesis Tests

Suppose we want to test $\mathrm{H}_0 : \theta = \theta_0$ against the alternative $\mathrm{H}_1 : \theta < \theta_0$ at size α. We would set $Z_T = \left(\widehat{\theta} - \theta\right) / \sqrt{\mathcal{V}\left(\widehat{\theta}\right)}$ and reject H_0 in favor of H_1 if $Z_T < c$, where c would be selected so that

$$\Pr\left[Z_T < c\right] = \alpha.$$

Thus $c = q_T(\alpha)$. Since this is unknown, a bootstrap test replaces $q_T(\alpha)$ with the bootstrap estimate $\widetilde{q}_T(\alpha)$, and the test rejects if $Z_T < \widetilde{q}_T(\alpha)$.[13]

Computationally, these critical values can be estimated from a bootstrap simulation by sorting the bootstrap t-statistics $\widetilde{Z}_T = \left(\widetilde{\theta} - \widehat{\theta}\right) / \sqrt{\mathcal{V}\left(\widetilde{\theta}\right)}$. Note that the bootstrap test is centered at the estimate $\widehat{\theta}$, and that the variance is calculated on the bootstrap sample. After sorting the t-statistics, we find the estimated quantiles $\widetilde{q}_T(\alpha)$ and/or $\widetilde{q}_T(1-\alpha)$.

Two-Sided Hypothesis Tests

Suppose we want to test $\mathrm{H}_0 : \theta = \theta_0$ against the alternative $\mathrm{H}_1 : \theta \neq \theta_0$ at size α. We would set $Z_T = \left(\widehat{\theta} - \theta\right) / \sqrt{\mathcal{V}\left(\widehat{\theta}\right)}$ and reject H_0 in favor of H_1 if $|Z_T| > c$, where c would be selected so that

$$\Pr\left[|Z_T| > c\right] = \alpha.$$

[13] Similarly, if the alternative is $\mathrm{H}_1 : \theta > \theta_0$, the bootstrap test rejects if $Z_T > \widetilde{q}_T(1-\alpha)$.

Thus $c = q_T(1-\alpha)$ is the $1-\alpha$ quantile of $|Z_T|$. Since this is unknown, a bootstrap test replaces $q_T(1-\alpha)$ with the bootstrap estimate $\widetilde{q}_T(1-\alpha)$, and the test rejects if $|Z_T| > \widetilde{q}_T(1-\alpha)$.

Computationally, $\widetilde{q}_T(1-\alpha)$ is estimated from a bootstrap simulation by sorting the bootstrap t-statistics $\left|\widetilde{Z}_T\right| = \left|\widetilde{\theta} - \widehat{\theta}\right| / \sqrt{\mathcal{V}\left(\widetilde{\theta}\right)}$, and taking the upper $\alpha\%$ quantile. The bootstrap test rejects if $\left|\widetilde{Z}_T\right| > \widetilde{q}_T(1-\alpha)$.

Vector Tests

If θ is a vector, then to test $\mathrm{H}_0 : \theta = \theta_0$ against the alternative $\mathrm{H}_1 : \theta \neq \theta_0$ at size α, we would use a Wald statistic

$$W_T = \left(\widehat{\theta} - \theta\right)' \widehat{\mathcal{V}}_\theta^{-1} \left(\widehat{\theta} - \theta\right),$$

or some asymptotically chi-square statistic. Thus here $Z_T = W_T$. The ideal test rejects if $W_T \geq q_T(1-\alpha)$, where $q_T(1-\alpha)$ is the $(1-\alpha)\%$ quantile of the distribution of W_T. The bootstrap test rejects if $W_T \geq \widetilde{q}_T(1-\alpha)$, where $\widetilde{q}_T(1-\alpha)$ is the $(1-\alpha)\%$ quantile of the distribution of

$$\widetilde{W}_T = \left(\widetilde{\theta} - \widehat{\theta}\right)' \widetilde{\mathcal{V}}_\theta^{-1} \left(\widetilde{\theta} - \widehat{\theta}\right).$$

Computationally, $\widetilde{q}_T(1-\alpha)$ is found as the quantile from simulated values of $\widetilde{W}_T$. Note that the simulation considers a quadratic form in $\left(\widetilde{\theta} - \widehat{\theta}\right)$, not $\left(\widehat{\theta} - \theta_0\right)$.

2.4.8 Bootstrap Procedures for Regression Analysis

Here, we consider extensions of the simple nonparametric bootstrap presented above for cases that feature dependence. We focus on procedures that can be used to resample data for the linear regression model.[14]

[14] Procedures for the nonlinear regression model, qualitative choice, and simultaneous equations models may be constructed in similar fashion.

Simple Bootstrap

Consider the case in which X is fixed (deterministic) and resample the residuals from the OLS model. The steps of the Monte Carlo approximation procedure are:

a. Compute the OLS estimate of β $(\widehat{\beta})$, and the OLS noise component estimate $(\widehat{u})$.

b. Draw T pseudo-random errors $(\widetilde{u}_t^j)$ from $\widehat{u}$ with replacement $(t = 1, \cdots, T)$.

c. Form the T bootstrap sample observations $\widetilde{y}_t^j = x_t'\widehat{\beta} + \widetilde{u}_t^j$ $(t = 1, \cdots, T)$.

d. Compute and save the OLS estimate $\widetilde{\beta}^j$, obtained from the replicated sample $\left(\widetilde{Y}, X\right)$. Compute any other statistic of interest $(\widetilde{Z}^j)$.

e. Repeat steps b, c, and d J times.

We can use the set of replicated OLS estimates to approximate the sampling distribution of $\widehat{\beta}$ and Z_T. Given that we treated X as fixed, all inferential statements are made conditionally on the observed values of the regressors.

This approach has two problems. The first is that although the OLS residuals have the same expected value as the unobserved errors, their variance-covariances matrices are different because

$$\mathcal{E}\left[\widehat{u}\widehat{u}'\right] = \sigma^2 M \neq \sigma^2 I = \mathcal{E}\left[uu'\right].$$

Thus, the OLS residuals are dependent and heteroskedastic. We may overcome the dependence issue by i.i.d. sampling in the Monte Carlo simulation procedure. The variance issue may be addressed in a variety of ways. First, we may simply choose to ignore the heteroskedasticity by noting that $\widehat{u} \xrightarrow{p} u$, and thus the problem is insignificant in large samples. Alternatively, we could rescale the OLS residuals to homoskedasticity before generating the bootstrap samples in this manner:

$$\widehat{u}_t^* = \frac{\widehat{u}_t}{\sqrt{1 - h_t}} - \frac{1}{T}\sum_{s=1}^{T}\frac{\widehat{u}_s}{\sqrt{1 - h_s}},$$

where $h_t = x_t \left(X'X\right)^{-1} x_t'$. The second of these expressions is placed in order to assure that the mean of $\widehat{u}^*$ is zero. As the resampled residuals may not have a zero sample mean, they must be recentered before proceeding with bootstrapping.

The second problem with this approach is that it assumes that the error terms u are i.i.d. and the regressors are deterministic (thus, ignoring the sampling variation in X).

Paired Resampling

To accommodate the LRM with potentially stochastic regressors and potential heteroskedasticity, we can use a procedure known as paired resampling. The steps of the Monte Carlo approximation procedure under paired resampling are as follow:

a. Compute the OLS estimate of β $(\widehat{\beta})$.

b. Draw T pseudo-random pairs $(\widetilde{y}_t^j, \widetilde{x}_t^j)$ from (Y, X) $(t = 1, \cdots, T)$.

c. Compute and save the OLS estimate $\widetilde{\beta}^j$, obtained from the replicated sample $\left(\widetilde{Y}, \widetilde{X}\right)$. Compute any other statistic of interest $(\widetilde{Z}^j)$.

d. Repeat steps b and c J times.

Thus, the paired sampling procedure is similar to the simple nonparametric bootstrap discussed earlier, and it reflects the joint stochastic character of (Y, X). In addition, the paired bootstrap is less sensitive to model specification errors than is the residual sampling algorithm and is routinely used with cross section data. A potential drawback is that the bootstrap design $\widetilde{X}$ may not have full rank.

Wild Bootstrap

When heteroskedasticity is present but it's form is unknown, it can be imitated using the so-called wild bootstrap; thus providing refinements for the LRM with heteroskedastic errors. In this case the bootstrap error term is defined as

$$\widetilde{u}_t^j = \widehat{u}_t^* v_t,$$

where v_t are mutually independent drawings, completely independent of the original data, from some auxiliary distribution such that $\mathcal{E}(v_t) = 0$ and $\mathcal{V}(v_t) = 1$. Two popular choices for the distribution of v are:

$$F_1 = \begin{cases} \left(1-\sqrt{5}\right)/2 & \text{with probability } p_1 \\ \left(1+\sqrt{5}\right)/2 & \text{with probability } 1-p_1 \end{cases}$$

and

$$F_2 = \begin{cases} 1 & \text{with probability } p_2 \\ -1 & \text{with probability } 1-p_2 \end{cases},$$

where $p_1 = \left(1+\sqrt{5}\right)/\left(2\sqrt{5}\right)$ and $p_2 = 1/2$.

2.4.9 Dependent Data

With dependent data, asymptotic refinements cannot be obtained by using independent bootstrap samples. Bootstrap sampling must be carried out in a way that suitably captures the dependence of the data-generation process. At present, higher-order asymptotic approximations and asymptotic refinements are available only when the data-generation process is stationary and strongly geometrically mixing. Except when stated otherwise, it is assumed here that this requirement is satisfied. Non-stationary DGPs are discussed in Horowitz (2001).

Parametric Models

Bootstrap sampling that captures the dependence of the data can be carried out relatively easily if there is a parametric model, such as an ARMA model, that reduces the DGP to a transformation of independent random variables. For example, suppose that the series $\{x_t\}$ is generated by the stationary, invertible, finite-order ARMA model

$$A(L,\alpha)\, x_t = B(L,\beta)\, u_t$$

where A and B are known functions, L is the lag operator, α and β are vectors of parameters, and $\{u_t\}$ is a sequence of i.i.d. random variables. Let

$\widehat{\alpha}$ and $\widehat{\beta}$ be consistent estimators of α and β, and let $\{\widehat{u}_t\}$ be the centered residuals. A bootstrap sample $\{\widetilde{x}_t\}$ can be generated as

$$A\left(L,\widehat{\alpha}\right)\widetilde{x}_t = B\left(L,\widehat{\beta}\right)\widetilde{u}_t$$

where $\{\widetilde{u}_t\}$ is a random sample from the empirical distribution of the residuals $\{\widehat{u}_t\}$.

Block Bootstrap

When there is no parametric model that reduces the DGP to independent sampling from some probability distribution, the bootstrap can be implemented by dividing the data into blocks and sampling the blocks randomly with replacement. The blocks may be non-overlapping or overlapping. To describe these blocking methods more precisely, let the data consist of observations $\{x_t : t = 1, \ldots, T\}$. With non-overlapping blocks of length l, block 1 is observations $\{x_j : j = 1, \ldots, l\}$, block 2 is observations $\{x_{l+j} : j = 1, \ldots, l\}$, and so forth. With overlapping blocks of length l, block 1 is observations $\{x_j : j = 1, \ldots, l\}$, block 2 is observations $\{x_{j+1} : j = 1, \ldots, l\}$, and so forth. The bootstrap sample is obtained by sampling blocks randomly with replacement and laying them end-to-end in the order sampled. Overlapping blocks provide somewhat higher estimation efficiency than non-overlapping ones.

Regardless of the blocking method that is used, the block length must increase with increasing sample size T to make bootstrap estimators of moments and distribution functions consistent. The asymptotically optimal block length is defined as the one that minimizes the asymptotic mean-square error of the block bootstrap estimator. The asymptotically optimal block length and its rate of increase with increasing T depend on what is being estimated. With either overlapping or non-overlapping blocks, the asymptotically optimal block-length is $l \backsim T^r$, where $r = 1/3$ for estimating bias or variance, $r = 1/4$ for estimating a one-sided distribution function, and $r = 1/5$ for estimating a symmetrical distribution function.

This procedure does not guarantee that the resulting series is stationary. A procedure that does is known as stationary bootstrap. It uses overlapping blocks with lengths that are sampled randomly from the geometric distribution $(<, >)$poro.

2.5 Further Reading

Students approaching simulation and resampling methods for the first time will find Hansen (2022) especially accessible, with clear examples and a modern treatment of Monte Carlo and bootstrap techniques. Efron and Tibshirani (1993) remains the seminal introduction to the bootstrap and is still one of the most cited and readable sources on the topic. Davidson and MacKinnon (1993) offers a bridge between theory and application, with particular emphasis on implementation in econometric settings.

For deeper theoretical insights, Horowitz (2001) provides a comprehensive and rigorous overview of the bootstrap, including asymptotic refinements and extensions for dependent data. Hendry (1984) discusses the role of simulation in econometric research design, focusing on how Monte Carlo studies can clarify the finite-sample properties of estimators and tests. Ullah (2004) complements these treatments with a full account of inference in finite samples, including both analytical and simulation-based approaches.

Judd (1998) extends the discussion beyond econometrics by presenting numerical methods widely used in economic modeling and policy analysis, including simulation and quadrature methods. Mittelhammer, Judge, and Miller (2000) provides a more technical foundation for computer-based inference, particularly useful for students working in GAUSS or other matrix-oriented languages.

In applied settings, Lam and Veill (2002) show how bootstrap methods can be used to construct prediction intervals for regression forecasts, and Politis and Romano (1994) introduce the stationary bootstrap, a valuable resampling technique for dependent time series. Finally, Sheskin (2000) is a practical reference for simulation diagnostics and distributional tests, especially when assessing the performance of pseudo-random number generators or bootstrap replications.

2.6 Workout Problems

1. Construct an algorithm for generating pseudo-random outcomes for the $\mathcal{U}(0,1)$ distribution based on the linear congruential rule.

2. How would you generate pseudo-random outcomes for the $\mathcal{U}(a,b)$ distribution?

3. Generate pseudo-random numbers from a logistic $(0,1)$ distribution. Recall that its c.d.f. is $\mathcal{F}(x) = e^{-x}/[1+e^{-x}]$.

4. How would you generate pseudo-random outcomes for the $\mathcal{N}(a,b)$ distribution?

5. Assume that you have an efficient way of generating $\mathcal{N}(0,1)$ pseudo-random numbers. Devise an algorithm for generating pseudo-random numbers from a multivariate normal distribution with mean A and variance-covariance Ω, with A being a $k \times 1$ vector of constants and Ω a $k \times k$ positive definite matrix.

6. Construct an algorithm to generate pseudo-random outcomes from a Cauchy distribution.

7. Construct an algorithm to generate pseudo-random outcomes from a Student's t distribution.

8. Construct an algorithm to generate pseudo-random outcomes from an F distribution.

9. Construct an algorithm to generate pseudo-random outcomes from a Chi-square distribution.

10. Using the rejection method generate an estimate of π.

11. Replicate the results of Section 2.3.3.

Chapter 3

Functional Form

3.1 Introduction

Having reviewed the basic principles of OLS estimation, here we will cover some topics that have to do with its practice (generically known as the choice of functional form).

The document is organized as follows: Section 3.2 describes the use of dummy variables. Section 3.3 discusses the issue of possible nonlinearities. Section 3.4 introduces a test for normality. Section 3.5 considers measurement errors in variables. Section 3.6 evaluates the effect of omitting relevant variables. Section 3.7 analyzes the properties of the OLS estimator when irrelevant variables are included. Section 3.8 discusses multicollinearity. Section 3.9 focuses on influential analysis. Section 3.10 discusses model selection strategies. Finally, Section 3.11 introduces the issue of specification searches.

3.2 Dummy Variables

In many applications, a bulk of the regressors are binary variables which take on the value 0 or 1. We call these dummy variables. Often the regressor is binary because this is the way the data was recorded. In other cases, the binary regressor has been constructed from other variables in the dataset.

For example, a dummy variable may be used to denote the gender (male/female) of an individual. We are interested in estimating $\mathcal{E}\left(Wage\,|Gender\right)$. There are several equivalent ways to write this down. One is to define a dummy variable[1]

$$d_{1,i} = \begin{cases} 1 & \text{female} \\ 0 & \text{male} \end{cases}$$

and the model is $W_i = \beta_0 + \beta_1 d_{1,i} + u_i$ ($W = Wage$). In this model, $\beta_0 = \mathcal{E}\left(Wage\,|Male\right)$ and $\beta_0 + \beta_1 = \mathcal{E}\left(Wage\,|Female\right)$. Second, we could define the variable

$$d_{2,i} = \begin{cases} 0 & \text{female} \\ 1 & \text{male} \end{cases}$$

and the model is $W_i = \beta_0 + \beta_1 d_{2,i} + u_i$. In this model, $\beta_0 = \mathcal{E}\left(Wage\,|Female\right)$ and $\beta_0 + \beta_1 = \mathcal{E}\left(Wage\,|Male\right)$. Third, we could define both $d_{1,i}$ and $d_{2,i}$ as above, and write the model as $W_i = \beta_1 d_{1,i} + \beta_2 d_{2,i} + u_i$. Here, $\beta_1 = \mathcal{E}\left(Wage\,|Female\right)$ and $\beta_2 = \mathcal{E}\left(Wage\,|Male\right)$. These three models are equivalent. In this sense, as dummy variables are essentially qualitative variables, it does not matter how we define them as long as we are consistent.

A standard mistake is to include an intercept, $d_{1,i}$ and $d_{2,i}$ all in the regression. These are clearly perfectly collinear ($d_{1,i} + d_{2,i} = 1$) so that this cannot be done (X would not have full rank).

If the equation of interest is $\mathcal{E}\left(Wage\,|Education, Gender\right)$, a typical regression model is

$$W_i = \beta_0 + \beta_1 d_{1,i} + \beta_2 E_i + u_i.$$

This model specifies an intercept effect for gender. That is, males and females have different intercepts, but the same return on education.

A regression model allowing for slope differences is[2]

$$W_i = \beta_0 + \beta_1 d_{1,i} + \beta_2 E_i + \beta_3 d_{1,i} E_i + u_i.$$

This allows for greater differences between groups. When there are several continuous regressors, it may be desirable to have a slope effect for some, but not all, of the regressors.

From the standpoint of our regression theory, we think of d_i as a random variable, from the sampling process which generated the other variables. The idea is that if we sample from the entire population of individuals, some random draws will be women and some will be men.

[1] Following convention, with cross-sectional observations, we denote the regressors as x_i instead of x_t and the sample size as N instead of T.

[2] Slope differences are often referred to as interactions.

It is interesting to see how our estimators handle dummy variables algebraically. Take the simple model

$$W_i = \beta_1 d_{1,i} + \beta_2 d_{2,i} + u_i,$$

we can write this in matrix notation as

$$W = D_1\beta_1 + D_2\beta_2 + u.$$

By construction $D_1' D_2 = 0$. Thus,

$$\begin{aligned}
\widehat{\beta}_1 &= \left(D_1' D_1\right)^{-1} D_1' W \\
&= \frac{\sum_{i=1}^N d_{1,i} W_i}{\sum_{i=1}^N d_{1,i}} \\
&= \frac{1}{N_1} \sum_{i=1}^N d_{1,i} W_i \\
&= \overline{W}_1,
\end{aligned}$$

where N_1 is the number of observations with $d_{1,i} = 1$, and $\overline{W}_1$ is the sample mean among those observations. Similarly, $\widehat{\beta}_2 = \overline{W}_2$.

The variance-covariance estimate is

$$\begin{aligned}
\widehat{\mathcal{V}}\left(\widehat{\beta}\right) &= \widehat{\sigma}^2 \begin{bmatrix} \left(D_1' D_1\right)^{-1} & 0 \\ 0 & \left(D_2' D_2\right)^{-1} \end{bmatrix} \\
&= \begin{bmatrix} \frac{\widehat{\sigma}^2}{N_1} & 0 \\ 0 & \frac{\widehat{\sigma}^2}{N_2} \end{bmatrix},
\end{aligned}$$

where

$$\widehat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^N \widehat{u}_i^2.$$

is the estimate of σ^2 based on the full sample.

Another candidate for the variance-covariance estimate is

$$\begin{aligned}\overline{\mathcal{V}}\left(\widehat{\beta}\right) &= \begin{bmatrix} \widehat{\sigma}_1^2\left(D_1'D_1\right)^{-1} & 0 \\ 0 & \widehat{\sigma}_2^2\left(D_2'D_2\right)^{-1}\end{bmatrix} \\ &= \begin{bmatrix} \frac{\widehat{\sigma}_1^2}{N_1} & 0 \\ 0 & \frac{\widehat{\sigma}_2^2}{N_2}\end{bmatrix},\end{aligned}$$

where

$$\widehat{\sigma}_j^2 = \frac{1}{N_j}\sum_{i=1}^{N} d_{j,i}\widehat{u}_i^2 \quad \text{for} \quad j = 1, 2$$

are the estimates of σ^2 based on observations with $d_{1,i} = 1$ and $d_{2,i} = 1$, respectively. Thus the conventional estimate imposes the restriction that the variance of u is the same for the two groups, while the second considers the possible presence of heteroskedasticity.

Another frequent application of dummy variables is when dealing with series that present seasonality. Take for example the (log of) Chilean quarterly GDP. Figure 3.1 shows that it has a strong seasonal component. Let

$$y_t = \beta_0 + \beta_1 q_{1,t} + \beta_2 q_{2,t} + \beta_3 q_{3,t} + \beta_4 t + u_t, \tag{3.1}$$

where

$$q_{1,t} = \begin{cases} 1 & \text{first quarter} \\ 0 & \text{otherwise}\end{cases}, \quad q_{2,t} = \begin{cases} 1 & \text{second quarter} \\ 0 & \text{otherwise}\end{cases}, \quad q_{3,t} = \begin{cases} 1 & \text{third quarter} \\ 0 & \text{otherwise}\end{cases}.$$

As there are four possible quarters and we introduced a constant in the model, we defined only three dummy variables. Thus, $\mathcal{E}\left(y_t\,|\text{First quarter}\right) = \beta_0 + \beta_1 + \beta_4 t$ and $\mathcal{E}\left(y_t\,|\text{Fourth quarter}\right) = \beta_0 + \beta_4 t$.

Some researchers prefer to work with series that are seasonally adjusted; that is series from which the seasonal component was removed. If the pattern of seasonality were to correspond to (3.1), an easy way to recover the seasonally adjusted series would be:

$$y_t^* = y_t - \left(\widehat{\beta}_1 q_{1,t} + \widehat{\beta}_2 q_{2,t} + \widehat{\beta}_3 q_{3,t}\right), \tag{3.2}$$

where $\widehat{\beta}_i$ is the OLS estimate of β_i in (3.1). There are other ways to deal with seasonality, and if present, removing it from the original series prior to working with it isn't usually the best practice. The dashed line in Figure 3.1 presents the result of applying (3.2) to the log of GDP.

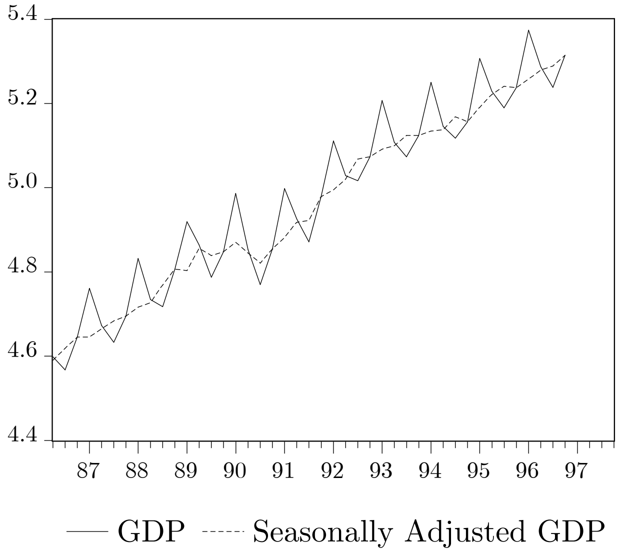

Figure 3.1: Use of Quarterly Dummies

3.3 Nonlinearities

The models that we presented until now considered y to be a linear function of the regressors and the error term. In this section we will discuss how restrictive is this assumption.

3.3.1 Linearization of Nonlinear Models

Even the simplest economic model usually considers functional forms that are nonlinear. A simple example comes from a Cobb-Douglas production function,

$$Q_i = U_i K_i^{\alpha} L_i^{\gamma}, \tag{3.3}$$

where Q_i denotes the output of firm i, K_i its capital stock, and L_i its labor force. This specification is nonlinear and may lead us to think that OLS cannot be applied. However, if we take logs to (3.3) we obtain (lower case letters denote logs):

$$q_i = \alpha k_i + \gamma l_i + u_i.$$

If the Solow residual (not observed by the econometrician) is considered to be the corresponding error term, this transformation allows us to apply OLS.

Another example in which a simple transformation leads to linear models is:

$$Q_i = e^{x_i'\beta + u_i}.$$

Of course, there are instances in which there is no transformation that can be used to transform a nonlinear model into a linear one. Consider a CES (Constant Elasticity of Substitution) production function:

$$Q_i = \left[\alpha K_i^{\theta} + (1-\alpha) L_i^{\theta}\right]^{\frac{1}{\theta}} + u_i;$$

in such a case, OLS cannot be applied and we must estimate the model with other methods (such as Nonlinear Least Squares).

3.3.2 Nonlinearities in the Regressors

Suppose we are interested in $\mathcal{E}\left(y_t \left| x_t \right.\right) = m\left(x_t\right), x \in \mathbb{R}$, and the form of m is unknown. A common approach is to consider a polynomial approximation:

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 x_t^2 + \cdots + \beta_j x_t^j + u_t.$$

Letting $\beta = \left(\beta_0, \beta_1, \cdots, \beta_j\right)$ and $z_t = \left(1, x_t, x_t^2, \cdots, x_t^j\right)$, this is $y_t = z_t'\beta + u_t$, which is a linear regression model. Typically, the polynomial order j is kept quite small.

Now suppose that $x \in \mathbb{R}^2$. A simple quadratic approximation is

$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \beta_3 x_{1,t}^2 + \beta_4 x_{2,t}^2 + \beta_5 x_{1,t} x_{2,t} + u_t.$$

As the dimensionality of x increases, such approximations can become quite non-parsimonious. Most applications do not use more than quadratic terms or cubics without interactions:[3]

$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \beta_3 x_{1,t}^2 + \beta_4 x_{2,t}^2 + \beta_5 x_{1,t} x_{2,t} + \beta_6 x_{1,t}^3 + \beta_7 x_{2,t}^3 + u_t.$$

As these nonlinear models are linear in parameters, they can be estimated by OLS, and inference is conventional. However, the model is nonlinear, so

[3] Nonlinear approximations can also be made using alternative basis functions, such as Fourier series (sines and cosines), splines, “neural nets”, or “wavelets”.

its interpretation must take this into account. For example, in the cubic model given above, the slope with respect to $x_{1,t}$ is

$$\frac{\partial \mathcal{E}\left(y_t \left| x_t\right.\right)}{\partial x_{1,t}} = \beta_1 + 2\beta_3 x_{1,t} + \beta_5 x_{2,t} + 3\beta_6 x_{1,t}^2,$$

which is a function of $x_{1,t}$ and $x_{2,t}$, making reporting of the "slope" difficult. In many applications, it will be important to report the slopes for different values of the regressors, carefully chosen to illustrate the point of interest. In other applications, an average slope may be sufficient. There are two obvious candidates: The derivative evaluated at the sample averages

$$\left.\frac{\partial \mathcal{E}\left(y_t \left| x_t\right.\right)}{\partial x_{1,t}}\right|_{x_t=\overline{x}} = \beta_1 + 2\beta_3 \overline{x}_{1,t} + \beta_5 \overline{x}_{2,t} + 3\beta_6 \overline{x}_{1,t}^2$$

and the average derivative:

$$\frac{1}{T}\sum_{t=1}^{T}\frac{\partial \mathcal{E}\left(y_t \left| x_t\right.\right)}{\partial x_{1,t}} = \beta_1 + 2\beta_3 \overline{x}_{1,t} + \beta_5 \overline{x}_{2,t} + 3\beta_6 \frac{1}{T}\sum_{t=1}^{T} x_{1,t}^2.$$

A wider class of polynomial approximations use orthogonal polynomials that use recursive structures to obtain the n-th degree expansion of each family of polynomials. Their orthogonality properties make them especially useful for ensuring numerical stability in various computational methods.

Family	Domain	$P_0(x)$	$P_1(x)$	$P_n(x)$, $n \geq 2$
Chebyshev	$[-1,1]$	1	1	$2xP_{n-1}(x) - P_{n-2}(x)$
Legendre	$[-1,1]$	1	1	$\frac{2n-1}{n}xP_{n-1}(x) - \frac{n-1}{n}P_{n-2}(x)$
Laguerre	$[0,\infty)$	1	$1-x$	$\frac{2n-1-x}{n}P_{n-1}(x) - \frac{n-1}{n}P_{n-2}(x)$
Hermite	$(-\infty,\infty)$	1	$2x$	$2xP_{n-1}(x) - 2(n-1)P_{n-2}(x)$

If, when using the Chebyshev and Legendre polynomials, variable x is not in the [-1,1] interval, it can be rescaled by following the formula:

$$x^* = \frac{2x - \min(x) - \max(x)}{\max(x) - \min(x)}.$$

Once the polynomial expansions have been obtained, they can be used to estimate nonlinear relations by regressing y on them.

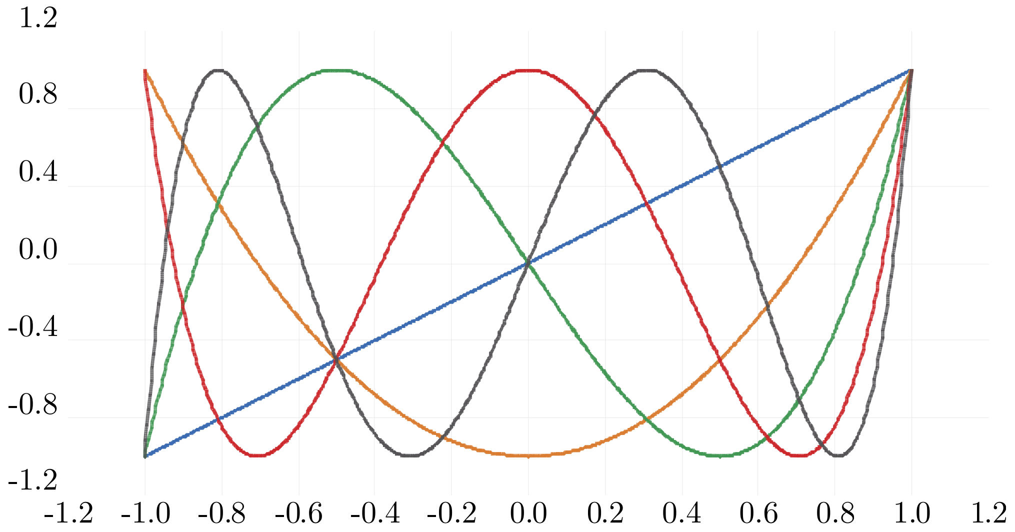

Figure 3.2: Chebyshev polynomials

3.3.3 Diagnosing Functional Form Misspecification

Once a linear model has been estimated, an important question is whether the functional form adequately captures the relationship between the regressors and the dependent variable. Functional form misspecification can lead to biased estimates, misleading inference, and poor predictive performance. In this section, we present several strategies for detecting such misspecification, with emphasis on simple diagnostics and formal testing procedures.

Augmented Regression Tests

One simple test for neglected nonlinearity is to add nonlinear functions of the regressors to the regression, and test their significance using a Wald test. If $y_t = x_t'\widehat{\beta}+\widehat{u}_t$ was estimated by OLS, let $z_t = h\left(x_t\right)$ denote nonlinear functions of x_t (perhaps squares of non-binary regressors). Fit $y_t = x_t'\widetilde{\beta} + z_t'\widetilde{\gamma} + \widetilde{u}_t$ by OLS, and form a Wald statistic for $\mathrm{H}_0 : \gamma = 0$. Rejection implies that the nonlinear terms help explain y_t and suggests functional form misspecification.

Ramsey's RESET Test

The RESET (Regression Equation Specification Error Test) is a general test for model misspecification, including functional form. The null model is

$$y_t = x_t'\beta + u_t,$$

which is estimated by OLS, yielding predicted values $\widehat{y}_t = x_t'\widehat{\beta}$. Now let

$$z_t = \begin{pmatrix} \widehat{y}_t^2 \\ \vdots \\ \widehat{y}_t^j \end{pmatrix}$$

be a $(j-1)$-vector of powers of $\widehat{y}_t$. Run the auxiliary regression

$$y_t = x_t'\widetilde{\beta} + z_t'\widetilde{\gamma} + \widetilde{u}_t \tag{3.4}$$

by OLS, and form the Wald statistic W_T for $\mathrm{H}_0 : \gamma = 0$. It is easy (although involves a somewhat lengthy derivation) to show that under the null hypothesis, $W_T \sim \chi^2_{j-1}$. Thus the null is rejected at the $\alpha\%$ level if W_T exceeds the upper $\alpha\%$ tail critical value of the χ^2_{j-1} distribution.

To implement the test, j must be selected in advance. Typically, small values such as $j = 2, 3,$ or 4 seem to work best.[4]

The RESET test appears to work well as a test of functional form against a wide range of smooth alternatives. It is particularly powerful at detecting single-index models of the form

$$y_t = G\left(x_t'\beta\right) + u_t,$$

where $G(\cdot)$ is a smooth "link" function. To see why this is the case, note that (3.4) may be written as

$$y_t = x_t'\widetilde{\beta} + \left(x_t'\widehat{\beta}\right)^2 \widetilde{\gamma}_1 + \left(x_t'\widehat{\beta}\right)^3 \widetilde{\gamma}_2 + \cdots + \left(x_t'\widehat{\beta}\right)^j \widetilde{\gamma}_{j-1} + \widetilde{u}_t,$$

which has essentially approximated $G(\cdot)$ by a j-th order polynomial.

[4] Some software packages also offer the so-called link test, which is a restricted version of the RESET test that includes only the square of the fitted values. Since the full RESET test is easy to implement and more flexible, it is preferable to use it.

Residual Diagnostics and Visual Inspection

Before conducting formal tests, it is often informative to visually inspect the residuals of the estimated model. Plotting residuals against fitted values or against individual regressors can reveal patterns that suggest functional form misspecification. For instance, if residuals fan out or exhibit a curved pattern rather than a random scatter around zero, this may indicate heteroskedasticity, omitted nonlinearities, or inappropriate variable transformations. Although residual plots are informal, they provide an intuitive and immediate sense of whether the linear specification is adequate.

At the same time, the use of residual plots demands caution. A common and problematic practice is to detect a single outlier or "weird" residual, introduce a dummy variable for that observation, and re-estimate the model. Unsurprisingly, the dummy will often appear highly significant and the fit of the model will improve substantially. However, this procedure is not valid: by identifying the outlier and constructing the dummy post hoc, we have already conditioned on the outcome. This introduces pre-testing bias — the specification search has contaminated the test. The p-value associated with the dummy's coefficient is no longer valid because the hypothesis was generated after examining the data. Furthermore, it is typically easy to construct a narrative that "explains" the dummy ex post, reinforcing a false sense of model validity.

This practice is closely related to the issue of influential observations discussed in Section 3.9, but it arises earlier — during model specification rather than robustness checking. It illustrates the broader point that specification searches based on the data can undermine inference unless properly accounted for. When residual analysis leads to changes in model specification, these changes should be viewed as exploratory, not confirmatory. If possible, the revised specification should be validated using a different sample or through cross-validation techniques to mitigate the risk of overfitting.

Nonparametric Comparisons

A more flexible and data-driven approach to checking model adequacy involves comparing the fitted values from the linear model with a nonparametric estimate of the conditional expectation. Techniques such as kernel regression, local polynomial regression, or spline smoothing allow for a flexible approximation of the relationship between the dependent variable and the regressors without imposing a specific functional form. If the nonparametric fit systematically deviates from the linear prediction, this may be evidence of misspecification. These methods serve as useful benchmarks, especially when theory does not strongly constrain the choice of functional form. While we do not develop these methods here, we return to them in Chapter [X], where they are presented as tools for both exploratory analysis and more formal specification testing.

Practical Advice

Tests for functional form are useful tools but should not be applied mechanically. The first step in any empirical exercise should be to rely on economic theory to motivate a plausible specification. Specification tests can complement this reasoning, but they cannot replace it. In small samples, the power of tests such as RESET or augmented regression tests can be limited, and failure to reject the null should not be interpreted as confirmation that the model is correct. Conversely, rejection of the null may result from factors other than functional form, including omitted variables, measurement error, or heteroskedasticity. The inclusion of nonlinear transformations or interaction terms should be guided by both theoretical plausibility and empirical performance, but care must be taken to avoid overfitting or inflating standard errors due to multicollinearity. Ultimately, model building involves trade-offs among robustness, interpretability, and tractability, and a well-specified model should balance these considerations while remaining faithful to the economic question at hand.

3.3.4 $\ln(Y)$ versus Y as Dependent Variable

An econometrician can estimate $Y = X\widehat{\beta} + \widehat{u}$ or $\ln(Y) = X\widehat{\beta} + \widehat{u}$ (or perhaps both). Which is preferable? There is a large literature on this subject, much of it quite misleading.

The plain truth is that either regression is fine, in the sense that both $\mathcal{E}(y_t \,|x_t)$ and $\mathcal{E}(\ln(y_t)\,|x_t)$ are well-defined (so long as $y_t > 0$). It is perfectly valid to estimate either or both regressions. They are different regression functions, neither is more nor less valid than the other. To test one specification versus the other, or select one specification over the other, requires the imposition of additional structure, such as the assumption that the conditional expectation is linear in x_t, and $u_t \sim \mathcal{N}(0, \sigma^2)$.[5]

There still may be good reasons for preferring the $\ln(Y)$ regression over the Y regression. First, it may be the case that $\mathcal{E}(\ln(y_t)|x_t)$ is roughly linear in x_t over the support of x_t, while the regression $\mathcal{E}(y_t|x_t)$ is nonlinear, and linear models are easier to report and interpret. Second, it may be the case that the errors in $u_t = \ln(y_t) - \mathcal{E}(\ln(y_t)|x_t)$ may be less heteroskedastic than the errors from the linear specification (although the reverse may be true!). Third, as long as $y_t > 0$, the range of $\widehat{\ln(y_t)}$ is well-defined in $\mathbb{R}$; of course, this is not the case for $\widehat{y}_t$ which for some values of x_t and $\widehat{\beta}$ may produce $\widehat{y}_t < 0$.[6] Finally, and this may be the most important reason, if the distribution of y_t is highly skewed, the conditional mean $\mathcal{E}(y_t|x_t)$ may not be a useful measure of central tendency, and estimates will be undesirably influenced by extreme observations ("outliers"). In this case, the conditional mean-ln $\mathcal{E}(\ln(y_t)|x_t)$ may be a better measure of central tendency, and hence more interesting to estimate and report. Nevertheless, we should be careful when the ln specification is used if we are interested in obtaining $\mathcal{E}(y_t|x_t)$; Jensen's inequality indicates that $\exp[\mathcal{E}(\ln(y_t)|x_t)] \neq \mathcal{E}[\exp(\ln(y_t)|x_t)]$.

3.4 Test for Normality

In econometric analysis, the normality of the error term u is not a prerequisite for the OLS estimator to be unbiased, consistent, or asymptotically normal. However, assuming normality offers certain advantages. Specifically, when residuals are normally distributed, finite sample tests (e.g., t-tests and F-tests) align with their asymptotic counterparts, allowing for direct application of these tests. In cases where residuals deviate from normality, reliance on asymptotic properties remains valid, though employing methods like the bootstrap can provide asymptotic refinements.

Moreover, if u is normally distributed, Y is also normally distributed, conditional on X. This characteristic is beneficial for constructing confidence intervals for conditional forecasts. Conversely, significant deviations from normality may indicate model misspecification. While the OLS estimator's asymptotic properties might still hold, alternative estimators could

[5] We will consider tests for non-nosted models such as these later.

[6] If that were the case $\widehat{\beta}$ would not satisfy the desirable properties we derived and we would have to use other estimation technique (for example, Tobit models).

offer greater efficiency in estimating the central tendency of the series. Understanding the nature of the non-normality can guide the selection of appropriate alternative distributions or highlight potential issues with using OLS.

A commonly employed test for assessing normality is the Jarque-Bera (JB) test.[7] This test evaluates whether the sample data exhibit skewness and kurtosis matching a normal distribution. Under the null hypothesis of normality, the JB statistic follows a chi-squared distribution with two degrees of freedom. The test statistic is calculated as:

$$JB = \frac{T}{6}\left[S^2 + \frac{(K-3)^2}{4}\right] \xrightarrow{D} \chi^2_2.$$

S represents skewness, measuring the asymmetry of the distribution around the mean. It is estimated as:

$$S = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{z_t - \overline{z}}{s_z}\right)^3,$$

where z denotes the observed values, $\overline{z}$ is the sample mean, and s_z is the sample standard deviation. A normal distribution has $S = 0$, indicating symmetry. Positive skewness (S>0) suggests a longer right tail, while negative skewness (S<0) indicates a longer left tail.

K denotes kurtosis, assessing the peakedness or flatness of the distribution:

$$K = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{z_t - \overline{z}}{s_z}\right)^4,$$

where K=3 when z is normal. Values greater than 3 (K>3) indicate a leptokurtic distribution with sharper peaks and fatter tails, while values less than 3 (K<3) suggest a platykurtic distribution with flatter peaks and thinner tails.

3.5 Measurement Errors

Consider the model

$$Y^* = X^*\beta + u.$$

[7] Other tests that can be used are the Anderson-Darling test, the Shapiro-Wilk test, and the Kolmogorov-Smirnov test.

Suppose that the econometrician does not observe Y^* or X^*, but observes $Y = Y^* + v$ and $X = X^* + w$ instead, where $v \sim (0, \sigma_v^2 I)$, $w \sim (0, \sigma_w^2 I)$.[8]

Consider first the case in which only Y^* is observed with error. Then

$$\begin{aligned} Y &= X^*\beta + u + v \\ &= X^*\beta + u^*, \end{aligned}$$

where $u^* = u + v$. This model satisfies the assumptions of the LRM, thus $\widehat{\beta}$ will be unbiased and efficient (of course, not as efficient as when Y^* is observed). Thus, when the dependent variable is measured with error, the properties of the OLS estimator are not modified.

Next consider the case in which X^* is measured with error,

$$\begin{aligned} Y^* &= (X - w)\,\beta + u \\ &= X\beta + u^*, \end{aligned}$$

where $u^* = u - w\beta$. Since $X = X^* + w$, the regressor is correlated with the disturbance, given that

$$\text{Cov}\,(X, u^*) = \text{Cov}\,(X^* + w, u - w\beta) = -\beta\sigma_w^2,$$

which violates the assumption of no correlation between the regressor and the error term. Thus, $\widehat{\beta}$ will be biased and inconsistent.

Our assumption about the source of the measurement errors is somewhat naive, as we took them as unsystematic. In general, measurement errors tend to be systematic and $\widehat{\beta}$ may be biased and inconsistent even in the first case analyzed.

3.6 Omitted Variables

Consider the model

$$\begin{aligned} &\text{Correct Model:} \quad Y = X_1\beta_1 + X_2\beta_2 + u \\ &\text{Estimated Model:} \ Y = X_1\beta_1 + u \cdot \end{aligned}$$

[8] Here, we consider $k = 1$.

If we estimate the "incorrect" model we obtain:

$$\begin{aligned}\widehat{\beta}_1 &= \left(X_1'X_1\right)^{-1} X_1'Y \\ &= \beta_1 + \left(X_1'X_1\right)^{-1} X_1'X_2\beta_2 + \left(X_1'X_1\right)^{-1} X_1'u.\end{aligned}$$

Then

$$\mathcal{E}\left(\widehat{\beta}_1\right) = \beta_1 + \underbrace{\left(X_1'X_1\right)^{-1} X_1'X_2}_{Z}\beta_2.$$

Each column of Z is the column of the slopes of the regression of X_2 on X_1. Thus the estimator of $\widehat{\beta}_1$ will generally be biased and will not be able to provide a consistent measure of $\partial Y/\partial X_1$. The estimator will be unbiased if either $Z = 0$ (which states that X_1 and X_2 are orthogonal) or if $\beta_2 = 0$ in which case, the estimated model would indeed be the correct model and this section would not have this title.

The direction of the bias from omitting relevant variables is difficult to assess in the general case; nevertheless, it can better be understood when β_1 and β_2 are scalars. In such case,

$$\mathcal{E}\left(\widehat{\beta}_1\right) = \beta_1 + \frac{\operatorname{Cov}\left(X_1, X_2\right)}{\mathcal{V}\left(X_1\right)}\beta_2.$$

The direction of the bias will depend on how X_1 and X_2 are correlated and on the sign of β_2. For example, if $\operatorname{sgn}(\operatorname{Cov}\left(X_1, X_2\right)\beta_2) > 0$, then $\mathcal{E}\left(\widehat{\beta}_1\right) > \beta_1$ and our estimator will overestimate the effect of X_1 on Y.

Let us see what happens with the variance that would be attributed to $\widehat{\beta}_1$ if the incorrect model were estimated

$$\mathcal{V}\left(\widehat{\beta}_1 | X_1\right) = \sigma^2 \left(X_1'X_1\right)^{-1}.$$

On the other hand, if we had estimated the "correct" model, $\mathcal{E}\left(\widehat{\beta}_1^*\right) = \beta_1$ and $\mathcal{V}\left(\widehat{\beta}_1^* | X_1, X_2\right)$ would have been equal to the upper left block of $\sigma^2 \left(X'X\right)^{-1}$, with $X = \left[\begin{array}{cc} X_1 & X_2 \end{array}\right]$.[9] It can be shown (in fact you proved this in our review of OLS) that

$$\mathcal{V}\left(\widehat{\beta}_1^* | X\right) = \sigma^2 \left(X_1'M_2X_1\right)^{-1},$$

[9] We denote by $\widehat{\beta}_1^*$ the estimator that would have been obtained with the correct model.

where $M_2 = I - X_2 \left(X_2' X_2\right)^{-1} X_2'$.

To compare both variance-covariance matrices, let's analyze their inverses

$$\left[\mathcal{V}\left(\widehat{\beta}_1 \left| X_1\right.\right)\right]^{-1} - \left[\mathcal{V}\left(\widehat{\beta}_1^* \left| X\right.\right)\right]^{-1} = \sigma^{-2}\left[X_1' X_2 \left(X_2' X_2\right)^{-1} X_2' X_1\right], \quad (3.5)$$

which is positive definite.

Thus, we may be inclined to conclude that although $\widehat{\beta}_1$ is biased it has a smaller variance than $\widehat{\beta}_1^*$. Nevertheless, recall that σ^2 is not known and needs to be estimated. Proceeding as usual (thinking that the estimated model is correct) we would obtain

$$\widetilde{\sigma}^2 = \frac{\widehat{u}'\widehat{u}}{T - k_1},$$

but $\widehat{u} = M_1 Y = M_1 \left(X_1\beta_1 + X_2\beta_2 + u\right) = M_1 X_2 \beta_2 + M_1 u$. Then

$$\begin{aligned}\mathcal{E}\left(\widehat{u}'\widehat{u}\right) &= \beta_2' X_2' M_1 X_2 \beta_2 + \sigma^2 \text{tr}\left(M_1\right) \\ &= \beta_2' X_2' M_1 X_2 \beta_2 + \sigma^2 \left(T - k_1\right).\end{aligned}$$

The first term is the population counterpart to the increase in the SSR due to dropping X_2 from the regression. As this term is positive, $\widetilde{\sigma}^2$ will be biased upward (the true variance is smaller). Unfortunately, to take into account this bias we would require to know β_2.

In conclusion, if we omit a relevant variable from the regression, both $\widehat{\beta}_1$ and $\widetilde{\sigma}^2$ are biased. Even when $\widehat{\beta}_1$ may be more precise than $\widehat{\beta}_1^*$, this should provide us little comfort since we cannot estimate σ^2 consistently. The only case in which $\widehat{\beta}_1$ would be unbiased is if X_1 and X_2 were orthogonal.

3.7 Irrelevant Variables

Consider the model

$$\begin{aligned}&\text{Correct Model:} \quad Y = X_1\beta_1 + u \\ &\text{Estimated Model:} \; Y = X_1\beta_1 + X_2\beta_2 + u \cdot\end{aligned}$$

If we estimate the "incorrect" model we obtain:

$$\begin{aligned}\widehat{\beta}_1 &= \left(X_1' M_2 X_1\right)^{-1} X_1' M_2 Y \\ &= \beta_1 + \left(X_1' M_2 X_1\right)^{-1} X_1' M_2 u.\end{aligned}$$

Then

$$\mathcal{E}\left(\widehat{\beta}_1\right) = \beta_1.$$

In fact,

$$\mathcal{E}\left(\widehat{\beta}\right) = \mathcal{E}\begin{bmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{bmatrix} = \begin{bmatrix} \beta_1 \\ 0 \end{bmatrix}.$$

By the same reasoning we can prove that

$$\mathcal{E}\left(\widetilde{\sigma}^2\right) = \mathcal{E}\left(\frac{\widehat{u}'\widehat{u}}{T - k_1 - k_2}\right) = \sigma^2.$$

Then what is the problem? It would seem that it is preferred to "overfit" the model. The cost of "overfitting" is the reduction in the precision of the estimators. Recall that

$$\widehat{\beta}_1 = \beta_1 + \left(X_1' M_2 X_1\right)^{-1} X_1' M_2 u,$$

then

$$\mathcal{V}\left(\widehat{\beta}_1 \left| X\right.\right) = \sigma^2 \left(X_1' M_2 X_1\right)^{-1}.$$

As we proved in (3.5), the variance of $\widehat{\beta}_1$ is larger than if the correct model were estimated, because in such a case

$$\mathcal{V}\left(\widehat{\beta}_1^* \left| X_1\right.\right) = \sigma^2 \left(X_1' X_1\right)^{-1}.$$

Both estimators would have equal asymptotic efficiency if X_1 and X_2 were orthogonal. On the other hand, if X_1 and X_2 were highly correlated, including X_2 would greatly inflate the variance of the estimator.

3.8 Multicollinearity

Multicollinearity arises when the measured variables are too highly intercorrelated to allow for precise analysis of the individual effects of each one. In this section we will discuss its nature, possible ways to detect it, its effects, and "remedies".

3.8.1 Perfect Collinearity

If $\text{rank}(X'X) < k$, then $\widehat{\beta}$ is not defined. This is defined as multicollinearity. This happens if and only if the columns of X are linearly dependent. Most commonly, this arises when sets of regressors which are included are identically related. For example, let X include logs of two prices and the log of the relative price $\ln(p_1), \ln(p_2)$ and $\ln(p_1/p_2)$. When this happens, the applied researcher quickly discovers the error, as the statistical software will be unable to construct $(X'X)^{-1}$. Since the error is quickly discovered, this is rarely a problem of applied econometric practice. Thus, the problem with multicollinearity is not with the data, but with a bad specification.

3.8.2 Near Multicollinearity

It is often argued that in contrast to perfect collinearity (where the problem arises from specification), near multicollinearity is a statistical "problem". The problem in estimation is not identification but precision. Indeed, the higher the correlation between regressors, the less precise will be the estimates. What is troubling about this definition of a "problem" is that our complaint is with the sample that was given to us!

The usual "symptoms" of the "problem" are:

- Small changes in the data produce wide swings in parameter estimates.
- While the t statistics of the parameters estimates are low (not significant), the R^2 is high.[10]
- The coefficients have the wrong sign or implausible magnitudes.

The problem arises when $X'X$ is "near singular" and the columns of X are close to linear dependence.[11] One implication of near singularity of matrices is that the numerical reliability of the calculations is reduced. It is more likely that the reported calculations will be in error due to floating-point calculation difficulties.

[10] A pervasive practice is to use near multicollinearity as an excuse for bad specifications when t statistics are low.

[11] This definition is not precise as we have not said what it means for a matrix to be near singular.

As the problem is with $(X'X)^{-1}$, let us take a closer look to it. The j-th element of the diagonal of $(X'X)^{-1}$ is (we let $j = 1$ for convenience):

$$\begin{aligned}
(x_1' M_2 x_1)^{-1} &= \left(x_1'x_1 - x_1'X_2\left(X_2'X_2\right)^{-1}X_2'x_1\right)^{-1} \\
&= \left(x_1'x_1\left(1 - \frac{x_1'X_2\left(X_2'X_2\right)^{-1}X_2'x_1}{x_1'x_1}\right)\right)^{-1} \\
&= \left(x_1'x_1\left(1 - R_1^2\right)\right)^{-1} \\
&= \frac{1}{x_1'x_1\left(1 - R_1^2\right)},
\end{aligned}$$

where X_2 is the $T \times (k-1)$ matrix of X that excludes x_1 and R_1^2 is the (uncentered) R^2 of the regression of x_1 on the other regressors. Thus,

$$\mathcal{V}\left(\widehat{\beta}_1\right) = \frac{\sigma^2}{x_1'x_1\left(1 - R_1^2\right)}.$$

If we have a set of regressors that is highly correlated to x_1, then R_1^2 will tend to 1 and $\mathcal{V}\left(\widehat{\beta}_1\right) \to \infty$.

Detection

A rule of thumb that has been suggested is that we should be concerned with multicollinearity when the overall R^2 in the regression is lower than any R_j^2. This rule is of course suggestive as it does not tell us how to proceed.

An alternative measure of collinearity has been proposed by Belsley and is based on the conditioning number (γ), which is defined as:

$$\gamma = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}},$$

where the λ's are the eigenvalues of $B = S\left(X'X\right)S$, with $S = \text{diag}\left(1/\sqrt{x_j'x_j}\right)$. That is:

$$S = \begin{bmatrix}
\frac{1}{\sqrt{x_1'x_1}} & 0 & \cdots & 0 \\
0 & \frac{1}{\sqrt{x_2'x_2}} & 0 & \vdots \\
\vdots & 0 & \ddots & 0 \\
0 & \cdots & 0 & \frac{1}{\sqrt{x_k'x_k}}
\end{bmatrix}.$$

If the regressors are orthogonal ($R_j^2 = 0 \; \forall j$), γ will equal one. The higher the intercorrelation, the higher the conditioning number will be.[12] Belsley suggests that values of γ in excess of 20 indicate potential problems.[13]

If we conclude that there is a potential "problem" of multicollinearity, how do we deal with it? Three approaches are usually suggested:

- Reduce the dimension of X (drop variables). The obvious problem is that as the variables that are omitted were "relevant", $\widehat{\beta}$ will be biased (as we already discussed). Thus, this practice makes explicit the trade-off between variance reduction and bias.
- Principal components
- Ridge regression

Principal Components

Take the model

$$Y = X\beta + u.$$

Consider the transformation

$$\begin{aligned} Y &= XPP'\beta + u \\ &= XP\theta + u \\ &= Z\theta + u, \end{aligned}$$

where

$$\underset{k\times k}{P} = \begin{bmatrix} p_1 & \cdots & p_k \end{bmatrix}$$

and p_j is the jth orthogonal eigenvector (characteristic vector) of $X'X$. These eigenvectors are ordered by the order of magnitude of the corresponding eigenvalues.[14] Thus,

$$\underset{T\times k}{Z} = \begin{bmatrix} z_1 & \cdots & z_k \end{bmatrix}$$

is the matrix of principal components, i.e. $z_j = Xp_j$ is called the jth principal component, where $z_j' z_j = \lambda_j$.[15]

[12] If perfect collinearity is present, $\lambda_{\min} = 0$, and $\gamma \to \infty$.

[13] GAUSS tip: γ^2 of the matrix B can be obtained using the command `cond(B)`.

[14] The matrix P satisfies the condition $PP' = P'P = I$.

[15] λ_j denotes the jth largest eigenvalue of $X'X$.

The principal components estimator of β is obtained by "deleting" one or more of the z_j, applying OLS to the reduced model, and transforming the estimator obtained to the original parameter space.

This is, partition $X \left[P_1 \;\; P_2 \right] = \left[Z_1 \;\; Z_2 \right]$, then

$$\begin{aligned} Y &= XP_1\theta_1 + XP_2\theta_2 + u \\ &= Z_1\theta_1 + Z_2\theta_2 + u. \end{aligned}$$

If we omit Z_2 from the model, we obtain

$$\widehat{\theta}_1 = \left(Z_1'Z_1\right)^{-1} Z_1'Y.$$

As Z_1 and Z_2 are orthogonal, $\widehat{\theta}_1$ is unbiased. Furthermore, $\mathcal{V}\left(\widehat{\theta}_1\right) = \sigma^2 \left(Z_1'Z_1\right)^{-1}$.

This estimator has desirable properties for θ_1 but not for the actual parameters of interest (β). Now we'll discuss a transformation of θ_1 back into β that is usually proposed.

Notice that $\beta = P\theta = P_1\theta_1 + P_2\theta_2$. By omitting Z_2 we implicitly assumed that θ_2 is equal to zero in which case $\widehat{\beta}^* = P_1\widehat{\theta}_1$ would be the principal components estimator of β.

As discussed in Section 3.6, it is easy to prove that $\mathcal{V}\left(\widehat{\beta}^*\right) < \mathcal{V}\left(\widehat{\beta}\right)$. However, this estimator will be biased unless $P_2\theta_2 = 0$.

Until now we have remained silent with respect to how to choose Z_2. Two approaches have been suggested:

- Include in Z_2 the components with the smallest eigenvalues. This amounts to assuming that near collinearity is equivalent to perfect collinearity, which may not be a good strategy.

- Test for $P_2'\theta_2 = 0$ (which is not a trivial task).

Ridge Regression

Let $\Lambda = P'X'XP =$diag$(\lambda_1, \ldots, \lambda_k)$ be the diagonal matrix of eigenvalues of $X'X$ (as before, P is the matrix of eigenvectors of $X'X$). The Generalized Ridge Regression estimator (GRR) is defined by

$$\widetilde{\theta} = \left(\Lambda + W\right)^{-1} Z'Y = \left(\Lambda + W\right)^{-1} \Lambda\widehat{\theta},$$

where

$$W = \operatorname{diag}(w_1, \ldots, w_k), \quad w_i > 0,$$

and

$$\widehat{\theta} = \Lambda^{-1} Z'Y$$

is the OLS estimator of θ. Recalling that $\theta = P'\beta$, the GRR estimator of β is

$$\widetilde{\beta} = P\widetilde{\theta}.$$

The GRR estimator depends on the choice of W. It can be shown that the values of w_i that minimize the MSE of $\widetilde{\beta}$ are given by

$$w_i = \frac{\sigma^2}{\theta_i^2},$$

where θ_i is the i-th element of θ. An operational estimator can be obtained by replacing σ^2 and θ_i with their OLS estimates:

$$\widehat{w}_i = \frac{\widetilde{\sigma}^2}{\widehat{\theta}_i^2}.$$

A simpler version of the estimator, called the Ordinary Ridge Regression estimator (ORR), is obtained by setting $W = wI$:

$$\widetilde{\beta}_{ORR} = \left(X'X + wI\right)^{-1} X'Y.$$

While no explicit optimum value for w can be found, several stochastic choices have been proposed.[16] Among the most popular we have:

$$\widehat{w}^I = \frac{k\widetilde{\sigma}^2}{\widehat{\beta}'\widehat{\beta}} \quad \text{and} \quad \widehat{w}^{II} = \frac{k\widetilde{\sigma}^2}{\widehat{\beta}'X'X\widehat{\beta}}.$$

Even though the Ridge Regression estimators may have smaller MSE than OLS, it is important to mention several drawbacks:

- There is no consensus with respect to the choice of the shrinkage parameter.
- The shrinkage parameter does not have a standard distribution.
- Because of this, the distribution of the Ridge Regression estimator of β will also be non-standard in which case inference cannot be conducted with the usual tests (specially in small samples).

[16] w is referred to as the shrinkage parameter.

3.8.3 Bottom Line

If you are uncomfortable with the reasoning above, you are not alone. There is no pair of words that is more misused both in econometrics texts and in the applied literature than the pair "multicollinearity problem". That many of the explanatory variables used in econometrics are highly collinear is a fact of life. It is perfectly clear that there are realizations of $X'X$ which would be much preferred to the actual data. But a complaint about the apparent malevolence of nature is not at all constructive, and the *ad-hoc* cures for a "bad" sample, such as the ones outlined, can be disastrously inappropriate. It would be better if we'd accept the fact that our non-experimental data is sometimes not very informative about the parameters of interest.

An example may clarify what we are really talking about. Consider the two-variable linear model $y_t = \beta_1 x_{1,t} + \beta_2 x_{2,t} + u_t$ and suppose that a regression of x_2 on x_1 yields the result $x_{2,t} = \widehat{\theta} x_{1,t} + \widehat{u}_t$, where $\widehat{u}$ is, by construction, orthogonal to x_1. Substitute this auxiliary relationship into the original one to obtain the model

$$\begin{aligned} y_t &= \beta_1 x_{1,t} + \beta_2 \left(\widehat{\theta} x_{1,t} + \widehat{u}_t\right) + u_t \\ &= \left(\beta_1 + \beta_2 \widehat{\theta}\right) x_{1,t} + \beta_2 \widehat{u}_t + u_t \\ &= \delta_1 z_{1,t} + \delta_2 z_{2,t} + u_t, \end{aligned}$$

where $\delta_1 = \left(\beta_1 + \beta_2 \widehat{\theta}\right), \delta_2 = \beta_2, z_1 = x_1$, and $z_2 = x_2 - \widehat{\theta} x_1$. A researcher who used the variables x_1 and x_2 and the parameters β_1 and β_2 might report that β_2 is estimated inaccurately because of the collinearity problem. But a researcher who happened to stumble on the model with variables z_1 and z_2 and parameters δ_1 and δ_2 would report that there is no collinearity problem because z_1 and z_2 are orthogonal (recall that x_1 and $\widehat{u}$ are orthogonal). This researcher would nonetheless report that $\delta_2 (= \beta_2)$ is estimated inaccurately, not because of collinearity, but because z_2 does not vary adequately.[17]

What this example illustrates is that collinearity as a cause of weak evidence is indistinguishable from inadequate variability as a cause of weak evidence. In light of that fact, it is surprising that all econometrics texts have sections dealing with the "collinearity problem" but none has a section on the "inadequate variability problem".

[17] Recall that if x_1 and x_2 are highly collinear, $\widehat{u}$ will not fluctuate that much given that $\widehat{u}'\widehat{u}$ (the SSR of the regression of x_2 on x_1) would be small.

In summary, collinearity is bound to be present in applied econometric practice. If we use principal components we will usually encounter problems in interpreting the results, given that they come from a combination of parameters. On the other hand, Ridge Regression estimators have a non-standard distribution for the parameters of interest. Thus, is there a simple solution to this "problem"? Basically, no. Fortunately, multicollinearity does not lead to errors in inference. The asymptotic distribution is still valid. OLS estimates are asymptotically normal, and estimated standard errors are consistent. So reported confidence intervals are not inherently misleading. They will be large, correctly indicating the inherent uncertainty about the true parameter values.

3.9 Influential Analysis

OLS seeks to prevent a few large residuals at the expense of incurring into many relatively small residuals, only a few observations can be extremely influential in the sense that dropping them from the sample changes some elements of $\widehat{\beta}$ substantially. There is a systematic way to find those influential observations. Let $\widehat{\beta}^{(t)}$ be the OLS estimate of β that would be obtained if OLS were used on a sample from which the t-th observation were omitted. The key equation is

$$\widehat{\beta}^{(t)} - \widehat{\beta} = -\left(\frac{1}{1-p_t}\right)(X'X)^{-1} x_t \widehat{u}_t, \tag{3.6}$$

where p_t is defined as

$$p_t \equiv x_t'(X'X)^{-1} x_t,$$

which is the t-th diagonal element of the projection matrix P. It is easy to show that

$$0 \leq p_t \leq 1 \text{ and } \sum_{t=1}^{T} p_t = k, \tag{3.7}$$

so p_t equals k/T on average.

To illustrate the use of (3.6) in a specific example, consider the relationship between the monetary policy rate and economic growth in Chile between 1986 and 1999. Figure **??** plots the quarterly GDP growth rate against to policy rate. It is clear from the first panel that the position of the estimated

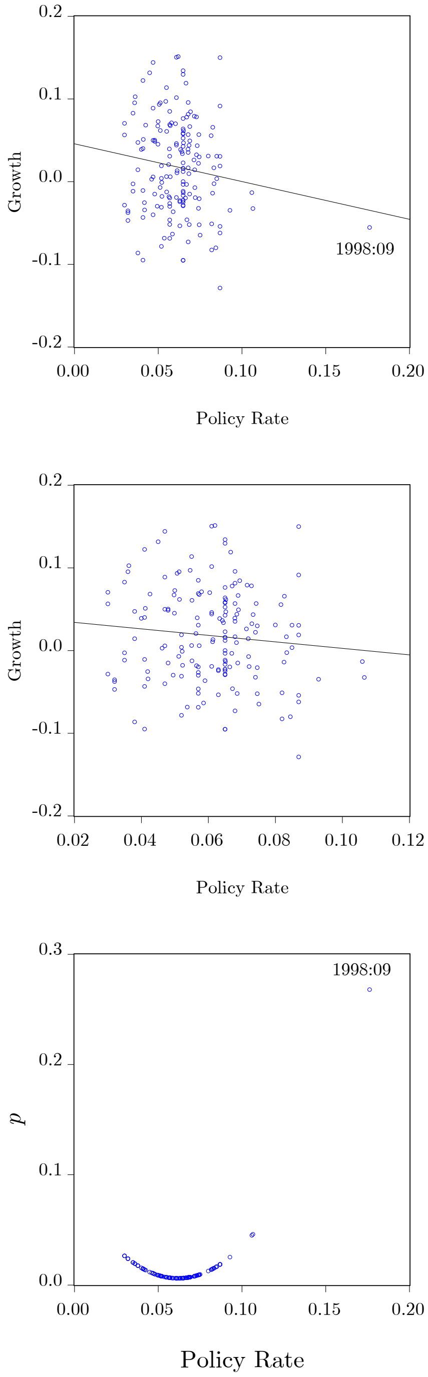

Figure 3.3: Monetary Policy Rate, Growth, and p_t

regression line depends very much on the single outlier (September of 1998, when the interest rate increased to almost 18%!). Indeed if this observation is dropped from the sample (second panel of Figure 3.3), the estimated slope coefficient drops (in absolute value) from -0.46 to -0.39.[18] In the case of a simple regression, it is easy to spot outliers by visually inspecting a plot such as Figure 3.3. This strategy would not work if there was more than one nonconstant regressor. Analysis based on (3.6) is obviously not restricted to simple regressions. The third panel of Figure 3.3 displays the association between the policy rate and p_t. As is evident from that figure, the value of p_t for September of 1998 is 0.268; a value which is well above the average of 0.012 ($= k/T = 2/168$) and is highly influential.[19] Note that we could not have detected the influential observation by looking at the residuals, which is not surprising because the algebra of OLS is designed to avoid large residuals at the expense of many small residuals for other observations.

What should be done with influential observations? It depends. If the influential observations satisfy the regression model, they provide valuable information about the regression function unavailable from the rest of the sample and should definitely be kept in the sample. In case the influential observations are atypical of the rest of the sample some practitioners suggest to drop them from the sample. Yet others prefer to use other estimates than OLS to measure central tendency.

3.10 Model Selection

We have discussed the costs and benefits of inclusion/exclusion of variables. How does a researcher go about selecting an econometric specification, when economic theory does not provide complete guidance? This is the question of model selection. It is important that the model selection question be well-posed. For example, the question: "What is the right model for y?" is not well posed, because it does not make clear the conditioning set. In contrast, the question, "Which subset of $(x_1, \cdots, x_k)$ enters the regression function $\mathcal{E}\left(y_t \,|x_{1,t}, \cdots, x_{k,t}\right)$?" is well posed.

In many cases the problem of model selection can be reduced to the comparison of two nested models, as the larger problem can be written as a sequence of such comparisons. We thus consider the question of the inclusion

[18] However, neither of this coefficients is statistically significant at standard levels.

[19] In fact, this value is 22.5 times higher than the average!

of X_2 in the linear regression

$$Y = X_1\beta_1 + X_2\beta_2 + u,$$

where X_1 is $T \times k_1$ and X_2 is $T \times k_2$. This is equivalent to the comparison of two models

$$\mathcal{M}_1 : Y = X_1\beta_1 + u$$
$$\mathcal{M}_2 : Y = X_1\beta_1 + X_2\beta_2 + u \cdot$$

Note that $\mathcal{M}_1 \subset \mathcal{M}_2$. To be concrete, we say that $\mathcal{M}_2$ is true if $\beta_2 \neq 0$. To fix notation, models 1 and 2 are estimated by OLS, with residual vectors $\widehat{u}_1$ and $\widehat{u}_2$, estimated variances $\widehat{\sigma}_1^2$ and $\widehat{\sigma}_2^2$, etc., respectively.

A model selection procedure is a data-dependent rule which selects one of the models. We can write this as $\widehat{\mathcal{M}}$. There are many possible desirable properties for a model selection procedure. One useful property is consistency, that it selects the true model with probability one if the sample is sufficiently large. A model selection procedure is consistent if

$$\Pr\left[\widehat{\mathcal{M}} = \mathcal{M}_1 \left| \mathcal{M}_1 \right.\right] \rightarrow 1$$
$$\Pr\left[\widehat{\mathcal{M}} = \mathcal{M}_2 \left| \mathcal{M}_2 \right.\right] \rightarrow 1.$$

We now discuss a number of possible model selection methods.

3.10.1 Selection Based on Fit

Natural measures of the fit of a regression are SSR $(\widehat{u}'\widehat{u})$, $R^2 = 1 - (\widehat{u}'\widehat{u})/\widehat{\sigma}_y^2$ or Gaussian log-likelihood $\ell\left(\widehat{\beta}, \widehat{\sigma}^2\right) = -(T/2)\ln\widehat{\sigma}^2 + a$ (where a is a constant). It might therefore be thought attractive to base a model selection procedure on one of these measures of fit. The problem is that each of these measures are necessarily monotonic between nested models, namely $\widehat{u}_1'\widehat{u}_1 \geq \widehat{u}_2'\widehat{u}_2$, $R_1^2 \leq R_2^2$, and $\ell_1 \leq \ell_2$, so model $\mathcal{M}_2$ would always be selected, regardless of the actual data and probability structure. This is clearly an inappropriate decision rule!

3.10.2 Selection Based on Testing

A common approach to model selection is to base the decision on a statistical test such as the Wald W_T

$$W_T = T\left(\frac{\widehat{\sigma}_1^2 - \widehat{\sigma}_2^2}{\widehat{\sigma}_2^2}\right).$$

The model selection rule is as follows; for some critical level α, let c_α satisfy $\Pr\left[\chi^2_{k_2} > c_\alpha\right]$. Then select $\mathcal{M}_1$ if $W_T \leq c_\alpha$, else select $\mathcal{M}_2$.

The major problem with this approach is that the critical level α is indeterminate. The reasoning which helps guide the choice of α in hypothesis testing (controlling Type I error) is not relevant for model selection. That is, if α is set to be a small number, then $\Pr\left[\widehat{\mathcal{M}} = \mathcal{M}_1 \left| \mathcal{M}_1\right.\right] \approx 1-\alpha$ but $\Pr\left[\widehat{\mathcal{M}} = \mathcal{M}_2 \left| \mathcal{M}_2\right.\right]$ could vary dramatically, depending on the sample size, etc. Another problem is that if α is held fixed, then this model selection procedure is inconsistent, as

$$\Pr\left[\widehat{\mathcal{M}} = \mathcal{M}_1 \left| \mathcal{M}_1\right.\right] \to 1-\alpha < 1.$$

3.10.3 Selection Based on Adjusted R-squared

Since R^2 is not a useful model selection rule, as it always "prefers" the larger model, Theil proposed an adjusted coefficient of determination

$$\overline{R}^2 = 1 - \frac{\left(\widehat{u}'\widehat{u}\right)/(T-k)}{\widehat{\sigma}_y^2} = 1 - \frac{\widetilde{\sigma}^2}{\widehat{\sigma}_y^2}.$$

At one time, it was popular to pick between models based on $\overline{R}^2$. This rule is to select $\mathcal{M}_1$ if $\overline{R}_1^2 > \overline{R}_2^2$, else select $\mathcal{M}_2$. Since $\overline{R}^2$ is a monotonically decreasing function of $\widetilde{\sigma}^2$, this rule is the same as selecting the model with the smaller $\widetilde{\sigma}^2$, or equivalently, the smaller $\ln\left(\widetilde{\sigma}^2\right)$. It is helpful to observe that

$$\begin{aligned}\ln\left(\widetilde{\sigma}^2\right) &= \ln\left(\widehat{\sigma}^2\frac{T}{T-k}\right) \\ &= \ln\left(\widehat{\sigma}^2\right)+\ln\left(1+\frac{k}{T-k}\right) \\ &\simeq \ln\left(\widehat{\sigma}^2\right)+\frac{k}{T-k} \\ &\simeq \ln\left(\widehat{\sigma}^2\right)+\frac{k}{T}\end{aligned}$$

(the first approximation is $\ln(1+w)\simeq w$ for small w). Thus selecting based on $\overline{R}^2$ is the same as selecting based on $\ln\left(\widehat{\sigma}^2\right)+\frac{k}{T}$, which is a particular choice of penalized likelihood criteria. It turns out that model selection based on any criterion of the form

$$\ln\left(\widehat{\sigma}^2\right)+c\frac{k}{T}, \qquad c>0 \tag{3.8}$$

is inconsistent, as the rule tends to overfit. Indeed, since under $\mathcal{M}_1$,

$$T\left(\ln\widehat{\sigma}_1^2-\ln\widehat{\sigma}_2^2\right)\simeq W_T\sim\chi^2_{k_2}, \tag{3.9}$$

$$\begin{aligned}\Pr\left[\widehat{\mathcal{M}}=\mathcal{M}_1\,|\mathcal{M}_1\right] &= \Pr\left[\overline{R}_1^2>\overline{R}_2^2\,|\mathcal{M}_1\right] \\ &\simeq \Pr\left[T\ln\left(\widetilde{\sigma}_1^2\right)<T\ln\left(\widetilde{\sigma}_2^2\right)|\mathcal{M}_1\right] \\ &\simeq \Pr\left[T\ln\left(\widehat{\sigma}_1^2\right)+ck_1<T\ln\left(\widehat{\sigma}_2^2\right)+c\left(k_1+k_2\right)|\mathcal{M}_1\right] \\ &= \Pr\left[W_T<ck_2\,|\mathcal{M}_1\right] \\ &\to \Pr\left[\chi^2_{k_2}<ck_2\right]<1.\end{aligned}$$

3.10.4 Selection Based on Information Criteria

Akaike Information Criterion

Akaike proposed an information criterion which takes the form

$$AIC=-\frac{2\ell}{T}+2\frac{k}{T},$$

which with a Gaussian log-likelihood can be approximated by (3.8) with $c = 2$:

$$AIC \simeq \ln\left(\widehat{\sigma}^2\right) + 2\frac{k}{T}.$$

This imposes a larger penalty on overparameterization than does $\overline{R}^2$. Akaike's motivation for this criterion is that a good measure of the fit of a model density $f(Y|X, \mathcal{M})$ to the true density $f(Y|X)$ is the Kullback distance $K(\mathcal{M}) = \mathcal{E}[\ln f(Y|X) - \ln f(Y|X, \mathcal{M})]$. The log-likelihood function provides a decent estimate of this distance, but it is biased, and a better, less-biased estimate can be obtained by introducing the penalty $2k$. The actual derivation is not very enlightening, and the motivation for the argument is not fully satisfactory, so we omit the details. Despite these concerns, the AIC is a popular method for model selection. The rule is to select $\mathcal{M}_1$ if $AIC_1 < AIC_2$, else select $\mathcal{M}_2$.

Since the AIC takes the form (3.8), it is an inconsistent model selection criterion, and tends to overfit.

Schwarz Criterion

While many modifications of the AIC have been proposed, the most popular appears to be one proposed by Schwarz, based on Bayesian arguments. His criterion, known as the BIC (for Bayesian Information Criterion), is

$$BIC = -\frac{2\ell}{T} + \frac{k}{T}\ln(T),$$

which with a Gaussian log-likelihood can be approximated by

$$BIC \simeq \ln\left(\widehat{\sigma}^2\right) + \frac{k}{T}\ln(T).$$

Since $\ln(T) > 2$ (if $T > 8$), the BIC places a larger penalty than the AIC on the number of estimated parameters and is more parsimonious.

In contrast to the other methods discussed above, BIC model selection is consistent. Indeed, since (3.9) holds under $\mathcal{M}_1$,

$$\frac{W_T}{\ln(T)} \xrightarrow{p} 0,$$

so

$$\begin{aligned}\Pr\left[\widehat{\mathcal{M}}=\mathcal{M}_1\left|\mathcal{M}_1\right.\right] &= \Pr\left[BIC_1<BIC_2\left|\mathcal{M}_1\right.\right]\\ &= \Pr\left[W_T<k_2\ln\left(T\right)\left|\mathcal{M}_1\right.\right]\\ &= \Pr\left[\frac{W_T}{\ln\left(T\right)}<k_2\left|\mathcal{M}_1\right.\right]\\ &\to \Pr\left(0<k_2\left|\mathcal{M}_1\right.\right)=1.\end{aligned}$$

Also under $\mathcal{M}_2$, one can show that

$$\frac{W_T}{\ln\left(T\right)}\xrightarrow{p}\infty,$$

thus

$$\begin{aligned}\Pr\left[\widehat{\mathcal{M}}=\mathcal{M}_2\left|\mathcal{M}_2\right.\right] &= \Pr\left[BIC_2<BIC_1\left|\mathcal{M}_2\right.\right]\\ &= \Pr\left[\frac{W_T}{\ln\left(T\right)}>k_2\left|\mathcal{M}_2\right.\right]\\ &\to 1.\end{aligned}$$

Hannan-Quinn Criterion

Yet another popular model selection criterion is known as the HQC that is defined as

$$HQC=-\frac{2\ell}{T}+2\frac{k}{T}\ln\left(\ln\left(T\right)\right),$$

which with a Gaussian log-likelihood can be approximated by

$$HQC\simeq\ln\left(\widehat{\sigma}^2\right)+2\frac{k}{T}\ln\left(\ln\left(T\right)\right).$$

Since $\ln\left(\ln\left(T\right)\right)>1$ (for $T>15$), the HQC places a larger penalty than the AIC on the number of estimated parameters and is more parsimonious. In turn, as $2\ln\left(\ln\left(T\right)\right)<\ln\left(T\right)$ $(\forall T>0)$, the BIC places a larger penalty than the HQC and selects more parsimonious models. As is the case with BIC, HQC is consistent.

A Final Word of Caution

All the results derived were obtained in the OLS context with Gaussian innovations. Although the conclusions at which we arrived concerning the model selection criteria will not be affected, in more general cases, the exact formulas for each criterion will depend on ℓ and not just $\widehat{\sigma}^2$.

Another important point that cannot be ignored is that in order to compare different models with any of these criteria, both the dependent variable and the sample size need to be the same.

Which model selection criterion is the "best" is still an open question and an active field of research. While consistency is a desirable property, there may be cases in which more parsimonious models run the risk of excluding relevant variables and that is why some researchers prefer HQC which is consistent and not as parsimonious as BIC. From a practical standpoint, it is important to look at the three criteria. Who knows, they may all choose the same the model!

3.10.5 Selection Among Multiple Regressors

We have discussed model selection between two models. The methods extend readily to the issue of selection among multiple regressors. The general problem is the model

$$y_t = \beta_1 x_{1,t} + \beta_2 x_{2,t} + \cdots + \beta_k x_{k,t} + u_t,$$

and the question is which subset of the coefficients are non-zero (equivalently, which regressors enter the regression).

There are two leading cases: ordered and unordered regressors. In the ordered case, the models are:

$$\begin{aligned}
\mathcal{M}_1 &: \ \beta_1 \neq 0, \beta_2 = \beta_3 = \cdots = \beta_k = 0 \\
\mathcal{M}_2 &: \ \beta_1 \neq 0, \beta_2 \neq 0, \beta_3 = \cdots = \beta_k = 0 \\
&\ \ \vdots \\
\mathcal{M}_k &: \ \beta_1 \neq 0, \beta_2 \neq 0, \beta_3 \neq 0, \cdots, \beta_k \neq 0,
\end{aligned}$$

which are nested. The selection criterion uses the estimates of the k models by OLS, stores the residual variances $\widehat{\sigma}^2$ for each model, and then selects the model that minimizes it.

In the unordered case, a model consists of any possible subset of the regressors $\{x_{1,t}, \cdots, x_{k,t}\}$, and the selection criterion can be implemented by estimating all possible subset models. However, there are 2^k such models, which can be a very large number. For example, $2^{10} = 1024$, and $2^{20} = 1,048,576$. In the latter case, a full-blown implementation of the chosen model selection criterion would be computationally demanding.

3.11 Specification Searches

Economic theory often is vague about the relationship between economic variables. As a result, many economic relations have been initially established from apparent empirical regularities and had not been predicted *ex ante* by theory. In the limited sample sizes typically encountered in economic studies, systematic patterns and apparently significant relations are bound to occur if the data are analyzed with sufficient intensity.[20] If not accounted for, this practice, referred to as data mining, can generate serious biases in statistical inference.[21]

The data miner's strategy is revealed by considering some typical quotations in applied research:

> "Because of space limitations, only the best of a variety of alternative models are presented here."
>
> "The precise variables included in the regression were determined on the basis of extensive experimentation (on the same body of data)."
>
> "Since there is no firmly validated theory, we avoided *a priori* specification of the functions we wished to fit."
>
> "We let the data specify the model."

The estimation and hypothesis testing procedures discussed so far, are valid only when *a priori* considerations rather than exploratory data mining

[20] A colorful example known as the newsletter scam is instructive: One selects a large number of individuals to receive a free copy of a stock market newsletter; to half the group one predicts the market will go up next week; to the other half, that the market will go down. The next week, one sends the free newsletter only to those who received the correct prediction; again, half are told the market will go up and half down. The process is repeated several times and a few months later, the group that received perfect predictions is asked to pay for such "good" forecasts.

[21] Other names for it are: data snooping, data grubbing, and data fishing.

determine the set of variables to be included in a regression. When the data miner uncovers t-statistics that appear significant at the 0.05 level by running a large number of alternative regressions on the same body of data, the probability of a Type I error of rejecting the null hypothesis when it is true is much greater than the claimed 5%.

When all the candidate explanatory variables are orthogonal and the variance of the innovation is known, Lovell (1983) presents a rule of thumb for assessing the true significance level when data mining has taken place. When a search has been conducted for the best k out of c candidate explanatory variables, a regression coefficient that appears to be significant at the level $\widehat{\alpha}$ should be regarded as significant at only level

$$\alpha = 1 - (1 - \widehat{\alpha})^{c/k}$$

or, as a short cut guide, the significance level is approximately

$$\alpha \approx \frac{c}{k}\widehat{\alpha}.$$

As an example, assume that we are interested in estimating the demand for real money holdings and consider the following model:

$$m_t = a + bi_t + u_t$$

where m is the log of real money holdings and i is "the" interest rate. As there are several candidates for i we search among $c = 10$ of them. If we find one of them to be significant at $\widehat{\alpha}$, the true level should be approximately $10{\cdot}\widehat{\alpha}$.

This approximation assumes that the variance of u is known and that the candidate interest rates are orthogonal between each other. Neither of these assumptions are realistic. White (2000) presents a general strategy for analyzing models that were subject to data mining.

3.12 Further Reading

For classical treatments of functional form, omitted variables, and specification analysis, foundational discussions can be found in Amemiya (1985), Greene (1993), and Hayashi (2000). Hansen (2022) and Davidson and MacKinnon (2004) provide modern textbook coverage with clarity and precision, particularly on nonlinearity diagnostics and model selection.

Leamer (1983a) offers a sharp critique of uncritical specification search, a theme developed more formally in his contribution to the Handbook of Econometrics (1983). Lovell (1983) and White (2000) are essential readings on data mining and pre-testing bias, while Sullivan, Timmermann, and White (1998) illustrate these dangers empirically in financial applications.

Ruud (2000), Mittelhammer, Judge, and Miller (2000), and Baltagi (1999) include discussions on model selection and multicollinearity from a classical perspective. For computational approaches, including principal components and ridge regression, Thisted (1988) and Judd (1998) offer valuable algorithmic insight.

3.13 Workout Problems

1. On average, is it more or less likely to incur in Type I errors when a relevant variable is omitted?

2. In the case of omitted variables, prove that even if X_1 and X_2 were orthogonal, $\widetilde{\sigma}^2$ would be biased.

3. In the case of inclusion of irrelevant variables, prove that if X_1 and X_2 were orthogonal, $\widehat{\beta}_1$ and $\widehat{\beta}_1^*$ would be equally efficient.

4. Prove that if the regressors are orthogonal, Belsley's γ will equal one.

5. Prove (3.6) and (3.7).

6. Show that the Gaussian log-likelihood is $\ell_T\left(\widehat{\beta},\widehat{\sigma}^2\right) = -\left(T/2\right)\ln\widehat{\sigma}^2 + a$. Find the value of a.

7. Prove that HQC is consistent.

Chapter 4

Generalized Least Squares

4.1 Introduction

The *HLRM* assumes that the noise components are i.i.d., or at least independent with identical first and second order moments. Either assumption implies the special case of the noise covariance matrix given by $\mathcal{E}\left[uu'\right] = \sigma^2\Omega = \sigma^2 I_T$. By assuming that the covariance matrix is proportional to the identity matrix, we reduced the number of unknowns in the covariance matrix from $T\left(T+1\right)/2$ to 1. Recognizing that this model specification will not be consistent with all DGPs, we now generalize the noise covariance specification and consider the problem of recovering point estimates of β, σ^2, and Ω when $\Omega \neq I_T$. This generalization allows for the errors to be heteroskedastic $\left[\mathcal{V}\left(u_t\right) \neq \mathcal{V}\left(u_s\right) \text{ for } t \neq s\right]$, autocorrelated $\left[\text{Cov}\left(u_t u_s\right) \neq 0 \text{ for } t \neq s\right]$, or both and thereby broadens our modelling basis.

This document is organized as follows: Section 4.2 considers the estimation and inference problems when Ω is a known positive definite symmetric matrix. Section 4.3 considers the case in which the elements of Ω are unknown. We also expand our analysis to examine the problem of attempting to identify and estimate the precise form of Ω via diagnostic testing. Finally, Section **??** summarizes the main conclusions.

4.2 Efficient Estimation

Efficient estimation of β requires knowledge of Ω. In this section we assume that Ω is a known, symmetric, positive definite matrix. We will first discuss the effects of ignoring that $\Omega \neq I_T$. Then we will present the efficient estimation of parameters when Ω is known.

4.2.1 Applying OLS when $\Omega \neq I_T$

What would happen if we ignore that $\Omega \neq I_T$ and proceed as if the assumptions of the $HLRM$ applied? What are the properties of $\widehat{\beta}$ under these circumstances? How are testing and confidence interval estimation procedures affected?

First, note that the OLS estimator is

$$\begin{aligned} \widehat{\beta} &= \left(X'X\right)^{-1} X'Y \\ &= \beta + \left(X'X\right)^{-1} X'u. \end{aligned}$$

As the assumption that $\mathcal{E}\left(u\,|X\right) = 0$ is maintained, we have that $\mathcal{E}\left(\widehat{\beta}\,|X\right) = \beta$ and $\mathcal{E}\left(\widehat{\beta} - \beta\right) = 0$. Thus, the OLS estimator of β is still unbiased (and as we will discuss later, it is also consistent). On the other hand, given that $\mathcal{E}\left[uu'\right] = \sigma^2\Omega$,

$$\begin{aligned} \mathcal{V}\left(\widehat{\beta}\,|X\right) &= \mathcal{E}\left[\left(\widehat{\beta} - \beta\right)\left(\widehat{\beta} - \beta\right)'|X\right] \\ &= \mathcal{E}\left[\left(X'X\right)^{-1} X'uu'X\left(X'X\right)^{-1}|X\right] \\ &= \sigma^2\left(X'X\right)^{-1}\left(X'\Omega X\right)\left(X'X\right)^{-1}. \end{aligned}$$

It follows that $\sigma^2\left(X'X\right)^{-1}$ represents the covariance matrix of the OLS estimator only if $\Omega = I_T$. Thus, if the variance of OLS is computed as $\sigma^2\left(X'X\right)^{-1}$, any inference based on it or on $\widetilde{\sigma}^2\left(X'X\right)^{-1}$ will be misleading. Not only is this the wrong matrix to use, but $\widetilde{\sigma}^2$ will be a biased estimator of σ^2. Furthermore, F and t tests as the ones discussed in the OLS context will no longer be valid.

Fortunately, when Ω is known, it is trivial to make the corrections needed in order to make valid inference with the OLS estimator. In particular, $\overline{\sigma}^2 = (T-k)\,\widetilde{\sigma}^2/\text{tr}[M\Omega]$ is an unbiased estimator of σ^2 and it could be used in the "correct" variance-covariance matrix (i.e. $\sigma^2\left(X'X\right)^{-1}\left(X'\Omega X\right)\left(X'X\right)^{-1}$).[1]

[1] Exercise 1 of this chapter asks you to prove this.

F and t tests as the ones considered in our OLS overview can be trivially modified. In particular, if we wanted to test the hypothesis $\mathrm{H}_0 : Q'\beta = c$ we would use an F test of the form:

$$\frac{1}{q}\frac{\left(Q'\widehat{\beta}-c\right)'\left[Q'\left(X'X\right)^{-1}\left(X'\Omega X\right)\left(X'X\right)^{-1}Q\right]^{-1}\left(Q'\widehat{\beta}-c\right)}{\overline{\sigma}^2}\sim F_{q,T-k}.$$

4.2.2 GLS Estimator

As Ω is positive definite, it can be factored into $\Omega = C\Lambda C'$, where the columns of C are the eigenvectors of Ω and the eigenvalues of Ω are arrayed in the diagonal matrix Λ. Let $\Lambda^{1/2}$ be the diagonal matrix with ith element $\sqrt{\lambda_i}$, where λ_i is the ith eigenvalue of Ω. Define $R = C\Lambda^{1/2}$ and $S' = C'\Lambda^{-1/2}$, then $\Omega = RR'$ and $\Omega^{-1} = S'S$.[2]

If we premultiply $Y = X\beta + u$ by S we obtain

$$SY = SX\beta + Su \quad \text{or} \quad Y_* = X_*\beta + u_*. \tag{4.1}$$

Note that

$$\begin{aligned}\mathcal{V}\left(u_*\right) &= \mathcal{E}\left(u_*u_*'\right)\\ &= \sigma^2 S\Omega S'\\ &= \sigma^2 I.\end{aligned}$$

As all the assumption that led us to the derivation of the OLS estimator are satisfied in the transformed model (4.1), we have:

$$\begin{aligned}\widehat{\beta}_{GLS} &= \left(X_*'X_*\right)^{-1}X_*'Y_*\\ &= \left(X'S'SX\right)^{-1}X'S'SY\\ &= \left(X'\Omega^{-1}X\right)^{-1}\left(X'\Omega^{-1}Y\right).\end{aligned}$$

It is easy to verify that $\mathcal{E}\left(\widehat{\beta}_{GLS}\left|X\right.\right) = \beta$ and

$$\begin{aligned}\mathcal{V}\left(\widehat{\beta}_{GLS}\left|X\right.\right) &= \sigma^2\left(X_*'X_*\right)^{-1}\\ &= \sigma^2\left(X'\Omega^{-1}X\right)^{-1}.\end{aligned}$$

[2] This follows from the orthogonality of C, in which case $I = C'C = CC'$ and thus $C' = C^{-1}$. Thus, $R = S^{-1}$.

The term GLS stands for the Generalized Least Squares Estimator (also known as the Aitken estimator). The unbiased estimator of σ^2 based on the GLS estimator is

$$\widetilde{\sigma}^2_{GLS} = \frac{\widehat{u}_*'\widehat{u}_*}{T-k} = \frac{\left(Y - X\widehat{\beta}_{GLS}\right)' \Omega^{-1} \left(Y - X\widehat{\beta}_{GLS}\right)}{T-k}.$$

Because the GLS estimator is identical to the OLS estimator applied to the transformed linear model (4.1), which adheres to the classical assumptions of the LRM, $\widehat{\beta}_{GLS}$ is BLUE. This can be demonstrated by showing that the OLS estimator is inefficient relative to the GLS estimator. In comparing the variance-covariance matrices of $\widehat{\beta}$ and $\widehat{\beta}_{GLS}$, note that

$$\mathcal{V}\left(\widehat{\beta}\,|X\right) - \mathcal{V}\left(\widehat{\beta}_{GLS}\,|X\right) = \sigma^2 D\Omega D',$$

where $D = (X'X)^{-1} X' - (X'\Omega^{-1}X)^{-1} X'\Omega^{-1}$, and thus the difference is at least positive semidefinite and nonzero if $D \neq 0$ because Ω is positive definite. Therefore, $\widehat{\beta}$ is less efficient than the GLS estimator $\widehat{\beta}_{GLS}$ when Ω is known. However, we emphasize for future reference that the definition of the OLS estimator $\widehat{\beta}$ does not require knowledge of Ω whereas $\widehat{\beta}_{GLS}$ does.

4.2.3 Hypothesis Testing

Because we are effectively applying the OLS principle to the transformed LRM (4.1), all the hypothesis-testing and confidence interval estimation procedures discussed relative to the LRM apply to the GLS estimator. For example when we test the hypothesis $\mathrm{H}_0 : Q'\beta = c$, we obtain the F-distributed test statistic

$$\frac{1}{q}\frac{\left(Q'\widehat{\beta}_{GLS} - c\right)' \left[Q' \left(X'\Omega^{-1}X\right)^{-1} Q\right]^{-1} \left(Q'\widehat{\beta}_{GLS} - c\right)}{\widetilde{\sigma}^2_{GLS}} \sim F_{q,T-k}.$$

For a scalar hypothesis, be it an equality or inequality hypothesis, the t-statistic

$$\frac{Q'\widehat{\beta}_{GLS} - c}{\left[\widetilde{\sigma}^2_{GLS} Q' \left(X'\Omega^{-1}X\right)^{-1} Q\right]^{1/2}} \sim S_{T-k}.$$

can be used. Confidence intervals are generated in the usual way, but we have to replace the variance-covariance component of $\widehat{\beta}_{GLS}$.

4.3 Estimation when Ω is Unknown

Previously, we assumed that Ω was known, in which case a simple transformation of the LRM yielded a noise variance-covariance matrix that was proportional to the identity matrix. In practice, Ω is unknown and here we focus on the problem of estimating the unknown parameters of Ω.

Given the GLS estimator, intuition might suggest that the way to proceed is to replace the unknown Ω with an estimator $\widehat{\Omega}$. This would lead to the Feasible Generalized Least Squares (FGLS) estimator of β defined by

$$\widehat{\beta}_{FGLS} = \left(X'\widehat{\Omega}^{-1}X\right)^{-1}\left(X'\widehat{\Omega}^{-1}Y\right). \tag{4.2}$$

Although this might seem a reasonable thing to do, what are the statistical implications of this approach? This question becomes specially intriguing when we realize that there are more unknowns, $T\left(T+1\right)/2$ in Ω than observations, for $T > 1$. To achieve a solution to the estimation of Ω, one must make restrictive assumptions concerning the number of unknowns involved in representing its structure.

In practice, the elements of Ω are assumed to be functions, $\Omega\left(\theta\right)$, of a reduced and fixed number of unknown parameters θ that remain unchanged as the sample size increases. The problem then reduces to obtaining $\widehat{\theta}$ and use it to compute $\widehat{\Omega} = \Omega\left(\widehat{\theta}\right)$ that is then replaced in (4.2). As we need to impose a structure, we will focus on the most common applications of FGLS, namely heteroskedasticity and autocorrelation.

4.3.1 Heteroskedasticity

Heteroskedasticity arises when, even though the Cov$(u_t u_s) = 0$ for $t \neq s$, $\mathcal{V}\left(u_t\right) = \sigma_t^2$ for $t = 1, \cdots, T$, in which case the innovations are not i.i.d. (they are independent, but not identically distributed). The variance-covariance matrix is:

$$\mathcal{E}\left(uu'\right) = \sigma^2\Omega = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_T^2 \end{bmatrix} = \sigma^2\begin{bmatrix} w_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_T \end{bmatrix}. \tag{4.3}$$

Given that this is an arbitrary scaling, we shall often use the normalization

$$\operatorname{tr}(\Omega) = \sum_{t=1}^{T} w_t = T,$$

which in the $HLRM$ implies that $w_t = 1\ \forall t$.

Heteroskedasticity arises in numerous applications, particularly when dealing with cross-sections. Figure 4.1 presents a typical example. In this case we consider the demand of a good conditional on the income of the individual, where the dispersion from the OLS line is increasing on income. That is, when income raises we observe that the deviations from the conditional mean increase with it (in absolute value).

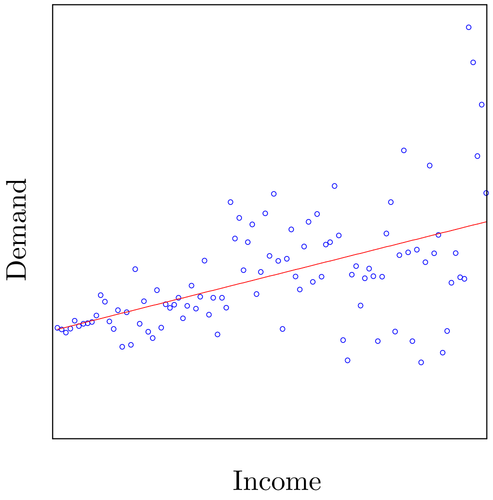

Figure 4.1: Example of Heteroskedasticity

Consequences for OLS

As discussed above, even in the presence of heteroskedasticity, the OLS estimator $\widehat{\beta}$ is unbiased and consistent; however, from the Gauss-Markov theorem we know that, in this case, it is inefficient. The efficient estimator is the GLS estimator, but given that Ω is not known it cannot be obtained. We also know that

$$\mathcal{V}\left(\widehat{\beta}\,|X\right) = \left(X'X\right)^{-1}\left(X'\sigma^2\Omega X\right)\left(X'X\right)^{-1}.$$

As Ω is unknown, it would appear that to compute an estimate of the variance-covariance matrix of the OLS estimator we need to estimate $\sigma^2\Omega$. This is not the case, as in the presence of heteroskedasticity we have

$$\begin{aligned}\Sigma &= T^{-1}\sigma^2 X'\Omega X \\ &= T^{-1}\sum_{t=1}^{T}\sigma_t^2 x_t x_t'.\end{aligned}$$

White (1980) demonstrated that under quite general conditions

$$\widehat{\Sigma} = T^{-1}\sum_{t=1}^{T}\widehat{u}_t^2 x_t x_t' \xrightarrow{p} \Sigma.$$

Therefore, a consistent estimator of the variance-covariance matrix of $\widehat{\beta}$ would be[3]

$$\widehat{\mathcal{V}}\left(\widehat{\beta}\,|X\right) = T\left(X'X\right)^{-1}\widehat{\Sigma}\left(X'X\right)^{-1}. \tag{4.4}$$

This result is extremely useful, given that we do not need to know the precise nature of the pattern of heteroskedasticity. Almost every statistical package reports (optionally) White's heteroskedasticity consistent variance-covariance matrix. In practice, we are not sure if heteroskedasticity is present but (4.4) provides a "vaccine" that allows us to make correct inference with the OLS estimator even when heteroskedasticity is present.

Testing for Heteroskedasticity

Usually, we are not certain when the data has heteroskedasticity, and if present, in what form. Several tests have been proposed to detect the presence of heteroskedasticity and they usually follow this strategy: As the OLS estimator $\widehat{\beta}$ is unbiased and consistent, $\widehat{u}$ will mimic (imperfectly) the true pattern of heteroskedasticity. Therefore, the tests are applied to $\widehat{u}$.

[3] Some statistical packages use $\widetilde{\Sigma} = \widehat{\Sigma}T/\left(T-k\right)$ instead of $\widehat{\Sigma}$ which has better finite sample properties. At any rate, both estimators are consistent.

White's Test As in all heteroskedasticity tests, the null hypothesis is that there is homoskedasticity. That is, $\mathrm{H}_0 : \sigma_t^2 = \sigma^2 \ \forall t$. Under the null hypothesis, the estimator of the covariance matrix of $\widehat{\beta}$ is $\widehat{\mathcal{V}}\left(\widehat{\beta}\,|X\right) = \widetilde{\sigma}^2 \left(X'X\right)^{-1}$, while under the alternative it is (4.4). Based on this observation, White (1980) devised a test that can be carried out by computing the TR^2 of the regression of $\widehat{u}_t^2$ on a constant and the unique variables in $x_t \otimes x_t$.[4] If we denote by J the number of regressors (excluding the constant), it can be show that $TR^2 \sim \chi_J^2$.[5]

An example may clarify this simple procedure. Consider the model $y_t = \beta_0 + \beta_1 x_t + \beta_2 z_t + u_t$. The steps needed to conduct White's test are:

- Obtain $\widehat{\beta}$ and the OLS residuals $\{\widehat{u}_t\}_{t=1}^T$.
- Regress $\widehat{u}_t^2$ on a constant, x_t, z_t, x_t^2, z_t^2, and $x_t z_t$.
- Compute TR^2 from the previous regression.
- For a given significance level, compare TR^2 with the critical value of a Chi-square distribution with 5 degrees of freedom. If TR^2 exceeds the critical value reject the null (homoskedasticity) in favor of the alternative (heteroskedasticity). Otherwise, the null is not rejected.

This test is extremely general as we do not need to make assumptions with respect to the nature of heteroskedasticity. Although this is a virtue, it is also a shortcoming, because it may hide other specification problems instead of heteroskedasticity (such as nonlinearities) and lacks of power (presents important Type II errors). Finally, this test is not constructive, because even if homoskedasticity is rejected it does not tell us how to proceed.

Goldfeld-Quandt's Test Assume that the observations can be divided in two groups in such a way that under the null both groups have the same variance, while under the alternative, the variances differ systematically. If we rank the observations according to some criterion, we can separate the observations in those with high and low variances. The test is applied by dividing the sample into two groups with T_1 and T_2 observations. To obtain

[4] The Kroneker product of matrices y and z is denoted by $y \otimes z$.

[5] We will prove this later.

statistically independent variance estimators, the regression is then estimated separately with the two sets of observations. The test statistic is

$$\frac{\widehat{u}_1'\widehat{u}_1}{\widehat{u}_2'\widehat{u}_2}\frac{T_2 - k}{T_1 - k} \sim F_{T_1-k,T_2-k},$$

where we assume that the disturbance variance is larger in the first sample.[6]

In order to increase the power of the test, Goldfeld and Quandt suggest that a number of observations in the middle of the (sorted) sample be omitted. However, the more observations are dropped, the smaller will be the degrees of freedom for estimation in each group, which will tend to diminish the power of the test. As a consequence, the choice of how many central observations to drop is largely subjective. Empirical evidence suggests that no more than a third of the observations should be dropped. If u is normally distributed, the statistic is exactly distributed as F under the null, and the nominal size of the test is correct. If not, the F distribution is inappropriate and some alternative method with known asymptotic properties, such as White's test, is called for. Finally, recall that to divide the sample in two groups, some criterion is needed. In the case of Figure 4.1 such a criterion is obvious (Income), but, in general, this would rarely be the case.

Glesjer's Test Other tests for heteroskedasticity follow the same procedure outlined for White's test with only minor modifications. The modifications are usually with respect to the dependent variable of the auxiliary regression for the variance component. Some tests use $\widehat{u}_t^2$ as the dependent variable and others use variables such as $\ln(\widehat{u}_t^2)$ or $|\widehat{u}_t|$. The common feature of Glesjer-type of tests is that they regress the dependent variable on a constant and a vector of variables d which are not necessarily the same ones used on the original regression model.[7] All the tests that take $\widehat{u}_t^2$ as the dependent variable use the TR^2 statistic of the auxiliary regression and compare it with the Chi-square distribution with degrees of freedom equal to dimension of d (the number of regressors excluding the constant). When the dependent variable is $\ln(\widehat{u}_t^2)$ or $|\widehat{u}_t|$ the test statistic is not TR^2 (<, >)[pp. 537]Mitte.

These tests share the same problems of White's test, but when applied they usually provide an aid to specify Ω when heteroskedasticity is detected.

[6] It not, reverse the subscripts.

[7] In this sense, Glejser-type of tests are generalizations of White's test, given that d_t can always include the non-redundant vectors of $x_t \otimes x_t$.

FGLS under Heteroskedasticity

When Ω is known, the GLS estimator is

$$\widehat{\beta}_{GLS} = \left(X'\Omega^{-1}X\right)^{-1}\left(X'\Omega^{-1}Y\right).$$

As in this case Ω takes the form of (4.3), it is trivial to verify that when OLS is applied to the transformed model (4.1), we obtain

$$\widehat{\beta}_{GLS} = \left[\sum_{t=1}^{T}\theta_t x_t x_t'\right]^{-1}\left[\sum_{t=1}^{T}\theta_t x_t y_t\right],$$

where $\theta_t = 1/w_t$, which yields the well-known Weighted Least Squares (WLS) estimator. A typical example of such an estimator corresponds to the case when $\sigma_t^2 \propto x_{j,t}^2$ for a regressor j (for example, in Figure 4.1 $x_{j,t}$ would be Income). In such case, if the original model was

$$y_t = \beta_0 + \sum_{l=1}^{k}\beta_l x_{l,t} + u_t,$$

the transformed model would be

$$\frac{y_t}{x_{j,t}} = \beta_j + \beta_0\frac{1}{x_{j,t}} + \sum_{l\neq j}\beta_l\frac{x_{l,t}}{x_{j,t}} + \frac{u_t}{x_{j,t}},$$

where the new innovation $u_t^* = u_t/x_{j,t}$ now has variance σ^2.

Of course, when Ω is not known GLS is unfeasible. To obtain the FGLS estimator described in (4.2) we need to impose structure on Ω. In the case of heteroskedasticity, this is typically done through a formulation that is similar to the one described on Glesjer's test. That is, we impose a functional form for $\sigma_t^2 = f\left(d_t, \theta\right)$ where θ is a vector of parameters to be estimated. Common choices for $f\left(d_t, \theta\right)$ are $\theta' d_t$, $\left(\theta' d_t\right)^2$, or $\exp\left(\theta' d_t\right)$; the last two having the advantage that the range of $f\left(\cdot\right)$ is always non-negative which is a requirement that must be satisfied (remember that σ_t^2 can't be negative!).

Two steps are required to obtain the FGLS estimator in this case. First, estimate $\widehat{\beta}$ (the OLS estimator) and construct $\widehat{u}_t^2$ with it. Regress $\widehat{u}_t^2$ on d_t (which usually includes a constant) in a manner consistent with $f\left(\cdot\right)$. From this auxiliary regression obtain $\widehat{\theta}$ and with it compute $\widehat{\sigma}_t^2 = f\left(d_t, \widehat{\theta}\right)$. The second step is then to apply (4.2) and obtain

$$\widehat{\beta}_{FGLS} = \left[\sum_{t=1}^{T} \frac{1}{\widehat{\sigma}_t^2} x_t x_t'\right]^{-1} \left[\sum_{t=1}^{T} \frac{1}{\widehat{\sigma}_t^2} x_t y_t\right].$$

There may be cases in which θ (the parameters that characterize the conditional heteroskedasticity) includes some elements of β (the parameters that characterize the conditional mean) making FGLS not fully efficient. In that case, MLE is called for.[8]

Concluding, the two-step FGLS estimator is asymptotically equivalent to GLS if the form of heteroskedasticity is known, but Ω must be estimated. If the form of heteroskedasticity is unknown, OLS may have better statistical properties as $f(\cdot)$ may not be well specified. In such a case, some researchers advocate for considering the estimator just described as a quasi-FGLS estimator in the sense that $f(\cdot)$ is viewed as an approximation to the true conditional variance. If the model for the conditional variance is misspecified, $\left[\sum_{t=1}^{T} \frac{1}{\widehat{\sigma}_t^2} x_t x_t'\right]^{-1}$ won't be a consistent estimator of the variance of the FGLS estimator. An appropriate solution in such case is to use a White-type estimator of the covariance matrix that is robust to misspecification of the conditional variance. Such an estimator is

$$\widetilde{\mathcal{V}}\left(\widehat{\beta}_{FGLS}\right) = \left[\sum_{t=1}^{T} \frac{1}{\widehat{\sigma}_t^2} x_t x_t'\right]^{-1} \left[\sum_{t=1}^{T} \frac{\widehat{u}_t^2}{\widehat{\sigma}_t^4} x_t x_t'\right] \left[\sum_{t=1}^{T} \frac{1}{\widehat{\sigma}_t^2} x_t x_t'\right]^{-1}. \tag{4.5}$$

4.3.2 Autocorrelation

Briefly defined, autocorrelation is present in the *LRM* when the assumption that $\mathcal{E}(u_t u_s) = 0$ for $t \neq s$ is not satisfied. The problems of estimation and inference are similar (though more involved) to the ones discussed in the presence of heteroskedasticity. As before, OLS is inefficient and inference based on the usual covariance matrix is adversely affected. Depending on the underlying process, GLS and FGLS may be used to circumvent the problem. We shall emphasize, however, that the models we will examine are far

[8] ARCH (autoregressive conditional heteroskedasticity) models are popular examples of such a case.

removed from the classical regression framework. The exact or finite-sample properties of the estimators are rarely known and we must usually rely only on asymptotic properties.

Most econometric textbooks consider simple modifications to incorporate the presence of autocorrelation on the residuals. The usual setup considers a model like:

$$\begin{aligned} y_t &= \beta' x_t + u_t \\ u_t &= \sum_{s=1}^{m} \theta_s u_{t-s} + v_t, \end{aligned}$$

with v_t being i.i.d. As we mentioned, when heteroskedasticity is present, the u_t's are independent but not identically distributed. On the other hand, when there is autocorrelation on u, the u_t's cease to be independent.

When the OLS estimator was introduced, we showed that it has the property of decomposing Y into two orthogonal components, one that can be written as a linear combination of the column vector of X and another that is orthogonal to X. In principle, the first component can be interpreted as systematic and the second should be unpredictable. If $\widehat{u}$ can be predicted using the information at hand, it means that there still is a systematic part that should be accounted for and introduced in the systematic component.

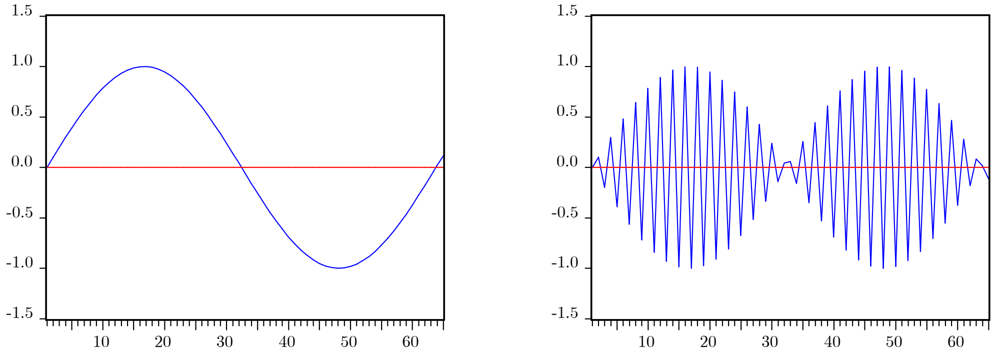

Figure 4.2: Examples of Autocorrelation

Figure 4.2 presents two (idealized) examples of residuals that still have important systematic components. The first panel shows an example of positive autocorrelation. That is, when the residual is above the mean on a given period, the probability that the following residual is above the mean exceeds 0.5. The second panel shows a typical pattern of negative autocorrelation as the residuals shift signs on a systematic fashion. In practice, if autocorrelation is present, it is not as easy to spot as in the examples just provided.

Even though, most econometric textbooks treat the presence of autocorrelation as a "problem" it actually must be interpreted as an opportunity to improve the specification of the underlying DGP. If, for example, we were using a model for forecasting, residuals such as the ones displayed on Figure 4.2 would be unacceptable, as there is a discernible pattern that tells us that we would be underestimating or overestimating Y on a predictable manner.

We will now provide a brief introduction to time series processes (given that, by definition, autocorrelation is present only on this type of data), discuss the effects that the presence of autocorrelation has on the OLS estimate, present testing procedures for autocorrelation, and the methods that are usually proposed to deal with it.

Preliminaries

Definition 15 *$\{y_t\}$ is weakly (covariance) stationary if $\mathcal{E}\left(y_t\right) = \mu$ is independent of t, and $Cov(y_t, y_{t-s}) = \gamma_s$ is independent of t for all s. γ_s is called the s-th autocovariance.*

Definition 16 *$\{y_t\}$ is strongly (strictly) stationary if the joint distribution of $\{y_{t-s}\}_{s=0}^{k}$ is independent of t for all k.*

Note that weak stationarity implies that the first two unconditional moments of y are time invariant, while strong stationarity implies that all unconditional moments (not only the first two) are time invariant. Thus, when a series is strictly stationary it is also covariance stationary, but the converse is not necessarily true.

Definition 17 *$\rho_s = \gamma_s/\gamma_0 = Corr\left(y_t, y_{t-s}\right)$ is the s-th autocorrelation.*

Definition 18 *y_t is a white-noise process if $\rho_s = 0$ $\forall s \neq 0$.*

Definition 19 *y_t is said to follow an $AR(j)$ process if $y_t = \alpha_0 + \sum_{i=1}^{j} \alpha_i y_{t-i} + u_t$, where u_t is a white-noise process.*

Definition 20 *y_t is said to follow an MA(j) process if $y_t = \alpha_0 + \sum_{i=1}^{j} \alpha_i u_{t-i} + u_t$, where u_t is a white-noise process.*

An algebraic construct which is useful for the analysis of autoregressive models is the lag operator.

Definition 21 *The lag operator L satisfies $Ly_t = y_{t-1}$.*

Defining $L^2 = LL$, we see that $L^2 y_t = y_{t-2}$. In general, $L^j y_t = y_{t-j}$. An AR(1) model of the form $y_t = \alpha y_{t-1} + u_t$ can then be expressed as $(1 - \alpha L) y_t = u_t$. Equivalently, the AR(j) model is

$$y_t = \alpha_1 y_{t-1} + \cdots + \alpha_j y_{t-j} + u_t$$

Using the lag operator,

$$\left(1 - \alpha_1 L - \cdots - \alpha_j L^j\right) y_t = u_t$$

or

$$\alpha(L) y_t = u_t$$

where

$$\alpha(L) = 1 - \alpha_1 L - \cdots - \alpha_j L^j$$

We call $\alpha(L)$ the autoregressive polynomial of y_t. The Fundamental Theorem of Algebra says that any polynomial can be factored as

$$\alpha(L) = \left(1 - \lambda_1^{-1} L\right)\left(1 - \lambda_2^{-1} L\right) \cdots \left(1 - \lambda_j^{-1} L\right)$$

where the $\lambda_1, \cdots, \lambda_j$ are the complex roots of $\alpha(L)$, which satisfy $\alpha(\lambda_i) = 0$. Let $|\lambda|$ denote the modulus of a complex number λ.

Theorem 22 *The AR(j) process is weakly stationary if and only if $|\lambda_i| > 1$ for all i.*

A usual way of stating this is "all roots lie outside the unit circle." If the previous conditions are not satisfied, we say that the series is non-stationary. There is no need to present conditions such as these for a MA(j) process, because any MA(j) process is always stationary.

As the purpose of this introduction is simply to provide a few definitions that will be useful to address the presence of autocorrelation in the *LRM* and not a formal treatment of time series models (we will do this later), we will now discuss how to compute Ω for different types of processes for u_t.

The models that are considered usually take the form of AR(j) processes for the error term that is assumed to be weakly stationary. The simplest of which is an AR(1) $u_t = \theta u_{t-1} + v_t$, with $v_t \sim (0, \sigma_v^2)$. It is trivial to verify that in this case, Theorem 22 is satisfied if $|\theta| < 1$. Assuming that this is the case, we can easily derive the autocorrelations and autocovariances of this process. As $\mathcal{E}(u_t) = 0$,

$$\begin{array}{ll} \gamma_0 = \sigma_v^2 / \left(1 - \theta^2\right) & \rho_0 = 1 \\ \gamma_1 = \theta \gamma_0 & \rho_1 = \theta \\ \vdots & \vdots \\ \gamma_j = \theta^j \gamma_0 & \rho_j = \theta^j . \end{array} \tag{4.6}$$

The autocorrelations and autocovariances of a MA(1) process such as $u_t = \theta v_{t-1} + v_t$, with $v_t \sim (0, \sigma_v^2)$ are:

$$\begin{array}{ll} \gamma_0 = \sigma_v^2 \left(1 + \theta^2\right) & \rho_0 = 1 \\ \gamma_1 = \theta \sigma_v^2 & \rho_1 = \theta / \left(1 + \theta^2\right) \\ \gamma_j = 0 \text{ for } j > 1 & \rho_j = 0 \text{ for } j > 1. \end{array}$$

Thus, if the stochastic process that follows u_t is known, the construction of Ω is trivial once we obtain the autocovariances and autocorrelations.

Consequences for OLS

As with the case of heteroskedasticity, when autocorrelation on the residuals is present $\sigma^2 \left(X'X\right)^{-1}$ is not the covariance matrix of the OLS estimator. The covariance matrix is still

$$\mathcal{V}\left(\widehat{\beta}\,|X\right) = \left(X'X\right)^{-1} \left(X'\sigma^2 \Omega X\right) \left(X'X\right)^{-1}.$$

Given that Ω is unknown, our problem is still one of consistently estimating the matrix $\Sigma = T^{-1}\sigma^2 X'\Omega X$ except that now Ω is no longer diagonal. In fact now

$$\begin{aligned}\Sigma &= T^{-1}\sigma^2 X'\Omega X \qquad (4.7)\\ &= T^{-1}\sum_{t=1}^{T}\sum_{s=1}^{T}\operatorname{Cov}(u_t,u_s)\,x_t x_s'.\end{aligned}$$

The purely heteroskedastic case is a special case of (**??**), where σ_t^2 =Cov(u_t, u_t) and Cov$(u_t, u_s) = 0$, $\forall t \neq s$.

Analogous to White's approach, suppose we replace Cov(u_t, u_s) with the estimate $\widehat{u}_t\widehat{u}_s$, defining the estimator

$$\widetilde{\Sigma} = T^{-1}\sum_{t=1}^{T}\sum_{s=1}^{T}\widehat{u}_t\widehat{u}_s x_t x_s'.$$

Unfortunately, without further adjustment, $\widetilde{\Sigma}$ is not useful since the orthogonality of $\widehat{u}$ and X implies $\widetilde{\Sigma} = 0$. Newey and West propose to use the following expression as a consistent estimator of Σ:

$$\widehat{\Sigma} = T^{-1}\sum_{\substack{t=1\\|t-s|\le K}}^{T}\sum_{s=1}^{T} w(t-s)\,\widehat{u}_t\widehat{u}_s x_t x_s' \xrightarrow{p} \Sigma, \qquad (4.8)$$

where K is a finite positive number and $w(t-s) = 1 - \frac{|t-s|}{K}$ is a weighting scheme that ensures that $\widehat{\Sigma}$ is positive definite.

A consistent estimator of the variance-covariance matrix of $\widehat{\beta}$ would be

$$\widehat{\mathcal{V}}\left(\widehat{\beta}\,|X\right) = T\left(X'X\right)^{-1}\widehat{\Sigma}\left(X'X\right)^{-1}.$$

This is even more useful than White's result as (4.8) provides an estimator that is consistent with the presence of autocorrelation, heteroskedasticity, or both. Matrices of the form of (4.8) are often referred to as Heteroskedasticity-Autocorrelation Consistent (HAC) matrices. Almost every statistical package reports (optionally) different types of HAC matrices that differ on the choice of w and K among other factors. In fact, Newey-West's HAC matrix is not the most efficient estimator of Σ and the search of HAC matrices is still an active field of research in econometrics. What is important is that HAC matrices provide a "vaccine" that allows us to make correct inference with the OLS estimator even when there may be heteroskedasticity and/or autocorrelation present in the data. Of course, when the only "problem" is with heteroskedasticity and not autocorrelation, White's matrix is usually preferred.

In some applications—especially when data is organized by groups such as individuals, firms, or regions—it is reasonable to expect that errors may be correlated within clusters but uncorrelated across clusters. In such cases, cluster-robust standard errors provide consistent inference even when both heteroskedasticity and within-cluster autocorrelation are present. A full discussion of clustered inference, including how it generalizes White and HAC estimators, will be presented in the context of panel data models.

Testing for Autocorrelation

As in the case of heteroskedasticity, there are two basic reasons why analysts test for autocorrelation. The first is to discern if there is a reason for using the robust estimate of Ω to perform inference, without the need for a particular specification of Ω. The other is to identify a particular structure of the autocorrelation for use in defining a FGLS estimator of the parameters of the model.

Unlike the case of heteroskedasticity, there is more agreement among applied econometricians regarding the form of the alternative hypothesis if the null of no autocorrelation in the noise component is rejected. In particular, the alternative hypothesis is generally taken to be some autocorrelation process of order m, AR(m), which is represented by

$$u_t = \sum_{s=1}^{m} \theta_s u_{t-s} + v_t, \tag{4.9}$$

where v_t's are i.i.d. random variables with $\mathcal{E}\left(v_t\right) = 0$ and $\mathcal{V}\left(v_t\right) = \sigma_v^2$.

Once again, the tests are usually applied to $\widehat{u}$.

Durbin-Watson's Test This test is applicable when the alternative hypothesis is that m in (4.9) is equal to 1. That is, under the alternative hypothesis the noise component follows an AR(1) process. The Durbin-Watson (DW) test is based on the test statistic

$$\mathrm{DW} = \frac{\sum_{t=2}^{T} \left(\widehat{u}_t - \widehat{u}_{t-1}\right)^2}{\sum_{t=1}^{T} \widehat{u}_t^2} = \frac{\widehat{u}' Z \widehat{u}}{\widehat{u}' \widehat{u}} \tag{4.10}$$

and assumes normality for the noise component. The matrix Z is a Toeplitz-like $(T \times T)$ banded matrix with -1's on the first off-diagonal bands, 2's on the diagonal band except for the unit values in the (1,1) and (T, T) positions, and zeros everywhere else:

$$Z \equiv \begin{bmatrix} 1 & -1 & & & & & \\ -1 & 2 & -1 & & & & \\ & -1 & 2 & \ddots & & & \\ & & -1 & \ddots & -1 & & \\ & & & \ddots & 2 & -1 & \\ & & & & -1 & 2 & -1 \\ & & & & & -1 & 1 \end{bmatrix}. \tag{4.11}$$

The finite sample properties of DW under the null hypothesis (no autocorrelation) is quite complicated and depends on the value of the X matrix, which must not contain lagged values of the dependent variable. The dependence on X is easy to verify as $\widehat{u} = Mu$. Thus (4.10) can be expressed as

$$\text{DW} = \frac{(u/\sigma)' \, MZM \, (u/\sigma)}{(u/\sigma)' \, M \, (u/\sigma)}.$$

Given this dependence, Durbin and Watson provided tables of bounds for the statistic to facilitate applications of the test. However, these table values include inconclusive ranges of the values for DW in which no decision regarding the null hypothesis can be made. The DW is two-sided and is presented in Figure 4.3. It rejects the null of no first-order autocorrelation if DW is outside of the range $(dl, 4 - dl)$, where the interval bounds are available in DW tables.[9] For example, the test rejects the null hypothesis under the alternative of positive autocorrelation if DW<dl, and it rejects the null under the alternative of negative autocorrelation if DW>$4 - dl$. The test has two inconclusive regions that are represented by the intervals (dl, du) and $(4 - du, 4 - dl)$; if DW is inside one of these intervals no decision can be made regarding the null. Finally, if DW is inside the interval $(du, 4 - du)$

[9] These interval bounds depend not only on the number of observations, but also on the number of explanatory variables (excluding the constant).

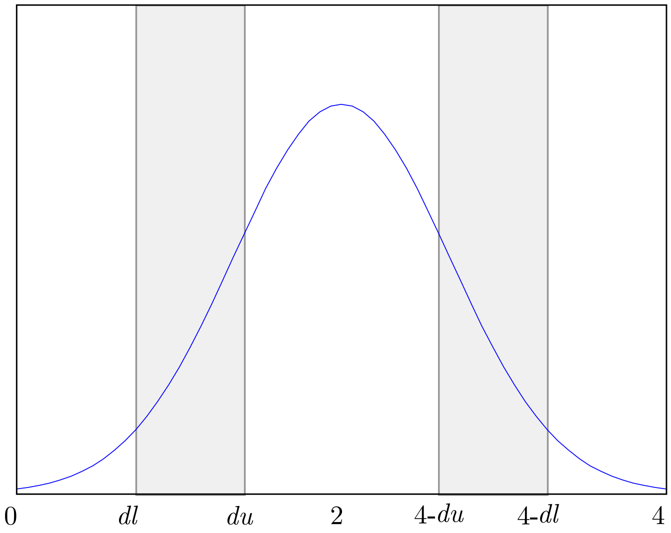

Figure 4.3: Durbin-Watson Statistic

the null hypothesis is not rejected. The asymptotic central tendency of DW is 2.

While popular a few decades ago, this test has several important limitations that make it rarely useful in practice. First, this test is only valid when the alternative hypothesis is that the noise component follows an AR(1) process. Second, it can not be applied if lags of the dependent variable are included as regressors. Third, the test implicitly assumes that the noise component is normal. Fourth, "inconclusive" regions are bound to be present as the exact distribution of DW depends on the data matrix X.

Given the power of modern computers, it is now possible to calculate cumulative distribution function values for DW under the normality assumption and thus determine the exact probability values of DW tests. In particular, for a given value of $d \in [0, 4]$ (this represents the asymptotically admissible range of DW),

$$\Pr\left(\mathrm{DW} < d\right) = \frac{1}{2} - \frac{1}{\pi} \int_0^\infty \frac{\sin\left[e\left(w\right)\right]}{w \cdot g\left(w\right)} dw, \tag{4.12}$$

where

$$\begin{aligned} e(w) &\equiv \frac{1}{2}\sum_{t=1}^{T}\left[\tan^{-1}(\lambda_t w)\right] \\ g(w) &\equiv \left[\prod_{t=1}^{T}\left(1+\lambda_t^2 w^2\right)\right]^{1/4} \end{aligned}$$

and the λ_t's are the T eigenvalues of the matrix $H = MZM - dM$, with Z as defined in (4.11). In the numerical calculation of (4.12), one can truncate ∞ to the value

$$v = \left[\frac{\zeta \sum_{t=1}^{T^*} |\lambda_t|^{1/2} \pi T^*}{2}\right]^{-2/T^*},$$

where T^* is the number of nonzero λ_t's with the assurance that $\Pr(\mathrm{DW} < d)$ is calculated to within ζ (e.g. $\zeta = 0.001$) of the true value; that is, less than ζ of the value of the integral has been truncated. In calculating the integrand values of (4.12), it is also useful to note that

$$\lim_{w\to 0}\frac{\sin[e(w)]}{w\cdot g(w)} = \frac{\sum_{t=1}^{T^*}\lambda_t}{2}.$$

Remember that the tabled bounds of Figure 4.3 and the preceding exact c.d.f. calculations for DW are based on the assumption that under the null, $u_t \sim \mathcal{N}(0,\sigma^2)$. Thus, even the exact c.d.f. calculations become approximate when the normality for u does not hold. The accuracy in such cases is unknown.

However, (4.10) can be written as

$$\mathrm{DW} = \frac{\sum_{t=2}^{T}\left(\widehat{u}_t^2 + \widehat{u}_{t-1}^2 - 2\widehat{u}_t\widehat{u}_{t-1}\right)}{\sum_{t=1}^{T}\widehat{u}_t^2},$$

because $\sum_{t=2}^{T}\widehat{u}_t^2/\sum_{t=1}^{T}\widehat{u}_t^2$ and $\sum_{t=2}^{T}\widehat{u}_{t-1}^2/\sum_{t=1}^{T}\widehat{u}_t^2$ each converge to 1 as $T \to \infty$, and because $\sum_{t=2}^{T}\widehat{u}_t\widehat{u}_{t-1}/\sum_{t=1}^{T}\widehat{u}_t^2 \to \widehat{\rho}$. Then DW$\approx 2(1-\widehat{\rho})$, where $\widehat{\rho}$ is the OLS estimate of ρ in the auxiliary regression

$$\widehat{u}_{t-1} = \rho\widehat{u}_t + v_t. \tag{4.13}$$

Asymptotically, testing the null hypothesis that $\rho = 0$ is equivalent to testing if DW is significantly different from 2. Note that testing the significance of ρ does not require normality (asymptotically), which suggests that an asymptotic Z-test may be preferred when the normality assumption is not appropriate.

Durbin's h Test A variation of the DW test that can be applied when the model contains lagged dependent variables as regressors is the Durbin h test. In this case, the auxiliary regression (4.13) is still valid, but the most familiar implementation of this test takes the form:

$$\mathrm{h} = \left(1 - \frac{\mathrm{DW}}{2}\right)\sqrt{\frac{T}{1 - T\widehat{\sigma}^2_{\widehat{\alpha}}}} \xrightarrow{D} \mathcal{N}(0,1),$$

where $\widehat{\sigma}^2_{\widehat{\alpha}}$ is the estimator of the variance of the parameter associated with the first lag of the dependent variable included in the original estimation. A few features of this test are worth commenting on: First, it does not matter how many regressors (including lags of the dependent variable) are included in the regression; to compute the h test, we need consider only the variance of the coefficient associated with the first lag of the dependent variable. Second, the test is not applicable when $T\widehat{\sigma}^2_{\widehat{\alpha}} > 1$, but in practice this does not happen often; when that is the case it is better to stick with the results from (4.13) or the other tests described below. Finally, the properties of the test are known only asymptotically, thus care should be taken when applying it with small samples.

Breusch-Godfrey's Test This test is similar in spirit to the heteroskedasticity tests discussed above in the sense that it uses the TR^2 statistic of an auxiliary regression. It is more general than the DW and h test because it can be applied to evaluate alternative hypotheses that are not restricted to first-order autocorrelation. As all the tests discussed, the null hypothesis continues to be that there is no autocorrelation in the residuals, while the alternative of this test is that the residuals follow an AR(m) or MA(m) process.

Operationally, the test is carried out be regressing $\widehat{u}_t$ on $x_t, \widehat{u}_{t-1}, \cdots, \widehat{u}_{t-m}$ and referring TR^2 to the tabled critical value of a Chi-squared distribution with m degrees of freedom. Since $X'\widehat{u} = 0$, the test is equivalent to regressing $\widehat{u}_t$ on the part of the lagged residuals that is unexplained by X. There is therefore a compelling logic to it; if any fit is found, it is due to correlation between the current and lagged residuals. The test is a joint test for the first m autocorrelations of u, not just the first.

Q **Statistics** Q tests are variants that are asymptotically equivalent to the Breusch-Godfrey test. These tests use estimates of the autocorrelations of $\widehat{u}$. The s-th order autocorrelation is approximated by:

$$r_s = \frac{\sum_{t=s+1}^{T} \widehat{u}_t \widehat{u}_{t-s}}{\sum_{t=1}^{T} \widehat{u}_t^2}.$$

Box and Pierce propose to use

$$Q_{BP} = T \sum_{s=1}^{m} r_s^2 \xrightarrow{D} \chi_m^2$$

to test the null hypothesis of no autocorrelation, against the alternative of an AR(m) or MA(m) process.

Ljung and Box introduced a modified test for the null hypothesis whose finite sample distribution is more closely approximated by the central Chi-square distribution. This test is defined as

$$Q_{LB} = T\left(T+2\right) \sum_{s=1}^{m} \frac{r_s^2}{T-s} \xrightarrow{D} \chi_m^2.$$

These, as well as the Breusch-Godfrey test, are extremely general. Some textbooks complain that these tests are silent with respect to the choice of m thus making them subjective. This criticism does not make sense. Recall that what we look for is for the residuals to be best described (at least) as white-noise processes; thus these tests should be performed for as many values of m as the sample permits. Failures to reject the null signal specification problems that should be addressed.

FGLS under Autocorrelation

When Ω is known, the GLS estimator is

$$\widehat{\beta}_{GLS} = \left(X'\Omega^{-1}X\right)^{-1}\left(X'\Omega^{-1}Y\right).$$

The form that Ω will take depends on the process that the residuals follow. Here we will concentrate our attention to the simple case of the AR(1) model for u_t. We derived the autocorrelations and autocovariances for this process in (4.6); thus in this case we have

$$\sigma^2\Omega = \frac{\sigma_v^2}{1-\theta^2}\begin{bmatrix} 1 & \theta & \cdots & \theta^{T-2} & \theta^{T-1} \\ \theta & 1 & \ddots & \theta^{T-3} & \theta^{T-2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \theta^{T-2} & \theta^{T-3} & \ddots & 1 & \theta \\ \theta^{T-1} & \theta^{T-2} & \cdots & \theta & 1 \end{bmatrix}.$$

Recalling that we need to find the matrix P such that $\Omega^{-1} = P'P$ to estimate the transformed model by OLS, it is easy to verify that in this case

$$P = \begin{bmatrix} \sqrt{1-\theta^2} & 0 & \cdots & 0 & 0 \\ -\theta & 1 & \ddots & 0 & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & -\theta & 1 & 0 \\ 0 & \cdots & 0 & -\theta & 1 \end{bmatrix}, \tag{4.14}$$

in which case the variables used in the transformed model (4.1) are:

$$Y_* = \begin{bmatrix} \sqrt{1-\theta^2}y_1 \\ y_2 - \theta y_1 \\ \vdots \\ y_T - \theta y_{T-1} \end{bmatrix} \qquad X_* = \begin{bmatrix} \sqrt{1-\theta^2}x_1' \\ x_2' - \theta x_1' \\ \vdots \\ x_T' - \theta x_{T-1}' \end{bmatrix}.$$

Thus, the GLS estimator is obtained applying OLS to the transformed model.

When Ω is not known, GLS is unfeasible and we need to obtain an estimate for $\widehat{\theta}$ in (4.14) to estimate (4.2). A popular candidate is:

$$\widehat{\theta} = \frac{\sum_{t=2}^{T}\widehat{u}_t\widehat{u}_{t-1}}{\sum_{t=2}^{T}\widehat{u}_t^2}.$$

Theil suggested to use $\widetilde{\theta} = (T-k)\,\widehat{\theta}/\,(T-1)$ to obtain a better finite sample estimator. Yet another estimator is $\overline{\theta} = 1-\mathrm{DW}/2$. On the other hand Durbin suggested obtaining $\widehat{\theta}$ from the nonlinear regression model

$$y_t = \theta y_{t-1} + \beta' x_t + \theta\beta' x_{t-1} + v_t.$$

Once $\widehat{\theta}$ is obtained from any of these auxiliary procedures, the second step is to apply (4.2) and obtain the FGLS estimator. This estimator may be obtained following the procedures devised by Cochrane and Orcutt or Hildreth and Lu.

4.4 Bottom Line

In a regression model FGLS is asymptotically superior to OLS. There are at least five reasons why we do not exclusively estimate regression models by FGLS.

First, FGLS depends on the specification and estimation of Ω. Since its structure is unknown, and it may be estimated with considerable error, the estimated $\widehat{\Omega}$ may contain more noise than information about the true Ω. In this case FGLS will do worse than OLS. When the structure of heteroskedasticity is unknown or poorly specified, FGLS can perform worse than OLS. This reflects the classic bias–variance trade-off: FGLS reduces variance only if the weights are well estimated. Otherwise, it may introduce harmful bias. Robust OLS (e.g., White or HAC) may be preferable in these cases.

Second, in the case of heteroskedasticity, $\widehat{\sigma}_t^2$ may be negative for some t, and this requires trimming. This introduces an element of arbitrariness which is unsettling to empirical researchers.

Third, OLS is a more robust estimator of the parameter vector. The asymptotic properties of FGLS depend on the particular selection of Ω. The point is that the efficiency gains from FGLS are built on the stronger assumption of a correct Ω, and the cost is a reduction of robustness due to misspecification.

Fourth, the OLS estimator is consistent and correct inference can be made using a HAC matrix or White's heteroskedastic consistent covariance matrix.

Fifth, in the case of autocorrelation, not rejecting a particular form of autocorrelation does not necessarily imply its acceptance. An example of the empirical practice that was prevalent a few decades ago is instructive.

Assume that a researcher run a regression of the form

$$y_t = \beta x_t + u_t. \tag{4.15}$$

After the OLS results were obtained, the researcher typically looked at the DW statistic, if it was low, it was assumed that the residuals followed an AR(1) process of the form

$$u_t = \rho u_{t-1} + v_t, \tag{4.16}$$

with v_t being a white-noise process. If we substitute (4.15) on (4.16) and rearrange, we obtain

$$y_t = \beta x_t - \rho\beta x_{t-1} + \rho y_{t-1} + v_t. \tag{4.17}$$

Compare (4.17) with the following model

$$y_t = \alpha_0 x_t + \alpha_1 x_{t-1} + \alpha_2 y_{t-1} + v_t. \tag{4.18}$$

Of course, (4.17) is a special case of (4.18), as for (4.17) to hold, the following restriction on (4.18) must be satisfied:

$$\alpha_0\alpha_2 + \alpha_1 = 0. \tag{4.19}$$

This nonlinear restriction (known as the common factors restriction) is never tested when we automatically "correct" for first order autocorrelation. Thus, it is perfectly possible (and indeed frequent) for the DW to be far from 2 and model (4.15)-(4.16) not to be true as (4.19) may not be satisfied.

A pervasive practice with the tradition of the literature of departures from standard assumptions on the *LRM* is that it usually begins with simple models, and after a "problem" is detected, a "remedy" such as the Cochrane-Orcutt procedure is applied immediately. The problem with this philosophy is that it goes from specific-to-general and the risk of ending up estimating a misspecified model is undesirably large. A healthier practice is to use tests as tools not for detecting "problems" but opportunities, in the sense that the tests provide us with guidelines of directions towards uncovering the true DGP. Therefore, it is always better to begin estimating a model such as (4.18) and not (4.15), and test whether or not a restriction such as (4.19) is supported by the data. This type of modelling goes from general-to-specific because it allows us to end with a more parsimonious model only when the evidence points in that direction, in which case the risk of misspecification is reduced.

4.5 Further Reading

Several references offer treatments of generalized least squares and robust inference similar in scope and rigor to the one presented in this chapter. Amemiya (1985) develops the theory of GLS and FGLS in a general framework, emphasizing efficiency and the implications of misspecification. Hayashi (2000) discusses GLS and heteroskedasticity-consistent inference early on, using moment-based arguments that connect naturally with GMM. Ruud (2000) provides detailed derivations and emphasizes the consequences of model misspecification for both estimation and inference. Mittelhammer, Judge, and Miller (2000) examine GLS and FGLS from a decision-theoretic perspective, including the role of prior information and loss functions. Hansen (2022) gives a modern and clean treatment of GLS, FGLS, and robust standard errors, including simulation-based illustrations that echo the exercises presented in this chapter.

Introductory discussions of GLS can be found in several standard texts. Johnston and DiNardo (1997) offer a clear and pedagogical treatment of GLS and its variants, with intuitive examples. Baltagi (1999) covers the basics of GLS and prepares the ground for its application in panel data settings. Greene (1993) provides an accessible overview of GLS and FGLS procedures and discusses various practical issues that arise in applied work.

For advanced treatments, White (1980) remains the foundational reference on heteroskedasticity-consistent inference, introducing the now-standard "White standard errors." Newey and West (1987) extend this approach to accommodate both heteroskedasticity and autocorrelation, providing the basis for HAC estimators used extensively in time series econometrics. Davidson and MacKinnon (2004) offer an extensive and applied account of robust inference methods and their finite sample properties, including simulations and implementation details.

4.6 Workout Problems

1. Prove that $\widetilde{\sigma}^2$ is biased when $\Omega \neq I_T$.

2. Construct a 95% confidence interval for the OLS estimator when Ω is known.

3. Prove (4.5).

4. Derive the autocovariances and autocorrelations of an AR(2) process.

5. Derive the autocovariances and autocorrelations of an MA(2) process.

6. Derive the autocovariances and autocorrelations of an ARMA(1,1) process.

7. Prove (4.14).

8. Describe the Cochrane-Orcutt procedure.

9. Describe the Hildreth-Lu procedure.

10. Consider the following Data Generating Process (DGP):

$$y_t = 2x_t + u_t,$$

where $x_t \sim \mathcal{N}(0, 1)$ and $u_t \sim \mathcal{N}(0, \sigma_t^2)$, where:

$$\sigma_t^2 = 1 + 2x_t^2.$$

An econometrician (incorrectly) postulates that:

$$\sigma_t^2 = \exp(\gamma z_t),$$

where $z_t = t/T$. Conduct a Monte Carlo experiment with $J = 10000$ replications, for sample sizes $T = 25, 100, 1000$. For each sample obtain the OLS, GLS, and FGLS of the incorrectly specified model. Compare the mean, bias, standard deviation, RMSE, and empirical coverage of a 95% confidence interval of each estimator. In the case of OLS consider the case in which the conventional (incorrect) and robust variance is used.

Chapter 5

Causal Identification

5.1 Introduction

Econometric models often aim to answer causal questions: What is the effect of education on earnings? How do interest rates affect investment? Does access to a social program improve health outcomes? In these and countless other examples, our goal is not simply to describe associations in the data, but to isolate the effect of one variable on another—to estimate causal parameters.

Yet regression, by its very nature, is a tool of correlation. The causal content of a regression coefficient does not emerge from the formula, but from the assumptions we impose on the data-generating process. Chief among these is the assumption that the regressors are exogenous—that is, uncorrelated with the error term. When this assumption fails, the Ordinary Least Squares (OLS) estimator is no longer consistent, and the coefficients lose their causal meaning. This is the central issue of endogeneity.

Endogeneity arises from many sources. Some are familiar: omitted variable bias, measurement error, or simultaneity. Others are subtler: behavioral responses, self-selection, or institutional feedback. Regardless of its origin, the presence of endogeneity undermines the credibility of estimated effects. The challenge for the econometrician is therefore to find or create exogenous variation—variation in explanatory variables that is plausibly uncorrelated with unobserved determinants of the outcome.

This chapter examines the principal strategies available to identify causal effects in the presence of endogeneity. Broadly speaking, these strategies fall into two complementary traditions. The first is structural: it relies on economic theory to justify the exclusion of certain variables or the existence of valid instruments. Methods such as instrumental variables and simultaneous equations models belong to this tradition. They operate within systems of equations, often using features of the economic structure to identify causal parameters. The second is design-based: it emphasizes the way data is generated or collected. Instead of seeking instruments, it looks for natural or actual experiments, policy discontinuities, or temporal variation that approximate randomized assignment. Methods such as randomized controlled trials, difference-in-differences, and regression discontinuity designs exemplify this approach.

Both traditions are united by a common purpose: to isolate variation that is as-if random and to use that variation to make credible causal claims. Understanding when and how such identification is possible is the cornerstone of applied econometrics.

The rest of this chapter is organized as follows. Section 5.2 discusses the problem of endogeneity in the linear model and explains why OLS fails when regressors are endogenous. Section 5.3 introduces instrumental variables as a method for recovering consistent estimates, develops the two-stage least squares estimator, and presents diagnostic tools for assessing instrument validity. Section 5.4 turns to systems of simultaneous equations, showing how structural relationships give rise to endogeneity and how 2SLS can be used within identified systems. Section 5.5 shifts to design-based approaches and reviews the logic of randomized experiments, the pitfalls of observational data, and the principles behind natural experiments. Finally, Section 5.6 presents two central quasi-experimental methods—difference-in-differences and regression discontinuity—and discusses their assumptions, implementation, and threats to validity.

5.2 Understanding Endogeneity

The Ordinary Least Squares (OLS) estimator is often introduced as the gold standard of regression analysis: it is linear, unbiased, consistent, and efficient—so long as the regressors are exogenous. But in the real world of economic data, this exogeneity assumption is more often violated than satisfied. When regressors are endogenous—meaning they are correlated with the error term—OLS no longer delivers reliable estimates of causal effects.

The resulting estimators are not merely imprecise; they are systematically misleading.

Why does this happen? Endogeneity typically arises in three familiar situations: omitted variables, measurement error, and simultaneity. Each of these mechanisms causes the regressors to pick up not only their own intended variation, but also unobserved components that belong in the error term. When this happens, the clean separation between "explained" and "unexplained" in the model collapses.

To see how this plays out in a stylized setting, consider a simple linear model where the true regressor is unobserved:

$$y_t = \beta x_t^* + u_t,$$

but $x^* \sim (0, \sigma_{x^*}^2)$ is not observed. Instead, we observe a noisy measurement:

$$x_t = x_t^* + v_t,$$

with $v \sim (0, \sigma_v^2)$ independent of everything else. If we estimate the model using x_t, the OLS estimator becomes:

$$\widehat{\beta} = \frac{\sum xy}{\sum x^2} = \beta \frac{\sum xx^*}{\sum x^2} + \frac{\sum xu}{\sum x^2} \xrightarrow{p} \beta \left(\frac{\sigma_{x^*}^2}{\sigma_{x^*}^2 + \sigma_v^2} \right).$$

This is the textbook case of attenuation bias: the estimate is systematically biased toward zero. The reason is straightforward. The measurement error v adds noise to the regressor, which reduces the signal-to-noise ratio. The regression interprets the weakened correlation between x and y as a weaker causal effect of x^*, even though the true effect β hasn't changed.

More generally, consider the multivariate linear model:

$$Y = X\beta + u,$$

If the regressors X are correlated with the disturbance term u, then the probability limit of the OLS estimator is no longer β, but:

$$\widehat{\beta} \xrightarrow{p} \beta + \Sigma_{xx}^{-1} \Sigma_{xu} \neq \beta.$$

This inconsistency stems from the term $\Sigma_{xu} = T^{-1} \mathcal{E}\left(X'u\right) \neq 0$, which reflects the degree to which the regressors are endogenous. In this case, OLS not only fails to deliver a meaningful estimate—it delivers a biased estimate with high confidence.

The same logic applies to systems of equations. In simultaneous equations models, the dependent variable in one equation appears as a regressor in another. This creates a mechanical correlation between the "regressor" and the "error," since both variables are jointly determined. As we will see in Section 5.4, this structural endogeneity invalidates OLS as a method for recovering the true parameters of the system.

In all these cases, what is needed is a strategy to separate the endogenous part of the regressor from the exogenous variation that drives causal effects. In the next section, we introduce the instrumental variables method—a powerful solution to this problem that allows us to restore consistency by leveraging external sources of variation.

Before we proceed, it's important to emphasize a subtle point: the inconsistency of OLS is not a technicality—it is a fundamental obstacle to causal inference. When a regressor is endogenous, its coefficient no longer has a well-defined interpretation as a marginal effect, let alone a policy-relevant parameter. Any statistical procedure that fails to address this issue risks drawing conclusions from statistical artifacts rather than economic relationships.

5.3 Instrumental Variables

When faced with endogeneity, the core problem is the contamination of a regressor by unobserved shocks that also affect the outcome. This contamination breaks the exogeneity condition and renders OLS inconsistent. This is not a failure of the model per se, but a mismatch between what the data reveal and what the structural equation requires. The question now becomes: how can we recover consistent estimates when some regressors are endogenous?

The idea behind instrumental variables (IV) is both powerful and elegant: if we cannot trust the regressor, perhaps we can find another variable—an instrument—that is correlated with it but uncorrelated with the error term. In this way, we recover a clean source of variation in the regressor, one that mimics the randomness of a controlled experiment.

The idea is simple in principle: if a regressor X is correlated with the structural error term u, we seek a variable Z—a valid instrument—that is required to satisfy two conditions:

- Instrument relevance [Z is correlated with X]: $T^{-1}Z'X \xrightarrow{p} \Sigma_{ZX}$

- Instrument exogeneity [Z is uncorrelated with u]: $T^{-1}Z'u \xrightarrow{p} 0$

Such a Z allows us to extract from X only the part of its variation that is unrelated to the error term and hence suitable for estimating β. In other words, we use Z as a proxy for the exogenous component of X.

The success of this strategy hinges entirely on whether these two properties hold. If Z is weakly correlated with X, the instrument is said to be weak, and the resulting estimator becomes unreliable. If Z is correlated with u, then IV estimation reintroduces bias through the back door.

Instrument relevance implies that the instruments provide enough variation to identify the endogenous component of the regressor. This is a quantitative property: the instruments must shift the endogenous regressors in a way that is strong and detectable in the sample. Instruments that fail to do this are referred to as weak, and they pose serious problems. When instruments are weak, the IV estimator can be biased in small samples (sometimes worse than OLS) and its distribution may be far from normal, making inference unreliable. This is especially problematic when the first-stage relationship is only marginally significant.

In contrast, instrument exogeneity is a qualitative property: the instrument must affect the outcome only through its correlation with the endogenous regressor, not directly or through omitted variables. This assumption cannot be tested directly in just-identified models but becomes testable (though not fully verifiable) in overidentified settings. Violations of exogeneity introduce bias akin to that of the original endogeneity problem IV was designed to solve.

In what follows, we assume valid instruments for estimation and then return to formal tools for evaluating these conditions.

5.3.1 From Moment Conditions to Estimation

Let us now formalize the IV estimator, beginning with the just-identified case and then generalizing to allow more instruments than endogenous regressors. In both cases, the logic is grounded in a system of moment conditions: we assume that the instrument is correlated with the endogenous regressor but uncorrelated with the structural error term.

The Just-Identified Case: $\mathbf{dim}(X) = \mathbf{dim}(Z) = k$

Suppose we are estimating the linear model:

$$Y = X\beta + u, \tag{5.1}$$

and we have a matrix of k instruments Z satisfying $\mathcal{E}[Z'u] = 0$ and $\text{rank}(Z'X) = k$.

Then, in large samples, the moment condition $\mathcal{E}[Z'(Y - X\blacksquare)] = 0$ leads to the estimator::

$$\widehat{\beta}_{SIV} = \left(Z'X\right)^{-1} Z'Y.$$

This is the standard instrumental variables estimator (SIV), which exists and is consistent as long as $X'Z$ is nonsingular. The intuition is straightforward: we are using variation in X that comes from Z—which we believe to be uncorrelated with u—to estimate the structural parameter β.

The estimator of the variance-covariance matrix of $\widehat{\beta}_{SIV}$ is:

$$\widehat{\mathcal{V}}\left(\widehat{\beta}_{SIV}\right) = \widehat{\sigma}^2 \left(Z'X\right)^{-1} Z'Z \left(X'Z\right)^{-1},$$

where:

$$\widehat{\sigma}^2 = T^{-1}\left(Y - X\widehat{\beta}_{SIV}\right)'\left(Y - X\widehat{\beta}_{SIV}\right).$$

The Overidentified Case: $\mathbf{dim}(X) < \mathbf{dim}(Z) = k$

When we have more instruments than endogenous variables, we can no longer directly invert $Z'X$, as it is now a tall matrix. However, if we premultiply (5.1) by Z' we obtain:

$$Z'Y = Z'X\beta + Z'u. \tag{5.2}$$

Note that $\mathcal{E}[Z'u] = 0$ (because of the property of exogeneity of Z) and:

$$\mathcal{V}\left(Z'u\right) = \sigma^2 \mathcal{E}\left(Z'Z\right).$$

We see that $Z'Z$ plays the same rule of Ω in the GLS framework. Thus, applying the same logic that we used in (4.1) we obtain the GLS estimator:

$$\widehat{\beta}_{IV} = \left(X'P_Z X\right)^{-1} X'P_Z Y, \tag{5.3}$$

where $P_Z = Z\left(Z'Z\right)^{-1} Z'$.

As derived in Section 4.2,

$$\widehat{\mathcal{V}}\left(\widehat{\beta}_{IV}\right) = \widehat{\sigma}^2 \left(X' P_Z X\right)^{-1}$$

$$\widehat{\sigma}^2 = T^{-1}\left(Y - X\widehat{\beta}_{IV}\right)'\left(Y - X\widehat{\beta}_{IV}\right).$$

Furthermore, as:

$$\widehat{\beta}_{IV} = \beta + \left[T^{-1}X'Z\left(Z'Z\right)^{-1}Z'X\right]^{-1}T^{-1}X'Z\left(Z'Z\right)^{-1}Z'u,$$

the estimator is consistent as:

$$\begin{aligned} T^{-1}X'Z\left(Z'Z\right)^{-1}Z'X &\xrightarrow{p} \Sigma_{XZ}\Sigma_{ZZ}^{-1}\Sigma_{ZX} \\ T^{-1}X'Z\left(Z'Z\right)^{-1}Z'u &\xrightarrow{p} \Sigma_{XZ}\Sigma_{ZZ}^{-1}\Sigma_{Zu} = 0. \end{aligned}$$

This derivation reinforces a central theme of this chapter: even though OLS fails under endogeneity, the broader principle of orthogonal projection—which underpins both OLS and GLS—can be salvaged by shifting from X to instruments Z.

Viewed this way, IV emerges as a sequence of nested estimators. Since GLS is the OLS estimator applied to a transformed model, and IV is the GLS estimator of a moment condition system pre-multiplied by Z, the IV estimator can be seen as the OLS estimator of a transformation of a transformed model.

5.3.2 Two-Stages Least Squares (2SLS)

While the IV estimator arises from moment conditions and has a GLS interpretation, it can also be viewed as the outcome of applying OLS twice.

- Stage 1: Regress each variable in X on Z to obtain $\widehat{X}$:

$$\widehat{X} = Z\left(Z'Z\right)^{-1}Z'X = P_Z X,$$

- Stage 2: Regress Y on $\widehat{X}$ to obtain the 2SLS estimator:

$$\begin{aligned} \widehat{\beta}_{2SLS} &= \left(\widehat{X}'\widehat{X}\right)^{-1}\widehat{X}'Y \\ &= \left(X'P_Z X\right)^{-1}X'P_Z Y = \widehat{\beta}_{IV}. \end{aligned}$$

This two-stage perspective is not just computationally convenient. It reinforces the intuition behind the method: IV uses only the part of the regressor that is predictable from exogenous sources. The rest is discarded, as it is potentially contaminated by endogeneity.

5.3.3 Assessing Instrument Validity

The credibility of the instrumental variables estimator rests on two key requirements: relevance and exogeneity. These conditions must be satisfied for the estimator to identify the causal effect of interest consistently. Although neither condition is automatically guaranteed in applied work, both can be examined—relevance through formal diagnostics, and exogeneity through overidentification tests in appropriate settings.

Relevance: Are the Instruments Strong Enough?

An instrument must explain a meaningful share of the variation in the endogenous regressor. If it does not, the first-stage regression will be weak, and the resulting IV estimates will be unreliable, even in large samples. Formally, suppose the structural model is given by:

$$Y = X\beta + u,$$

and the m instruments Z are used to address the endogeneity of X. The first-stage regression is:

$$X = Z\Pi + v.$$

In practice, we distinguish between regressors that are suspected to be endogenous and those that are believed to be exogenous. Instrumentation should only be applied to the former. For exogenous regressors, the best instrument is the regressor itself: it is perfectly correlated with itself and, by definition, uncorrelated with the error term. Instrumenting such variables is not only unnecessary—it introduces inefficiency and, if the instruments are weak or invalid, may contaminate the estimation.

As $m \geq k$, to satisfy the relevance condition, the matrix $Z'X$ must have rank k^*, where k^* refers specifically to the number of endogenous regressors. If this condition fails, β is not identified. Even when it holds in large samples, poor finite-sample performance can result when the instruments explain little of the variation in X. These cases are referred to as involving weak instruments.

When there is only one endogenous regressor, instrument strength is typically assessed using the F-statistic from the first-stage regression. A value below 10 is commonly interpreted as evidence of weak instruments. This rule of thumb originates from the critical values derived by Stock and Yogo (2005), who showed that small F-statistics are associated with large biases in the IV estimator and size distortions in conventional inference.

In the case of multiple endogenous variables, the simple F-statistic is no longer appropriate. Instead, one can use the Cragg-Donald Wald F-statistic, which tests the joint significance of the excluded instruments under the assumption of homoskedasticity. When heteroskedasticity or clustering is present, the Kleibergen-Paap rk statistic is preferred. These statistics assess whether the instruments are sufficiently informative to identify the endogenous components of X, and are interpreted relative to critical values that depend on the number of instruments and endogenous regressors. See Stock and Yogo (2005) and Kleibergen and Paap (2006) for details.

Exogeneity: Is the Instrument Uncorrelated with the Error?

Exogeneity is a qualitative assumption: instruments must affect the outcome only through the endogenous variables. In just-identified models, this is untestable. In overidentified models, it can be assessed using the overidentifying restrictions test.

Let $\widehat{u}_{IV} = Y - X\widehat{\beta}_{IV}$ be the residuals from the IV regression. The test proceeds as follows:

- Regress $\widehat{u}_{IV}$ on the instruments Z.
- Compute the statistic TR^2.

Under the null that all instruments are valid:

$$TR^2 \sim \chi^2_{m-k^*},$$

where m is the number of instruments and k^* is the number of endogenous regressors. This is the Sargan test under homoskedasticity, or the Hansen J-test when using robust standard errors. A rejection implies that at least one instrument is invalid.

Choosing Instruments: Art and Theory

The search for valid instruments is often the most conceptually demanding part of empirical work using IV methods. There is no general algorithm for constructing them; success depends on a careful understanding of the economic setting and a credible argument for both relevance and exogeneity.

Instruments are typically drawn from variables that influence the endogenous regressor but are plausibly excluded from the structural equation. Common sources include predetermined variables such as lagged values, policy rules or thresholds that induce variation across individuals or regions, or shocks that shift supply without directly affecting demand. For example, in models where price is endogenous, cost shifters that do not influence demand directly can serve as potential instruments. Similarly, institutional features—such as eligibility criteria, legal mandates, or administrative assignment—can introduce exogenous variation that is both strong and arguably unrelated to the unobserved determinants of the outcome.

But even when an instrument satisfies statistical diagnostics, its credibility depends on the exclusion restriction. This is not a property that can be tested in just-identified models, and only partially tested in overidentified ones. It must be argued on theoretical, institutional, or structural grounds. A compelling IV strategy is always built on a plausible economic mechanism that justifies both the presence and exclusion of the instrument.

5.3.4 When Should We Use IV? Testing for Endogeneity

Instrumental variables estimation is a powerful tool—but one that should be used sparingly and only when justified. Because valid instruments are difficult to find, introducing IV estimation without strong evidence of endogeneity can lead to inefficiency and unnecessary complexity. In most applications, the maintained hypothesis is that the regressors are exogenous. The relevant question is therefore not whether we can estimate the model using IV, but whether we must.

To formalize this, suppose we are interested in estimating the structural equation:

$$Y = X\beta + u,$$

but we suspect that X may be endogenous. This endogeneity may arise because X is not an exogenous variable, but rather the outcome of another process. Specifically, assume:

$$X = Z\gamma + v,$$

where Z is a matrix of instruments and v captures the unobserved determinants of X. This equation reflects the marginal behavior of X.

The key insight is this: if v is correlated with u, then a change in X that arises from a change in v will also be associated with a change in u. This creates a mechanical correlation between X and u, rendering the OLS estimator inconsistent and undermines its causal interpretation. If a shift in X is not exogenous but is instead driven by v, and v is correlated with u, then a change in X also entails a change in u. In that case, we cannot attribute the full change in Y to a change in X via β, since part of the change in Y comes from the simultaneous movement in the error term. It is precisely this concern that motivates testing for endogeneity.

The Hausman test provides a formal statistical procedure to test the null hypothesis that X is exogenous. Under the null, $\text{Cov}(X, u) = 0$, and both OLS and IV estimators are consistent, but OLS is more efficient. Under the alternative, $\text{Cov}(u, v) \neq 0$, OLS is inconsistent and IV is preferred.

The test can be implemented in two equivalent ways. One approach proceeds as follows:

- Estimate the first-stage regression:

  $$X = Z\gamma + v,$$

  and obtain the residuals $\widehat{v}$.

- Include $\widehat{v}$ as an additional regressor in the structural equation:

  $$Y = X\beta + \delta\widehat{v} + u.$$

- Test the null hypothesis $\text{H}_0 : ■ = 0$. A statistically significant $\widehat{■}$ implies that v and u are correlated, confirming endogeneity.

An alternative test, which delivers the same conclusion, consists in evaluating whether the residuals from the structural model estimated by OLS are correlated with the residuals $\widehat{v}$ from the first-stage regression. Specifically, one can test whether $\mathrm{Cov}(\widehat{u}, \widehat{v}) = 0$, where $\widehat{u}$ is the OLS residual. A significant correlation between these residuals suggests that the component of X driven by v is also associated with the outcome error u, providing evidence of endogeneity.[1]

Yet another approach compares the OLS and IV estimates of β directly. Under exogeneity, both are consistent, and their difference should vanish asymptotically. A significant difference between the two estimators suggests that at least one is inconsistent—typically OLS.

The Hausman test, then, is not merely a comparison of two estimation procedures. It formalizes the economic question of whether we can safely ignore the marginal process generating X. If the answer is yes—if the variation in X is orthogonal to the structural disturbance—we proceed with OLS. If not, we turn to IV, not merely as a statistical fix, but as a strategy to isolate the exogenous variation in X and recover the causal effect of interest.

5.4 Simultaneous Equations

Endogeneity does not always arise from omitted variables or measurement error. In many cases, it is a consequence of how the model is structured. Economic theory often implies that variables are determined jointly rather than sequentially. In such settings, endogeneity is not an empirical nuisance—it is intrinsic to the system. This motivates the framework of simultaneous equations models, in which some regressors are endogenous by design.

5.4.1 Structural and Reduced-Form Systems

Consider the classical example of a competitive market with supply and demand:

$$\begin{aligned} \text{Demand: } & Q = \alpha_1 P + \alpha_2 X + u_d \\ \text{Supply: } & Q = \beta_1 P + u_s \end{aligned} . \tag{5.4}$$

[1] Note that one cannot test whether $\mathrm{Cov}(X, u) = 0$ by regressing the OLS residuals on X. By construction, OLS ensures $\widehat{u} \perp X$, regardless of whether X is exogenous. The residuals are orthogonal to the regressors by the algebra of OLS, not by assumption about the underlying data-generating process. Thus, such a test would be uninformative.

Here, Q (quantity) and P (price) are endogenous variables; X (e.g., income) is exogenous. The terms u_d and u_s capture demand and supply shocks, respectively. Under the standard assumption that agents take prices as given, but prices and quantities are determined simultaneously in equilibrium, we cannot treat P or Q as exogenous in either equation.[2]

The two equations above are structural equations: they describe behavioral or technological relationships, typically derived from economic theory. The parameters $(\alpha_1, \alpha_2, \beta_1)$ are structural parameters. They have a direct economic interpretation—elasticities, marginal effects, or policy parameters—and are the primary object of interest in structural analysis.

By solving the system for the endogenous variables in terms of the exogenous variables, we obtain the reduced-form equations:

$$\begin{aligned}
\beta_1 P + u_s &= \alpha_1 P + \alpha_2 X + u_d \\
P &= \frac{\alpha_2}{\beta_1 - \alpha_1} X + \frac{u_d - u_s}{\beta_1 - \alpha_1} = \delta_1 X + v_1 \\
Q &= \frac{\beta_1 \alpha_2}{\beta_1 - \alpha_1} X + \frac{\beta_1 u_d - \alpha_1 u_s}{\beta_1 - \alpha_1} = \delta_2 X + v_2,
\end{aligned}$$

where the coefficients (δ_1, δ_2) are functions of the structural parameters, and v_1, v_2 are linear combinations of u_d and u_s. These reduced-form equations describe how the endogenous variables respond to exogenous variation in equilibrium.

Importantly, the reduced-form parameters are not devoid of content. Although they do not admit a structural interpretation, they summarize the equilibrium response of the endogenous variables to changes in the exogenous environment. They are particularly useful for forecasting and for describing the joint dynamics of the system. Moreover, unlike structural parameters, the reduced-form parameters can be estimated consistently by OLS, because the right-hand side variables (X) are exogenous.

[2] Even if the structural shocks u_d and u_s are uncorrelated, a shift in u_d affects both Q and P in equilibrium. An increase in u_d raises demand, which leads to a new equilibrium with higher quantity and price. As a result, the equilibrium value of P reflects information about u_d, violating the exogeneity condition required for OLS. Therefore, P and u_d are correlated, and OLS estimation of the demand equation yields an inconsistent estimator of α_1.

5.4.2 The Identification Problem

The central problem is whether the structural parameters can be recovered from the reduced-form estimates. This is the identification problem. A structural equation is said to be identified if its parameters can be expressed uniquely as functions of the reduced-form parameters. If not, no consistent estimation is possible—regardless of sample size or estimation method.

Two standard conditions are used to assess identification:

- The order condition is a necessary condition. Let G be the number of endogenous variables and K the number of exogenous variables in the system. In a particular equation, let g_n denote the number of endogenous variables and k_n the number of included exogenous variables in equation n. Equation n is identified if:

  $$K - k_n \geq g_n - 1.$$

  If the inequality is strict, the equation is overidentified; if equality holds, it is exactly identified. Otherwise, it is underidentified.

- The rank condition is necessary and sufficient but more difficult to verify directly. It requires that the excluded exogenous variables exert enough variation to affect the included endogenous regressors through the rest of the system.

Returning to our example in (5.4): we have $G = 2$ endogenous variables (P, Q) and $K = 1$ exogenous variable (X).

In the demand equation, $g_d = 2$, $k_d = 1$, so $K - k_d = 0 < 1 = g_d - 1$.

In the supply equation, $g_s = 2$, $k_s = 0$, so $K - k_s = 1 = g_s - 1$.

This means that while the demand equation is underidentified, the supply equation is exactly identified and can be estimated consistently. The key is that the exogenous variable X, which shifts the demand curve, induces movements along the supply curve. As changes in X shift demand, they generate observable equilibrium adjustments in both P and Q, tracing out the supply relationship. Since X is excluded from the supply equation but affects P and Q through its impact on demand, it serves as a valid instrument for identifying the structural parameter β_1, the supply elasticity.

5.4.3 Estimating Identified Equations

When identification is secured, the structural equation can be estimated using IV techniques. Consider the structural equation n :

$$Y_n = Y_{\overline{n}}\beta_n + X_n\gamma_n + u_n = Z_n\alpha_n + u_n,$$

where Y_n is the $T \times 1$ vector of observations of the endogenous variable n, $Y_{\overline{n}}$ is the $T \times (g-1)$ matrix of the other endogenous variables in equation n, X_n is the $T \times k_n$ matrix of exogenous in equation n, and $\alpha_n' = [\beta_n' \ \gamma_n']$ is the vector of structural parameters of equation n. Finally, Let X be the collection of all the exogenous variables in the system denote the matrix of all exogenous variables in the system (included and excluded in equation n).

Two-Stage Least Squares (2SLS) proceeds as follows:

- First-stage: Regress each of the variables in $Y_{\overline{n}}$ on X and compute the fitted values:
$$\widehat{Y_{\overline{n}}} = X\left(X'X\right)^{-1}X'Y_{\overline{n}} = P_X Y_{\overline{n}}.$$

- Second-stage: Regress Y_n on $\widehat{Y_{\overline{n}}}$ and X_n. The resulting estimator is:
$$\widehat{\alpha}_n = \left(Z_n' P_X Z_n\right)^{-1} Z_n' P_X Y_n,$$
where $Z_n = \left[\widehat{Y_{\overline{n}}} \ X_n\right]$ is the full regressor matrix for the structural equation.

As with any IV method, the consistency of the 2SLS estimator depends on the validity of the instruments (i.e., the exogeneity of X) and the identification of equation n.

As already discussed in Section 5.3,

$$\widehat{\mathcal{V}}\left(\widehat{\alpha}_n\right) = \widehat{\sigma}_n^2\left(Z_n' P_X Z_n\right)^{-1}$$

$$\widehat{\sigma}^2 = T^{-1}\left(Y_n - Z_n\widehat{\alpha}_n\right)'\left(Y - Z_n\widehat{\alpha}_n\right).$$

5.4.4 Why Structure Matters

Simultaneous equations models illustrate that endogeneity is not merely a statistical nuisance to be corrected with instruments. Rather, it is a consequence of structural interdependence: the variables are jointly determined in equilibrium. Estimating a single equation as if it were isolated ignores the logic of the system and typically leads to biased and inconsistent estimates.

The reduced-form equations play a valuable complementary role. They reveal how the endogenous variables move in response to exogenous shocks and allow for consistent forecasting. But they do not yield interpretable economic parameters unless identification is achieved. Structural equations, by contrast, allow us to isolate behavioral relationships, perform counterfactual analysis, and guide policy—provided we are able to identify and estimate them correctly.

5.5 Design-Based Causal Inference

In recent decades, empirical economics has undergone a methodological shift. Rather than specifying a fully articulated structural model and estimating its parameters, researchers increasingly focus on identifying causal effects through the exploitation of exogenous variation. This design-based approach places less emphasis on modeling the full system of economic behavior and more on uncovering local causal relationships under minimal assumptions.

The movement has been fueled by dissatisfaction with the limits of observational data and the complexity of structural models. But it has also been enabled by a realization: under certain conditions, credible causal inference can be achieved with surprisingly simple tools. In many cases, identification rests not on functional form or distributional assumptions, but on clever research design and transparent assumptions about the source of variation.

5.5.1 Randomized, Natural, and Quasi-Experiments

The most credible source of causal identification is the randomized controlled trial (RCT), in which treatment is assigned randomly and independently of potential outcomes. Randomization ensures that, on average, treated and control groups differ only by treatment status. As a result, the difference in their outcomes consistently estimates the average treatment effect. No control variables, structural assumptions, or complex econometric models are needed.

But randomization is often infeasible, unethical, or both. Economists therefore look for natural experiments, in which variation in treatment arises from policy shifts, institutional rules, or other factors that mimic random assignment. For example, eligibility thresholds, lotteries, geographic boundaries, or staggered implementation schedules can generate plausibly exogenous variation.

Let $D \in \{0,1\}$ denote a binary treatment indicator, where $D = 1$ if the unit receives the treatment and $D = 0$ otherwise. Under ideal conditions, the average treatment effect can be estimated using a simple difference in means:

$$\delta = \mathcal{E}\left[Y \,|\, D = 1\right] - \mathcal{E}\left[Y \,|\, D = 0\right].$$

This estimator is consistent if the treatment indicator D is uncorrelated with the structural error term, which is satisfied under random assignment. The parameter δ is often referred to as the difference estimator, since in this setting it coincides with the sample difference in means between treated and control units.

In observational settings, the same logic applies if treatment assignment is independent of potential outcomes conditional on observable characteristics X. In that case, we can consistently estimate the treatment effect by controlling for X in a regression:

$$Y = X\beta + D\delta + u.$$

Here, δ still represents a difference in outcomes between treated and untreated units, but conditional on covariates X. Provided that all relevant confounders are included, the difference estimator remains consistent.

In practice, randomization is rarely feasible. Economists therefore look for natural experiments, in which some external force—policy, geography, institutions—creates variation in treatment that approximates random assignment. Examples include eligibility thresholds, lotteries, court rulings, or staggered policy implementation. When these forces affect treatment but not the potential outcomes directly, they can generate valid causal inferences.

More broadly, quasi-experiments refer to empirical strategies where treatment is not randomly assigned, but researchers attempt to identify causal effects by controlling for observables or exploiting discontinuities, trends, or timing. Matching methods, difference-in-differences, and regression discontinuity, discussed in Section 5.6 all fall under this umbrella. The logic remains the same: isolate variation in treatment that is uncorrelated with the unobservables driving the outcome.

What distinguishes these approaches from classical structural econometrics is not the use of simple linear models—they often rely on regressions with dummies and few covariates—but the conceptual foundation. The key assumption is not that the model is correctly specified in a structural sense, but that the variation in treatment can be viewed as exogenous. This shift in emphasis—from modeling the system to designing the source of variation—is the defining feature of the design-based approach.

5.5.2 Internal and External Validity

The appeal of design-based approaches lies in their focus on internal validity: the extent to which a causal effect is identified for the sample at hand. If the identifying assumptions are credible, the estimated effect has a causal interpretation for some subpopulation (e.g., compliers, units near a threshold, or matched pairs).

But internal validity is not the same as external validity. Many design-based estimates are local: they capture causal effects for a narrow set of units under specific conditions. Extrapolating beyond this context—across groups, time periods, or policies—requires additional assumptions, often unstated and rarely tested.

Moreover, even internally valid estimates may fail to capture economically meaningful mechanisms. A policy change may produce an effect, but the absence of a structural framework means we may not know why the effect occurs, how persistent it is, or whether it generalizes.

5.5.3 A Healthy Dose of Skepticism

Design-based methods represent a major advance in empirical rigor. But they also risk encouraging a superficial understanding of causality. The danger is not that these methods are wrong, but that they are overinterpreted. If every policy change or discontinuity is treated as a natural experiment, and every shift in D is regressed on Y without deeper modeling, causal inference risks becoming a dummy-variable festival.

These strategies identify partial effects, not systems. They often abstract away from equilibrium responses, behavioral margins, and structural mechanisms. Without theoretical discipline, they can produce correct but economically uninformative estimates.

Clever mining for interventions is not a substitute for economic theory. The ability to detect a statistically significant change after a policy shock does not, on its own, reveal economic structure or guide decision-making. It is all too easy to mistake identification for explanation.

Moreover, when agents form rational expectations, the identifying variation must not only be exogenous in the econometric sense (uncorrelated with the error term), but also unanticipated by forward-looking individuals. A policy reform, even if not implemented—or not certain to be implemented—can still affect current behavior through expectations. As a result, what appears to be a clean, sharp intervention may already have been priced into decisions. This undermines the notion that observed outcomes reflect only the realized change, and complicates both identification and interpretation.

Even when a given intervention reveals a causal effect in a specific context, extrapolating that effect to other settings—different populations, magnitudes, or information structures—is hazardous. The very act of being informed and forward-looking makes agents' responses sensitive to the broader environment in which the intervention occurs.

Causal inference requires both design and theory. One without the other is either naïve or vacuous. Recognizing this tension is the first step toward using design-based methods responsibly—not as a substitute for economic reasoning, but as a complement to it.

5.6 Quasi-Experimental Methods

All quasi-experimental methods aim to estimate the effect of a treatment or intervention. To define such effects rigorously, we adopt the potential outcomes framework, also known as the Rubin causal model. For each unit i, define $Y_i(1)$ as the potential outcome if treated,$Y_i(0)$ as the potential outcome if untreated.

The individual treatment effect is given by:

$$\tau_i = Y_i(1) - Y_i(0).$$

However, we never observe both outcomes for the same unit. If unit i is treated ($D_i = 1$), we observe $Y_i\,(1)$, but not $Y_i\,(0)$; if untreated ($D_i = 0$), we observe $Y_i\,(0)$, but not $Y_i\,(1)$. The missing outcome is the counterfactual: what would have happened to the same unit had treatment status been different.

Observed outcomes can be written as:

$$Y_i = D_i Y_i\,(1) + (1 - D_i)\,Y_i\,(0)\,,$$

and the fundamental problem of causal inference is that one of the two potential outcomes is always unobserved. Because we cannot identify τ_i for any individual, we focus on population-level causal parameters, denoted here by δ.

The key parameters of interest include:

- The Average Treatment Effect (ATE):

  $$\delta_{ATE} = \mathcal{E}\left[Y\,(1) - Y\,(0)\right],$$

  which represents the mean causal effect across the population.

- The Average Treatment Effect on the Treated (ATT):

  $$\delta_{ATT} = \mathcal{E}\left[Y\,(1) - Y\,(0)|\,D = 1\right],$$

  the average effect for those who actually received treatment.

- The Average Treatment Effect on the Untreated (ATU):

  $$\delta_{ATT} = \mathcal{E}\left[Y\,(1) - Y\,(0)|\,D = 0\right],$$

the average effect for those who were not treated.
The Local Average Treatment Effect (LATE):

$$\delta_{LATE} = \mathcal{E}\left[Y\,(1) - Y\,(0)|\text{ compliers}\right],$$

the average effect for the subpopulation whose treatment status is changed by an instrument or threshold.

Each quasi-experimental method identifies one of these parameters under a specific set of assumptions. Matching methods aim to approximate the counterfactual outcome by comparing observationally similar units. Difference-in-differences uses time variation to account for unobserved fixed differences.

Regression discontinuity exploits sharp changes in treatment probability at a known cutoff.

In all cases, the goal is to estimate how outcomes would have differed for the same unit under different treatment assignments. That is, to use observed variation to approximate the missing counterfactual. Understanding what parameter is being identified, for whom, and under what assumptions, is essential for any valid causal interpretation.

5.6.1 Propensity Score Matching

When treatment assignment is not random, but is driven by observable characteristics, one strategy to recover causal effects is to compare treated and untreated units that are similar along those dimensions. Propensity Score Matching (PSM) operationalizes this idea. It constructs a counterfactual outcome for each treated unit by identifying comparable control units with similar probabilities of receiving treatment, conditional on observables.

Let $D \in \{0,1\}$ denote the treatment indicator, and X a vector of observed covariates. The key identifying assumption is conditional independence (or unconfoundedness):

$$(Y(1), Y(0)) \perp D | X.$$

That is, once we condition on X, treatment assignment is independent of potential outcomes.[3] If this assumption holds, we can recover causal parameters by comparing treated and control units that are observationally equivalent.

However, matching on the full vector X may be impractical in high dimensions. Rosenbaum and Rubin (1983) showed that it suffices to match on the propensity score:

$$p(X) = \Pr(D = 1 \mid X).$$

If conditional independence holds given X, it also holds given $p(X)$. Matching on this scalar index simplifies the problem and facilitates implementation.

[3] As an example, suppose a government offers a business training program to a subset of micro-entrepreneurs. Participation is voluntary, and those who opt in may differ systematically from those who don't. If we observe detailed data on education, prior experience, and capital, we can estimate the probability of participation and match treated and untreated firms with similar scores. Comparing their post-program profits allows us to estimate the effect of training, assuming all relevant differences are captured in observables.

Yet matching is only valid if treated and untreated units share support:

$$0 < p(X) < 1 \text{ for all } X.$$

This overlap (or common support) condition is essential. Without it, some treated units have no comparable controls, and no reliable counterfactuals can be constructed. Diagnosing overlap—and pruning unmatched observations where necessary—is an integral part of any credible PSM implementation.

In practice, PSM proceeds in three steps:

- Estimate $p(X)$, typically via logit or probit, techniques that we will discuss in Chapter 10.
- Match treated and untreated units based on $p(X)$, using nearest-neighbor, radius, or kernel methods.
- Compare outcomes across matched groups to estimate causal effects—usually δ_{ATT}, the average treatment effect on the treated.

This is not an estimation procedure in the usual sense—it is a design stage, intended to approximate the conditions of a randomized experiment by balancing covariates. Matching diagnostics are essential: covariate balance must be checked after matching, and poor balance should prompt reconsideration of the propensity score specification or matching algorithm.

PSM does not eliminate bias from unobserved confounding. It assumes that all relevant differences across groups are captured in X, and that the functional form used to estimate $p(X)$ is adequate. If these assumptions fail, so will the method. Furthermore, matching discards information—especially when poor overlap requires trimming—and estimates may be sensitive to the choices made in implementation.

Still, when the assumptions are credible and overlap is strong, PSM offers a transparent way to construct counterfactuals and to interpret treatment effects in observational data. It reminds us that causal identification begins with comparability, not computation.

5.6.2 Difference-in-Differences

Difference-in-Differences (DiD) is one of the most widely used quasi-experimental designs for estimating causal effects when treatment varies across groups and over time. It relies on the idea that, in the absence of treatment, outcomes for treated and control groups would have evolved in parallel. Under that assumption, the difference in outcome changes across groups captures the causal effect of treatment.

The classical DiD design involves two groups—treated and control—and two periods—before and after treatment. Let Y_{it} be the outcome for unit i at time t, and D_{it} indicate treatment status. The difference-in-differences estimator is given by:

$$\delta = \left[\mathcal{E}\left(Y_1^T\right) - \mathcal{E}\left(Y_0^T\right)\right] - \left[\mathcal{E}\left(Y_1^C\right) - \mathcal{E}\left(Y_0^C\right)\right],$$

where T and C index treated and control groups, and 1 and 0 indicate post- and pre-treatment periods, respectively.[4]

This estimator is consistent for the average treatment effect on the treated (ATT), provided the parallel trends assumption holds:

$$\mathcal{E}\left[Y_1^T - Y_0^T \middle| D = 1\right] - \mathcal{E}\left[Y_1^C - Y_0^C \middle| D = 0\right],$$

in the absence of treatment.

That is, the treated group would have followed the same trend as the control group if the treatment had not occurred. This is a strong assumption—and a crucial one.

DiD is often estimated using a two-way fixed effects model:

$$Y_{it} = \alpha_i + \lambda_t + \delta D_{it} + u_{it},$$

where α_i absorbs time-invariant group effects, λ_t controls for time shocks common to all units, and D_{it} indicates treatment exposure. Here, δ recovers the DiD effect under the maintained assumptions.

Beyond the Regression: Visual Diagnostics and Pre-Trends

A responsible DiD implementation is never just a regression. Since the parallel trends assumption is untestable in the post-treatment period, researchers must diagnose its plausibility in the pre-treatment period. A well-established

[4] For example, assume that a minimum wage increase is implemented in State A but not in neighboring State B. By comparing the change in employment levels before and after the reform across the two states, we can estimate the causal effect of the policy, assuming that in the absence of the change, employment trends in both states would have evolved similarly.

approach is visual inspection of trends across groups before treatment. If the groups moved together before the intervention, the parallel trends assumption becomes more credible.

In more flexible specifications, one can estimate event study or dynamic treatment effects:

$$Y_{it} = \alpha_i + \lambda_t + \sum_{j \neq -1} \delta_j \cdot I\left\{t = j\right\} \cdot D_i + u_{it},$$

where $I\{t = j\}$ is an indicator function equal to 1 if period t is j, and 0 otherwise. The coefficients δ_j capture the evolution of the treatment effect over time relative to a baseline period (often normalized to $j = -1$). These specifications allow researchers to detect anticipation effects, delayed responses, and—critically—pre-treatment deviations, which violate the DiD identifying assumption.

Such deviations are not just statistical nuisances—they are conceptual red flags. If treatment and control groups diverge before the intervention, we cannot assume that their counterfactual paths would have been parallel afterward.

Extensions, Limitations, and Interpretation

Modern applications of DiD often involve staggered adoption or multiple treatment periods, where some units are treated earlier than others. These settings complicate interpretation: in two-way fixed effects models, treatment effects may be averaged across units with different timings, leading to estimators that do not recover any well-defined causal effect without further assumptions. Recent advances (e.g., Callaway & Sant'Anna, 2021; Sun & Abraham, 2021) provide methods to estimate group-time-specific treatment effects under weaker conditions.

Even in the canonical two-period case, DiD identifies only ATT: the effect for those exposed to treatment in the post-period. Generalizing beyond this group requires stronger, often implicit assumptions.

Ultimately, DiD is not just a method—it is a counterfactual logic. It compares outcome trends across groups that differ only in treatment exposure. When well-executed, with credible assumptions and transparent diagnostics, it offers a powerful tool for causal inference. But when applied mechanically, without attention to dynamics, heterogeneity, or anticipatory behavior, it can lead to misleading conclusions.

5.6.3 Regression Discontinuity Designs

Regression Discontinuity Designs (RDD) exploit situations where treatment assignment changes discontinuously at a known threshold of a continuous variable. When units just above and below the threshold are similar in all respects except treatment status, the discontinuity in outcomes at the cutoff can be attributed to the causal effect of the treatment.

Let R be a running variable that determines eligibility, and let r_0 denote a known threshold. In the sharp RDD case, treatment is assigned deterministically:

$$D = I\left\{R \geq r_0\right\}.$$

The outcome equation is:

$$Y = \delta D + f(R) + u,$$

where $f(R)$ is a smooth function of the running variable and u is the structural error term.

Under the key assumption that potential outcomes $Y(0)$ and $Y(1)$ are continuous in R at r_0, the jump in the conditional expectation of Y at the threshold identifies the local average treatment effect:

$$\delta_{LATE} = \lim_{r \downarrow r_0} \mathcal{E}[Y \mid R = r] - \lim_{r \uparrow r_0} \mathcal{E}[Y \mid R = r].$$

RDD is often interpreted as approximating a randomized experiment locally: near the threshold, units are assumed to be similar in both observable and unobservable characteristics, with treatment assignment essentially random.[5]

In fuzzy RDD, treatment assignment jumps discontinuously at the threshold, but not perfectly. That is,

$$\Pr(D = 1 \mid R = r)$$

has a discontinuity at r_0, but D is not a deterministic function of R. In this case, the treatment effect is still identified, but only for compliers—units

[5] For example, a scholarship is awarded to students who score above 85 on an entrance exam. Students with scores just above and just below the cutoff are likely to be similar in all other respects. By comparing GPA outcomes for these students, we can estimate the causal effect of receiving the scholarship, provided there is no manipulation of scores and other factors vary smoothly around the threshold.

whose treatment status changes because of crossing the threshold. Estimation proceeds using a Wald estimator:

$$\delta_{LATE} = \frac{\lim_{r \downarrow r_0} \mathcal{E}[Y \mid R = r] - \lim_{r \uparrow r_0} \mathcal{E}[Y \mid R = r]}{\lim_{r \downarrow r_0} \mathcal{E}[D \mid R = r] - \lim_{r \uparrow r_0} \mathcal{E}[D \mid R = r]}.$$

RDD is typically implemented as a local linear regression on either side of the cutoff, possibly with kernel weighting and bandwidth selection. Graphical analysis is essential: plotting average outcomes and treatment probabilities against the running variable offers intuitive diagnostics and helps detect manipulation or discontinuities in baseline characteristics.

RDD has strong internal validity but limited external validity. The estimated effect is local to the cutoff and may not generalize to units far from the threshold. Moreover, RDD requires that agents cannot precisely manipulate their value of R. If agents can sort around the threshold—by misreporting income, age, test scores, etc.—then the continuity assumption may fail, and the design breaks down.

When valid, however, RDD provides a transparent and compelling causal estimate under weak assumptions. It requires no instruments, no parametric specification of the treatment effect, and no modeling of selection. What it needs is a threshold—and units near it that are, for all practical purposes, comparable.

5.7 Further Reading

A formal treatment of instrumental variables and simultaneity is provided in Amemiya (1985) and Hayashi (2000). For structural interpretations and identification, see Ruud (2000). The econometric logic of simultaneous equations models is developed in depth in Greene (2012) and Johnston and DiNardo (1997).

The potential outcomes framework from an econometric perspective can be found in Abadie and Imbens (2006) and Imbens and Rubin (2015). Angrist and Pischke (2008) remains an classic reference for the entire class of quasi-experimental methods, blending economic intuition with practical implementation.

Methodological extensions for staggered adoption and heterogeneous effects in the Difference-in-differences framework are discussed in Callaway and

Sant'Anna (2021) and Sun and Abraham (2021).

An accessible applied treatment of RDD can be found in Cattaneo et al. (2019).

5.8 Workout Problems

1. Consider the linear structural model:

$$Y = X\beta + u, \text{ with } X = Z\gamma + v,$$

where Z is a matrix of instruments.

(**a**) State the conditions under which the IV estimator of β is consistent.

(**b**) Show that the OLS estimator is biased if $Cov(X, u) \neq 0$.

(**c**) Suppose that Z is weakly correlated with X but uncorrelated with u. Derive the asymptotic bias of the IV estimator in the presence of weak instruments.

2. Consider the following structural model for a specific good:

$$\begin{aligned} \text{Demand: } & Q = \alpha_1 P + \alpha_2 X + u_d \\ \text{Supply: } & Q = \beta_1 P + \beta_2 W + u_s \end{aligned}$$

where $\mathcal{E}(u_d) = \mathcal{E}(u_s) = 0 = Cov(u_d, u_s)$, $\mathcal{V}(u_d) = \sigma_d^2$, $\mathcal{V}(u_s) = \sigma_s^2$, with with P and Q representing price and quantity, X as income, and W as wages, both of which are considered exogenous.

(**a**) (5 points) Based on economic theory, what are the expected signs of the coefficients in this model? Justify your response in terms of the relationship between the dependent and independent variables.

(**b**) (10 points) Derive the reduced-form equations for P and Q. Describe how these equations respond to changes in the exogenous variables.

(**c**) (10 points) Use the order conditions for identification to assess which structural parameters are identified.

(**d**) (5 points) Describe how you would estimate the parameters of this model.

3. Assume treatment $D \in \{0,1\}$, covariates X, and potential outcomes $Y(1), Y(0)$.

(**a**) Define the conditional independence assumption (CIA) and overlap condition formally.

(**b**) Show that under CIA and overlap, the average treatment effect on the treated can be written as:

$$\delta_{ATT} = \mathcal{E}[\mathcal{E}[Y \mid D = 1, p(X)] - \mathcal{E}[Y \mid D = 0, p(X)] \mid D = 1].$$

(**c**) Discuss the role of balancing properties and diagnostics in evaluating whether CIA is plausible.

4. Suppose you have two groups A and B and two periods $t = 0, 1$. Only group A is treated in period 1.

(**a**) Derive the DiD estimator and interpret it as an estimator of δ_{ATT}.

(**b**) Suppose group A's outcome would have increased faster than group B's even without treatment. What is the sign of the bias in the DiD estimator?

(**c**) Discuss how event study plots and placebo tests can help assess the credibility of the parallel trends assumption.

5. Suppose treatment is assigned based on a cutoff in a running variable R, such that:

$$D = I\left\{R \geq r_0\right\}, Y = \delta D + f(R) + u.$$

(**a**) State the continuity assumption required for identification of δ.

(**b**) Show how the treatment effect can be estimated using local linear regression.

(**c**) Discuss what may go wrong if agents can manipulate R, and how to detect this empirically.

Chapter 6

Nonlinear Least Squares

6.1 Introduction

We say that the regression function $m(x_t, \beta) = \mathcal{E}(y_t | x_t, \beta)$ is nonlinear in the parameters if it cannot be written as $m(x_t, \beta) = f(x_t)' \beta$ for some function $f(\cdot)$.

Examples of nonlinear regression functions include:

$$\begin{aligned} m(x, \beta) &= \beta_1 + \beta_2 \frac{x}{1 + \beta_3 x} \\ m(x, \beta) &= \beta_1 + \beta_2 x^{\beta_3} \\ m(x, \beta) &= \beta_1 + \beta_2 \exp(\beta_3 x) \\ m(x, \beta) &= \beta_1 + \beta_2 x_1 + \beta_4 x I(x > \beta_3). \end{aligned}$$

In the first three examples, $m(x, \beta)$ is (generically) differentiable in the parameters β. In the final example, $m(\cdot)$ is not differentiable with respect to β_3.

Nonlinear regression is frequently adopted because the functional form $m(x, \beta)$ is suggested by an economic model. In other cases, it is adopted as a flexible approximation to an unknown regression function.

This document provides a brief introduction to the nonlinear least squares problem and is organized as follows: Section 6.2 presents the nonlinear least squares (NLLS) estimator. Section 6.3 discusses issues related to numerical methods that can be used to obtain it. Section 6.4 addresses the issue of inference in the NLLS context. Section 6.5 generalizes the NLLS estimator when homoskedasticity and/or autocorrelation are present. Finally, Section

6.6 presents a particular class of nonlinear models denominated artificial neural networks (ANN).

6.2 NLLS Estimation

The NLLS estimator can be defined as the estimator that solves the following optimization problem:

$$\widehat{\beta}_{NLLS} = \arg\min_{\beta} S_T(\beta),$$

where

$$\begin{aligned} S_T(\beta) &= \sum_{t=1}^{T} (y_t - m(x_t, \beta))^2 \\ &= [Y - m(X, \beta)]' [Y - m(X, \beta)]. \end{aligned}$$

For notational convenience, let $m_t = m(x_t, \beta)$. When it exists, we form the Jacobian matrix Z as the stacked gradient vectors for each observation

$$\underset{T \times k}{Z(\beta)} = \frac{\partial m}{\partial \beta'} = \begin{bmatrix} \frac{\partial m_1}{\partial \beta_1} & \cdots & \frac{\partial m_1}{\partial \beta_k} \\ \vdots & \ddots & \vdots \\ \frac{\partial m_T}{\partial \beta_1} & \cdots & \frac{\partial m_T}{\partial \beta_k} \end{bmatrix}.$$

A more general characterization of the NLLS estimator can be defined as the set of parameters that solve the FONC:

$$\left.\frac{\partial S_T(\beta)}{\partial \beta}\right|_{\widehat{\beta}} = -2Z\left(\widehat{\beta}\right)' \left[Y - m\left(X, \widehat{\beta}\right)\right] = 0. \tag{6.1}$$

At least three features of (6.1) are worth discussing: First, the OLS estimator is a particular case of the NLLS estimator, given that in the former $m(X, \beta) = X\beta$ and $Z(\beta) = X$ in which case (6.1) has an analytical solution that coincides with OLS. Second, when (6.1) is a nonlinear function of β, it cannot be solved analytically and numerical methods are required. Finally, given that the FONC (6.1) are nonlinear there may be more than one root that solves the problem, and a general evaluation for a global minimum may be needed.

The following section discusses how to use numerical methods to obtain the NLLS estimator.

6.3 Numerical Methods

An algorithm that is commonly used for NLLS estimation problems is the Gauss-Newton method. Here we present the algorithm and its properties.

6.3.1 Gauss-Newton Method

Numerical algorithms are often used in order to solve nonlinear problems for which an analytical solution is not available. In the case of NLLS estimation, it is important to consider that the source of nonlinearity comes from the function $m(x_t, \beta)$.

Consider a first-order Taylor series approximation of $m(x_t, \beta)$ about a starting value β_0:

$$m(x_t, \beta) \cong m(x_t, \beta_0) + \left.\frac{\partial m_t}{\partial \beta'}\right|_{\beta_0} (\beta - \beta_0). \tag{6.2}$$

If we substitute the first-order approximation (6.2) in the objective function we obtain

$$S_T(\beta) \cong \sum_{t=1}^{T} \left[y_t - m(x_t, \beta_0) - \left.\frac{\partial m_t}{\partial \beta'}\right|_{\beta_0} (\beta - \beta_0) \right]^2. \tag{6.3}$$

Note that (6.3) is not the original problem but an approximation near β_0. This new objective function is now quadratic in β (given that β_0 is known), thus the FONC will be linear in β and can be solved analytically.

Consequently, if we consider the FONC, the problem of estimating β in a neighborhood of β_0 is a linear least squares problem. Conditional on the starting value β_0, the solution to the problem is

$$\beta_1 = \beta_0 + \left[Z(\beta_0)' Z(\beta_0) \right]^{-1} \left[Z(\beta_0)' (Y - m(X, \beta_0)) \right]. \tag{6.4}$$

We repeat the Gauss-Newton iterations until convergence ($\beta_j \cong \beta_{j-1}$) and set $\beta_j = \widehat{\beta}_{NLLS}$.[1] Once the NLLS estimator for β is obtained, we can derive consistent estimators for σ^2 using $\widehat{\sigma}^2 = T^{-1} S_T\left(\widehat{\beta}\right)$ or $\widetilde{\sigma}^2 = (T-k)^{-1} S_T\left(\widehat{\beta}\right)$.

[1] Criteria for evaluating the convergence of the iterative solution are discussed below.

Alternative Motivation for Gauss-Newton

If instead of substituting (6.2) into the objective function, we replace it in the model we have

$$y_t - m_t(\beta_0) + \left.\frac{\partial m_t}{\partial \beta'}\right|_{\beta_0} \beta_0 \cong \left.\frac{\partial m_t}{\partial \beta'}\right|_{\beta_0} \beta + u_t$$

or, in matrix notation

$$\begin{aligned} \underbrace{Y - m(X, \beta_0) + Z(\beta_0)\beta_0}_{Y^*} &\cong \underbrace{Z(\beta_0)}_{X^*}\beta + u \\ Y^* &\cong X^*\beta + u. \end{aligned}$$

As this specification satisfies the OLS principles, β_1 can be obtained from a linear regression of Y^* on X^*. That is,

$$\begin{aligned} \beta_1 &= (X^{*\prime}X^*)^{-1} X^{*\prime}Y^* \\ &= \left[Z(\beta_0)' Z(\beta_0)\right]^{-1} \left[Z(\beta_0)'(Y - m(X, \beta_0) + Z(\beta_0)\beta_0)\right] \\ &= \beta_0 + \left[Z(\beta_0)' Z(\beta_0)\right]^{-1} \left[Z(\beta_0)'(Y - m(X, \beta_0))\right], \end{aligned}$$

which is precisely the estimator derived in (6.4).

The principal advantage of the Gauss-Newton algorithm is convenience. First, note that the step taken to compute β_j from β_{j-1} is simply the linear regression of the approximation errors $(Y - m)$ on the columns of the Jacobian matrix (evaluated at β_{j-1}). Second, the method only requires the user to evaluate the nonlinear regression function and to compute the Jacobian matrix for each step, while other methods (that we will discuss later) require the computation of a Hessian matrix. Finally, it is important to remember that the Gauss-Newton method is only applicable to problems that may be equivalently solved by minimizing the nonlinear least-squares objective function.

Two important technical details with respect to the Gauss-Newton algorithm are important. First, in order for the algorithm to work $Z(\beta_j)' Z(\beta_j)$ must be invertible when evaluated at β_j; there may be instances in which this condition is not satisfied and variants of the method have to be considered to ensure that the matrix is positive definite.[2] Second, given that the

[2] In fact, good commercial softwares never use the Gauss-Newton algorithm as described above. For example, GAUSS generally uses a wide variety of methods known as quasi-Newton algorithms.

algorithm requires the computation of the Jacobian matrix $Z\left(\beta_j\right)$ for each iteration, whenever possible, it is desirable to obtain analytical expressions for the derivatives and not rely on numerical derivatives which may not be as accurate.

Starting Values and Convergence Criteria

Gauss-Newton is an iterative solution algorithm. A fully specified algorithm based on an iterative scheme should have three components: a method for deciding on a starting value for the iteration, a method for obtaining the next iterate from its predecessors, and a method for deciding when to stop the iterative process. The preceding subsection examined how to generate the iterative sequence itself, but treated as given the starting value (β_0) and stopping criteria. Next we discuss the practicalities of selecting initial values and determining convergence.

Starting values matter in two ways. First, if the initial value for the iteration is too far away from the solution, the iteration can diverge. Second, it is possible that the FONC $\left.\frac{\partial S_T(\beta)}{\partial \beta}\right|_{\widehat{\beta}} = 0$ have multiple roots. In this case, the root to which the sequence converges will depend on the starting value for the iteration. Unfortunately, there is little constructive theory about choosing starting values, however a few words of general advice can be given. Most of the time, the equations being solved are similar to others whose roots are easily obtained, so that a root of the latter can be used as a starting point for the equations of actual interest. However, sometimes there is no alternative for starting than with a guess or two and to observe the progress of the iteration, hoping that it will be possible to adjust the starting point or iteration method as needed to achieve convergence. A useful approach, when possible, is to graph the functions. This can often provide not only good starting values, but also some insight concerning an appropriate form for the iteration.

The problem of detecting multiple roots, and of settling on the appropriate one when more than one root is found, is a difficult one. Probably the most successful general approach for discovering whether there are multiple roots is to start the iteration several times, from the vicinity of possible solutions if enough is known about the function, or using randomly chosen starting values otherwise. Figure 6.1 provides an example of such a case, where the objective function has three critical values, two of each are minima and one a maximum. If our objective is to minimize $S_T(\beta)$ and start the algorithm with a initial value such as a, the algorithm will eventually converge to $\widehat{\beta}_1$ which is not a global minimum, while if b is used as the initial value, the algorithm will converge to $\widehat{\beta}_2$.

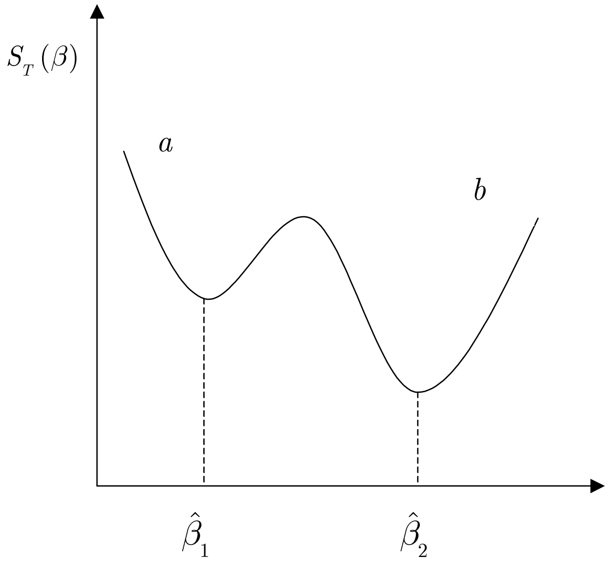

Figure 6.1: Multiple Roots and Initial Values

There is more to say about stopping an iteration than starting one. There are two reasons for bringing an iteration to a halt: either the iteration has converged or it has not. Since the solution of the equation is not known explicitly, the decision as to whether an iteration has converged is based on monitoring either the sequence of iterates to see if β_j is sufficiently close to β_{j-1}, or the sequence of function evaluations $\left.\frac{\partial S_T(\beta)}{\partial \beta}\right|_{\beta_j}$ to see if these become sufficiently close to zero. If β is a scalar, the two most common definitions for successive iterates to be "sufficiently close" are embodied in the *absolute converge criterion*, which asserts convergence when $\left|\beta_j - \beta_{j-1}\right| < tol$, and the *relative converge criterion*, which asserts convergence when $\left|\left(\beta_j - \beta_{j-1}\right)/\beta_{j-1}\right| < tol$, where *tol* is a preselected tolerance. The absolute convergence criterion is most suitable when the solution is close to zero; in this case the denominator of the relative criterion can foster numerical difficulties. On the other hand, when the solution is large (far away from zero) the relative criterion is generally more satisfactory. Once the convergence criterion is satisfied, we ask if $\left.\frac{\partial S_T(\beta)}{\partial\beta}\right|_{\beta_j}$ is "nearly" zero. More precisely we stop if $\left|\left.\frac{\partial S_T(\beta)}{\partial\beta}\right|_{\beta_j}\right| \leq \delta$ for some prespecified δ. If we want high precision, we will choose small δ, but that choice must be reasonable. Choosing $\delta = 0$ is nonsense, since it is unachievable; equally pointless is choosing $\delta = 10^{-20}$ on a 12-digit machine where $\frac{\partial S_T(\beta)}{\partial\beta}$ can be calculated with at most 12 digits of accuracy.[3]

When β is a vector, it is usually satisfactory to define relative and absolute convergence in terms of some norm on β such as the l_2 (euclidean) norm (sum of squares) or the sup-norm (maximum component magnitude). Always remember that even if β_j satisfies the stopping rule, we want to check that $\left.\frac{\partial S_T(\beta)}{\partial\beta}\right|_{\beta_j}$ is close to zero.

6.3.2 Concentration

A major simplification can be achieved through concentration. This can be done when we partition $\beta = (\gamma, \delta)$ so that

$$m(x_t, \beta) = \gamma' x_t(\delta),$$

where $x_t(\delta)$ is a $k \times 1$ function of x_t and δ. In all the examples presented on the Introduction, this can be done with δ of much smaller dimension than γ. In many cases, δ is a scalar.

The SSR function $S_T(\beta) = S_T(\gamma, \delta)$ and thus

$$\min_{\beta} S_T(\beta) = \min_{\delta}\, \min_{\gamma} S_T(\gamma, \delta).$$

[3] The range of numbers that are machine-representable varies greatly across machines; one should always have a good idea of their value when working on a computer. *Machine epsilon* is the smallest *relative* quantity that is machine-representable. Formally this is the smallest ε such that the machine knows that $1-\varepsilon < 1 < 1+\varepsilon$. It is also important to know *machine infinity*, that is, the largest number such that both it and its negative are representable. *Overflow* occurs when an operation takes machine representable numbers but wants to produce a number which exceeds machine infinity in magnitude. A *machine zero* is any quantity that is equivalent to zero on the machine. *Underflow* occurs when an operation takes nonzero quantities but tries to produce a nonzero magnitude less than machine zero. The analyst must either know these important constants for his machine or the more conservative guesses. Much of the software contains a section where the user must specify these arithmetic constants.

Since γ enters the model linearly, we see that

$$\begin{aligned}\widehat{\gamma}(\delta) &= \arg\min_{\gamma} S_T(\gamma, \delta) \\ &= \left[X(\delta)' X(\delta)\right]^{-1} X(\delta)' Y,\end{aligned}$$

where $X(\delta)$ is the $T \times k$ matrix of the stacked $x_t(\delta)'$.

Now set

$$S_T(\delta) = S_T(\widehat{\gamma}(\delta), \delta),$$

which is the concentrated sum of squared residuals. We have $\widehat{\delta} = \arg\min_{\delta} S_T(\delta)$ and $\widehat{\gamma} = \widehat{\gamma}\left(\widehat{\delta}\right)$. The pair $\left(\widehat{\gamma}, \widehat{\delta}\right)$ are the joint NLLS estimates of (γ, δ).

The main benefit of concentration is that the dimension of the numerical optimization is typically reduced dramatically. When δ is scalar, the final minimization over δ can be done by a grid search.

6.4 Inference with Linear Constraints

Given the interpretation that we provided for the NLLS estimator, it is not difficult to derive its asymptotic variance-covariance matrix. As motivated in section 6.3.1, the NLLS is the OLS estimator of the transformed model of Y^* on X^*; in which case

$$\begin{aligned}\mathcal{V}\left(\widehat{\beta}_{NLLS} \,|X\right) &= \mathcal{E}\left[\left(\widehat{\beta}_{NLLS} - \beta\right)\left(\widehat{\beta}_{NLLS} - \beta\right)'\right] \\ &= \sigma^2 \left(X^{*\prime} X^*\right)^{-1},\end{aligned}$$

or, if we define $\widehat{Z} = Z\left(\widehat{\beta}_{NLLS}\right)$,

$$\mathcal{V}\left(\widehat{\beta}_{NLLS} \,|X\right) = \sigma^2 \left(\widehat{Z}' \widehat{Z}\right)^{-1}. \tag{6.5}$$

As σ is not known, a consistent estimator of the variance-covariance matrix of $\widehat{\beta}$ can be derived by replacing σ with $\widehat{\sigma}$ or $\widetilde{\sigma}$. It is important mention, that in contrast to the OLS estimator, we never claimed that the NLLS estimator was unbiased. The only property that we attach to the NLLS estimator is that under certain circumstances (that we will discuss latter), it is consistent. For that reason, all the tests that we conduct are only asymptotically valid and may perform poorly in small samples.

Once an estimator for the variance-covariance matrix of $\widehat{\beta}_{NLLS}$ is obtained, inference and hypothesis testing can be conducted as usual. Next, we discuss some of these tests.

6.4.1 The t Test

As was the case with the OLS estimator, if we are interested in testing $\mathrm{H}_0 : Q'\beta = c$ when the null hypothesis corresponds to a single linear combination of parameters; that is, when $q = 1$, we construct

$$\frac{Q'\widehat{\beta} - c}{\widetilde{\sigma}\left[Q'\left(\widehat{Z}'\widehat{Z}\right)^{-1}Q\right]^{1/2}} \overset{a}{\backsim} S_{T-k},$$

where $\overset{a}{\backsim}$ is meant to imply "is approximately distributed as". We will show later that as was the case with the OLS estimator, this test is asymptotically distributed as a $\mathcal{N}(0,1)$. In fact, given that the properties of $\widehat{\beta}$ are only known asymptotically, it is preferable to take the critical values of the normal distribution, or when small sample deviations appear to be important, to obtain critical values using the bootstrap or a similar method.

With these tools we can construct confidence intervals C_T of β_i. Given that C_T is a function of the data, we must always recall that it is random. Its objective is to cover β_i with high probability. The coverage probability is $\Pr(\beta \in C_T)$. We say that C_T has $(1-\alpha)\,\%$ coverage for β if $\Pr(\beta \in C_T) \to (1-\alpha)$. We construct a confidence interval as follows:

$$\Pr\left[\widehat{\beta}_i - z_{\alpha/2}\sqrt{\widehat{\mathcal{V}}_{i,i}} < \beta_i < \widehat{\beta}_i + z_{\alpha/2}\sqrt{\widehat{\mathcal{V}}_{i,i}}\right] = 1 - \alpha,$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the distribution being considered (asymptotically it should be the normal distribution). The most common choice of α is 0.05.

6.4.2 The F Test

When $q > 1$ we cannot apply the t test described above. In that case we can use a simple pseudo-Likelihood Ratio Test which we will discuss at length later. For now suffices to state the basic results. Under the null hypothesis, it can be shown that

$$\frac{T-k}{q}\frac{S_T\left(\overline{\beta}\right)-S_T\left(\widehat{\beta}\right)}{S_T\left(\widehat{\beta}\right)}\overset{a}{\sim}F_{q,T-k}.$$

The asymptotically valid test takes the form

$$\frac{S_T\left(\overline{\beta}\right)-S_T\left(\widehat{\beta}\right)}{\widehat{\sigma}^2}\xrightarrow{D}\chi^2_q. \tag{6.6}$$

In contrast to inference in OLS, (6.6) is not equivalent to the Wald test when the model is estimated in the NLLS context. In this case, the Wald test takes the form

$$\frac{T-k}{q}\frac{\left(Q'\widehat{\beta}-c\right)'\left[Q'\left(\widehat{Z}'\widehat{Z}\right)^{-1}Q\right]^{-1}\left(Q'\widehat{\beta}-c\right)}{S_T\left(\widehat{\beta}\right)}\overset{a}{\sim}F_{q,T-k}.$$

The asymptotically valid Wald test can be computed as

$$\frac{\left(Q'\widehat{\beta}-c\right)'\left[Q'\left(\widehat{Z}'\widehat{Z}\right)^{-1}Q\right]^{-1}\left(Q'\widehat{\beta}-c\right)}{\widehat{\sigma}^2}\xrightarrow{D}\chi^2_q. \tag{6.7}$$

Once again, as in the case of t tests, we reject the null hypothesis when the value computed exceeds the critical value.

Gallant provides Monte Carlo evidence that suggests that (6.6) has higher power than (6.7) in small samples, and it should be used when possible. Nevertheless, notice that in order to obtain (6.6) two optimizations have to be considered in order to obtain the SSR of the constrained and unconstrained model, and even with linear constraints, the constrained model may induce the more nonlinearities than the unconstrained model, making its computation more demanding. As usual, there is no free lunch!

6.4.3 Identification

Identification is often tricky in nonlinear regression models. Suppose that

$$m\left(x_t,\beta\right)=\beta_1'z_t+\beta_2'x_t\left(\delta\right).$$

This model is linear when $\beta_2 = 0$, and this is often a useful hypothesis to consider. Thus, we want to test

$$\mathrm{H}_0 : \beta_2 = 0.$$

However, under H_0, the model is:

$$y_t = \beta_1' z_t + u_t$$

and both, β_2 and δ, have dropped out. This means that under H_0, δ is not identified. This renders the usual distribution theory invalid. Thus, when the truth is that $\beta_2 = 0$, the parameter estimates are not asymptotically normally distributed. Furthermore, tests of H_0 do not have asymptotic normal or chi-square distributions.

The asymptotic theory of such tests is non-standard and is subject of recent developments in econometrics. Simulation techniques (similar to bootstrap) are usually used to construct critical values in a given application.

6.5 GLS and NLLS

The GLS and FGLS procedures can be extended to estimate the parameters of nonlinear regression models. As was the case, with linear models when $\Omega \neq I_T$, the NLLS estimator derived in (6.4) will still be consistent, but not efficient. Once again, we discuss how to proceed in the case where Ω is known, when it is unknown, and how to conduct inference if the pattern of heteroskedasticity and/or autocorrelation is unknown.

6.5.1 Nonlinear GLS Estimator

As was the case with the linear model, as Ω is a known positive definite matrix, it can be factored into $\Omega = C\Lambda C'$ where the columns of C are the eigenvectors of Ω and the eigenvalues of Ω are arrayed in the diagonal matrix Λ. Let $\Lambda^{1/2}$ be the diagonal matrix with ith element $\sqrt{\lambda_i}$, where λ_i is the ith eigenvalue of Ω. Define $R = C\Lambda^{1/2}$ and $S' = C\Lambda^{-1/2}$, then $\Omega = RR'$ and $\Omega^{-1} = S'S$.

If we consider the transformed model described in section 6.3.1 and define $Y_* = SY^*, X_* = SX^*$, and $u_* = Su$ we obtain

$$Y_* = X_*\beta + u_*, \tag{6.8}$$

where

$$\begin{aligned}\mathcal{V}(u_*) &= \mathcal{E}(u_* u_*') \\ &= \sigma^2 S\Omega S' \\ &= \sigma^2 I.\end{aligned}$$

Thus, all the assumption that led us to the derivation of the OLS estimator are satisfied in the transformed model (6.8). In this case we have:

$$\begin{aligned}\beta_1 &= (X_*' X_*)^{-1} X_*' Y_* \\ &= \left(X^{*\prime}\Omega^{-1}X^*\right)^{-1} X^{*\prime}\Omega^{-1}Y^* \\ &= \beta_0 + \left[Z(\beta_0)'\Omega^{-1}Z(\beta_0)\right]^{-1}\left[Z(\beta_0)'\Omega^{-1}(Y - m(X,\beta_0))\right],\end{aligned}$$

which is simply a generalization of Gauss-Newton algorithm that incorporates information regarding Ω. We denote by $\widehat{\beta}_{GNLLS}$ to the value of β at which this iterative process converges.

The variance-covariance of $\widehat{\beta}_{GNLLS}$ is

$$\begin{aligned}\mathcal{V}\left(\widehat{\beta}_{GNLLS}\,|X\right) &= \sigma^2 (X_*' X_*)^{-1} \\ &= \sigma^2\left(\widehat{Z}'\Omega^{-1}\widehat{Z}\right)^{-1}.\end{aligned}$$

An estimator of σ^2 based on the GNLLS estimator is

$$\widetilde{\sigma}^2_{GNLLS} = \frac{\widehat{u}_*'\widehat{u}_*}{T-k}.$$

Once the estimator of the variance-covariance matrix is obtained, inference can be conducted exactly as we showed for the GLS estimator of the linear model.

6.5.2 Nonlinear FGLS

Previously, we assumed that Ω was known, in which case a simple transformation yielded a noise variance-covariance matrix that was proportional to the identity matrix. When Ω is unknown, we simply replace the unknown Ω with an estimator $\widehat{\Omega}$. This would lead to the Feasible Generalized Nonlinear Least Squares (FGNLLS) estimator of β defined by

$$\widehat{\beta}_{FGNLLS} = \left(X^{*\prime}\widehat{\Omega}^{-1}X^*\right)^{-1}\left(X^{*\prime}\widehat{\Omega}^{-1}Y^*\right). \tag{6.9}$$

As was the case with the FGLS estimator in the linear context, we must impose a structure in order to obtain the estimator of Ω. In practice, the elements of Ω are assumed to be functions, $\Omega(\theta)$, of a reduced and fixed number of unknown parameters θ that remain unchanged as the sample size increases. The problem then reduces to obtaining $\widehat{\theta}$ and use it to compute $\widehat{\Omega} = \Omega\left(\widehat{\theta}\right)$ that is then replaced in (6.9). All the problems and procedures that we discussed for the linear case are valid here.

6.5.3 Applying NLLS: Ignoring that $\Omega \neq I_T$

As was the case with the linear model, if we ignore that $\Omega \neq I_T$, the NLLS estimator will still be consistent but inefficient. More importantly, inference based on the variance-covariance matrix described in (6.5) will be misleading, given that in this case it will be given by

$$\mathcal{V}\left(\widehat{\beta}_{NLLS}\left|X\right.\right) = \sigma^2\left(\widehat{Z}'\widehat{Z}\right)^{-1}\left(\widehat{Z}'\Omega\widehat{Z}\right)\left(\widehat{Z}'\widehat{Z}\right)^{-1}. \tag{6.10}$$

In the presence of heteroskedasticity, we can construct a consistent estimator of this covariance matrix using a suitable modification of White's (1980) matrix Σ.

$$\widehat{\Sigma} = T^{-1}\sum_{t=1}^{T}\widehat{u}_t^2\widehat{z}_t\widehat{z}_t' \xrightarrow{p} \Sigma.$$

In the case of autocorrelation and/or heteroskedasticity we can compute Newey and West's HAC matrix as:

$$\widehat{\Sigma} = T^{-1}\sum_{\substack{t=1 \\ |t-s|\leq K}}^{T}\sum_{s=1}^{T} w\left(t-s\right)\widehat{u}_t\widehat{u}_s\widehat{z}_t\widehat{z}_s' \xrightarrow{p} \Sigma,$$

where K is a finite positive number and $w\left(t-s\right) = 1 - \frac{|t-s|}{K}$ is a weighting scheme that ensures that $\widehat{\Sigma}$ is positive definite.

In either case, a consistent estimator of the variance-covariance matrix of $\widehat{\beta}_{NLLS}$ would be

$$\widehat{\mathcal{V}}\left(\widehat{\beta}_{NLLS}\left|X\right.\right) = T\left(\widehat{Z}'\widehat{Z}\right)^{-1}\widehat{\Sigma}\left(\widehat{Z}'\widehat{Z}\right)^{-1}.$$

Once again, these results are extremely useful given that we do not need to know the precise nature of the pattern of heteroskedasticity and/or autocorrelation. Once a consistent estimator of the covariance matrix of the NLLS estimator is obtained, inference can be conducted as usual.

6.6 Artificial Neural Networks

ANN are a class of input-output models developed by cognitive scientists interested in understanding how computation is performed by the brain. Much is still unknown about how the brain trains itself to process information, so theories abound. The human brain consists of a large number (more than a billion) of neural cells that process informations. Each cell works like a simple processor and only the massive interaction between all cells and their parallel processing makes the brain's abilities possible.

As figure 6.2 indicates, a neuron consists of a core, dendrites for incoming information and an axon with dendrites for outgoing information that is passed to connected neurons. Information is transported between neurons in form of electrical stimulations along the dendrites. Incoming informations that reach the neuron's dendrites is added up and then delivered along the neuron's axon to the dendrites at its end, where the information is passed to other neurons if the stimulation has exceeded a certain threshold. In this case, the neuron is said to be activated. If the incoming stimulation had been too low, the information will not be transported any further. In this case, the neuron is said to be inhibited.[4]

Despite the fact that ANN are far from anything close to a realistic description of how brains actually work, they have shown their remarkable ability to derive meaning from complicated or imprecise data. They can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. A trained neural network can be thought of as an "expert" in the category of information it has been given to analyze.

[4] See Stergiou and Siganos (1996) and Fröhlich (1996) for definitions and non technical introductions to ANN.

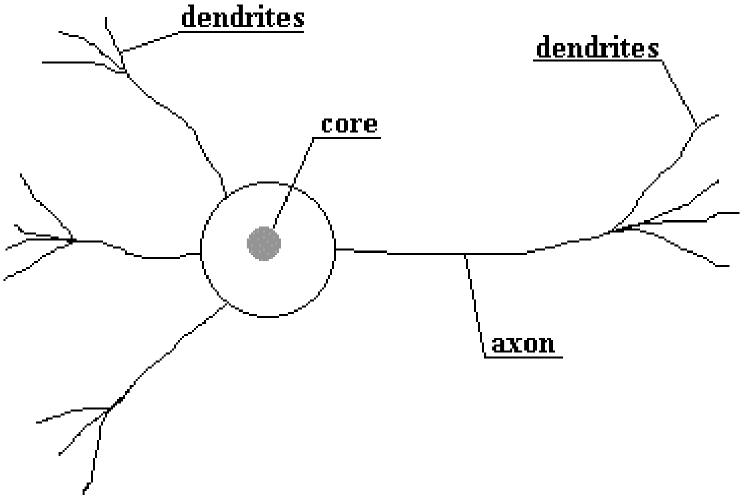

Figure 6.2: Structure of a Neural Cell in the Human Brain

6.6.1 Types of ANN

Like the human brain, ANN consists of neurons and connections between them. The neurons are transporting incoming information on their outgoing connections to other neurons. In neural net terms these connections are called weights. The information is simulated with specific values stored in those weights.

As shown in figure 6.3, an artificial neuron looks similar to a biological neural cell. And it works in the same way. Information (called the input) is sent to the neuron on its incoming weights. This input is processed by a propagation function that adds up the values of all incoming weights. The resulting value is compared with a certain threshold value by the neuron's activation function. If the input exceeds the threshold value, the neuron will be activated, otherwise it will be inhibited. If activated, the neuron sends an output on its outgoing weights to all connected neurons and so on.

In a neural net, the neurons are grouped in layers, called neuron layers. Usually each neuron of one layer is connected to all neurons of the preceding and the following layer (except the input layer and the output layer of the net). The information given to a neural net is propagated layer-by-layer from input layer to output layer through either none, one or more hidden layers.

Several types of neural nets exist depending on how the layers are con-

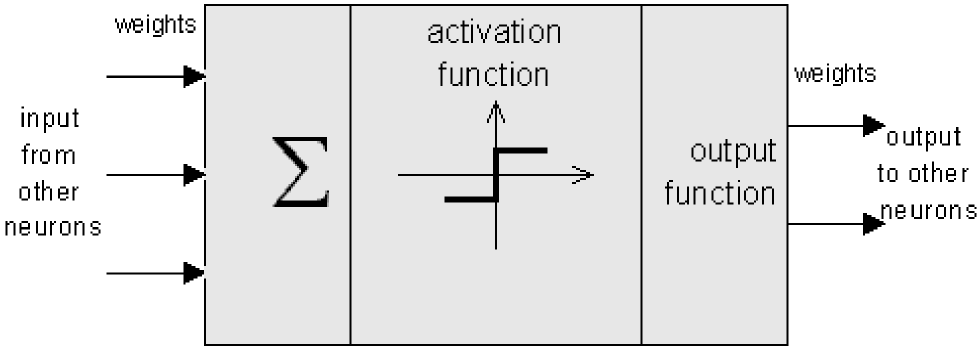

Figure 6.3: Structure of a Neuron in a Neural Net

nected among each other (see figure 6.4). For example, a perceptron is a very simple neural network with two layers (input and output).[5] Feedforward neural nets allow only neuron connections between two different layers, while nets of the feedback type have also connections between neurons of the same layer.

6.6.2 Feedforward Neural Networks

We are interested in approximating $\mathcal{E}\left(y_t\left|x_t\right.\right) = m\left(x_t\right)$, but we do not know $m\left(\cdot\right)$. By far, the artificial neural network most commonly used to approximate it is the single hidden layer, q hidden units, feedforward neural network feedforward network (or ANN(q) for short):

$$m\left(x_t\right) \simeq F\left(x_t, \beta\right) = \sum_{j=1}^{q} G\left(x_t'\gamma_j\right)\theta_j,$$

where β is the vector of parameters that characterizes the network. Associated with the input layer (x), there is a single hidden layer with q values of $x_t'\gamma_j$ (hidden units) that are "connected" through the activation function $G\left(\cdot\right)$ and the θ_j weights to the output layer (which in this case is y). The activation function most commonly used is the logistic activation function $G\left(v\right) = \left(1 + e^{-v}\right)^{-1}$.[6]

[5] In its simplest version, a perceptron can be recognized as a linear model.

[6] See Kuan and White (1994) for a detailed discussion.

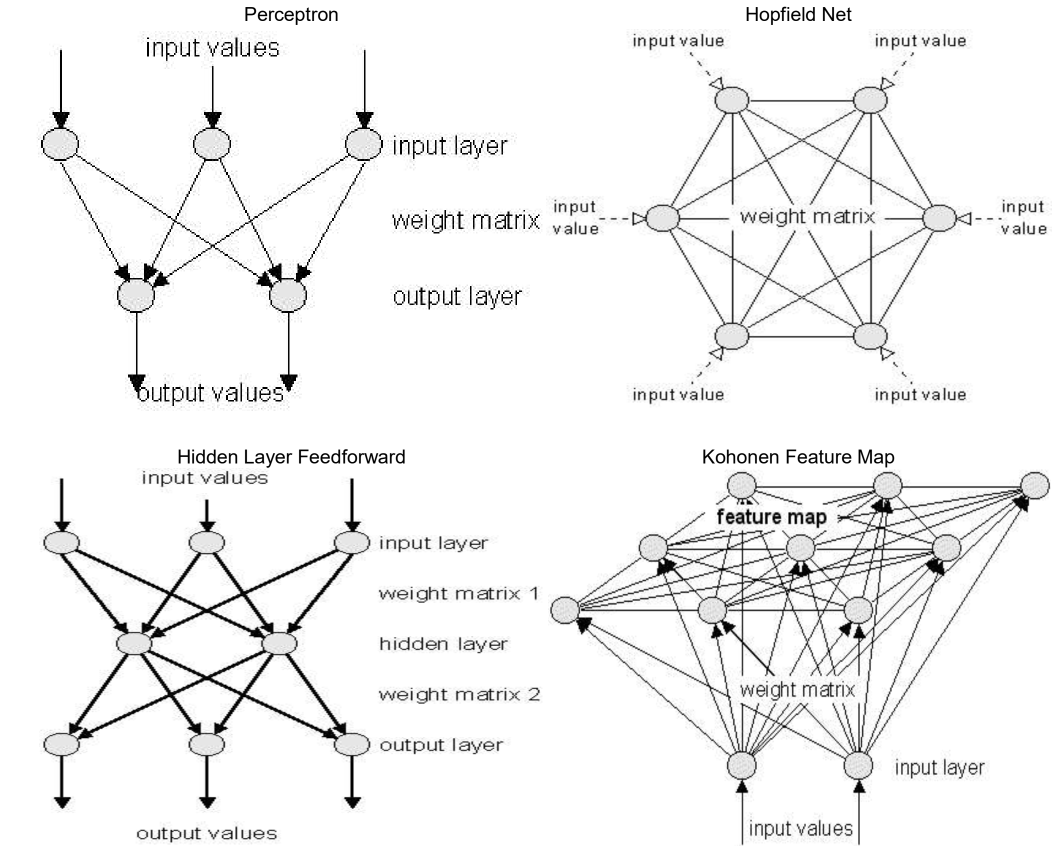

Figure 6.4: Types of Neural Nets

A variant of the single hidden layer network, that is particularly relevant in econometric applications, has direct connections from the input to the output layer as well as the hidden layer. Output of this network (known as the Augmented Hidden Layer Feedforward network) can be expressed as:

$$m\left(x_t\right) \simeq F\left(x_t, \beta\right) = x_t'\alpha + \sum_{j=1}^{q} G\left(x_t'\gamma_j\right)\theta_j. \tag{6.11}$$

Specification (6.11) nests the linear model as a special case (with $\theta_j = 0$ for $j > 0$). Furthermore, due to its parallelism and intrinsic nonlinearity, functions of the form of (6.11) can be viewed as "universal approximators," that is, as a flexible functional form that, provided with sufficiently many hidden units and properly adjusted parameters, can approximate $m(\cdot)$ arbitrarily well.[7]

Estimation and Inference

As ANN constitute a particular class of NLLS models, the parameters can be estimated by minimizing:

$$S_T(\beta) = \sum_{t=1}^{T} (y_t - F(x_t, \beta))^2$$

using numerical method such as the one discussed above.

As ANN were first used in fields other than econometrics, ANN practitioners often use an estimation strategy that differs from Gauss-Newton. One method (known as backpropagation) relies on a recursive estimation of β and is referred to as "learning." In practice it is simply a numerical method also known as steepest descent and is not as efficient as Gauss-Newton.

A way to improve the properties of numerical methods, is to transform y and x such that they are on a comparable scale. Two of the transformation methods commonly used are:

$$z_t^* = \frac{z_t - \min(z_t)}{\max(z_t) - \min(z_t)} \quad \text{or} \quad z_t^* = \frac{z_t - \overline{z}}{\sigma(z)},$$

where, in the first case the variable is mapped to the [0,1] interval and in the second it is standardized.[8]

Finally, as suggested by Franses and van Dijk (2000), estimation of the parameters may benefit from preventing estimates from becoming unduly large. This can be achieved by augmenting the objective function with a penalty term (commonly referred to as weight decay) to obtain:

$$S_T(\beta) = \sum_{t=1}^{T} (y_t - F(x_t, \beta))^2 + \lambda_\alpha \alpha'\alpha + \lambda_\theta \theta'\theta + \lambda_\gamma \gamma'\gamma,$$

[7] Although not routinely used, multiple hidden layer networks can further increase nonlinear features of the data and may in cases require fewer hidden units (q). For example, a two layer network can uniformly approximate certain mappings containing discontinuities that a single layer network can´t.

[8] That is, $\overline{z}$ and $\sigma(z)$ are estimates of the mean and standard deviation of z.

where $\lambda_\alpha, \lambda_\theta$, and λ_γ are small penalty factors specified in advance.

Inference is conducted by using either (6.5) (or (6.10) if required). One word of caution; the standard linear model occurs as the special case in which $\theta_1 = \theta_2 = \ldots = \theta_q$. A moment's reflection reveals an interesting obstacle to the application of standard statistical inference. The γ_j parameters are not identified under the null hypothesis, constituting another example of what was discussed on section 6.4.3. Nonstandard tools have to be used to test this hypothesis.

An Example

Chaotic behavior is interesting because complex nonlinear dynamics that appear to be random and unpredictable can be generated from systems that are deterministic. Moreover, the specific pattern followed by a particular realizations depends crucially on the initial conditions.[9]

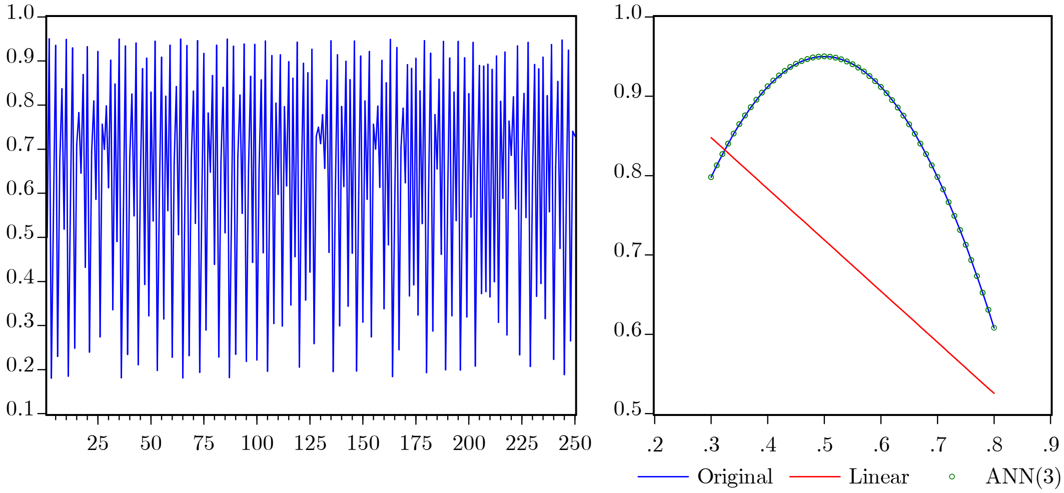

Figure 6.5: Logistic Map: $y_t = 3.8y_{t-1}\,(1 - y_{t-1})$

For example, consider the process known as logistic map and depicted on the first panel of figure 6.5. The power of ANN models can be demonstrated

[9] See Nychka et al (1990) for further details on chaos and statistics.

by comparing the fits of a linear and a nonlinear approximation fitted by a single hidden layer feedforward network to a time series generated with the logistic map.

The second panel of figure 6.5 shows that while the linear approximation behaves rather poorly (as expected), even a small ANN (with one hidden layer and 3 hidden units) can fit the relationship almost perfectly.

6.7 Further Reading

For a formal treatment of nonlinear least squares estimation, Amemiya (1985) presents foundational results on consistency and asymptotic normality. Gallant (1987) offers a comprehensive development of nonlinear statistical models, including identification, estimation, and inference. Ruud (2000) situates NLLS within the broader class of extremum estimators and discusses the conditions under which standard asymptotic results apply.

Mittelhammer, Judge, and Miller (2000) emphasize the connections between nonlinear estimation, information theory, and numerical implementation. Hansen (2022) provides a modern and intuitive introduction to nonlinear methods, with clear examples of Gauss–Newton and related algorithms. Davidson and MacKinnon (1994) discuss estimation under heteroskedasticity, numerical optimization techniques, and bootstrap-based inference for nonlinear models.

Artificial neural networks (ANN), as a flexible class of nonlinear approximators, are examined from an econometric perspective in Kuan and White (1994), who formalize their consistency, convergence rates, and universal approximation properties. Franses and van Dijk (2000) complement this with empirical applications and regularization techniques useful in practice.

6.8 Workout Problems

1. Which of the following regression models can be transformed into linear models and which cannot?

   a. $y_t = \beta_1 \left[\beta_2 x_t^{-\beta_3} + (1-\beta_2)\, z_t^{-\beta_3}\right]^{-\beta_4/\beta_3} \exp(u_t)$

   b. $y_t = \beta_1 + \beta_2 \left[\frac{x_t^{\beta_3}-1}{\beta_3}\right] + u_t$

c. $y_t = \beta_1 x_t^{\beta_2} + u_t$

d. $y_t = \beta_1 x_t^{\beta_2} \exp(u_t)$

e. $y_t = [1 + \exp(\beta' x_t) + u_t]^{-1}$

2. Prove that β_1 derived in (6.4) is the solution to $\min S_T(\beta)$ in (6.3).

3. Write a program that determines your machine's epsilon.

4. Write a program that determines your machine's infinity.

5. Using (6.4), derive (6.5).

6. Derive (6.10).

Chapter 7

Maximim Likelihood

7.1 Introduction

Around 1925, R. A. Fisher introduced the principle of maximum likelihood as a general method for statistical inference. In likelihood theory, inference about unknown parameters begins with the specification of the probability function that governs the observations in a sample of random variables. This function fully characterizes the probabilistic behavior of the data that one will use to learn about the parameters of the data-generating process. Given a sample drawn from a distribution with a specified probability density function, the maximum likelihood estimator (MLE) is defined as the value of the parameters that maximizes this probability—or likelihood—function. The idea is conceptually straightforward and appealing: one chooses the parameter values that make the observed data more probable than any other possible configuration of parameters.

The likelihood approach provides a unifying framework that encompasses the estimation methods presented in earlier chapters. As discussed in Section 1.5.2, the ordinary least squares (OLS) estimator can be interpreted as the MLE that arises when the regression disturbances are assumed to be independently and normally distributed with zero mean and constant variance. Likewise, the nonlinear least squares (NLLS) estimator of Chapter 6 corresponds to the MLE obtained when the model is nonlinear in its parameters but the disturbances still follow a normal distribution. In both cases, estimation proceeds by maximizing a function of the sample that is proportional to the probability of observing the data under a particular parameter configuration. From this perspective, least squares methods appear as special cases of maximum likelihood estimation, which generalizes them by exploiting the full probabilistic specification of the model rather than relying solely on its first two moments. This generalization will prove useful in the analysis of models where the distributional structure itself carries essential information, such as in models with limited dependent variables or in dynamic systems where regimes or states change over time.

Under mild regularity conditions, the MLE possesses several desirable properties that account for its central role in econometrics. It is consistent, asymptotically normal, and asymptotically efficient, meaning that in large samples it achieves the smallest possible variance among a broad class of estimators. It is also invariant to reparameterizations and, in many practical situations, computationally tractable. These properties make the MLE a natural benchmark against which other estimators are evaluated. Nevertheless, the method is not free from drawbacks. Its performance in small samples may differ substantially from its asymptotic behavior, and the accuracy of its conclusions depends on the correctness of the assumed distribution. In misspecified models, the MLE may remain consistent for certain pseudo-parameters but generally loses its efficiency. Later in the chapter we will see that these issues can be mitigated by considering the quasi-maximum likelihood approach and robust inference methods.

The remainder of this chapter is organized as follows. Section 7.2 defines the likelihood function and the MLE. Section 7.3 develops the Cramér–Rao lower bound, which establishes a benchmark for the variance of unbiased estimators. Section 7.4 discusses inference based on the likelihood principle. Section 7.5 introduces the invariance property. Section 7.6 presents the quasi-maximum likelihood estimator (QMLE) and its properties. Section 7.7 examines numerical methods used to obtain MLEs, and Section 7.8 illustrates the main results with applications.

7.2 The Likelihood Function

In order to define the maximum likelihood estimator formally, it is necessary to specify how the probability of the observed sample depends on the unknown parameters of the model. Suppose that for a random variable y the probability density function (p.d.f.) is denoted by $f(y;\theta_0)$, where θ_0 represents the true but unknown value of the parameter vector. In practice, we observe a particular realization of a random sample $\{y_1, \cdots, y_T\}$, but the parameter vector that generated the sample remains unknown. The likelihood function describes this situation by reversing the roles of the arguments in the p.d.f.: while the p.d.f. expresses the probability of the data given the parameters, the likelihood treats the observed data as fixed and views the function as a mapping from possible parameter values to their plausibility, conditional on the observed sample.

Definition 23 *The likelihood function of θ for a random variable y with p.d.f. $f(y;\theta_0)$ is defined to be*

$$L(\theta; y) \equiv f(y;\theta).$$

We will denote the log-likelihood function by

$$\ell(\theta; y) = \ln L(\theta; y).$$

This change in perspective is more than notational: it reflects the fundamental idea that the likelihood function measures how compatible different parameter values are with the observed data. For this reason, the MLE is defined as the value of θ that maximizes the log-likelihood function over the set of admissible parameter values.

When the sample consists of T independent and identically distributed observations of y, the joint p.d.f. of the sample is the product of the individual densities, and the sample log-likelihood function becomes:

$$\ell(\theta; Y) = \ln \prod_{t=1}^{T} f(y_t;\theta) = \sum_{t=1}^{T} \ell(\theta; y_t). \tag{7.1}$$

This additive representation will prove useful in both analytical and numerical work. It is often preferable to work with the log-likelihood rather than the likelihood itself, since the latter may take extremely large or extremely small values as the sample size grows, while the former remains well scaled and is numerically more stable.

The same reasoning extends naturally to conditional models (when the density of y_t depends on a vector of explanatory variables x_t).

Definition 24 *The conditional likelihood function of θ for a random variable y with conditional p.d.f. $f(y|x;\theta_0)$ given the random variable x is*

$$L(\theta;y|x) \equiv f(y|x;\theta).$$

We will denote the conditional log-likelihood function by

$$\ell(\theta;y|x) = \ln L(\theta;y|x).$$

The p.d.f., conditional or not, may not be defined over all possible values of the real parameter vector θ. For example, the variance parameter of a normal distribution must be positive. We will denote by Θ the set of parameter values of θ permitted by the probability model. This set is called the parameter space.

Our interest in the log-likelihood function derives from its relationship to the unknown θ_0. A special feature of the log-likelihood function is that its expectation is maximized at the parameter value θ_0, when this expectation exists.

Lemma 25 *If $\mathcal{E}\left[\sup_{\theta\in\Theta}\{\ell(\theta;y|x)\}\right]$ exists, then*

$$\mathcal{E}[\ell(\theta;y|x)|x] \leq \mathcal{E}[\ell(\theta_0;y|x)|x].$$

Proof. Left as an exercise. ■

Because the true value θ_0 maximizes the expectation of the log-likelihood function, it is natural to construct an estimator of θ_0 from the value of θ that maximizes the sample (or empirical) counterpart: the average log-likelihood function of the T observations.

Definition 26 *The MLE is a value of the parameter vector that maximizes the sample average log-likelihood function. We will denote this estimator by $\widehat{\theta}_{MLE}$:*

$$\widehat{\theta}_{MLE} = \arg\max_{\theta\in\Theta}\frac{1}{T}\sum_{t=1}^{T}\ell(\theta;y_t|x_t). \tag{7.2}$$

Because the logarithm is a monotonic transformation, maximizing the average log-likelihood is equivalent to maximizing the product of the individual likelihoods instead of (7.2):

$$\prod_{t=1}^{T} L(\theta;y_t|x_t).$$

In large samples, these two criteria lead to the same estimator, although numerical maximization is far more stable when applied to the log-likelihood.[1]

It is instructive to view the sample log-likelihood as a measure of fit: parameter values yielding higher log-likelihoods produce model-implied distributions under which the observed data are more probable. In this sense, the log-likelihood plays the same role in maximum likelihood estimation that the negative sum of squared residuals plays in least squares estimation. Whereas least squares methods select parameter values that minimize the discrepancy between observed and predicted means, the likelihood principle chooses parameters that make the entire observed sample most plausible according to the model's probabilistic structure.

The theoretical justification for this procedure stems from the fact that, under suitable regularity conditions, the expected value of the log-likelihood function is maximized at the true parameter value. This property provides the basis for the consistency of the MLE and for the asymptotic theory developed later in the book. For now, it suffices to note that the MLE can be understood as the empirical counterpart of this population maximization problem: it seeks the parameter value that maximizes the observed log-likelihood, thereby mimicking the behavior of its expected value.

7.3 Cramer-Rao Lower Bound

An important aspect of estimation theory is to understand how precise an estimator can be. When several estimators of a given parameter exist, it is natural to ask whether there is a theoretical lower bound on the variance that any unbiased estimator can achieve. The Cramér–Rao lower bound provides such a benchmark and plays a central role in defining the notion of efficiency in estimation. It shows that the covariance matrix of any unbiased estimator must be at least as large as the inverse of the Fisher information matrix, a quantity that measures the amount of information the sample conveys about the parameters of interest.

Theorem 27 *Let z be an n-component vector of random variables (not necessarily independent) the likelihood function of which is given by $L(\theta; z)$, where θ is a k-component vector of parameters in some parameter space Θ.*

[1] GAUSS tip: GAUSS uses a library called MAXLIK in order to obtain the MLE. As input, the user must provide a procedure with the $T \times 1$ vector of $\ell(\theta; y_t | x_t)$.

Let $\widetilde{\theta}(z)$ *be an unbiased estimator of* θ_0 *with a finite variance-covariance matrix. Furthermore, assume that* $L(\theta; z)$ *and* $\widetilde{\theta}(z)$ *satisfy*

$$(A)\ \mathcal{E}\ \tfrac{\partial\ell}{\partial\theta}\big|_{\theta_0} = 0$$

$$(B)\ \mathcal{E}\ \tfrac{\partial^2\ell}{\partial\theta\partial\theta'}\Big|_{\theta_0} = -\mathcal{E}\left[\tfrac{\partial\ell}{\partial\theta}\tfrac{\partial\ell}{\partial\theta'}\right]\big|_{\theta_0}$$

$$(C)\ \mathcal{E}\left[\tfrac{\partial\ell}{\partial\theta}\tfrac{\partial\ell}{\partial\theta'}\right]\big|_{\theta_0} > 0$$

$$(D)\ \textstyle\int \tfrac{\partial L}{\partial\theta}\big|_{\theta_0}\widetilde{\theta}' dz = I_k.$$

Then, for $\theta \in \Theta$

$$\mathcal{V}\left(\widetilde{\theta}\right) \geqq \left[-\mathcal{E}\ \frac{\partial^2\ell}{\partial\theta\partial\theta'}\bigg|_{\theta_0}\right]^{-1}. \tag{7.3}$$

As $\partial\ell/\partial\theta$ is a k-vector, 0 in assumption A is a vector of k zeroes. Assumption C means that the left-hand side is a positive definite matrix. The integral in assumption D is an n-tuple integral over the whole domain of the Euclidian n-space because z is an n-vector. We will now present a proof of this theorem.

Proof. Define $P = \mathcal{E}\left(\widetilde{\theta}-\theta_0\right)\left(\widetilde{\theta}-\theta_0\right)'$, $Q = \mathcal{E}\left(\widetilde{\theta}-\theta_0\right)\left(\partial\ell/\partial\theta'|_{\theta_0}\right)$, and $R = \mathcal{E}\left[(\partial\ell/\partial\theta)(\partial\ell/\partial\theta')\right]|_{\theta_0}$. Then

$$\begin{bmatrix} P & Q \\ Q' & R \end{bmatrix} \geqq 0, \tag{7.4}$$

because P is a variance-covariance matrix. Premultiply both sides of (7.4) by $[I, -QR^{-1}]$ and postmultiply them by $[I, -QR^{-1}]'$ to obtain:

$$\begin{aligned} & \begin{bmatrix} I & -QR^{-1} \end{bmatrix} \begin{bmatrix} P & Q \\ Q' & R \end{bmatrix} \begin{bmatrix} I \\ -R^{-1}Q' \end{bmatrix} \\ = & \begin{bmatrix} P - QR^{-1}Q' & Q - Q \end{bmatrix} \begin{bmatrix} I \\ -R^{-1}Q' \end{bmatrix} \\ = & \ P - QR^{-1}Q' \geqq 0, \end{aligned}$$

where R^{-1} can be defined because of assumption C. But we have

$$\begin{aligned}
Q &\equiv \mathcal{E}\left[\left(\widetilde{\theta}-\theta_0\right)\left.\frac{\partial \ell}{\partial \theta'}\right|_{\theta_0}\right] \\
&= \mathcal{E}\left[\widetilde{\theta}\left.\frac{\partial \ell}{\partial \theta'}\right|_{\theta_0}\right] \qquad \text{by assumption A} \\
&= \mathcal{E}\left[\widetilde{\theta}\left.\left[\frac{1}{L}\frac{\partial L}{\partial \theta'}\right]\right|_{\theta_0}\right] \qquad \text{because } \frac{\partial \ell}{\partial \theta'} = \frac{\partial \ln L}{\partial \theta'} = \frac{1}{L}\frac{\partial L}{\partial \theta'} \\
&= \int\left[\widetilde{\theta}\left.\left[\frac{1}{L}\frac{\partial L}{\partial \theta'}\right]L\right|_{\theta_0}\right]dz \\
&= \int \widetilde{\theta}\left.\frac{\partial L}{\partial \theta'}\right|_{\theta_0} dz \\
&= I_k \quad \text{by assumption D.}
\end{aligned}$$

As $P - QR^{-1}Q' \geqq 0$ and $Q = I_k$, we have that $P - R^{-1} \geqq 0$, or $P \geqq R^{-1}$. Therefore (7.3) follows from assumption B and from noting that $P = \mathcal{V}\left(\widetilde{\theta}\right)$.
■

The matrix

$$\mathcal{I}\left(\theta_0\right) = -\mathcal{E}\left[\left.\partial^2 \ell/\partial\theta\partial\theta'\right|_{\theta_0}\right]$$

is known as the Fisher information matrix. It quantifies how sensitive the likelihood function is to small changes in the parameter vector. When the curvature of the log-likelihood around its maximum is sharp, the information content is high, implying that the parameters can be estimated precisely. The Cramér–Rao lower bound therefore establishes that the covariance matrix of any unbiased estimator cannot be smaller than the inverse of this information matrix.

This result provides a general efficiency benchmark. If an estimator attains the bound in (7.3), it is said to be efficient, meaning that no other unbiased estimator can have a smaller variance. The theorem generalizes the Gauss–Markov result of classical linear regression: whereas the Gauss–Markov theorem identifies the best linear unbiased estimator within a particular model, the Cramér–Rao theorem applies to any unbiased estimator, linear or not, and within any properly specified probabilistic model. The two results share the same spirit—efficient use of information—but the Cramér–Rao bound offers a universal measure of achievable precision.

In the context of maximum likelihood estimation, the importance of this theorem is twofold. First, it establishes the theoretical lower limit that serves as the asymptotic variance of the MLE under correct specification. Second, it shows that the inverse of the expected negative Hessian of the log-likelihood function—the information matrix—plays a central role in inference. Under suitable conditions, the MLE attains the Cramér–Rao bound asymptotically, becoming the most efficient estimator among a wide class of consistent estimators. In this sense, the Cramér–Rao theorem not only provides a mathematical result but also underpins the practical justification for likelihood-based estimation and inference.

Assumptions A, B, and D seem arbitrary at first glance. We shall now inquire into their significance.

Assumption A is equivalent to

$$\begin{aligned}\mathcal{E}\left.\frac{\partial \ell}{\partial \theta}\right|_{\theta_0} &= \mathcal{E}\left.\left[\frac{1}{L}\frac{\partial L}{\partial \theta}\right]\right|_{\theta_0} \\ &= \int \left[\frac{1}{L}\frac{\partial L}{\partial \theta}\right] L\Big|_{\theta_0}\, dz \\ &= \int \left.\frac{\partial L}{\partial \theta}\right|_{\theta_0} dz.\end{aligned}$$

If the operations of differentiation and integration can be interchanged, the last expression is equivalent to

$$\int \left.\frac{\partial L}{\partial \theta}\right|_{\theta_0} dz = \left.\left[\frac{\partial}{\partial \theta}\int L\, dz\right]\right|_{\theta_0} = 0,$$

because $\int L\, dz$ integrates to 1.

Assumption B follows from

$$\begin{aligned}\mathcal{E}\left.\frac{\partial^2 \ell}{\partial \theta \partial \theta'}\right|_{\theta_0} &= \mathcal{E}\left.\left\{\frac{\partial}{\partial \theta}\left[\frac{1}{L}\frac{\partial L}{\partial \theta'}\right]\right\}\right|_{\theta_0} \\ &= \mathcal{E}\left.\left[-\frac{1}{L^2}\frac{\partial L}{\partial \theta}\frac{\partial L}{\partial \theta'} + \frac{1}{L}\frac{\partial^2 L}{\partial \theta \partial \theta'}\right]\right|_{\theta_0} \\ &= -\mathcal{E}\left.\left[\frac{\partial \ell}{\partial \theta}\frac{\partial \ell}{\partial \theta'}\right]\right|_{\theta_0} + \int \left.\frac{\partial^2 L}{\partial \theta \partial \theta'}\right|_{\theta_0} dz.\end{aligned}$$

Once again, if differentiation and integration are interchangeable, the last expression (always evaluated at θ_0) is equivalent to

$$\begin{aligned}
\mathcal{E}\frac{\partial^2 \ell}{\partial\theta\partial\theta'} &= -\mathcal{E}\left[\frac{\partial\ell}{\partial\theta}\frac{\partial\ell}{\partial\theta'}\right] + \int \frac{\partial^2 L}{\partial\theta\partial\theta'}dz \\
&= -\mathcal{E}\left[\frac{\partial\ell}{\partial\theta}\frac{\partial\ell}{\partial\theta'}\right] + \frac{\partial}{\partial\theta}\int \frac{\partial L}{\partial\theta'}dz \\
&= -\mathcal{E}\left[\frac{\partial\ell}{\partial\theta}\frac{\partial\ell}{\partial\theta'}\right],
\end{aligned}$$

because $\int \frac{\partial L}{\partial\theta'}\Big|_{\theta_0} dz = 0$ as maintained by assumption A.

Finally, Assumption D can be motivated in similar fashion if interchangeability is allowed. That is,

$$\begin{aligned}
\int \frac{\partial L}{\partial\theta}\bigg|_{\theta_0} \widetilde{\theta}' dz &= \left[\frac{\partial}{\partial\theta}\int L\widetilde{\theta}' dz\right]\bigg|_{\theta_0} \\
&= I_k \quad \text{because } \mathcal{E}\left(\widetilde{\theta}\right) = \theta.
\end{aligned}$$

Given that assumptions A, B, and D rely on the interchangeability between differentiation and integration, the next theorem presents the sufficient conditions for this to hold in the case where θ is a scalar. The case for a vector θ can be treated in essentially the same manner, although the notation gets more complex.

Theorem 28 *If (i) $\partial f(z,\theta)/\partial\theta$ is continuous in $\theta \in \Theta$ and z, where Θ is an open set, (ii) $\int f(z,\theta)\,dz$ exists, and (iii) $\int |\partial f(z,\theta)/\partial\theta|\,dz < M < \infty$ for all $\theta \in \Theta$, then the following condition holds*

$$\frac{\partial}{\partial\theta}\int f(z,\theta)\,dz = \int \frac{\partial f(z,\theta)}{\partial\theta}dz.$$

Proof. Using assumptions (i) and (ii), we have

$$\begin{aligned}
&\left|\int\left[\frac{f(z,\theta+h)-f(z,\theta)}{h} - \frac{\partial f(z,\theta)}{\partial\theta}\right]dz\right| \\
\leq &\int\left|\frac{f(z,\theta+h)-f(z,\theta)}{h} - \frac{\partial f(z,\theta)}{\partial\theta}\right|dz \\
= &\int\left|\frac{\partial f(z,\theta^*)}{\partial\theta} - \frac{\partial f(z,\theta)}{\partial\theta}\right|dz,
\end{aligned}$$

where θ^* is between θ and $\theta + h$. Next, we can write the last integral as $\int = \int_A + \int_{\overline{A}}$, where A is a sufficiently large compact set in the domain of z and $\overline{A}$ is its complement. But we can make $\int_A$ sufficiently small because of (i) and $\int_{\overline{A}}$ sufficiently small because of (iii). ■

7.4 Inference

Once an estimator has been obtained, the next task is to conduct inference about the parameters of the model. In the maximum likelihood framework, inference rests on the large-sample distribution of the estimator, which allows constructing confidence intervals and testing hypotheses. Although the asymptotic properties of the MLE will be derived rigorously in Chapter 9, it is useful at this stage to present their main implications and to discuss how they are applied in practice.

Under suitable regularity conditions, the first-order condition that defines the MLE can be written as:

$$\frac{1}{T}\sum_{t=1}^{T} \left.\frac{\partial \ell_t(\theta)}{\partial \theta}\right|_{\widehat{\theta}} = 0,$$

where $\ell_t(\theta)$ denotes the contribution of observation t to the log-likelihood. A linearization of this condition around the true parameter value θ_0, together with a central limit theorem for the score function, implies that the estimator is asymptotically normal. Specifically, it can be shown (as will be demonstrated in Chapter 9) that:

$$\sqrt{T}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{D} \mathcal{N}\left(0, H_0^{-1} O_0 H_0^{-1}\right)$$

where

$$O_0 = \mathcal{E}\left[\left.\frac{\partial \ell}{\partial \theta}\frac{\partial \ell}{\partial \theta'}\right|_{\theta_0}\right] = \mathcal{E}\left[g_0 g_0'\right], \quad H_0 = \mathcal{E}\left[\left.\frac{\partial^2 \ell}{\partial \theta \partial \theta'}\right|_{\theta_0}\right].$$

The matrices O_0 and H_0 are, respectively, the expected outer product of gradients (g_0) and the expected Hessian of the log-likelihood function. The product $H_0^{-1} O_0 H_0^{-1}$ provides the asymptotic covariance of the estimator. If the model is correctly specified, the information matrix equality ensures that $O_0 = -\, H_0$, in which case the expression simplifies to:

$$\sqrt{T}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{D} \mathcal{N}\left(0, -H_0^{-1}\right)$$

and the MLE is said to be asymptotically efficient, since it achieves the Cramér–Rao lower bound.

In empirical work, the matrices O_0 and H_0 are unknown and must be estimated from the data. Consistent sample analogs are obtained by replacing expectations with averages and evaluating them at the estimated parameter values:

$$\widehat{O} = \frac{1}{T}\sum_{t=1}^{T}\left[\left.\frac{\partial \ell_t}{\partial \theta}\frac{\partial \ell_t}{\partial \theta'}\right|_{\widehat{\theta}}\right], \qquad \widehat{H} = \frac{1}{T}\sum_{t=1}^{T}\left[\left.\frac{\partial^2 \ell_t}{\partial \theta \partial \theta'}\right|_{\widehat{\theta}}\right].$$

Wald tests can be conducted as usual once we have an estimator of its variance-covariance matrix. It turns out (we will show this later) that, in general, this matrix is:

$$\widehat{\theta} \overset{a}{\backsim} \mathcal{N}\left(\theta_0, H_0^{-1} O_0 H_0^{-1}\right), \tag{7.5}$$

Given that the variance covariance matrix depends on the unknown value θ_0, and if $\widehat{\theta}$ is consistent, we obtain a consistent estimator of the variance covariance matrix by replacing O_0 and H_0 with $T\widehat{O}$ and $\widehat{H}$ respectively, where:

$$\widehat{O} = \frac{1}{T}\sum_{t=1}^{T}\left[\left.\frac{\partial \ell_t}{\partial \theta}\right|_{\widehat{\theta}} \left.\frac{\partial \ell_t}{\partial \theta'}\right|_{\widehat{\theta}}\right], \qquad \widehat{H} = \left[\left.\frac{\partial^2 \ell}{\partial \theta \partial \theta'}\right|_{\widehat{\theta}}\right].$$

Replacing O_0 and H_0 with these estimates yields a consistent estimator of the asymptotic covariance matrix:

$$\widehat{\mathcal{V}}\left(\widehat{\theta}\right) = T^{-1}\widehat{H}^{-1}\widehat{O}\widehat{H}^{-1}, \tag{7.6}$$

and, under correct specification where $\widehat{O} \approx -\ \widehat{H}$,

$$\widehat{\mathcal{V}}\left(\widehat{\theta}\right) = -T^{-1}\widehat{H}^{-1},$$

or

$$\widehat{\mathcal{V}}\left(\widehat{\theta}\right) = \left(T\widehat{O}\right)^{-1}.$$

The second estimator, based on the outer product of gradients (OPG), is convenient when second derivatives are difficult or costly to compute, while the first, based on the inverse of the information matrix, is typically more precise when the model is well behaved.[2]

[2] There are several tests on the literature to evaluate whether or not $-\widehat{H}$ is statistically different from $\widehat{O}$ and are generically known as Information Matrix Tests.

Both estimators rely on the assumption that the model is correctly specified. When the true data-generating process differs from the assumed probability model, the information matrix equality no longer holds, and only the more general form in (7.6) remains consistent. This version is therefore robust to mild forms of misspecification and constitutes the basis for the sandwich variance estimator, which will be formally justified in Chapter 9 when the asymptotic distribution of the MLE is derived rigorously.

Once an estimator of the covariance matrix is obtained, inference can be conducted as usual. In particular, if we want to test the null $\mathrm{H}_0 : Q'\theta = c$ we have:

$$\left(Q'\widehat{\theta} - c\right)' \left[-Q'\widehat{\mathcal{V}}\left(\widehat{\theta}\right)Q\right]^{-1} \left(Q'\widehat{\theta} - c\right) \xrightarrow{D} \chi^2_q,$$

where q is the number of restrictions tested (rank of Q).

A more powerful test, known as the likelihood ratio test (LRT) can be constructed as

$$2\left[\ell\left(\widehat{\theta}\right) - \ell\left(\overline{\theta}\right)\right] \xrightarrow{D} \chi^2_q,$$

where, $\overline{\theta}$ is the MLE of the model in which the constraints are imposed.

The results presented in this section summarize the essential ingredients of inference under maximum likelihood estimation. Their theoretical foundations—consistency, asymptotic normality, and asymptotic efficiency—will be formally developed in Chapter 9, where these results are derived as special cases of a more general large-sample theory for extremum estimators.

7.5 Invariance Property of the MLE

One of the most useful features of maximum likelihood estimation is that it is invariant under reparameterization.

Unlike many other estimation procedures, the MLE of a transformed parameter can be obtained directly from the MLE of the original parameter, without re-estimating the model.

This property reflects the fact that the likelihood principle depends only on the probabilistic structure of the model and not on the particular way in which the parameters are expressed.

Theorem 29 *Let $\widehat{\theta}$ denote the maximum likelihood estimator of a parameter vector $\theta \in \Theta$ based on the likelihood function $L(\theta; Y)$. Let $\psi = g\left(\theta\right)$ be a one-to-one and differentiable transformation from Θ into a new parameter space*

Ψ. *Then the MLE of ψ is*

$$\widehat{\psi} = g\left(\widehat{\theta}\right).$$

Proof. Maximizing $L(\theta; Y)$ with respect to θ is equivalent to maximizing $L(g^{-1}(\psi); Y)$ with respect to ψ, since g is one-to-one and the likelihood function is defined through the same probability model.

Therefore, the value of ψ that maximizes the likelihood is the image of the maximizer of $L(\theta; Y)$ under g. Thus:

$$\widehat{\psi} = g\left(\widehat{\theta}\right).$$

■

The invariance property ensures that the likelihood principle yields results that are independent of the chosen parameterization. In practice, once the MLE of a parameter is obtained, any smooth transformation of it—such as a variance instead of a standard deviation, a rate parameter instead of a mean, or an elasticity derived from structural coefficients—is automatically the MLE of the transformed quantity.

This result simplifies both analytical and numerical work: there is no need to redefine or re-maximize the likelihood after a change of variables.

7.6 Quasi-Maximum Likelihood

Is it possible for an estimator based on the likelihood function associated with a parametric p.d.f. family to possess good asymptotic properties even if the p.d.f. family is misspecified in the sense of not encompassing the true p.d.f.? This question frames the problem area of Quasi-Maximum Likelihood (QML) estimation (also referred to as pseudo-ML estimation), which is concerned with methods of generating consistent and asymptotically normal estimators for model parameters of potentially misspecified parametric likelihood functions or models.

For example, the OLS and NLLS estimators are equivalent to the MLE that imposes normality on u. As we will show later, consistency and asymptotic normality of these estimators do not require normality. Thus, OLS and NLLS are examples of QML estimators. In particular, we act as if the p.d.f. of Y conditional on X is in the normal parametric family, even if we are not confident that it really is, and we derive the MLE of the parameters from the normal-based likelihood function. Within the regression model context, we achieve consistency and asymptotic normality under general conditions even if the likelihood function is misspecified. If the p.d.f. specification happens to be correct, we generally gain the ML property of asymptotic efficiency. This observation grants us wide latitude in invoking the normality assumption in regression-type models.

Furthermore, one sacrifices ML asymptotic efficiency only if one actually knows the correct parametric family of p.d.f.s but does not use it in the ML estimation problem and instead uses a QML estimator. In the case where the analyst does not know the correct parametric p.d.f., then for all practical purposes the question of ML asymptotic efficiency is moot (one need not lament the loss of ML asymptotic efficiency when there is no basis for achieving it).

More formally, let $\ell_Q(\theta; Y)$ denote a quasi-likelihood function based on a postulated joint p.d.f. for a random sample of Y that may or may not coincide with the true p.d.f. of Y. The QML estimator (QMLE) of θ_0 is defined as

$$\widehat{\theta}_{QMLE} = \arg\max_{\theta\in\Theta} \frac{1}{T}\sum_{t=1}^{T} \ell_Q(\theta; y_t).$$

Under quite general conditions, this estimator will be consistent and asymptotically normal, and will have the asymptotic distribution presented in (7.5).

However, unless the ℓ_Q is in fact a correctly specified likelihood, the covariance matrix is not given by the inverse of the negative of the Hessian or the outer product of gradients given that $-H_0$ will not coincide with O_0. In this case, a consistent estimator of the variance is:

$$\widehat{\mathcal{V}}\left(\widehat{\theta}_{QMLE}\right) = T^{-1}\widehat{H}^{-1}\widehat{O}\widehat{H}^{-1}. \tag{7.7}$$

Thus, unless the analyst is confident of the functional specification of the likelihood function, the asymptotic covariance matrix that simplifies $-H_0 = O_0$ such be avoided.

To provide an overview of the sense in which the QMLE approximates the true p.d.f. of a random sample, next we introduce the Kullback-Leibler discrepancy, which is a measure of the proximity of one p.d.f. to another.

The discrepancy of $p(y)$ relative to $f(y)$ is defined by

$$\mathrm{KL}(p, f) = \mathcal{E}_p \left[\ln \left[\frac{p(y)}{f(y)} \right] \right],$$

where $\mathcal{E}_p[\cdot]$ denotes the expectation taken with respect to the distribution $p(y)$. The KL discrepancy measure has several useful statistical properties. Of principal interest is that $\mathrm{KL}(p, f) \geq 0$, $\mathrm{KL}(p, f)$ is a strictly convex function of $p \geq 0$, and $\mathrm{KL}(p, f) = 0$ iff $p(y) = f(y)$ for every y. The closer $\mathrm{KL}(p, f)$ is to zero, the more similar is the distribution p relative to f. The greater the value of $\mathrm{KL}(p, f)$, the greater the discrepancy between both distributions.

Regarding the QMLE, suppose that the true p.d.f. underlying the DGP is given by $p(y)$ and that the analyst bases the definition of the quasi-likelihood function on the p.d.f. family $f(y;\theta)$ for $\theta \in \Theta_Q$. Define θ^* to be the value of θ that minimizes the Kullback-Leibler discrepancy of $p(y)$ relative to $f(y;\theta)$, that is,

$$\begin{aligned}\theta^* &= \arg\min_{\theta \in \Theta_Q} \mathrm{KL}(p, f; \theta) \\ &= \arg\min_{\theta \in \Theta_Q} \mathcal{E}_p \left[\ln \left[\frac{p(y)}{f(y;\theta)} \right] \right].\end{aligned}$$

Thus, θ^* is the parameter value associated with the quasi-likelihood specification that results in the least Kullback-Leibler discrepancy between the true p.d.f. underlying the DGP of y and the set of p.d.f. candidates available from within the analyst's quasi-likelihood specification of the probability model. Under general regularity conditions, it can be shown that $\widehat{\theta}_{QMLE} \xrightarrow{p} \theta^*$. In other words, the QMLE consistently estimates the particular value of θ that provides the closest match possible between the analyst's probability model and the true p.d.f. of y.

While extremely useful, this result also imposes the requirement of discipline for the analyst, in the sense that in order to minimize the Kullback-Leibler discrepancy he must use a candidate p.d.f. that closely resembles p. For example, it may be the case that even when the OLS estimator is a QMLE that is consistent and asymptotically normal, it may not minimize the Kullback-Leibler discrepancy if some features of the data are not well captured by imposing normality. At any rate, inference can always be conducted using (7.7).

7.7 Numerical Methods

In many econometric models, the likelihood function does not yield a closed-form solution for its maximizer. Numerical methods are then required to obtain the maximum likelihood estimator. Among the various optimization algorithms available, the Newton–Raphson and quasi-Newton methods are the most frequently used because of their generality and convergence speed. The ideas behind these methods are straightforward: by approximating the log-likelihood function with a quadratic expansion around a current guess, one can update the parameter vector iteratively until the procedure converges to the value that maximizes the function.

7.7.1 Newton-Raphson Method

Numerical algorithms are often used in order to solve nonlinear problems for which an analytical solution is not available. As most numerical methods, the Newton-Raphson algorithm solves an optimization problem that is not the one under consideration, but an approximation of it.

Consider a second-order (quadratic) Taylor series approximation of $\ell(\theta)$ about a starting value θ_1

$$\ell(\theta) \cong \ell(\theta_1) + g_1'(\theta - \theta_1) + \frac{1}{2}(\theta - \theta_1)' H_1 (\theta - \theta_1), \tag{7.8}$$

where g and H are defined as

$$g_1 = \left.\frac{\partial \ell}{\partial \theta}\right|_{\theta_1} \qquad H_1 = \left.\frac{\partial^2 \ell}{\partial \theta \partial \theta'}\right|_{\theta_1}.$$

This approximation is now linear-quadratic in θ (given that θ_1 is known) and is strictly concave if the Hessian matrix is negative definite. As the objective function is linear-quadratic, the FONC will be linear in θ and can be solved analytically. Conditional on the starting value θ_1, the next iterate is

$$\theta_2 = \theta_1 - H_1^{-1} g_1. \tag{7.9}$$

By repeating the process, we solve a sequence of conditionally linear-quadratic problems in which the current solution θ_j is derived from the previous solution θ_{j-1}. We repeat the Newton-Raphson iterations until convergence ($\theta_j \cong \theta_{j-1}$) and set $\theta_j = \widehat{\theta}_{ML}$.[3]

Under the Newton-Raphson algorithm, the step taken during iteration j, $\delta_j = \theta_j - \theta_{j-1}$, is called the *direction*. The direction is the vector describing a segment of a path from the starting point to the next step in the iterative solution. The inverse of the Hessian matrix determines the *angle* of the direction, and the gradient determines the *size* of the direction.

Inserting iteration (7.9) back into the approximation (7.8) yields

$$\ell\left(\theta_2\right) \cong \ell\left(\theta_1\right) - \frac{1}{2}\left(\theta_2 - \theta_1\right)' H_1 \left(\theta_2 - \theta_1\right). \tag{7.10}$$

Equation (7.10) shows a weakness of this method: Even if (7.10) holds exactly, $\ell\left(\theta_2\right) > \ell\left(\theta_1\right)$ is not guaranteed unless H_1 is a negative matrix (or equivalently, unless ℓ is locally concave). Another weakness is that even if H_1 is negative definite, $\theta_2 - \theta_1$ may be too large or too small; if it is too large, it overshoots the target; if it is too small, the speed of convergence is slow.

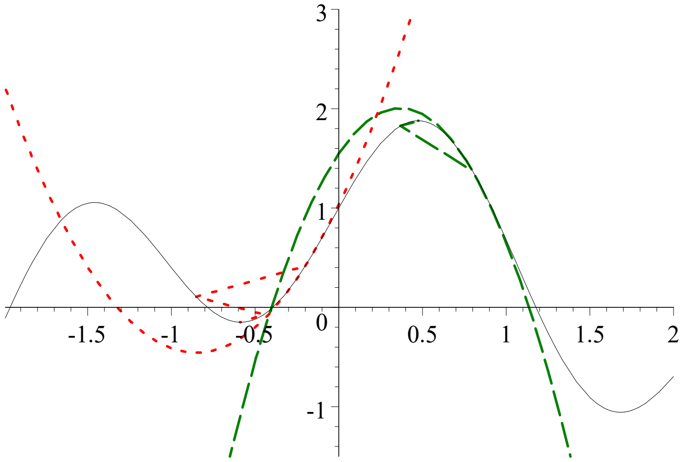

Figure 7.1: Newton-Raphson Method

[3] The same criteria that we discussed for evaluating the convergence of the iterative solution for the Gauss-Newton method can be applied here.

Figure 7.1 presents an example of an objective function (the continuous line) that is not globally concave. If 0.8 is chosen as a starting value, the dashed line displays the second order Taylor approximation of the objective function which corresponds to equation (7.8). From that point, the optimization of this approximation leads to the value of the next iterate that maximizes this approximation. Then, a new approximation is performed in this point, and the algorithm converges to the global maximum. On the other hand, if -0.2 is chosen as the starting value, the approximation (dotted line) renders a convex function and the algorithm converges to a minimum instead of a maximum.[4]

If we are certain that the problem has a unique solution, we can use a simplified version of the Newton-Raphson iteration, known as the method of steepest ascent. In this case, the Hessian matrix is replaced with the negative of an identity matrix

$$\theta_2 = \theta_1 + g_1.$$

As the term "steepest ascent" implies, the updating of θ_j in the algorithm is solely determined by the slope of $\ell\left(\theta_j\right)$. In general, this method converges to the solution more slowly than the Newton-Raphson method, but the computing time may be reduced if the Hessian matrix is especially costly to obtain.

When the Hessian fails to be negative definite, the search direction may point toward decreases of ℓ. This is simple enough to check and the response is to check in the opposite direction by changing the sign of the search direction in (7.9) as

$$\theta_2 = \theta_1 - \left(H_1 - \alpha_1 I\right)^{-1} g_1,$$

where I is the identity matrix and α_1 is a scalar to be appropriately chosen by the researcher subject to the condition that $H_1 - \alpha_1 I$ is negative definite. This modification is called quadratic hill-climbing. Choosing a large value of α makes the search direction similar to that of steepest ascent.

7.7.2 Quasi-Newton Methods

In many cases, the Hessian matrix used to compute the direction of the Newton-Raphson algorithm is difficult to state or computationally expensive

[4] Once again, the importance of the choice of the starting value is crucial. The same considerations that were discussed when presenting the Gauss-Newton method are valid here.

to form. Consequently, researchers have sought ways to produce an inexpensive estimate of the Hessian matrix. The quasi-Newton methods are based on using the information of the current iteration to compute the new Hessian matrix. Let $\delta_j = \theta_j - \theta_{j-1}$ be the change in the parameters in the current iteration and $\eta_j = g_j - g_{j-1}$ be the change in the gradient vector. Then, a natural estimate of the Hessian matrix at the next iteration, h_{j+1}, would be the solution of the system of linear equations

$$h_{j+1}\delta_j = \eta_j.$$

That is, h_{j+1} is effectively the ratio of the change in the gradient to the change in the parameters. This is called the quasi-Newton condition. There are many solutions to this set of linear equations. One of the most popular solutions is based on secant updates and is known as BFGS (for Broyden, Fletcher, Goldfarb, and Shanno) which is usually regarded as the best-performing method. The iterative steps used to update the Hessian matrix estimate take the form

$$h_{j+1} = h_j + \frac{\eta_j \eta_j'}{\eta_j' \delta_j} - \frac{g_j g_j'}{g_j' g_j}.$$

7.7.3 Numerical Tips

Oftentimes, the parameter space is precluded from adopting some values. One simple way to ensure that a numerical search always stays with certain specified boundaries is to reparameterize the likelihood function in terms of a vector that incorporates the desired restrictions.

For example, to ensure that a parameter a is always between ±1, we take

$$a = \frac{l}{1 + |l|}.$$

The goal is to find the value l that maximizes the log likelihood. No matter what value l takes, the value of a will always be less than 1 in absolute value and the likelihood function will be well defined. Once we have the value of $\widehat{l}$ that maximizes the likelihood function, the MLE of a is then given by

$$\hat{a} = \frac{\hat{l}}{1 + \left|\hat{l}\right|}.$$

Reparameterizing the likelihood function so that the estimates satisfy any necessary constraints is often easy to implement. However, if a standard error is calculated from the matrix of second derivatives of the log likelihood, it will correspond to the standard error of $\hat{l}$ and not of $\hat{a}$. To obtain the standard errors for $\hat{a}$, the best approach is first to parameterize the likelihood function in terms of l to find the MLE, and then to reparameterize it in terms of a to calculate the matrix of second derivatives evaluated at $\hat{a}$ and obtain the final standard error for $\hat{a}$. Alternatively, one can calculate an approximation to the standard error for $\hat{a}$ from the standard error for $\hat{l}$ based on the formula for a Wald test of a nonlinear hypothesis.

Another common restriction one needs to impose is for the variance parameter σ^2 be positive. An obvious way to achieve this is to parameterize the likelihood in terms of l which represents ±1 times the standard deviation. The procedure to evaluate the log likelihood then begins by squaring this parameter l:

$$\sigma^2 = l^2$$

and if the standard deviation σ is itself called, it is calculated as

$$\sigma = \sqrt{l^2}.$$

Other times, some of the unknown parameters are probabilities $\pi_1, \cdots, \pi_k$ which must satisfy the restrictions

$$\begin{aligned} 0 \ &\leq \ \pi_i \leq 1 \quad \text{for } i = 1, 2, \cdots, k \\ \sum_{i=1}^{k} \pi_i \ &= \ 1. \end{aligned}$$

In this case, one common approach is to parameterize the probabilities in terms of $l_1, l_2, \cdots, l_{k-1}$, where

$$\begin{aligned} \pi_i \ &= \ l_i^2 / \left(1 + l_1^2 + l_2^2 + \cdots + l_{k-1}^2\right) \quad \text{for } i = 1, 2, \cdots, k-1 \\ \pi_k \ &= \ 1 / \left(1 + l_1^2 + l_2^2 + \cdots + l_{k-1}^2\right). \end{aligned}$$

For more complex inequality constraints that do not admit a simple reparameterization, an approach that sometimes works is to put a branching statement in the procedure to evaluate the log likelihood function. The procedure first checks whether the constraint is satisfied. If it is, then the likelihood function is evaluated in the usual way. If it is not, the procedure returns a large negative number in place of the value of the log likelihood function. Sometimes such an approach will allow an MLE satisfying the specified conditions to be found with simple numerical search procedures. This approach may not be recommended when the values of the parameters that maximize the objective function are actually close to the boundaries. For such cases, more complex algorithms are available.[5]

7.8 Applications

The following examples illustrate the application of the results obtained in the previous sections. In both cases, the steps are explicit so that the reader can connect the theoretical definitions of the likelihood function, the MLE, and the information matrix with their practical computation. The first example concerns a simple univariate distribution, while the second revisits the classical normal regression model.

7.8.1 MLE of the Exponential Distribution

Let y be distributed exponential with parameter θ_0. Its p.d.f. is given by

$$f\left(y_t;\theta_0\right)=\frac{1}{\theta_0}e^{-\frac{y_t}{\theta_0}};\quad y_t>0,\theta_0>0.$$

It is easy to verify that $\mathcal{E}\left(y\right)=\theta_0$ given that

$$\begin{aligned}\mathcal{E}\left(y\right) &= \int_0^\infty z\,f\left(z;\theta_0\right)\,dz\\ &= \int_0^\infty z\,\frac{1}{\theta_0}e^{-\frac{z}{\theta_0}}\,dz\\ &= -ze^{-\frac{z}{\theta_0}}\Big|_0^\infty+\int_0^\infty e^{-\frac{z}{\theta_0}}\,dz\\ &= 0-\theta_0e^{-\frac{z}{\theta_0}}\Big|_0^\infty\\ &= \theta_0,\end{aligned}$$

and $\mathcal{V}\left(y\right)=\mathcal{E}\left(y^2\right)-\left[\mathcal{E}\left(y\right)\right]^2=\theta_0^2$ given that

[5] GAUSS has two libraries that can be used. CML is a library that specializes in constrained maximum likelihood estimation. CO is a library for general purpose constrained optimization problems.

$$\begin{aligned}
\mathcal{E}\left(y^2\right) &= \int_0^\infty z^2\, f\left(z;\theta_0\right)\, dz \\
&= \int_0^\infty z^2\, \frac{1}{\theta_0} e^{-\frac{z}{\theta_0}}\, dz \\
&= -z^2 e^{-\frac{z}{\theta_0}}\Big|_0^\infty + 2\int_0^\infty z e^{-\frac{z}{\theta_0}}\, dz \\
&= 0 - 2\theta_0^2 e^{-\frac{z}{\theta_0}}\Big|_0^\infty \\
&= 2\theta_0^2.
\end{aligned}$$

An alternative way to derive the central moments of this distribution is to use its Moment Generating Function $(M\left(t\right))$ which is:

$$M\left(t\right) = \left(1-\theta_0 t\right)^{-1}$$

in which case,

$$\begin{aligned}
\mathcal{E}\left(y\right) &= \left.\frac{\partial M\left(t\right)}{\partial t}\right|_{t=0} \\
&= \left.\frac{\theta_0}{\left(1-\theta_0 t\right)^2}\right|_{t=0} \\
&= \theta_0
\end{aligned}$$

and

$$\begin{aligned}
\mathcal{E}\left(y^2\right) &= \left.\frac{\partial^2 M\left(t\right)}{\partial t^2}\right|_{t=0} \\
&= \left.\frac{2\theta_0^2}{\left(1-\theta_0 t\right)^3}\right|_{t=0} \\
&= 2\theta_0^2.
\end{aligned}$$

The sample log-likelihood function for T observations of y is:

$$\begin{aligned}
\ell\left(\theta;Y\right) &= \sum_{t=1}^{T} \ell\left(\theta;y_t\right) \\
&= -T\ln\theta - \frac{1}{\theta}\sum_{t=1}^{T} y_t.
\end{aligned}$$

The FONC is:

$$\begin{aligned}\frac{\partial \ell(\theta;Y)}{\partial\theta} &= -\frac{T}{\theta}+\frac{1}{\theta^2}\sum_{t=1}^{T} y_t \\ &\Rightarrow \left.\frac{\partial \ell(\theta;Y)}{\partial\theta}\right|_{\widehat{\theta}} = 0 \\ &\Rightarrow \widehat{\theta} = \frac{\sum_{t=1}^{T} y_t}{T},\end{aligned}$$

and the SOSC is:

$$\begin{aligned}\frac{\partial^2 \ell(\theta;Y)}{\partial\theta^2} &= \frac{T}{\theta^2}-\frac{2}{\theta^3}\sum_{t=1}^{T} y_t \\ &\Rightarrow \left.\frac{\partial^2 \ell(\theta;Y)}{\partial\theta^2}\right|_{\widehat{\theta}} = -\frac{T}{\widehat{\theta}^2} < 0,\end{aligned}$$

which tells us that $\widehat{\theta}$ indeed is the $\arg\max_{\theta} \ell(\theta;Y)$.

Notice that $\widehat{\theta}$ is unbiased given that

$$\mathcal{E}\left(\widehat{\theta}\right) = \mathcal{E}\left(\frac{\sum_{t=1}^{T} y_t}{T}\right) = \theta_0$$

and

$$\begin{aligned}\mathcal{V}\left(\widehat{\theta}\right) &= \mathcal{V}\left(\frac{\sum_{t=1}^{T} y_t}{T}\right) \\ &= \frac{1}{T^2}\sum_{t=1}^{T}\mathcal{V}(y_t) \\ &= \frac{\theta_0^2}{T}.\end{aligned}$$

Next, we verify that $\mathcal{V}\left(\widehat{\theta}\right)$ attains the Cramer-Rao lower bound. To do so, we verify that the assumptions of Theorem 27 are satisfied.

$$(A) \quad \mathcal{E}\left.\frac{\partial \ell}{\partial\theta}\right|_{\theta_0} = \mathcal{E}\left[-\frac{T}{\theta_0}+\frac{1}{\theta_0^2}\sum_{t=1}^{T} y_t\right] = 0$$

$$
\begin{aligned}
(B)\ \ \mathcal{E}\left.\frac{\partial^2 \ell}{\partial\theta\partial\theta'}\right|_{\theta_0} &= \mathcal{E}\left[\frac{T}{\theta_0^2}-\frac{2}{\theta_0^3}\sum_{t=1}^{T} y_t\right] = -\frac{T}{\theta_0^2} \\
&= -\mathcal{E}\left.\left[\frac{\partial\ell}{\partial\theta}\frac{\partial\ell}{\partial\theta'}\right]\right|_{\theta_0} \\
&= -\,\mathcal{E}\sum_{t=1}^{T}\left.\left[-\frac{1}{\theta}+\frac{1}{\theta^2}y_t\right]^2\right|_{\theta_0} \\
&= -\frac{T}{\theta_0^2}
\end{aligned}
$$

$$
(C)\ \ \mathcal{E}\left.\left[\frac{\partial\ell}{\partial\theta}\frac{\partial\ell}{\partial\theta'}\right]\right|_{\theta_0} = \frac{T}{\theta_0^2} > 0
$$

$$
\begin{aligned}
(D)\ \ \int\left.\frac{\partial L}{\partial\theta}\right|_{\theta_0}\widehat{\theta}' dz &= \int\left[\partial\left(\theta^{-T}e^{-\frac{\sum y}{\theta}}\right)/\partial\theta\right]_{\theta_0}\widehat{\theta}dy \\
&= \int\left[-\frac{\sum y}{\theta_0}+\frac{\left(\sum y\right)^2}{T\theta_0^2}\right]L\left(\theta_0\right)dy \\
&= -1+2=1.
\end{aligned}
$$

Given that the Cramer-Rao assumptions hold, we known that

$$
\mathcal{V}\left(\widehat{\theta}\right) \geq \left[-\mathcal{E}\left.\frac{\partial^2\ell}{\partial\theta\partial\theta'}\right|_{\theta_0}\right]^{-1},
$$

but

$$
\left[-\mathcal{E}\left.\frac{\partial^2\ell}{\partial\theta\partial\theta'}\right|_{\theta_0}\right]^{-1} = \frac{\theta_0^2}{T},
$$

which coincides with $\mathcal{V}\left(\widehat{\theta}\right)$. Then $\widehat{\theta}$ attains the Cramer-Rao lower bound.

Given that θ_0 is unknown, the estimator of $\mathcal{V}\left(\widehat{\theta}\right)$ is

$$
\widehat{\mathcal{V}}\left(\widehat{\theta}\right) = \frac{\widehat{\theta}^2}{T}.
$$

As we mentioned in Section 7.5, one desirable property of MLE is that a reparameterization of the model leads to the same results (invariance to reparameterization). In this case it can be verified that if we define the p.d.f. of the exponential distribution in terms of $\vartheta_0 = 1/\theta_0$ we have:

$$f\left(y_t; \vartheta_0\right) = \vartheta_0 e^{-\vartheta_0 y_t}; \quad y_t > 0, \vartheta_0 > 0$$

and

$$\widehat{\vartheta} = \frac{T}{\sum_{t=1}^{T} y_t} = \frac{1}{\widehat{\theta}}.$$

7.8.2 MLE of the Normal Linear Regression Model

We already know that the MLE for β_0 and σ_0^2 in the NLRM are:

$$\begin{aligned} \widehat{\beta} &= \left(X'X\right)^{-1}\left(X'Y\right) \\ \widehat{\sigma}^2 &= T^{-1}\widehat{u}'\widehat{u} \end{aligned}$$

which are obtained from maximizing the log-likelihood function:

$$\begin{aligned} \ell\left(\theta; Y \left| X \right.\right) &= \sum_{t=1}^{T} \ell\left(\theta; y_t \left| x_t \right.\right) \\ &= -\frac{T}{2}\ln\left(2\pi\right) - \frac{T}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\left(Y - X\beta\right)'\left(Y - X\beta\right). \end{aligned}$$

To investigate the efficiency of the MLE we can obtain the Cramer-Rao lower bound by using the OPG and the Hessian matrix. They can be obtained from:

$$\begin{aligned} \frac{\partial \ell}{\partial \beta} &= \frac{1}{\sigma^2}\left(X'Y - X'X\beta\right) \\ \frac{\partial \ell}{\partial \sigma^2} &= -\frac{T}{2\sigma^2} + \frac{1}{2\sigma^4}\left(Y - X\beta\right)'\left(Y - X\beta\right) \\ \frac{\partial^2 \ell}{\partial \beta \partial \beta'} &= -\frac{1}{\sigma^2}X'X \\ \frac{\partial^2 \ell}{\partial \left(\sigma^2\right)^2} &= \frac{T}{2\sigma^4} - \frac{1}{\sigma^6}\left(Y - X\beta\right)'\left(Y - X\beta\right) \\ \frac{\partial^2 \ell}{\partial \beta \partial \sigma^2} &= -\frac{1}{\sigma^4}\left(X'Y - X'X\beta\right). \end{aligned}$$

As the Cramer-Rao assumptions are satisfied, the lower bound for any unbiased estimator is the inverse of information matrix which is:

$$\left[-\mathcal{E}\, \frac{\partial^2 \ell}{\partial\theta\partial\theta'}\bigg|_{\theta_0} \right]^{-1} = \begin{bmatrix} \sigma_0^2 \left(X'X\right)^{-1} & 0 \\ 0 & \frac{2\sigma_0^4}{T} \end{bmatrix}. \tag{7.11}$$

We known that $\mathcal{E}\left(\widehat{\beta}\right) = \beta_0$ and $\mathcal{V}\left(\widehat{\beta}\right) = \sigma_0^2 \left(X'X\right)^{-1}$, thus $\widehat{\beta}$ attains the lower bound, but $\widehat{\sigma}^2$ is biased and has $\mathcal{V}\left(\widehat{\sigma}^2\right) = 2\sigma_0^4 \left(T-k\right)/T^2$ and does not attain the lower bound. Even the unbiased estimator $\widetilde{\sigma}^2$ cannot attain this lower bound given that $\mathcal{V}\left(\widetilde{\sigma}^2\right) = 2\sigma_0^4/\left(T-k\right)$. In fact, no unbiased estimator for σ^2 attains this lower bound. Nevertheless, both $\widehat{\sigma}$ and $\widetilde{\sigma}$ are asymptotically efficient, in the sense that both are asymptotically unbiased and attain the lower bound asymptotically.

7.9 Further Reading

A detailed treatment of maximum likelihood estimation, its regularity conditions, and its asymptotic properties can be found in Amemiya (1985), Hansen (2022), Hayashi (2000), and Ruud (2000).

Comprehensive discussions of the Cramér–Rao lower bound and the information matrix equality are given in Rao (1973) and Mittelhammer, Judge, and Miller (2000).

White (1982) and Huber (1981) provide the foundations for the quasi–maximum likelihood estimator and robust inference under misspecification.

Gallant (1987) offers a detailed account of nonlinear likelihood models and numerical aspects of estimation, while Thisted (1988) presents computational algorithms for maximum likelihood procedures.

For further reading on numerical implementation and convergence properties, see Davidson and MacKinnon (1993) and Hamilton (1994).

7.10 Workout Problems

1. Prove Lemma 25.

2. Let y be a random variable that follows a $\mathcal{U}\left(0,\theta_0\right)$ distribution. Find the MLE.

3. (Normal Linear Regression) Let $y_t \,|x_t \backsim \mathcal{N}\left(x_t'\beta_0, \sigma_0^2\right)$

(**a**) Obtain the sample log-likelihood function for a sample of size T.

(**b**) Verify that $\mathcal{E}\left[\ell\left(\theta; Y\,|X\right)|X\right]$ is uniquely maximized at $\beta = \beta_0$ and $\sigma^2 = \sigma_0^2$.

(**c**) Verify that the assumptions for the Cramer-Rao lower bound in Section 7.8.2 are satisfied.

(**d**) Prove 7.11.

4. (Student t Linear Regression) Assume that the random variable $\left(y_t - x_t'\beta_0\right)/\sigma_0$ has a S_{v_0} distribution conditional on x.

(**a**) Obtain the sample log-likelihood function for a sample of size T.

(**b**) Verify that $\mathcal{E}\left[\ell\left(\theta; Y\,|X\right)|X\right]$ exists.

5. Let y be a random variable that follows an Erlang distribution (a special case of the Gamma distribution). Its p.d.f. is:

$$f\left(y; \alpha_0, \lambda_0\right) = \frac{\left(\alpha_0/\lambda_0\right)^{\alpha_0 - 1} e^{-y/\lambda_0}}{\lambda_0\left(\alpha_0 - 1\right)},\ y > 0,\ \lambda_0 > 0,\ \alpha_0 \in \mathbb{N}$$

(**a**) Find the mean, variance, skewness and kurtosis of y. [Tip: The moment generating function of this distribution is $M\left(t\right) = \left(1 - \lambda_0 t\right)^{-\alpha_0}$]

(**b**) Assuming that you know the value of α_0, find the MLE of λ_0 and verify that it is unbiased.

(**c**) Verify that the Cramer-Rao conditions are satisfied and obtain the variance of $\widehat{\lambda}$.

(**d**) If you observe the following sample of y: $(2, 8, 4, 7, 9)$ and you know that $\alpha_0 = 3$, test the null hypothesis $\mathrm{H}_0 : \lambda_0 = 2.5$ using Wald and LRT tests.

Chapter 8

Asymptotic Theory

8.1 Introduction

Once we leave the context of OLS with fixed regressors and normally distributed errors, it is frequently impossible (or at least impractical) to obtain exact results. By results, we mean that given a data set Y and a vector of parameter estimates $\widehat{\theta}$, we would like to know how 'far' is $\widehat{\theta}$ from θ_0 (true value of θ) and how to test an hypothesis about θ_0 with the observed sample (make inference).

Much of distribution theory uses to answer these questions with asymptotic theory, that is, the theory that describes the properties of estimators when the sample size is infinitely large.

In previous chapters, we derived the finite-sample properties of estimators under strong assumptions—fixed regressors, normality, and exact specification. However, outside these ideal conditions, exact distributional results are rarely available. Asymptotic theory provides a bridge between the tractable and the realistic: by studying the behavior of estimators as the sample size grows, we obtain approximations that justify inference when finite-sample formulas fail.

Of course, in reality we don't have infinitely large samples (statistics and econometrics would be meaningless sciences), thus asymptotic theory is used as an approximation (sometimes good, sometimes bad). Unfortunately, more accurate approximations are usually available in very simple and limited cases.[1]

[1] The main procedure used to get evidence in these cases is by means of Monte Carlo

In the previous chapters we relied mainly on finite-sample results derived for the classical linear model or specific likelihood structures. Those results depend on strong distributional assumptions that seldom hold exactly in applied work. Asymptotic theory provides the bridge between exact small-sample derivations and the general properties of modern estimators. By studying the behavior of estimators as the sample size grows, we can establish whether they are consistent and how they are distributed in large samples. These two properties—consistency and asymptotic normality—will underpin the analysis of extremum estimators such as NLLS and MLE in the following chapter.

The two questions posed are answered with the Law of Large Numbers (how 'far' is $\widehat{\theta}$ from θ_0) and the Central Limit Theorem (inference about θ_0) which we will discuss here.

The goal of this chapter is therefore twofold: to introduce the main tools that justify large-sample approximations, and to illustrate how these tools apply to linear regression estimators. The discussion will emphasize intuition as well as formal statements, keeping in mind that asymptotic results are approximations to finite-sample behavior.

This document is organized as follows: Section 8.2 provides some definitions that will be used on the rest of the document. Section 8.3 presents several Laws of Large Numbers (LLN). Section 8.4 introduces several Central Limit Theorems (CLT). Section 8.5 discusses the asymptotic properties of OLS, GLS, and FGLS. Finally, Section 8.6 ends be presenting some applications.

8.2 Preliminaries

Here we introduce a few concepts that are required in the subsequent analysis. In particular, the rigorous definition of a random variable and modes of converge of sequences of random variables.

8.2.1 Sequences, Limits, and Random Variables

Definition 30 *A sequence is defined as a countable infinite collection of ordered things.*

experiments, but as it is impractical to perform these experiments every time we want to perform a test, asymptotic theory is necessary.

Thus, we may construct sequences of numbers, vectors, matrices, etc. For asymptotic analysis we will be primarily interested in sequences that have finite limits.

Definition 31 *A real-valued sequence* $\{a_t\}$ *is said to converge to* a *if for any* $\varepsilon > 0$*, there exists an* N *such that* $\forall T > N$

$$\|a_T - a\| < \varepsilon,$$

where $\|\cdot\|$ *is the Euclidean distance. We write* $\lim_{T\to\infty} a_T = a$ *or* $a_T \to a$*.*

A sequence that converges is convergent.

In order to notice that a random variable is simply a function, we will first introduce some additional definitions. In concrete terms, a sample space Ω is the set of all possible outcomes of an experiment. Thus, in the experiment of throwing a die, the six faces of the die constitute the sample space. A subset of a sample space may be called an event. Thus, we speak of the event of an ace turning up or the event of an even number showing in the throw of a die. With each event, we associate a real number between 0 and 1 called the probability of the event.

When we think of a sample space, we often think of the other two concepts as well: the collection of its subsets (events) and the probabilities attached to them. The term probability space refers to all three concepts collectively.

Definition 32 *The collection* $\mathcal{A}$ *of subsets of* Ω *is called a* σ*-algebra if it satisfies the following properties:*

i) $\Omega \in \mathcal{A}$,

ii) $E \in \mathcal{A} \Rightarrow \overline{E} \in \mathcal{A}$ *(*$\overline{E}$ *refers to the complement of* E *with respect to* Ω*),*

iii) $E_j \in \mathcal{A}$, $j = 1, 2, \cdots \Rightarrow \cup_{j=1}^{\infty} E_j \in \mathcal{A}$.

Given a σ-algebra, we define over it a real-valued set function satisfying certain properties.

Definition 33 *A probability measure, denoted by* $\Pr(\cdot)$*, is a real-valued set function that is defined over a* σ*-algebra* $\mathcal{A}$ *and satisfies the following properties:*

i) $E \in \mathcal{A} \Rightarrow \Pr(E) \geq 0$,

ii) $\Pr(\Omega) = 1 \Rightarrow \Pr(\emptyset) = 0$,

iii) If $\{E_j\}$ *is a countable collection of disjoint sets in* $\mathcal{A}$*, then*

$$\Pr(\cup_j E_j) = \sum_j \Pr(E_j).$$

A probability space and a random variable are defined as follows:

Definition 34 *Given a sample space* Ω*, a* σ*-algebra* $\mathcal{A}$ *associated with* Ω*, and a probability measure* $\Pr(\cdot)$ *defined over* $\mathcal{A}$*, we call the triplet* $(\Omega, \mathcal{A}, \Pr)$ *a probability space.*

Definition 35 *A random variable on* $(\Omega, \mathcal{A}, \Pr)$ *is a real valued function defined over the sample space* Ω*, denoted by* $X(\omega)$ *for* $\omega \in \Omega$*, such that for any real valued number* x*,* $\{\omega \,|X(\omega) < x\} \in \mathcal{A}$*.*

Example 36 *In the sample space consisting of the six faces of a die, all the possible subsets (including the whole space and the null set) constitute a* σ*-algebra. A probability measure can be defined, for example, by assigning 1/6 to each face and extending probabilities to the other subsets according to the rules given by Definition 33. An example of a random variable defined over the space is a mapping of the even-numbered faces to one and the odd-numbered faces to zero.*

Definition 37 *The distribution function* $\mathcal{F}(x)$ *of a random variable* $X(\omega)$ *is defined by* $\mathcal{F}(x) = \Pr\{\omega\,|X(\omega) < x\}$*. A distribution function has the properties:*
i) $\mathcal{F}(-\infty) = 0$,
ii) $\mathcal{F}(\infty) = 1$,
iii) It is nondecreasing and continuous from the left.

Note that the distribution function can be defined for any random variable because a probability is assigned to every element of $\mathcal{A}$ and hence to $\{\omega\,|X(\omega) < x\}$ for any x. We shall write $\Pr\{\omega\,|X(\omega) < x\}$ more compactly as $\Pr\{X < x\}$.

From this point on, we treat random variables as measurable functions that map each outcome of the experiment into a real number. This formalism, though abstract, is what allows us to give rigorous meaning to convergence, expectations, and probability limits—concepts that underlie all asymptotic arguments.

8.2.2 Modes of Convergence

Here, we shall define four modes of convergence for a sequence of random variables and shall state relationships among them in the form of several theorems.

Definition 38 (Convergence in probability) *A sequence of real or vector valued random variables $\{x_t\}$ is said to converge to x in probability if*

$$\lim_{T\to\infty} \Pr\left(\|x_T - x\| > \varepsilon\right) = 0 \textit{ for any } \varepsilon > 0.$$

We write $x_T \xrightarrow{p} x$ or $p\lim x_T = x$.

Definition 39 (Convergence in mean square) *A sequence of real or vector valued random variables $\{x_t\}$ is said to converge to x in mean square if*

$$\lim_{T\to\infty} \mathcal{E}\left(x_T - x\right)^2 = 0.$$

We write $x_T \xrightarrow{M} x$.

Definition 40 (Almost sure convergence) *A sequence of real or vector valued random variables $\{x_t\}$ is said to converge to x almost surely if*

$$\Pr\left[\lim_{T\to\infty} x_T = x\right] = 1.$$

We write $x_T \xrightarrow{a.s.} x$.

Definition 41 (Convergence in distribution) *A sequence of real or vector valued random variables $\{x_t\}$ is said to converge to x in distribution if the distribution function $\mathcal{F}_T$ of x_T converges to the distribution $\mathcal{F}$ of x at every continuity point of $\mathcal{F}$. We write $x_T \xrightarrow{D} x$ and we call $\mathcal{F}$ the limiting distribution of $\{x_t\}$. If $\{x_t\}$ and $\{y_t\}$ have the same limiting distribution, we write $x_T \overset{LD}{=} y_T$.*

It is important to mention that in all cases, the limit may well be a random variable itself.

The distinction among the various convergence concepts may seem technical at first, yet it becomes essential when proving the properties of estimators. In econometrics we typically establish convergence in probability (to show consistency) and then convergence in distribution (to derive limiting distributions used for inference).

Theorem 42 (Chebyshev's inequality) *Let $\mathcal{E}(y) = 0$; then $\Pr(|y| \geq \xi) \leq \frac{1}{\xi^2}\mathcal{E}(y^2) \quad \forall \xi > 0$.*

Proof. We know that

$$\begin{aligned}
\Pr\left(|y| \geq \xi\right) &= \int_{|y|\geq\xi} d\mathcal{F}\left(y\right) \\
&\leq \int_{|y|\geq\xi} \frac{y^2}{\xi^2} d\mathcal{F}\left(y\right) \qquad \left(\text{because } \frac{|y|}{\xi} \geq 1\right) \\
&\leq \frac{1}{\xi^2}\int_{-\infty}^{\infty} y^2 d\mathcal{F}\left(y\right).
\end{aligned}$$

But $\frac{1}{\xi^2}\int_{-\infty}^{\infty} y^2 d\mathcal{F}\left(y\right) = \frac{1}{\xi^2}\mathcal{E}\left(y^2\right)$. ■

Theorem 43 (Chebyshev) *If* $\mathcal{E}x_t^2 = 0$*, then* $x_T \xrightarrow{p} 0$*.*

Proof. From Theorem 42 we know that

$$\begin{aligned}
\Pr\left(|x_T| \geq \xi\right) &\leq \frac{1}{\xi^2}\mathcal{E}\left(x_T^2\right) \\
&\Rightarrow \lim_{T\to\infty}\frac{1}{\xi^2}\mathcal{E}\left(x_T^2\right) = 0 \;\; \text{(by assumption)} \\
&\geq \lim_{T\to\infty}\Pr\left(|x_T| \geq \xi\right) \\
&\Rightarrow \lim_{T\to\infty}\Pr\left(|x_T| \geq \xi\right) = 0 \;\; \text{(because Pr can not be negative).}
\end{aligned}$$

Thus $x_T \xrightarrow{p} 0$. ■

Corollary 44 $x_T \xrightarrow{M} x \Rightarrow x_T \xrightarrow{p} x$

Proof. Left as an exercise. ■

Theorem 45 $x_T \xrightarrow{p} x \Rightarrow x_T \xrightarrow{D} x$

Proof. Left as an exercise. ■

Theorem 46 $x_T \xrightarrow{a.s.} x \Rightarrow x_T \xrightarrow{p} x$

Proof. Left as an exercise. ■

Having introduced the main modes of convergence, it is important to understand how they relate to one another. Some are stronger than others,

$$\begin{tikzcd}
 & a.s. \arrow[d] & \\
M \arrow[r] & p \arrow[r] & D
\end{tikzcd}$$

Figure 8.1: Logical relationships among modes of convergence

and these logical implications form the backbone of asymptotic reasoning in econometrics.

Figure 8.1 depicts the logical relationships among the four modes of convergence discussed. The converse of Theorem 45 is not generally true, but it holds in the special case x is equal to a constant α. We shall state it as a theorem.

Theorem 47 $x_T \xrightarrow{D} \alpha \Rightarrow x_T \xrightarrow{p} \alpha$ *(when α is a constant).*

Proof. Left as an exercise. ■

The next three convergence theorems are extremely useful in obtaining the asymptotic properties of estimators.

Theorem 48 (Mann and Wald) *Let x_t and x be $k-$vectors of random variables and let $g(\cdot)$ be a function from $\mathbb{R}^k$ to $\mathbb{R}$ such that the set S of discontinuity points of $g(\cdot)$ is closed and $\Pr(x \in S) = 0$. If $x_T \xrightarrow{D} x$, then $g(x_T) \xrightarrow{D} g(x)$.*

Theorem 49 (Continuous mapping theorem) *Let x_t be a vector of random variables and let $g(\cdot)$ be a real-valued function continuous at a constant vector point α. Then $x_T \xrightarrow{p} \alpha \Rightarrow g(x_T) \xrightarrow{p} g(\alpha)$.*

Proof. Continuity at α means that for any $\varepsilon > 0$ we can find δ such that $\|x_T - \alpha\| < \delta$ implies $|g(x_T) - g(\alpha)| < \varepsilon$. Therefore

$$\Pr[\|x_T - \alpha\| < \delta] \leq \Pr[|g(x_T) - g(\alpha)| < \varepsilon]$$

The theorem follows because the left-hand side converges to 1 by the assumption of the theorem. ■

Theorem 50 (Slutsky) *If $x_t \xrightarrow{D} x$ and $y_t \xrightarrow{p} \alpha$, then*
i) $x_t + y_t \xrightarrow{D} x + \alpha$,
ii) $x_t y_t \xrightarrow{D} \alpha x$,
iii) $x_t / y_t \xrightarrow{D} x/\alpha$, provided $\alpha \neq 0$.

Remark 51 *(Deltha Method) Suppose $\sqrt{T}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{D} \mathcal{N}(0, V)$ and $g(.)$ is continuously differentiable at θ_0. Then*

$$\sqrt{T}\left(g\left(\widehat{\theta}\right) - g(\theta_0)\right) \xrightarrow{D} \mathcal{N}\left(0, g'(\theta_0) V g'(\theta_0)'\right).$$

The delta method formalizes the intuition that smooth transformations of asymptotically normal estimators remain asymptotically normal, with variance scaled by the gradient of the transformation. It will be repeatedly used when analyzing nonlinear and likelihood-based estimators.

Example 52 *Example. If $\widehat{\sigma}^2$ is a consistent estimator of σ_0^2 and $\sqrt{T}\left(\widehat{\sigma}^2 - \sigma_0^2\right) \xrightarrow{D} \mathcal{N}(0, w^2)$, then by the delta method*

$$\sqrt{T}\left(\widehat{\sigma} - \sigma_0\right) \xrightarrow{D} \mathcal{N}\left(0, \frac{w^2}{4\sigma_0^2}\right).$$

8.2.3 Rates of Convergence

Definition 53 *If $f(\cdot)$ and $g(\cdot)$ are two-real valued functions of the positive integer variable t, then the notation $f(t) = o(g(t))$ means that*

$$\lim_{t\to\infty}\left[\frac{f(t)}{g(t)}\right] = 0.$$

In this case, we say that $f(t)$ is of smaller order than $g(t)$.

It is not necessary for $g(t)$ to have a limit, what is important is the comparison performed by the ratio. For example, if $f(t) \to 0 \Rightarrow f(t) = o(1)$. If $f(t) = o(t^{-1})$ it means that $f(t)$ converges to zero faster than $\frac{1}{t}$. Finally, if $f(t) = o(t)$ we don't known if $f(\cdot)$ has a limit, but we know that if it tends to infinity it does less rapidly than t.

Definition 54 *If $f(\cdot)$ and $g(\cdot)$ are two-real valued functions of the positive integer variable t, then the notation $f(t) = \mathcal{O}(g(t))$ means that there exists a K>0 independent of t, and a positive integer T such that*

$$\left|\frac{f(t)}{g(t)}\right| < K \quad \forall t > T.$$

In this case, we say that $f(t)$ and $g(t)$ have the same order asymptotically.

Notice that this definition does not preclude $\lim_{t\to\infty} |f(t)/g(t)| = 0$.

The previous definitions are intended for nonstochastic sequences. The following definition generalizes this concept for the case of sequences of random variables.

Definition 55 *Let $\{x_t\}$ be a sequence of random variables and $g(t)$ a real valued function of the positive integer t. Then we can write $x_t = o_p(g(t))$ if $x_t/g(t) \overset{p}{\to} 0$ and $x_t = \mathcal{O}_p(g(t))$ if for any $\varepsilon > 0$ there exists an M such that*

$$\Pr\left[\left|\frac{x_t}{g(t)}\right| < M\right] \geq 1 - \varepsilon \quad \forall t.$$

Next, we present a few rules that can be applied:

$$\begin{aligned}
\mathcal{O}(t^q) \pm \mathcal{O}(t^r) &= \mathcal{O}\left(t^{\max(q,r)}\right), \\
o(t^q) \pm o(t^r) &= o\left(t^{\max(q,r)}\right), \\
\mathcal{O}(t^q) \pm o(t^r) &= \mathcal{O}(t^q) \qquad q \geq r, \\
\mathcal{O}(t^q) \pm o(t^r) &= o(t^r) \qquad r > q, \\
\mathcal{O}(t^q)\,\mathcal{O}(t^r) &= \mathcal{O}\left(t^{q+r}\right) \qquad q \geq r, \\
o(t^q)\,o(t^r) &= o\left(t^{q+r}\right), \\
\mathcal{O}(t^q)\,o(t^r) &= o\left(t^{q+r}\right).
\end{aligned}$$

These notations will reappear throughout the book. For instance, when we say that an estimator is root-T consistent, we mean that $\sqrt{T}\left(\widehat{\theta} - \theta_0\right) = \mathcal{O}_p(1)$. Rates of convergence allow us to compare estimators and to derive asymptotic expansions of their distributions and biases.

8.2.4 Relationships among $\lim \mathcal{E}$, $A\mathcal{E}$, and plim

Let $\mathcal{F}_T$ be the distribution function of x_T and $\mathcal{F}_T \to \mathcal{F}$ at continuity points of $\mathcal{F}$. We have defined plim in Definition 38. We define $\lim \mathcal{E}$ and $A\mathcal{E}$ as follows:

$$\lim_{T\to\infty} \mathcal{E} x_T = \lim_{T\to\infty} \int_{-\infty}^{\infty} x \, d\mathcal{F}_T(x)$$

and

$$A\mathcal{E} x_T = \int_{-\infty}^{\infty} x \, d\mathcal{F}(x).$$

In words, $A\mathcal{E}$, which reads asymptotic expectation or asymptotic mean, is the mean of the limit distribution.

These three limit operators are similar but different; we can construct examples of sequences of random variables such that any two of the three concepts either differ from each other or coincide with each other. We shall state relationships among the operators in the form of examples. But first, note the following obvious facts:

- Of the three concepts, only $\operatorname{plim} x_t$ can be a nondegenerate random variable; therefore, if it is, it must differ from $\lim \mathcal{E} x_T$ or $A\mathcal{E} x_T$.
- If $x_T \xrightarrow{p} \alpha$, a constant, then $A\mathcal{E} x_T = \alpha$. This follows immediately from Theorem 45.

Example 56 *Let x_T be defined by*

$$x_T = \begin{cases} Z \text{ with probability } (T-1)/T \\ T \text{ with probability } 1/T \end{cases},$$

where $Z \backsim \mathcal{N}(0,1)$. Then $x_T \xrightarrow{p} Z$, $\lim \mathcal{E} x_T = 1$, and $A\mathcal{E} x_T = \mathcal{E} Z = 0$.

Example 57 *Let x_T be defined by*

$$x_T = \begin{cases} 0 \text{ with probability } (T-1)/T \\ T^2 \text{ with probability } 1/T \end{cases}.$$

Then $x_T \xrightarrow{p} 0$, $\lim \mathcal{E} x_T = \infty$, and $A\mathcal{E} x_T = 0$.

Example 58 *Let $x \backsim \mathcal{N}(\alpha, 1)$ and $y_T \backsim \mathcal{N}(\beta, T^{-1})$, where $\beta \neq 0$. Then x/y_T is distributed Cauchy and does not have a mean. Therefore $\lim \mathcal{E}(x/y_T)$ cannot be defined either. But, because $y_T \xrightarrow{p} \beta$, $A\mathcal{E}(x/y_T) = \alpha/\beta$ by Theorem* **??** *iii).*

We are now in position to define three important concepts regarding the asymptotic properties of estimators, namely, asymptotic unbiasedness and consistency.

Definition 59 *The estimator $\widehat{\theta}_T$ of θ_0 is said to be asymptotically unbiased if $A\mathcal{E}\widehat{\theta}_T = \theta_0$. We call $A\mathcal{E}\widehat{\theta}_T - \theta_0$ the asymptotic bias.*

Some authors define asymptotic unbiasedness using $\lim \mathcal{E}\widehat{\theta}_T$ instead of $A\mathcal{E}\widehat{\theta}_T$. Then, it refers to a different concept.

Definition 60 *The estimator $\widehat{\theta}_T$ of θ_0 is said to be a weakly consistent estimator if $\widehat{\theta}_T \xrightarrow{p} \theta_0$.*

Definition 61 *The estimator $\widehat{\theta}_T$ of θ_0 is said to be a strongly consistent estimator if $\widehat{\theta}_T \overset{a.s.}{\rightarrow} \theta_0$.*

Some authors use the term consistent (without the weakly) to refer to an estimator that converges in probability.

In view of the preceding discussions, it is clear that a strongly consistent estimator is weakly consistent, but not necessarily vice versa. Likewise, a consistent estimator is asymptotically unbiased, but not necessarily vice versa.

The distinction among $\lim \mathcal{E}$, $A\mathcal{E}$, and plim is often subtle but conceptually important; plim concerns convergence of the random sequence itself; $\lim \mathcal{E}$ concerns the convergence of its expectation under the sequence's original law; and $A\mathcal{E}$ computes the mean under the limiting distribution. These need not coincide, which explains why asymptotic unbiasedness does not always imply finite-sample unbiasedness.

8.3 Laws of Large Numbers

In econometrics, we want to know the conditions under which our estimators converge to the true value of parameters. These conditions are evaluated using Laws of Large Numbers. Here, we present four of them that can be invoked depending on the structure of the model under consideration. To make a correspondence with definitions of consistent estimators, laws that imply converge in probability are usually referred to as Weak Laws of Large Numbers (WLLN). Laws that imply almost sure convergence are termed Strong Laws of Large Numbers (SLLN).

The Law of Large Numbers formalizes the idea that averages of random variables stabilize as the sample grows. It provides the asymptotic justification for consistency: if sample moments converge to their population counterparts, estimators built from them converge to the true parameter values.

Theorem 62 (WLLN1, Chebyshev) *Let* $\mathcal{E}(x_t) = \mu_t$, $\mathcal{V}(x_t) = \sigma_t^2$, $Cov(x_i, x_j) = 0\ \forall i \neq j$. *If* $\lim\limits_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \sigma_t^2 \leq M < \infty$, *then*

$$\overline{x}_T - \overline{\mu}_T \xrightarrow{p} 0.$$

Proof. By Chebyshev's inequality we known that

$$\begin{aligned}
\Pr\left(\left|\overline{x}_T - \overline{\mu}_T\right| > \xi\right) &\leq \frac{1}{\xi^2}\mathcal{V}\left(\overline{x}_T\right) \\
&= \frac{1}{\xi^2}\mathcal{V}\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) \\
&= \frac{1}{\xi^2 T^2}\mathcal{V}\left(\sum_{t=1}^{T} x_t\right) \\
&= \frac{1}{\xi^2 T^2}\sum_{t=1}^{T}\mathcal{V}\left(x_t\right) \\
&= \frac{1}{\xi^2 T^2}\sum_{t=1}^{T}\sigma_t^2.
\end{aligned}$$

Then

$$\lim_{T\to\infty}\frac{1}{\xi^2 T^2}\sum_{t=1}^{T}\sigma_t^2 \leq \lim_{T\to\infty}\frac{1}{\xi^2}\frac{M}{T} = 0.$$

Which implies that $\overline{x}_T - \overline{\mu}_T \xrightarrow{M} 0$, so $\overline{x}_T - \overline{\mu}_T \xrightarrow{p} 0$ (Corollary 44). ■

Theorem 63 (WLLN2, Kinchine) *Let $\{x_t\}$ be i.i.d. and $\mathcal{E}(x_t) = \mu < \infty$, then*

$$\overline{x}_T \xrightarrow{p} \mu.$$

Note that WLLN1 needs to restrict the second moments but does not require the i.i.d. assumption, while WLLN2 does not restrict second moments (explicitly) but requires i.i.d.

In econometric applications, independence is often too strong an assumption. Later chapters will discuss extensions of these results under dependence (mixing and martingale difference sequences) that are relevant for time-series models.

Theorem 64 (SLLN1, Kolmogorov) *Let $\{x_t\}$ be independent with finite variance $\mathcal{V}(x_t) = \sigma_t^2 < \infty$. If $\sum_{t=1}^{\infty} \frac{\sigma_t^2}{t^2} < \infty$, then*

$$\overline{x}_T - \overline{\mu}_T \xrightarrow{a.s.} 0.$$

Theorem 65 (SLLN2, Kolmogorov) *Let $\{x_t\}$ be i.i.d. Then, a necessary and sufficient condition for $\overline{x}_T \xrightarrow{a.s.} \mu$ is that $\mathcal{E}x_t$ exists and $\mathcal{E}x_t = \mu$.*

8.4 Central Limit Theorems

The previous section presented the conditions under which the sample average converges (either in probability or almost surely) to the population mean, thus also converging in distribution (as Figure 8.1 depicted). This result is not very useful if we want to conduct inference about the population moment. Here, we will present the conditions under which random variables converge to more interesting distributions.

While LLNs ensure convergence of sample moments to their expectations, they are silent about the shape of the sampling distribution around the limit. Central Limit Theorems fill this gap by showing that properly normalized deviations tend to a normal distribution, allowing approximate inference.

Theorem 66 (CLT1, Lindeberg-Lévy) *Let $\{x_t\}$ be i.i.d. with $\mathcal{E}x_t = \mu$ and $\mathcal{V}x_t = \sigma^2$. Then*

$$Z_T = \frac{\overline{x}_T - \mu}{\left[\mathcal{V}\overline{x}_T\right]^{1/2}} = \sqrt{T}\frac{\overline{x}_T - \mu}{\sigma} \xrightarrow{D} \mathcal{N}(0,1).$$

Theorem 67 (CLT2, Liapunov) *Let $\{x_t\}$ be independent with $\mathcal{E}x_t = \mu_t$, $\mathcal{V}x_t = \sigma_t^2$, and $\mathcal{E}\left[|x_t - \mu_t|^3\right] = m_{3t}$. If*

$$\lim_{T\to\infty} \frac{\left[\sum_{t=1}^{T} m_{3t}\right]^{1/3}}{\left[\sum_{t=1}^{T} \sigma_t^2\right]^{1/2}} = 0,$$

then $Z_T = \frac{\overline{x}_T - \overline{\mu}_T}{[\mathcal{V}\overline{x}_T]^{1/2}} \xrightarrow{D} \mathcal{N}(0,1)$.

Theorem 68 (CLT3, Lindeberg-Feller) *Let $\{x_t\}$ be independent with distribution functions $\{\mathcal{F}_t\}$ and $\mathcal{E}x_t = \mu_t$ and $\mathcal{V}x_t = \sigma_t^2$. Define $C_T = \left(\sum_{t=1}^{T} \sigma_t^2\right)^{1/2}$. If*

$$\lim_{T\to\infty} \frac{1}{C_T^2} \sum_{t=1}^{T} \int_{|x-\mu_t|>\varepsilon C_T} (x - \mu_t)^2 \, d\mathcal{F}_t(x) = 0,$$

for every $\varepsilon > 0$, then $Z_T = \frac{\overline{x}_T - \overline{\mu}_T}{[\mathcal{V}\overline{x}_T]^{1/2}} \xrightarrow{D} \mathcal{N}(0,1)$.

We shall conclude this section by stating a multivariate CLT.

Theorem 69 *Let $\{x_t\}$ be a sequence of k-dimensional vectors of random variables. If $c'x_T$ converges to a normal random variable for every k-dimensional vector $c \neq 0$, then x_T converges to a multivariate normal random variable.*

Note that showing convergence of each element of x_T separately is not sufficient.

The multivariate extension is indispensable in econometrics, where estimators are vector-valued. It guarantees that joint asymptotic normality of estimators can be deduced by verifying normality of all linear combinations.

8.5 Asymptotic Distribution of LS Estimators

The main purpose of this section is to prove the consistency and asymptotic normality of the OLS, GLS, and FGLS estimators.

8.5.1 Asymptotic Distribution of OLS

Consistency of OLS

Next we present several theorems that can be applied in order to demonstrate the conditions under which the OLS estimator is consistent. We begin by demonstrating the consistency of the OLS estimator for fixed regressors, and later we extend the scope of the consistency theorems for cases in which the regressors are stochastic.

Equipped with the LLN and CLT, we now revisit the linear regression model to study how OLS behaves as the sample size grows. Our objective is to establish two fundamental properties—consistency and asymptotic normality—and to interpret them in light of the large-sample results previously developed.

Theorem 70 (Consistency of $\widehat{\beta}$, $k = 1$, X nonstochastic) *Consider the HLRM with $k = 1$. If $\lim_{T\to\infty} X'X = \infty$, then $\widehat{\beta} \xrightarrow{p} \beta_0$.*

Proof. In order to prove that $\widehat{\beta} \xrightarrow{p} \beta_0$ we will demonstrate that $\widehat{\beta} \xrightarrow{M} \beta_0$ and then invoke Corollary 44. Recall that $\widehat{\beta} - \beta_0 = X'u/X'X$. Thus

$$\begin{aligned}
\mathcal{E}\left(\widehat{\beta} - \beta_0\right)^2 &= \mathcal{E}\left(\frac{X'u}{X'X}\right)^2 \\
&= \mathcal{E}\left(\frac{\sum x_t u_t}{\sum x_t^2}\right)^2 \\
&= \frac{1}{\left(\sum x_t^2\right)^2}\mathcal{E}\left(\sum x_t u_t\right)^2 \quad \text{(because } x \text{ is nonstochastic)} \\
&= \frac{1}{\left(\sum x_t^2\right)^2}\mathcal{V}\left(\sum x_t u_t\right) \quad \text{(because } \mathcal{E}x_t u_t = 0\text{)} \\
&= \frac{1}{\left(\sum x_t^2\right)^2}\sum \mathcal{V}\left(x_t u_t\right) \\
&= \frac{\sigma_0^2}{\sum x_t^2} \\
&\Rightarrow \lim_{T\to\infty} \mathcal{E}\left(\widehat{\beta} - \beta\right)^2 \\
&= \lim_{T\to\infty} \frac{\sigma_0^2}{\sum x_t^2} = 0.
\end{aligned}$$

Thus, $\widehat{\beta} \xrightarrow{M} \beta_0$ and by Corollary 44, $\widehat{\beta} \xrightarrow{p} \beta_0$. ■

Theorem 71 (Consistency of $\widehat{\beta}$, $k \geq 1$, X nonstochastic) *Consider the HLRM with $k \geq 1$. If $\lambda_s(X'X) \rightarrow \infty$ (where $\lambda_s(X'X)$ is the smallest eigenvalue of $(X'X)$), then $\widehat{\beta} \xrightarrow{p} \beta_0$.*

Proof. We will present the proof following the same strategy as before; that is, to prove that $\widehat{\beta} \xrightarrow{p} \beta$ we will demonstrate that $\widehat{\beta} \xrightarrow{M} \beta_0$ and then invoke Corollary 44. Recall that $\mathcal{V}\left(\widehat{\beta}\,|X\right) = \sigma_0^2 (X'X)^{-1}$. Thus for any parameter j,

$$\begin{aligned}
\mathcal{E}\left(\widehat{\beta}_j - \beta_{0,j}\,|X\right)^2 &= \mathcal{V}\left(\widehat{\beta}_j\,|X\right) \\
&= \operatorname{diag}_j\left[\sigma_0^2 (X'X)^{-1}\right] \quad (\,j\text{th element of the diagonal}) \\
&\leq \operatorname{tr}\left[\sigma_0^2 (X'X)^{-1}\right] \\
&= \sum_{i=1}^{k} \frac{\sigma_0^2}{\lambda_i (X'X)} \quad (\lambda_i A \text{ is the } i\text{th eigenvalue of matrix } A) \\
&\leq \frac{k\sigma_0^2}{\lambda_s (X'X)}
\end{aligned}$$

But $\lambda_s(X'X) \rightarrow \infty$, then $\lim_{T\rightarrow\infty} \mathcal{E}\left(\widehat{\beta}_j - \beta_{0,j}\right)^2 \leq 0$. Thus, $\widehat{\beta} \xrightarrow{M} \beta$ and by Corollary 44, $\widehat{\beta} \xrightarrow{p} \beta$. ■

Lemma 72 *The following four statements are equivalent:*
i) $\lambda_s(X'X) \rightarrow \infty$ $\quad(\lambda_s A =$ smallest eigenvalue of $A)$,
ii) $\lambda_l (X'X)^{-1} \rightarrow 0$ $\quad(\lambda_l A =$ largest eigenvalue of $A)$,
iii) $tr(X'X)^{-1} \rightarrow 0$,
iv) $diag_j (X'X)^{-1} \rightarrow 0$ for $j = 1, 2, .., k$.

Proof. Left as an exercise. ■

The results of this Lemma are important because we can check for any of them to be satisfied in order to prove consistency of the OLS estimator.

Several important features of Theorems 70 and 71 are worth mentioning: First, consistency of the OLS estimator does not require to impose the assumption of normality on u. Second, as the sample size increases, the variance of the estimator converges to zero. Finally, as Figure 8.2 shows, given that the estimator converges in mean square and probability to the true value β_0, it also converges in distribution to a degenerate distribution with mass point at β_0.

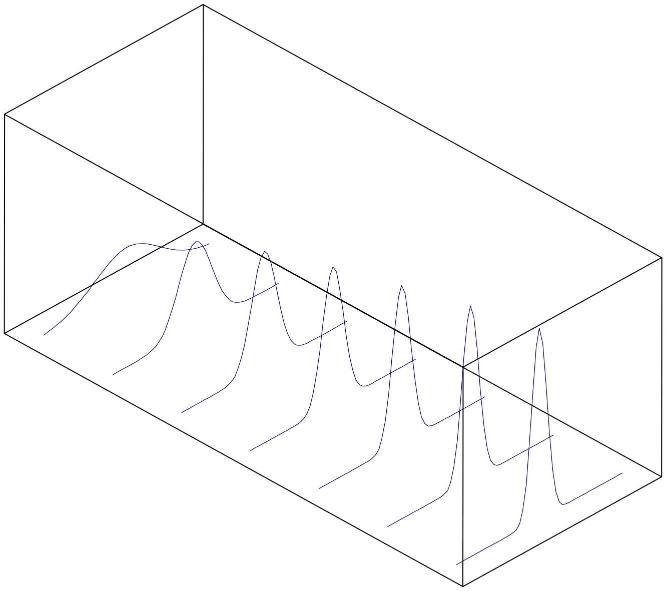

Figure 8.2: Convergence in distribution

Theorems 70 and 71 assume that $\mathcal{E}uu' = \sigma_0^2 I_T$, thus precluding the presence of heteroskedasticity and/or serial correlation on u. If $\mathcal{E}uu' = \sigma_0^2 \Omega_0$, with $\Omega_0 \neq I_T$, we know that the OLS estimator is unbiased. Next, we present the conditions under which the OLS estimator is also consistent.

The consistency of the OLS estimator is important for two reasons: First, if Ω_0 is unknown, we may prefer to obtain a consistent estimator of β_0 even though it is not as efficient as the GLS estimator (which is not obtainable anyway). Second, even if we knew the form of Ω_0 and decide to obtain the FGLS estimator, this estimator is obtained in a two-step procedure, where the first step uses the OLS estimator; it is clear that if the OLS estimator is not consistent, neither will be the FGLS estimator.

Before we present the conditions under which the OLS estimator is consistent when $\Omega_0 \neq I_T$, we present a Lemma that will be used in its demonstration.

Lemma 73 *Let A and B be nonnegative definite matrices of size T. Then $tr\,(AB) \leq \lambda_l\,(A)\,tr\,(B)$, where λ_l denotes the largest eigenvalue of A.*

Proof. Let H be a matrix such that $H'AH = D$, diagonal, and $H'H = I$. Then, tr(AB) =tr$(H'AHH'BH)$ =trDS, where $S = H'BH$. Let d_i be the ith diagonal element of D and s_{ii} be of S. Then

$$\text{tr}\,(DS) = \sum_{i=1}^{T} d_i s_{ii} \leq \max_i d_i \cdot \sum_{i=1}^{T} s_{ii}.$$

But $\max_i d_i \cdot \sum_{i=1}^{T} s_{ii} = \lambda_l\,(A)\,\text{tr}B$. ■

Theorem 74 (Consistency of $\widehat{\beta}$, $k \geq 1$, X nonstochastic, $\Omega_0 \neq I_T$) *Consider the LRM with $\mathcal{E}uu' = \sigma_0^2\Omega_0$. If :*
i) $\lambda_s\,(X'X) \to \infty$,
ii) $\lambda_l\Omega_0$ is bounded for all T,
Then $\widehat{\beta} \xrightarrow{p} \beta_0$.

Proof. We have

$$\begin{aligned}
\text{tr}\mathcal{V}\left(\widehat{\beta}\,|X\right) &= \text{tr}\left[\sigma_0^2\,(X'X)^{-1}\,X'\Omega_0 X\,(X'X)^{-1}\right] \\
&= \text{tr}\left[\sigma_0^2\Omega_0 X\,(X'X)^{-1}\,(X'X)^{-1}\,X'\right] \\
&\leq \sigma_0^2\lambda_l\,(\Omega_0)\;\text{tr}\left[X\,(X'X)^{-1}\,(X'X)^{-1}\,X'\right] \quad \text{(Lemma 73)} \\
&= \sigma_0^2\lambda_l\,(\Omega_0)\;\text{tr}\left[(X'X)^{-1}\right] \\
&= \sum_{i=1}^{k} \frac{\sigma_0^2\lambda_l\,(\Omega_0)}{\lambda_i\,(X'X)} \\
&\leq k\sigma_0^2\frac{\lambda_l\,(\Omega_0)}{\lambda_s\,(X'X)}.
\end{aligned}$$

But the last term converges to 0 because of assumption i) and ii), then $\widehat{\beta} \xrightarrow{M} \beta$ and by Corollary 44, $\widehat{\beta} \xrightarrow{p} \beta$. ■

It is trivial to verify that Theorem 74 generalizes the results of Theorem 71, given that if $\Omega_0 = I_T$, $\lambda_l\,(\Omega_0) = 1$.

Theorem 75 (Consistency of $\widehat{\beta}$, X stochastic) *In the LRM, assume that $\mathcal{E}u_t^2 < \infty$, $\mathcal{E}\left(x_t u_t\right) = 0$, $\mathcal{E}x_t'x_t < \infty$, and $S = \mathcal{E}x_t x_t'$. Then:*

i) $\frac{1}{T}X'X \xrightarrow{p} S$,

ii) $\frac{1}{T}X'u \xrightarrow{p} 0$,

iii) $\widehat{\beta} \xrightarrow{p} \beta_0$.

Proof. The assumptions that $\mathcal{E}x_t'x_t < \infty$ and $\mathcal{E}u_t^2 < \infty$ mean that all elements of x_t and u_t have finite second moments, and all cross-products are finite. To see this, first observe that since

$$\mathcal{E}x_t'x_t = \mathcal{E}x_{1,t}^2 + \cdots + \mathcal{E}x_{k,t}^2 < \infty,$$

then, it is the case that for all $j = 1, \cdots, k$, $\mathcal{E}x_{j,t}^2 < \infty$. By the Cauchy-Schwarz inequality, for each j and i,

$$\mathcal{E}\left|x_{j,t}x_{i,t}\right| \leq \mathcal{E}\left|x_{j,t}^2\right|^{1/2}\mathcal{E}\left|x_{i,t}^2\right|^{1/2} < \infty.$$

Using this result and the WLLN2 directly imply that for all j and i

$$\frac{1}{T}\sum_{t=1}^{T} x_{j,t}^2 \xrightarrow{p} \mathcal{E}x_{j,t}^2$$

and

$$\frac{1}{T}\sum_{t=1}^{T} x_{j,t}x_{i,t} \xrightarrow{p} \mathcal{E}x_{j,t}x_{i,t}.$$

Hence, $\frac{1}{T}X'X \xrightarrow{p} S$ which is the first result.

Furthermore, for each j

$$\mathcal{E}\left|x_{j,t}u_t\right| \leq \mathcal{E}\left|x_{j,t}^2\right|^{1/2}\mathcal{E}\left|u_t^2\right|^{1/2} < \infty.$$

Using this result and the WLLN2 imply that for all j

$$\frac{1}{T}\sum_{t=1}^{T} x_{j,t}u_t \xrightarrow{p} \mathcal{E}x_{j,t}u_t = 0.$$

Hence $\frac{1}{T}X'u \xrightarrow{p} 0$ which is the second result.

Finally,

$$\begin{aligned}\widehat{\beta} &= \beta_0 + (X'X)^{-1} X'u \\ &= \beta_0 + \left(\frac{1}{T}X'X\right)^{-1} \frac{1}{T}X'u \\ &\Rightarrow \widehat{\beta} \xrightarrow{p} \beta_0.\end{aligned}$$

as $\left(\frac{1}{T}X'X\right)^{-1} \xrightarrow{p} S^{-1}$ (Theorem 49) and $\frac{1}{T}X'u \xrightarrow{p} 0$. The result follows by applying Slutsky. ■

Theorem 76 (Consistency of $\widehat{\sigma}^2$) *Given the same assumptions of Theorem 71 plus $\mathcal{E}\left(u_t^4\right) = m_4$, then $\widehat{\sigma}^2 \xrightarrow{p} \sigma_0^2$.*

Proof. As $\widehat{\sigma}^2 = T^{-1}\widehat{u}'\widehat{u} = T^{-1}u'Mu = T^{-1}u'u - T^{-1}u'X\left(X'X\right)^{-1}X'u$. Let $H_1 = T^{-1}u'u$ and $H_2 = T^{-1}u'X\left(X'X\right)^{-1}X'u$. Then $\widehat{\sigma}^2 = H_1 - H_2$. It is trivial to verify that $H_1 \xrightarrow{p} \sigma_0^2$ (by WLLN2). Thus, we only have to prove that $H_2 \xrightarrow{p} 0$. By the (generalized) Chebyshev's inequality we know that

$$\Pr\left[H_2 \geq \varepsilon\right] \leq \frac{1}{\varepsilon}\mathcal{E}\left(H_2\right).$$

Then

$$\begin{aligned}\frac{1}{\varepsilon}\mathcal{E}\left(H_2\right) &= \frac{1}{\varepsilon T}\mathcal{E}\left(u'X\left(X'X\right)^{-1}X'u\right) \\ &= \frac{1}{\varepsilon T}\mathcal{E}\left[\mathrm{tr}\left(u'X\left(X'X\right)^{-1}X'u\right)\right] \quad \text{(given that } H_2 \text{ is a scalar)} \\ &= \frac{1}{\varepsilon T}\mathcal{E}\left[\mathrm{tr}\left(\left(X'X\right)^{-1}X'uu'X\right)\right] \quad \text{(given that tr}AB = \text{tr}BA\text{)} \\ &= \frac{1}{\varepsilon T}\mathrm{tr}\left[\mathcal{E}\left(\left(X'X\right)^{-1}X'uu'X\right)\right] \quad (\mathcal{E} \text{ and tr are linear operators}) \\ &= \frac{\sigma_0^2}{\varepsilon T}\mathrm{tr}\left[\mathcal{E}\left(\left(X'X\right)^{-1}X'X\right)\right] \\ &= \frac{k\sigma_0^2}{\varepsilon T}.\end{aligned}$$

Thus $\lim_{T\to\infty}\frac{k\sigma_0^2}{\varepsilon T} = 0$ which implies that $H_2 \xrightarrow{p} 0$. ■

A very important point that has to be considered is that consistency of the estimator requires that $k = o\left(T\right)$; that is, if the number of regressors increases, it must do so at a slower rate than the increase in the sample size.[2]

[2] This is very important, given that many applied researcher tend to include more explanatory variables every time their sample increases.

Corollary 77 $\widetilde{\sigma}^2 \xrightarrow{p} \sigma_0^2$.

Proof. Given that

$$\widetilde{\sigma}^2 = \frac{T}{T-k}\widehat{\sigma}^2$$

and $\frac{T}{T-k} \to 1$, $\widetilde{\sigma}^2 \xrightarrow{p} \sigma_0^2$ by Slutsky. ■

Asymptotic Normality of OLS

As was evident from Figure 8.2, $\widehat{\beta}$ does not have an interesting distribution if we want to test an hypothesis about β_0. This is so because the conditional variance of $\widehat{\beta}$ converges to zero as the sample increases. Except in special cases, the exact distribution of $\widehat{\beta}$ is unknown. Therefore we rely on approximations, and use a variety of techniques to assess the accuracy of these approximations. The dominant approximation technique relies on asymptotic theory, and is based on calculating the limiting distribution of a normalized version of $\widehat{\beta}$ using a CLT.

Theorem 78 (Normality of $\widehat{\beta}$; $k = 1$, X nonstochastic) *Consider the HLRM. Let*

$$r_T^2 = \frac{\max_{1\leq t\leq T} x_t^2}{X'X}.$$

If $\lim_{T\to\infty} r_T^2 = 0$, *then* $\sigma_0^{-1}\left(X'X\right)^{1/2}\left(\widehat{\beta} - \beta_0\right) \xrightarrow{D} \mathcal{N}(0,1)$.

Proof. Recall that in this case $\widehat{\beta} - \beta_0 = X'u/X'X$, $\mathcal{V}\left(\widehat{\beta}\,|X\right) = \sigma_0^2/X'X$, and $Z_T = X'u/\left(\sigma_0\left(X'X\right)^{1/2}\right)$; where Z_T corresponds to the standardized ratio of $\widehat{\beta} - \beta_0$ with respect to its standard error. We want to prove that $Z_T \xrightarrow{D} \mathcal{N}(0,1)$. Notice that the numerator $X'u = \sum x_t u_t$ is composed of T independent observations $w_t = x_t u_t$ with mean 0 and variance $\sigma_0^2 x_t^2$. As X is fixed, w is not i.i.d. given that the variance depends on x. For this reason, we cannot apply CLT1 directly. CLT2 is not an option, given that the assumptions of the theorem are silent with respect to the third moment of w. Thus, we will use CLT3. Notice that $C_T^2 = \sigma_0^2 \sum x_t^2$. We need to prove that

$$\lim_{T\to\infty} \frac{1}{C_T^2} \sum_{t=1}^{T} \int_{|w_t|>\varepsilon C_T} w^2 d\mathcal{F}_t(w) = 0.$$

Define H_T as

$$\begin{aligned}
H_T &= \frac{1}{C_T^2}\sum_{t=1}^{T}\int_{|x_t u_t|>\varepsilon C_T} w^2 d\mathcal{F}_t\left(w\right) \\
&= \frac{1}{C_T^2}\sum_{t=1}^{T}\int_{|x_t||u_t|>\varepsilon C_T} w^2 d\mathcal{F}_t\left(w\right) \quad \text{(given that } |x_t u_t| = |x_t|\,|u_t|\text{)} \\
&= \frac{1}{C_T^2}\sum_{t=1}^{T}\int_{|u_t|>\frac{\varepsilon C_T}{|x_t|}} w^2 d\mathcal{F}_t\left(w\right) \\
&= \frac{1}{C_T^2}\sum_{t=1}^{T}\int_{u_t^2>\frac{\varepsilon^2 C_T^2}{x_t^2}} w^2 d\mathcal{F}_t\left(w\right) \\
&\le \frac{1}{C_T^2}\sum_{t=1}^{T}\int_{u_t^2>\frac{\varepsilon^2 \sigma_0^2}{r_T^2}} w^2 d\mathcal{F}_t\left(w\right) \quad \text{(using the definitions of } r_T^2 \text{ and } C_T^2\text{)} \\
&= \frac{1}{C_T^2}\sum_{t=1}^{T}\int_{u_t^2>\frac{\varepsilon^2 \sigma_0^2}{r_T^2}} x^2u^2 d\mathcal{G}_t\left(u\right) \quad \begin{aligned}&\text{(as } \mathcal{F}_t\left(\alpha\right) = \Pr\left[x^2u^2<\alpha\right] \\ &= \Pr\left[u^2<\alpha/x^2\right] = \mathcal{G}_t\left(u\right)\text{)}\end{aligned} \\
&= \frac{1}{C_T^2}\sum_{t=1}^{T}x_t^2\int_{u_t^2>\frac{\varepsilon^2 \sigma_0^2}{r_T^2}} u^2 d\mathcal{G}_t\left(u\right) \quad \text{(as } x \text{ and } u \text{ are independent)} \\
&\le \frac{1}{C_T^2}X'X\int_{u_t^2>\frac{\varepsilon^2 \sigma_0^2}{r_T^2}} u^2 d\mathcal{G}_t\left(u\right) \\
&= \frac{1}{\sigma_0^2}\int_{u_t^2>\frac{\varepsilon^2 \sigma_0^2}{r_T^2}} u^2 d\mathcal{G}_t\left(u\right) \quad \text{(using the definition of } C_T^2\text{).}
\end{aligned}$$

As $\lim_{T\to\infty} r_T^2 = 0$, the area of integration shrinks as $T \to \infty$, in which case $\int(\cdot)$ converges to zero. Thus, $\lim_{T\to\infty} H_T = 0$ and $Z_T \overset{D}{\to} \mathcal{N}\left(0,1\right)$. The conclusion of the Theorem follows using Theorem **??**. ■

This theorem and the next ones, show that the asymptotic distribution that is relevant for testing hypothesis is the normal distribution. It is trivial to verify that given that σ_0^2 is not available, we can replace it with a consistent estimator (either $\widehat{\sigma}^2$ or $\widetilde{\sigma}^2$) given that

$$\frac{\widehat{\beta}-\beta_0}{\sqrt{\sigma_0^2\left(X'X\right)^{-1}}} = \frac{\widehat{\beta}-\beta_0}{\sqrt{\widehat{\sigma}^2\left(X'X\right)^{-1}}}\frac{\widehat{\sigma}}{\sigma_0}.$$

But $\frac{\widehat{\sigma}}{\sigma_0} \xrightarrow{p} 1$, then applying Slutsky, we have:

$$\frac{\widehat{\beta} - \beta_0}{\sqrt{\widehat{\sigma}^2 \left(X'X\right)^{-1}}} \xrightarrow{D} \mathcal{N}\left(0, 1\right).$$

Theorem 79 (Normality of $\widehat{\beta}$; $k \geq 1$, X nonstochastic) *Consider the HLRM. Let*

$$r_{i,T}^2 = \frac{\max_{1 \leq t \leq T} x_{i,t}^2}{x_i' x_i} \quad \textit{for } i = 1, \cdots, k.$$

Assume $\lim_{T\to\infty} r_{i,T}^2 = 0$ *for* $i = 1, \cdots, k$. *Define* $J = XS^{-1}$ *where* S *is the* $k \times k$ *diagonal matrix with elements* $\left(x_i' x_i\right)^{1/2}$ *and assume that* $\lim_{T\to\infty} J'J = R$ *exists and is nonsingular, then* $S\left(\widehat{\beta} - \beta_0\right) \xrightarrow{D} \mathcal{N}\left(0, \sigma_0^2 R^{-1}\right)$.

Proof. Notice that $S\left(\widehat{\beta} - \beta_0\right) = S\left(X'X\right)^{-1} X'u$ in which case we have that $\mathcal{E}\left[S\left(\widehat{\beta} - \beta_0\right)\right] = 0$ and $\mathcal{V}\left[S\left(\widehat{\beta} - \beta_0\right)|X\right] = \sigma_0^2 \left(J'J\right)^{-1}$. Thus, if $\lim_{T\to\infty} J'J = R$, we have that the two moments of the theorem coincide. We will now prove that $S\left(\widehat{\beta} - \beta_0\right) \xrightarrow{D} \mathcal{N}\left(0, \sigma_0^2 R^{-1}\right)$. In order to do so, we will first use the asymptotic distribution of a univariate transformation of $S\left(\widehat{\beta} - \beta_0\right)$ and then invoke Theorem 69. Let α be a k-vector of constants and define the scalar

$$H_T = \frac{\alpha' X' u}{\sqrt{\alpha' X' X \alpha \sigma_0^2}}.$$

Notice that $\mathcal{E}\left(H_T\right) = 0$ and $\mathcal{V}\left(H_T\right) = 1$. We will prove that $H_T \xrightarrow{D} \mathcal{N}\left(0, 1\right)$. Define $\widetilde{x} = X\alpha$, which is a T-component vector. Then

$$H_T = \frac{\widetilde{x}' u}{\sqrt{\widetilde{x}' \widetilde{x} \sigma_0^2}}.$$

We need to prove that if we define

$$\widetilde{r}_T^2 = \frac{\max_{1 \leq t \leq T} \widetilde{x}_t^2}{\widetilde{x}' \widetilde{x}}$$

and $\lim \widetilde{r}_T^2 = 0$, then invoking Theorem 78, $H_T \xrightarrow{D} \mathcal{N}(0,1)$. Let us analyze the numerator and denominator of $\widetilde{r}_T^2$ separately. The numerator is:

$$\begin{aligned}
\max_{1\leq t\leq T} \widetilde{x}_t^2 &= \max_{1\leq t\leq T} \left[\sum_{i=1}^{k} \alpha_i x_{i,t}\right]^2 \\
&\leq \max_{1\leq t\leq T} \left[\sum_{i=1}^{k} |\alpha_i| \, |x_{i,t}|\right]^2 \\
&\leq \left[\sum_{i=1}^{k} |\alpha_i| \, r_i s_i\right]^2 \quad \left(\text{where } s_i = \left(x_i' x_i\right)^{1/2}\right) \\
&\leq \left(\sum_{i=1}^{k} \alpha_i^2 s_i^2\right)\left(\sum_{i=1}^{k} r_i^2\right) \quad \text{(Cauchy-Schwarz inequality).}
\end{aligned}$$

On the other hand, the denominator can be expressed as:

$$\begin{aligned}
\widetilde{x}'\widetilde{x} &= \alpha' X' X \alpha = (\alpha' S)\, J' J\, (S\alpha) \\
&\geq \lambda_s(J'J)\, \alpha' S S \alpha \quad (\text{because } J'J \text{ is p.d.}) \\
&= \lambda_s(J'J)\left(\sum_{i=1}^{k} \alpha_i^2 s_i^2\right).
\end{aligned}$$

Combining both results, we have:

$$\begin{aligned}
\widetilde{r}_T^2 &= \frac{\max_{1\leq t\leq T} \widetilde{x}_t^2}{\widetilde{x}'\widetilde{x}} \\
&\leq \frac{\sum_{i=1}^{k} r_i^2}{\lambda_s(J'J)}.
\end{aligned}$$

Then

$$\begin{aligned}
\lim_{T\to\infty} \widetilde{r}_T^2 &\leq \lim_{T\to\infty} \frac{\sum_{i=1}^{k} r_i^2}{\lambda_s(J'J)} \\
&= \frac{0}{\lambda_s(R)} = 0 \quad \text{(given that } R \text{ is nonsingular).}
\end{aligned}$$

Thus, $H_T \xrightarrow{D} \mathcal{N}(0,1)$. The final result follows from applying Theorem 69. ■

Given that $S\left(\widehat{\beta}-\beta_0\right) \xrightarrow{D} \mathcal{N}\left(0, \sigma_0^2 R^{-1}\right)$ we conclude that

$$\left(\widehat{\beta}-\beta_0\right) \overset{a}{\backsim} \mathcal{N}\left(0, \sigma_0^2 S^{-1} R^{-1} S^{-1}\right).$$

But $S^{-1}\left(S^{-1}X'XS^{-1}\right)^{-1}S^{-1} = \left(X'X\right)^{-1}$. Then, $\left(\widehat{\beta}-\beta_0\right) \overset{a}{\backsim} \mathcal{N}\left(0, \sigma_0^2\left(X'X\right)^{-1}\right)$.

We used the assumption that $\lim_{T\to\infty} r_T^2 = 0$ in order to derive the asymptotic distribution of the OLS estimator. Intuitively, this assumption implies that in order to obtain an asymptotically normal distribution, we need to have a data set that does not have a very influential observation. If huge outliers are present, our results may be compromised. The following theorem shows, among other things, that $\lim_{T\to\infty} r_T^2 = 0$ is a less restrictive assumption than the commonly used assumption that $\lim T^{-1}X'X$ exists and is a nonzero constant. It follows that if $\lim T^{-1}X'X$ exists and is nonsingular, the condition of Theorem 79 is satisfied.

Theorem 80 *Given a sequence of constants* $\{x_t\}$*, consider the statements (for* $c_T = \sum_{t=1}^T x_t^2$ *and* $a < \infty, a \neq 0$*):*

i) $\lim_{T\to\infty} T^{-1}c_T = a$,

ii) $\lim_{T\to\infty} c_T = \infty$,

iii) $\lim_{T\to\infty} c_T^{-1}x_T^2 = 0$,

iv) $\lim_{T\to\infty} c_T^{-1}\max_{1\le t\le T} x_t^2 = 0$.

Then i)⇒ [ii) and iii)] ⇒iv).

Proof. Left as an exercise. ■

Next, we derive the asymptotic distribution of $\widehat{\beta}$ in the case where X is stochastic.

Theorem 81 (Normality of $\widehat{\beta}$; $k \geq 1$, X stochastic) *Assume that* $\mathcal{E}\left(x_t u_t\right) = 0$, $\mathcal{E}u_t^4 < \infty$, $\mathcal{E}\left|x_t^4\right| < \infty$, $S = \mathcal{E}x_t x_t' > 0$, *and* $\Sigma = \mathcal{E}\left(x_t x_t' u_t^2\right) > 0$. *Then:*

i) $v_t = x_t u_t$ *is i.i.d.,* $\mathcal{E}v_t = 0$, $\mathcal{E}v_t'v_t < \infty$, *and* $\mathcal{E}v_t v_t' = \Sigma$;

ii) $\frac{1}{\sqrt{T}}\sum_{t=1}^T v_t \xrightarrow{D} \mathcal{N}\left(0, \Sigma\right)$;

iii) $\sqrt{T}\left(\widehat{\beta}-\beta_0\right) \xrightarrow{D} \mathcal{N}\left(0, S^{-1}\Sigma S^{-1}\right)$.

Proof. Since (y_t, x_t) is i.i.d., so is $v_t = x_t (y_t - \beta_0' x_t)$. Furthermore $\mathcal{E} v_t = \mathcal{E} x_t u_t = 0$; by the Cauchy-Schwarz inequality we have

$$\begin{aligned} \mathcal{E} \left| v_t' v_t \right| &= \mathcal{E} \left| x_t x_t' u_t^2 \right| \\ &\leq \left(\mathcal{E} \left| x_t x_t' \right|^2 \right)^{1/2} \left(\mathcal{E} \left| u_t^4 \right| \right)^{1/2} \\ &= \left(\mathcal{E} \left| x_t \right|^4 \right)^{1/2} \left(\mathcal{E} \left| u_t^4 \right| \right)^{1/2} < \infty. \end{aligned}$$

Finally, $\mathcal{E} v_t v_t' = \mathcal{E} \left(x_t x_t' u_t^2 \right) = \Sigma$, which establishes the first result.
The second statement follows directly from CLT1.
Finally, as *ii*) shows,

$$\frac{1}{\sqrt{T}} X' u \xrightarrow{D} \mathcal{N} (0, \Sigma) .$$

But

$$\begin{aligned} \sqrt{T} \left(\widehat{\beta} - \beta_0 \right) &= \left(\frac{1}{T} X' X \right)^{-1} \left(\frac{1}{\sqrt{T}} X' u \right) \\ &\xrightarrow{D} \mathcal{N} \left(0, S^{-1} \Sigma S^{-1} \right), \end{aligned}$$

applying Theorems 75 and 49. ■

The asymptotic variance resembles "sandwich" formula we derived when u is heteroskedastic. If u were homoskedastic,

$$\Sigma = \mathcal{E} \left(x_t x_t' u_t^2 \right) = \mathcal{E} \left(x_t x_t' \sigma_0^2 \right) = \sigma_0^2 S$$

and the asymptotic variance matrix reduces to $S^{-1} \Sigma S^{-1} = \sigma_0^2 S^{-1}$.

The careful reader has probably noticed that Theorems 75 and 81 assume that x_t are i.i.d. These theorems cannot be invoked when x feature dependence (as in time series models) and a different CLT has to be invoked. We will present it later.

Theorem 82 (Normality of $\widehat{\sigma}^2$) *Consider the HLRM with the additional assumption that $\mathcal{E} u_t^4 = m_4$. Then*

$$\sqrt{T} \left(\widehat{\sigma}^2 - \sigma_0^2 \right) \xrightarrow{D} \mathcal{N} \left(0, m_4 - \sigma_0^4 \right) .$$

Proof. We can write

$$\sqrt{T}\left(\widehat{\sigma}^2-\sigma_0^2\right)=\frac{u'u-T\sigma_0^2}{\sqrt{T}}-\frac{1}{\sqrt{T}}u'Pu.$$

The second term converges to 0 in probability by the same reasoning as in the proof of Theorem 76, and the first term can be dealt with by applying CLT1. Therefore, the theorem follows from Theorem 50. ■

8.5.2 Asymptotic Distribution of GLS

We already showed that $\widehat{\beta}_{GLS}$ is the OLS estimator of a transformed model, and that $\widehat{\beta}_{GLS}=\left(X'\Omega_0^{-1}X\right)^{-1}\left(X'\Omega_0^{-1}Y\right)$ is unbiased and has $\mathcal{V}\left(\widehat{\beta}_{GLS}\left|X\right.\right)=\sigma_0^2\left(X'\Omega_0^{-1}X\right)^{-1}$, it is not surprising that the derivation of the asymptotic properties of GLS can be derived using the same arguments that we used for the OLS estimator.

Theorem 83 (Consistency of $\widehat{\beta}_{GLS}$) *Consider the LRM. If :*
i) $\lambda_s\left(X'X\right)\rightarrow\infty$,
ii) $\lambda_l\left(\Omega_0\right)$ is bounded for all T.
Then $\widehat{\beta}_{GLS}\xrightarrow{p}\beta_0$

Proof. We have

$$\begin{aligned}\operatorname{tr}\mathcal{V}\left(\widehat{\beta}_{GLS}\left|X\right.\right)^{-1} &= \frac{1}{\sigma_0^2}\operatorname{tr}\left(X'\Omega_0^{-1}X\right)\\ &= \frac{1}{\sigma_0^2}\operatorname{tr}\left[\Omega_0^{-1}XX'\right]\\ &\leq \frac{1}{\sigma_0^2}\lambda_l\left(\Omega_0^{-1}\right)\operatorname{tr}\left[XX'\right] \quad \text{(Lemma 73)}\\ &= \frac{\operatorname{tr}\left[XX'\right]}{\sigma_0^2\lambda_l\left(\Omega_0\right)}.\end{aligned}$$

But $\lambda_l\left(\Omega_0\right)\neq 0$ is bounded. Finally, as $\lambda_s\left(X'X\right)\rightarrow\infty$, $\operatorname{tr}[XX']\rightarrow\infty$. Thus by Lemma 72, $\mathcal{V}\left(\widehat{\beta}_{GLS}\left|X\right.\right)\xrightarrow{p}0$. ■

Theorem 84 (Normality of $\widehat{\beta}_{GLS}$) *Assume that $\mathcal{E}\left(x_tu_t\right)=0$, $\mathcal{E}u_t^4<\infty$, $\mathcal{E}\left|x_t^4\right|<\infty$, $\frac{1}{T}X'\Omega_0^{-1}X\xrightarrow{p}W>0$, and $\sigma_0^2\Omega_0=\mathcal{E}\left(uu'\right)>0$. Then:*

i) $\frac{1}{\sqrt{T}}X'\Omega_0^{-1}u \xrightarrow{D} \mathcal{N}\left(0, \sigma_0^2 W\right)$,
ii) $\sqrt{T}\left(\widehat{\beta}_{GLS} - \beta_0\right) \xrightarrow{D} \mathcal{N}\left(0, \sigma_0^2 W^{-1}\right)$.

Proof. The first result follows from CLT1. For the second result, notice that

$$\begin{aligned}\sqrt{T}\left(\widehat{\beta}_{GLS} - \beta_0\right) &= \left(\frac{1}{T}X'\Omega_0^{-1}X\right)^{-1}\left(\frac{1}{\sqrt{T}}X'\Omega_0^{-1}u\right) \\ &\xrightarrow{D} \mathcal{N}\left(0, \sigma_0^2 W^{-1}\right),\end{aligned}$$

applying Theorems 50 and 49. ■

8.5.3 Asymptotic Distribution of FGLS

The asymptotic properties of FGLS depend on the consistent estimation of Ω_0. Nevertheless, remember that in order to obtain the FGLS estimator we must have the correct specification for Ω_0.

For the FGLS estimator to be consistent we require that the number of the free parameters that characterize Ω should be either bounded or allowed to go to infinity at a slower rate than T.

As was noted earlier, the first step in obtaining FGLS is calculating OLS. Therefore the properties of FGLS depend on the properties of OLS. Theorem 74 described the conditions under which this is the case.

If Ω is correctly specified and OLS is consistent, then the FGLS estimator shares the same asymptotic distribution of GLS. In particular,

$$\sqrt{T}\left(\widehat{\beta}_{FGLS} - \beta_0\right) \xrightarrow{D} \mathcal{N}\left(0, \sigma_0^2 W^{-1}\right),$$

where $W =$plim $\left(\frac{1}{T}X'\Omega_0^{-1}X\right)$.

8.6 Applications

Next, we present some exercises that use the convergence theorems we discussed.

Exercise 85 *Consider the HLRM model with $k = 1$ with x nonstochastic and assume that* $\lim T^{-1}X'X = c \neq 0$. *Let us consider the probability limit of the estimator that is obtained by minimizing the sum of squares in the direction of the x-axis: $\widehat{\beta}_R = Y'Y/X'Y$.*

Solution 86 *We can write*

$$\widehat{\beta}_R = \frac{\beta_0^2 + 2\beta_0 \frac{X'u}{X'X} + \frac{u'u}{X'X}}{\beta_0 + \frac{X'u}{X'X}}.$$

We have that $X'u/T \xrightarrow{M} 0$ *and* $X'X/T \to c$; *hence* $\frac{X'u}{X'X} \xrightarrow{p} 0$ *(Slutsky). Also we have* $\frac{u'u/T}{X'X/T} \xrightarrow{p} \frac{\sigma_0^2}{c}$ *because of Theorem 49 and SLLN2. Therefore* $\widehat{\beta}_R \xrightarrow{p} \beta_0 + \frac{\sigma_0^2}{\beta_0 c}$. *Notice that* c *may be allowed to be* ∞, *in which case* $\widehat{\beta}_R$ *would be a consistent estimator of* β_0.

Exercise 87 *Consider the same model as in Example 85 except that now we assume that* $\lim T^{-2}X'X = \infty$. *Also assume that* $\lim r_T^2 = 0$. *Show that* $\widehat{\beta}$ *and* $\widehat{\beta}_R$ *have the same asymptotic distribution.*

Solution 88 *Clearly* $\widehat{\beta}$ *and* $\widehat{\beta}_R$ *are consistent. Therefore by Theorem 45, both estimators have the same degenerate limit distribution. But the question concerns the asymptotic distribution; therefore we next obtain the limit distribution of each estimator after a suitable normalization. We can write*

$$(X'X)^{1/2}\left(\widehat{\beta}_R - \beta_0\right) = \frac{\beta_0 (X'X)^{1/2}\left(\widehat{\beta} - \beta_0\right) + \frac{u'u}{(X'X)^{1/2}}}{\beta_0 + \frac{X'u}{X'X}}.$$

But by our assumptions $\frac{u'u}{(X'X)^{1/2}} \xrightarrow{p} 0$ *and* $\frac{X'u}{X'X} \xrightarrow{p} 0$. *Therefore* $(X'X)^{1/2}\left(\widehat{\beta}_R - \beta_0\right) \overset{LD}{=} (X'X)^{1/2}\left(\widehat{\beta} - \beta_0\right)$ *by repeated applications of Theorem 50.*

Exercise 89 *Consider the HLRM with* $k = 2$. *Assume that* $\lim T^{-1}X'X = A$, *where* A *is a* 2×2 *nonsingular matrix. Obtain the asymptotic distribution of* $\widehat{\beta}_1/\widehat{\beta}_2$ *assuming* $\beta_2 \neq 0$.

Solution 90 *We can write*

$$T^{1/2}\left[\frac{\widehat{\beta}_1}{\widehat{\beta}_2} - \frac{\beta_1}{\beta_2}\right] = T^{1/2}\left[\frac{\beta_2\left(\widehat{\beta}_1 - \beta_1\right) - \beta_1\left(\widehat{\beta}_2 - \beta_2\right)}{\widehat{\beta}_2 \beta_2}\right].$$

Because $\widehat{\beta}_2 \xrightarrow{p} \beta_2$, *the right-hand side has the same limit distribution as*

$$\beta_2^{-1}T^{1/2}\left(\widehat{\beta}_1 - \beta_1\right) - \beta_1\beta_2^{-2}T^{1/2}\left(\widehat{\beta}_2 - \beta_2\right).$$

But, our assumptions imply that $\left[T^{1/2}\left(\widehat{\beta}_1-\beta_1\right), T^{1/2}\left(\widehat{\beta}_2-\beta_2\right)\right]$ *converge in distribution to a bivariate normal variable (Theorems 80 and 79). Therefore, by Theorem 48 we have*

$$\sqrt{T}\left[\frac{\widehat{\beta}_1}{\widehat{\beta}_2}-\frac{\beta_1}{\beta_2}\right] \xrightarrow{D} \mathcal{N}\left(0, \sigma_0^2\delta' A^{-1}\delta\right),$$

where $\delta'=\left(\beta_2^{-1}, -\beta_1\beta_2^{-2}\right)$.

8.7 Further Reading

The asymptotic tools introduced in this chapter constitute the mathematical backbone of modern econometrics. Readers seeking a deeper or more general exposition will find complementary treatments in a number of classic and modern sources.

Amemiya (1985) provides a rigorous and elegant development of convergence concepts and asymptotic results for linear and nonlinear estimators. This chapter heavily draws from it. The treatment by Hansen (2022) is modern, intuitive, and tightly connected to econometric practice, whereas Hayashi (2000) develops asymptotic theory as a unifying framework for consistency and efficiency across estimation methods.

8.8 Workout Problems

1. Prove the Theorems and corollaries that were asked for in the text.

2. Let $\{a_t\}$, $t=1,2,\cdots$, be a nonnegative sequence such that $\left(\sum_{t=1}^{T} a_t\right)/T < M$ for some M and every T. Prove $\lim_{T\to\infty}\sum_{t=1}^{T}\left(a_t/t^2\right)<\infty$.

3. Let x_T be defined by

$$x_T=\begin{cases} 0 \text{ with probability } (T-1)/T \\ T^{1/2} \text{ with probability } 1/T \end{cases}.$$

Let $A\mathcal{E}\left(x_T\right)=a$, $\lim \mathcal{E}x_T=b$, $x_T \xrightarrow{p} c$, $x_T \xrightarrow{M} d$, $x_T \xrightarrow{a.s.} e$. Find a, b, c, d, and e.

4. Let x_T be defined by

$$x_T = \begin{cases} Z \text{ with probability } (T-1)/T \\ 2TZ \text{ with probability } 1/T \end{cases}.$$

Let $A\mathcal{E}(x_T) = a$, $\lim \mathcal{E} x_T = b$, $x_T \xrightarrow{p} c$, $x_T \xrightarrow{M} d$, $x_T \xrightarrow{a.s.} e$. Find $a, b, c, d,$ and e.

5. Show that $\lambda_s(X'X) \to \infty$ implies $x_i'x_i \to \infty$ for every i, where x_i is the ith column vector of X. Show that the converse does not hold.

6. Assume $k = 1$ in the HLRM. If there exist L and M such that $0 < L < X'X/T < M$ for all T, show $\widehat{\beta} \xrightarrow{a.s.} \beta_0$.

7. Suppose $Y = Y^* + v$ and $X = X^* + w$, where each variable is a vector of T components. Assume Y^* and X^* are nonstochastic and (v_t, w_t) is a bivariate i.i.d. random variable with mean 0 and constant variances σ_v^2, σ_w^2, respectively, and covariance σ_{vw}. Assume $Y^* = \beta_0 X^*$, but Y^* and X^* are not observable so that we must estimate β_0 on the basis of Y and X. Obtain the probability limit of $\widehat{\beta} = X'Y/X'X$ on the assumption that $\lim_{T\to\infty} X^{*\prime}X^*/T = M$.

8. Let ι be the vector of ones. Assuming $\lim_{T\to\infty} \iota' X^*/T = N \neq 0$ in the model of the previous exercise, obtain the asymptotic distribution of $\widetilde{\beta} = \iota'Y/\iota'X$ and compare it with $\widehat{\beta}$.

9. Consider the HLRM with $k = 1$. Obtain the asymptotic distribution of $\widetilde{\beta} = \iota'Y/\iota'X$ assuming $\lim_{T\to\infty} (\iota'X)^2/T = \infty$.

10. Consider the HLRM $Y = \alpha X + \beta Z + u$, where X and Z are T-component vectors of known constants, and $\sigma^2 = 1$. Suppose you are given an estimator $\widetilde{\beta}$ that is independent of u and $T^{1/2}\left(\widetilde{\beta} - \beta\right) \xrightarrow{D} \mathcal{N}(0,1)$. Define the estimator $\widetilde{\alpha}$ by

$$\widetilde{\alpha} = \frac{X'\left(Y - \widetilde{\beta} Z\right)}{X'X}.$$

Assuming $\lim T^{-1}X'X = c \neq 0$ and $\lim T^{-1}X'Z = d \neq 0$, obtain the asymptotic distribution of $\widetilde{\alpha}$.

11. Consider the HLRM $Y = \beta(X + \alpha\iota) + u$ where Y, X, ι, and u are T-vectors. Assume that $\lim T^{-1}X'\iota = 0$, and $\lim T^{-1}X'X = c \neq 0$. Supposing we have an estimate of α denoted by $\widetilde{\alpha}$ such that it is distributed independently of u and $T^{1/2}(\widetilde{\alpha} - \alpha) \xrightarrow{D} \mathcal{N}(0, \lambda^2)$, obtain the asymptotic distribution of $\widetilde{\beta}$ defined by

$$\widetilde{\beta} = \frac{(X + \widetilde{\alpha}\iota)' Y}{(X + \widetilde{\alpha}\iota)'(X + \widetilde{\alpha}\iota)}.$$

12. Consider the HLRM $Y = \alpha X + \beta Z + u$, where X and Z are T-component vectors such that $x_t = t^a$, $z_t = t^b$, $a, b \geq 0$. Verify that the OLS estimators are consistent. What happens if $a = b$? [Hint: recall that $T^{-1-p}\sum_{t=1}^{T} t^p \to 1/(p+1)$].

13. Consider the HLRM $Y = \alpha X + u$, where X is a T-component vector. Verify if the OLS estimator is consistent if

(**a**) $x_t = t^{-1/2}$,

(**b**) $x_t = a^t$ (for $0 < a < 1$, a known constant).

Chapter 9

Extremum Estimators

9.1 Introduction

Extremum estimators represent a class of estimators that encompass most of those used in econometric practice. By extremum estimators we mean estimators obtained by maximizing or minimizing a function defined over the parameter space.[1] Ordinary and generalized least squares (OLS, GLS), feasible GLS, nonlinear least squares (NLLS), maximum likelihood (ML), generalized method of moments (GMM) discussed latter, and least absolute deviations (LAD) estimators all belong to this family.

When an estimator can be expressed as an explicit function of the sample, its properties can often be studied directly, as in the case of OLS. More commonly, however, the estimator is defined only implicitly as the solution to an optimization problem with no closed-form expression. In such cases, general theorems on the asymptotic behavior of extremum estimators provide a unified approach for establishing large-sample results.

This chapter develops the large-sample theory of extremum estimators and illustrates its implications for the estimators most frequently used in econometrics.

Section 9.2 establishes conditions for existence, consistency, and asymptotic normality. Sections 9.3–9.5 apply these results to the OLS, NLLS, and MLE estimators. Section 9.6 analyzes the least-absolute-deviations es-

[1] Extremum estimators are also referred to as M-estimators. The term "M", stands for "maximum" and denotes estimators obtained by maximizing or minimizing a sample criterion function.

timator, whose nonsmooth criterion requires special treatment. Section 9.7 concludes with the asymptotic foundations of hypothesis testing through the Wald, Lagrange Multiplier, and Likelihood Ratio statistics.

9.2 General Results

Because there is no essential difference between maximization and minimization, we shall consider an estimator that maximizes a certain function of the parameters. Of course, we can minimize an objective function by maximizing the negative of the function. Let us denote the function by $Q_T(Y, \theta)$, where Y is a T-vector of random variables and θ is a k-vector of parameters.[2] Let us denote the domain of θ, or the parameter space, by Θ and the "true value" of θ by θ_0. The parameter space is the set of all the possible values that the true value θ_0 can take. When we take various operations on a function of Y, such as expectation or probability limit, we shall use the value θ_0.

An estimator $\widehat{\theta}$ is called an extremum estimator if there is a scalar objective function $Q_T(\theta)$ such that

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} Q_T(\theta). \tag{9.1}$$

This estimator may not exist if the maximization problem does not have a solution. However, if Θ is compact and the objective function is continuous, there exists a θ that solves (9.1) for any Y. In the event of multiple solutions, we would choose one from them. Strictly speaking, being a function of Y is not enough to make $\widehat{\theta}$ a well-defined random variable; $\widehat{\theta}$ needs to be a "measurable" function of Y.[3] The following lemma shows that $\widehat{\theta}$ is measurable if $Q_T(\theta)$ is.

Lemma 91 (Existence of Extremum Estimators) *Suppose that:*

i) Θ is a compact subset of $\mathbb{R}^k$,

ii) $Q_T(\theta)$ is continuous in θ for any data Y,

iii) $Q_T(\theta)$ is a measurable function for all $\theta \in \Theta$,

Then, there exists a measurable function $\widehat{\theta}$ of the data that solves (9.1).

[2] We shall write $Q_T(Y, \theta)$ more compactly as $Q_T(\theta)$, but always bare in mind that the objective function depends on the sample of Y and thus is itself a random variable.

[3] If a function is continuous, it is measurable.

Intuitively, compactness of the parameter space ensures the existence of a maximum, continuity guarantees that the criterion function behaves regularly, and measurability ensures that the estimator is a well-defined random variable. These are technical but minimal assumptions required for the estimator to exist as a measurable function of the data.

In most applications, we do not know the upper or lower bound for the true parameter vector. So the compactness assumption for Θ is something we may wish to avoid. In some of the asymptotic results that follow, we will replace the compactness assumption by some other conditions that are satisfied in many applications.

9.2.1 Consistency of Extremum Estimators

The first step in establishing large-sample properties is to show that the estimator converges in probability to the true parameter value.

As a preliminary to the consistency theorems, we shall define three modes of uniform convergence of a sequence of random variables.

Definition 92 *Let $g_T(\theta)$ be a nonnegative sequence of random variables depending on a parameter vector θ. Consider the three modes of uniform convergence of $g_T(\theta)$ to 0:*

i) $\Pr\left[\lim_{T\to\infty}\sup_{\theta\in\Theta} g_T(\theta)=0\right]=1$,

ii) $\lim_{T\to\infty}\Pr\left[\sup_{\theta\in\Theta} g_T(\theta)<\varepsilon\right]=1$ *for any* $\varepsilon>0$,

iii) $\lim_{T\to\infty}\inf_{\theta\in\Theta}\Pr\left[g_T(\theta)<\varepsilon\right]=1$ *for any* $\varepsilon>0$.

If i) holds, we say $g_T(\theta)$ converges to 0 almost surely uniformly in $\theta\in\Theta$. If ii) holds, we say $g_T(\theta)$ converges to 0 in probability uniformly in $\theta\in\Theta$. If iii) holds, we say $g_T(\theta)$ converges to 0 in probability semiuniformly in $\theta\in\Theta$. It is easy to see that i) implies ii) and iii) and ii) implies iii).

Uniform convergence strengthens the law of large numbers from pointwise to functional convergence. It guarantees that, with high probability, the entire function lies close to its limit for all parameter values.

Now we shall prove the consistency of extremum estimators. Because we need to distinguish the global maximum and a local maximum, we shall present three theorems to handle these cases.

Theorem 93 (Consistency of Extremum Estimators 1) *Suppose that:*

A) Θ is a compact subset of $\mathbb{R}^k$, *B) $Q_T(Y,\theta)$ is continuous in $\theta \in \Theta$ for all Y and is a measurable function of Y for all $\theta \in \Theta$,*

C) $T^{-1}Q_T(\theta)$ converges uniformly in probability to a nonstochastic function $Q(\theta)$,

i) (Identification) $Q(\theta)$ attains a unique global maximum at θ_0,

ii) $Q(\theta)$ is continuous in $\theta \in \Theta$.

Then $\widehat{\theta} \xrightarrow{p} \theta_0$.

Proof. Let N be an open neighborhood in $\mathbb{R}^k$ containing θ_0. Then $\overline{N} \cap \Theta$ is compact (where $\overline{N}$ is the complement of N in $\mathbb{R}^k$). Therefore $\max_{\theta \in \overline{N} \cap \Theta} Q(\theta)$ exists. Denote

$$\xi = Q(\theta_0) - \max_{\theta \in \overline{N} \cap \Theta} Q(\theta). \tag{9.2}$$

Let A_T be the event $|T^{-1}Q_T(\theta) - Q(\theta)| < \xi/2 \; \forall \theta$. Then[4]

$$A_T \Rightarrow Q\left(\widehat{\theta}\right) > T^{-1}Q_T\left(\widehat{\theta}\right) - \xi/2 \tag{9.3}$$

and

$$A_T \Rightarrow T^{-1}Q_T(\theta_0) > Q(\theta_0) - \xi/2. \tag{9.4}$$

But because $Q_T\left(\widehat{\theta}\right) \geq Q_T(\theta_0)$ by the definition of $\widehat{\theta}$, we have from (9.3)

$$A_T \Rightarrow Q\left(\widehat{\theta}\right) > T^{-1}Q_T(\theta_0) - \xi/2. \tag{9.5}$$

Therefore, adding both sides of the inequalities in (9.4) and (9.5), we obtain

$$A_T \Rightarrow Q\left(\widehat{\theta}\right) > Q(\theta_0) - \xi. \tag{9.6}$$

Therefore, from (**??**) and (9.6) we can conclude that

$$A_T \Rightarrow Q\left(\widehat{\theta}\right) > \max_{\theta \in \overline{N} \cap \Theta} Q(\theta),$$

which means that $A_T \Rightarrow \widehat{\theta} \in N$. This implies that $\Pr(A_T) \leq \Pr\left(\widehat{\theta} \in N\right)$. But, since $\lim_{T \to \infty} \Pr(A_T) = 1$ by assumption C, $\widehat{\theta} \xrightarrow{p} \theta_0$. ■

The intuition is straightforward: uniform convergence implies that the sample criterion mirrors the population criterion in large samples. Since $Q(\theta)$

[4] See Exercise 1.

attains a unique maximum at θ_0, the maximizer of $Q_T(\theta)$ must lie arbitrarily close to θ_0 with probability approaching one. Compactness guarantees that a maximizer exists within the parameter space.

Note that Theorem 93 requires the compactness of Θ. This is a weakness of the theorem in so far as the extremum estimators commonly used in practice are obtained by unconstrained maximization or minimization. This practice prevails because of its relative computational ease, even though constrained maximization or minimization would be desirable if a researcher believed that θ_0 lay in a proper subset of $\mathbb{R}^k$. The consistency of the unconstrained maximum $\widetilde{\theta}$ defined by

$$\widetilde{\theta} = \arg\max_{\theta \in \mathbb{R}^k} Q_T(\theta),$$

will follow from the additional assumption:

D) $\lim_{T\to\infty} \Pr\left[Q_T(\theta_0) > \sup_{\theta\notin\Theta} Q_T(\theta)\right] = 1$ because of the inequality

$$\Pr\left[Q_T(\theta_0) > \sup_{\theta\notin\Theta} Q_T(\theta)\right] \leq \Pr\left(\widetilde{\theta} \in \Theta\right).$$

Theorem 94 relaxes the requirements of compactness of Θ and of uniform convergence in probability of $T^{-1}Q_T(\theta)$. The conditions are replaced by the requirements that the objective function by concave in θ, exhibit ordinary convergence in probability, and that Θ be convex.

Theorem 94 (Consistency of Extremum Estimators 2) *Suppose that:*

A) Θ is a convex set, with θ_0 in the interior of Θ,

B) $Q_T(Y, \theta)$ is concave in $\theta \in \Theta$ for all Y and is a measurable function of Y for all $\theta \in \Theta$,

C) $T^{-1}Q_T(\theta)$ converges in probability to a nonstochastic function $Q(\theta)$,

i) (Identification) $Q(\theta)$ is uniquely maximized at θ_0,

ii) $Q(\theta)$ is concave in $\theta \in \Theta$.

Then $\widehat{\theta} \xrightarrow{p} \theta_0$.

Proof. Newey and McFadden (1994). ■

To motivate the precise conditions for consistency it is helpful to sketch the ideas on which the results are based. The basic idea is that if $T^{-1}Q_T(\theta)$ converges in probability to $Q(\theta)$ for every θ, and $Q(\theta)$ is maximized at the true parameter value θ_0, then the limit of the maximum $(\widehat{\theta})$ should be the maximum (θ_0) of the limit, under conditions for interchanging the maximization and limiting operations.

The estimator $\widehat{\theta}$ of Theorems 93 or 94 maximizes the function $Q_T(\theta)$ globally. However, in practice it is often difficult to locate a global maximum of $Q_T(\theta)$, for it means that we must look through the whole parameter space except in the fortunate situation where we can prove that $Q_T(\theta)$ is globally concave. Another weakness of both Theorems is that it is often difficult to prove that $Q(\theta)$ attains a global maximum at θ_0. Therefore we would also like to have a theorem regarding the consistency of a local maximum.

Theorem 95 (Consistency of Extremum Estimators 3) *Suppose that:*

A) Θ is a convex set, with θ_0 in the interior of Θ,

B) $Q_T(Y,\theta)$ is a measurable function of Y for all $\theta \in \Theta$, and $\partial Q_T/\partial\theta$ exists and is continuos in an open neighborhood $N_1(\theta_0)$ of θ_0,

C) There exists an open neighborhood $N_2(\theta_0)$ of θ_0 such that $T^{-1}Q_T(\theta)$ converges to a nonstochastic function $Q(\theta)$ in probability uniformly in θ in $N_2(\theta_0)$, and $Q(\theta)$ attains a strict local maximum at θ_0,

Let Θ_T be the set of roots of the equation

$$\frac{\partial Q_T}{\partial \theta} = 0, \tag{9.7}$$

corresponding to the local maxima. Then for any $\varepsilon > 0$,

$$\lim_{T\to\infty} \Pr\left[\inf_{\theta\in\Theta_T} (\theta-\theta_0)'(\theta-\theta_0) > \varepsilon\right] = 0.$$

Proof. Choose a compact set $S \subset N_1 \cap N_2$. Then the value of θ, say $\widetilde{\theta}$, that globally maximizes $Q_T(\theta)$ in S is consistent by Theorem 93. But because the probability that $T^{-1}Q_T(\theta)$ attains a local maximum at $\widetilde{\theta}$ approaches to 1 as T goes to ∞, $\lim_{T\to\infty}\Pr\left(\widetilde{\theta}\in\Theta_T\right) = 1$. ■

We sometimes state the conclusion of Theorem 95 simply as "there is a consistent root of (9.7)."

The usefulness of Theorem 95 is limited by the fact that it merely states that one of the local maxima is consistent and does not give any guide as to how to choose a consistent maximum. There are two ways we can gain some degree of confidence that a local maximum is a consistent root:

1. If the solution gives a reasonable value from an economic-theoretic viewpoint.

2. If the iteration by which the local maximum was obtained started from a consistent estimator.

This result allows for local identification of the true parameter. It shows that even if several roots exist, one of them must converge in probability to θ_0. In practice, it provides the theoretical justification for numerical optimization routines that converge to a consistent local optimum, provided that a consistent starting value is used.

To summarize, Theorem 93 establishes consistency under compactness and uniform convergence, Theorem 94 exploits concavity to relax these requirements, and Theorem 95 ensures that at least one stationary point is consistent when only local information is available. These three results together cover the settings most frequently encountered in econometric applications.

9.2.2 Asymptotic Normality of Extremum Estimators

Consistency establishes that the estimator converges to the true parameter value. To describe the distribution of the estimation error, we now consider the rate of convergence and the limiting distribution.

In this subsection we shall show that under certain conditions a consistent root of (9.7) is asymptotically normal. In order to derive the asymptotic properties of extremum estimators we will apply the Mean Value Theorem.

Theorem 96 (Mean Value Theorem) *Let $h : \mathbb{R}^p \to \mathbb{R}^q$ be continuously differentiable. Then $h(x)$ admits the mean value expansion*

$$\underset{q\times 1}{h(x)} = \underset{q\times 1}{h(x_0)} + \underset{q\times p}{\frac{\partial h(\overline{x})}{\partial x'}} \underset{p\times 1}{(x - x_0)},$$

where $\overline{x}$ is a mean value lying between x and x_0.

This result allows us to approximate the value of a differentiable function in a neighborhood of a point by its derivative evaluated at an intermediate point.

Applied to the first-order conditions of the criterion function, it provides the basis for deriving the asymptotic distribution of the estimator.

Theorem 97 (Asymptotic Normality of Extremum Estimators) *Suppose that the conditions of Theorem 93, 94, or 95 are satisfied, so that* $\widehat{\theta} \xrightarrow{p} \theta_0$*. Suppose, further, that*

A) $\partial^2 Q_T / \partial\theta\partial\theta'$ *exists and is continuous in an open, convex neighborhood of* θ_0*,*

B) $T^{-1}\left(\partial^2 Q_T / \partial\theta\partial\theta'\right)_{\widetilde{\theta}} \xrightarrow{p} A\left(\theta_0\right) = \lim_{T\to\infty} \mathcal{E}T^{-1}\left(\partial^2 Q_T / \partial\theta\partial\theta'\right)_{\theta_0}$ *for any* $\widetilde{\theta}$ *such that* $\widetilde{\theta} \xrightarrow{p} \theta_0$ *(with* $A\left(\theta_0\right)$ *being a finite nonsingular matrix),*

C) $T^{-1/2}\left(\partial Q_T / \partial\theta\right)_{\theta_0} \xrightarrow{D} \mathcal{N}\left[0, B\left(\theta_0\right)\right]$*, where* $B\left(\theta_0\right) = \lim_{T\to\infty} \mathcal{E}T^{-1}\left(\partial Q_T / \partial\theta\right)_{\theta_0} \times \left(\partial Q_T / \partial\theta'\right)_{\theta_0}$*.*

Then

$$\sqrt{T}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{D} \mathcal{N}\left[0, A\left(\theta_0\right)^{-1} B\left(\theta_0\right) A\left(\theta_0\right)^{-1}\right].$$

Proof. By a Taylor expansion and Theorem 96 we have

$$\left.\frac{\partial Q_T}{\partial\theta}\right|_{\widehat{\theta}} = \left.\frac{\partial Q_T}{\partial\theta}\right|_{\theta_0} + \left.\frac{\partial^2 Q_T}{\partial\theta\partial\theta'}\right|_{\widetilde{\theta}}\left(\widehat{\theta} - \theta_0\right),$$

where $\widetilde{\theta}$ lies between $\widehat{\theta}$ and θ_0. Noting that the left hand side is 0 by the definition of $\widehat{\theta}$, we obtain

$$\sqrt{T}\left(\widehat{\theta} - \theta_0\right) = -\left[\frac{1}{T}\left.\frac{\partial^2 Q_T}{\partial\theta\partial\theta'}\right|_{\widetilde{\theta}}\right]^{-1}\frac{1}{\sqrt{T}}\left.\frac{\partial Q_T}{\partial\theta}\right|_{\theta_0}. \tag{9.8}$$

Given that $\widetilde{\theta} \xrightarrow{p} \theta_0$, assumption B) implies

$$\frac{1}{T}\left.\frac{\partial^2 Q_T}{\partial\theta\partial\theta'}\right|_{\widetilde{\theta}} \xrightarrow{p} A\left(\theta_0\right)$$

and assumption C) implies

$$\frac{1}{\sqrt{T}}\left.\frac{\partial Q_T}{\partial\theta}\right|_{\theta_0} \xrightarrow{D} \mathcal{N}\left[0, B\left(\theta_0\right)\right].$$

The conclusion of the Theorem follows by repeated applications of Slutsky's Theorem. ■

The expression $A\left(\theta_0\right)^{-1} B\left(\theta_0\right) A\left(\theta_0\right)^{-1}$ is known as the asymptotic variance matrix. Keep in mind that this is not the variance matrix of $\widehat{\theta}$, which has as variance matrix $T^{-1}A\left(\theta_0\right)^{-1} B\left(\theta_0\right) A\left(\theta_0\right)^{-1}$. Given that in practice, we do not observe θ_0, we obtain an estimator for the variance matrix of $\widehat{\theta}$ by replacing θ_0 with its consistent estimator $\widehat{\theta}$ and obtain

$$\widehat{\mathcal{V}}\left(\widehat{\theta}\right) = T^{-1} A\left(\widehat{\theta}\right)^{-1} B\left(\widehat{\theta}\right) A\left(\widehat{\theta}\right)^{-1}.$$

The results established in this section form the backbone of modern asymptotic theory.

Uniform convergence ensures consistency, differentiability ensures asymptotic linearity, and the combination of both yields asymptotic normality.

The following sections illustrate how these general principles apply to specific estimators widely used in econometric analysis.

9.3 OLS as an Extremum Estimator

Ordinary least squares provides the simplest and most familiar example of an extremum estimator. It serves as a benchmark for understanding the general theory developed in the previous section. By deriving its large-sample properties directly from the extremum framework, we confirm that the consistency and asymptotic normality of OLS are special cases of the general results presented in Section 9.2.

As mentioned in the Introduction, the LS criterion for estimating θ is subsumed under the extremum estimation concept. Thus, the OLS estimator can be represented in extremum estimator form as

$$\widehat{\beta} = \arg\max_{\beta} Q_T\left(Y, X, \beta\right),$$

where

$$Q_T\left(Y, X, \beta\right) = \left[-\left(Y - X\beta\right)'\left(Y - X\beta\right)\right]. \tag{9.9}$$

Although in this case the estimator is available in closed form as $\widehat{\beta} = \left(X'X\right)^{-1}\left(X'Y\right)$, treating it as an extremum estimator will facilitate the later analysis of models for which closed-form solutions are not available.

9.3.1 Consistency of OLS

A demonstration that the OLS estimator is consistent can be based on an application of Theorem 94. First of all, note that we can express the objective

(9.9) as:

$$\begin{aligned}
Q_T(Y,X,\beta) &= -\left[(Y - X\beta_0) + (X\beta_0 - X\beta)\right]' \left[(Y - X\beta_0) + (X\beta_0 - X\beta)\right] \\
&= -\left[X\beta_0 - X\beta + u\right]' \left[X\beta_0 - X\beta + u\right] \\
&= -(\beta - \beta_0)' X'X (\beta - \beta_0) + 2(\beta - \beta_0)' X'u - u'u.
\end{aligned}$$

Clearly Q_T is concave (not only concave, but globally concave) in β, thus assumption B) of Theorem 94 is satisfied. Next we derive the function $Q(\theta)$. For that we need to either assume that the u_t's are i.i.d. or that their fourth moments are bounded. Furthermore, assume that $T^{-1}X'X \to S$, a finite positive definite matrix. Then:

$$\frac{1}{T} Q_T(\beta) = \frac{1}{T}\left[-(\beta - \beta_0)' X'X (\beta - \beta_0) + 2(\beta - \beta_0)' X'u - u'u\right]. \quad (9.10)$$

If u_t is i.i.d. $T^{-1}u'u \xrightarrow{p} \sigma_0^2$ by Kinchine's WLLN, while if the u_t's are not i.i.d. but have bounded fourth moments, we arrive to the same conclusion ($T^{-1}u'u \xrightarrow{p} \sigma_0^2$) applying Chebyshev's WLLN. Thus, the last term of (9.10) converges to $-\sigma_0^2$. The second term converges in probability to 0 because $\mathcal{E}\left[2T^{-1}(\beta - \beta_0)' X'u\right] = 0$ and $\mathcal{V}\left[2T^{-1}(\beta - \beta_0)' X'u\right] = 4T^{-1}\sigma_0^2 (\beta - \beta_0)' (T^{-1}X'X)(\beta - \beta_0)$; thus, as $T \to \infty$, $\mathcal{V}\left[2T^{-1}(\beta - \beta_0)' X'u\right] \to 0$; it follows that $2T^{-1}(\beta - \beta_0)' X'u \xrightarrow{M} 0$, so $2T^{-1}(\beta - \beta_0)' X'u \xrightarrow{p} 0$. Finally, as $T^{-1}X'X \to S$, we have that

$$\frac{1}{T} Q_T(\beta) \xrightarrow{p} -(\beta - \beta_0)' S (\beta - \beta_0) - \sigma_0^2 = Q(\beta).$$

Note that $Q(\theta)$ is a concave nonstochastic function of β. Finally, we have that the FONC and SOSC for maximizing Q are:

$$\frac{\partial Q}{\partial \beta} = -2(\beta - \beta_0)' S$$

and

$$\frac{\partial^2 Q}{\partial \beta \partial \beta'} = -2S < 0,$$

which confirms that Q is globally concave and that indeed $\beta_0 = \arg\max_\theta Q$ given that $\beta = \beta_0$ satisfies the FONC. Thus, $\widehat{\beta} \xrightarrow{p} \beta_0$.

Bare in mind that even though a closed form expression for $\widehat{\beta}$ is available, we did not use it in order to demonstrate the consistency of the OLS estimator. This is precisely the usefulness of the asymptotic theory for extremum estimators, given that we only need to concentrate on the properties of the objective function and do not need to even derive the extremum estimator.

If we wanted to use Theorem 93, we will need to prove that $T^{-1}Q_T(\beta)$ converges in probability uniformly to $Q(\beta)$ and we will also require the parameter space to be compact. The first requirement is easy to prove, given that

$$\frac{1}{T}Q_T(\beta)-Q(\beta)=-(\beta-\beta_0)'\left[\frac{X'X}{T}-M\right](\beta-\beta_0)+2(\beta-\beta_0)'\frac{X'u}{T}-\frac{u'u}{T}+\sigma_0^2.$$

Then,

$$\arg\max_{\beta}\left[\frac{1}{T}Q_T(\beta)-Q(\beta)\right]=\left[\frac{X'X}{T}-S\right]^{-1}\frac{X'u}{T}+\beta_0$$

and

$$\sup_{\beta}\left[\frac{1}{T}Q_T(\beta)-Q(\beta)\right]=\frac{u'X}{T}\left[\frac{X'X}{T}-S\right]^{-1}\frac{X'u}{T}-\frac{u'u}{T}+\sigma_0^2,$$

so

$$\lim_{T\to\infty}\Pr\left[\sup_{\beta}\left[\frac{1}{T}Q_T(\beta)-Q(\beta)\right]<\varepsilon\right]=1 \text{ for any } \varepsilon>0.$$

For the second requirement (compactness) we may choose arbitrarily large values (in absolute value) for each parameter.

The argument parallels the population and sample least-squares decomposition. The expected value of the sum of squared residuals is minimized at the true parameter, and the law of large numbers ensures that the sample criterion converges uniformly to its expectation. Concavity guarantees a unique maximizer, establishing that the OLS estimator is consistent.

9.3.2 Asymptotic Normality of OLS

In order to derive the asymptotic distribution of the OLS estimator, we need to verify that the conditions of Theorem 97 are satisfied. To do so, note that

$$\begin{aligned}\frac{\partial Q_T}{\partial\beta} &= -2\beta'X'X+2\beta_0'X'X+2X'u,\\ \frac{\partial^2 Q_T}{\partial\beta\partial\beta'} &= -2X'X.\end{aligned}$$

Therefore

$$\frac{1}{T}\frac{\partial^2 Q_T}{\partial\beta\partial\beta'} \to -2S = A\left(\theta_0\right),$$

which coincides with

$$A\left(\theta_0\right) = \lim_{T\to\infty} \mathcal{E}\frac{1}{T}\left.\frac{\partial^2 Q_T}{\partial\beta\partial\beta'}\right|_{\beta_0} = -2S.$$

In this case we did not need to take expectations on A given that we assumed that X was deterministic. Furthermore as S is positive definite, A is nonsingular; thus assumption B) of Theorem 97 is satisfied.

Next, we verify that assumption C) of Theorem 97 is also satisfied. First, let's derive an expression for $B\left(\theta_0\right)$:

$$\begin{aligned} B\left(\theta_0\right) &= \lim_{T\to\infty} \mathcal{E}\frac{1}{T}\left.\frac{\partial Q_T}{\partial\beta}\right|_{\beta_0}\left.\frac{\partial Q_T}{\partial\beta'}\right|_{\beta_0} \\ &= \lim_{T\to\infty} \mathcal{E}\frac{1}{T}4X'uu'X \\ &= \lim_{T\to\infty} 4\sigma_0^2\frac{1}{T}X'X \\ &= 4\sigma_0^2 S. \end{aligned}$$

Next, we need to verify that $T^{-1/2}\left(\partial Q_T/\partial\theta\right)_{\widetilde{\theta}} \xrightarrow{D} \mathcal{N}\left[0, B\left(\theta_0\right)\right]$. We have:

$$\frac{1}{\sqrt{T}}\left.\frac{\partial Q_T}{\partial\beta}\right|_{\widetilde{\beta}} \overset{LD}{=} \frac{1}{\sqrt{T}}\left(2X'u\right),$$

given that $\widetilde{\beta} \xrightarrow{p} \beta_0$. It is trivial to verify that

$$\frac{1}{\sqrt{T}}\left(2X'u\right) \xrightarrow{D} \mathcal{N}\left(0, 4\sigma_0^2 S\right). \tag{9.11}$$

Thus,

$$\sqrt{T}\left(\widehat{\theta}-\theta_0\right) \xrightarrow{D} \mathcal{N}\left[0, A\left(\theta_0\right)^{-1}B\left(\theta_0\right)A\left(\theta_0\right)^{-1}\right]$$

implies that

$$\sqrt{T}\left(\widehat{\beta}-\beta_0\right) \xrightarrow{D} \mathcal{N}\left[0, \left(-\frac{1}{2}S^{-1}\right)4\sigma_0^2 S\left(-\frac{1}{2}S^{-1}\right)\right],$$

or

$$\sqrt{T}\left(\widehat{\beta}-\beta_0\right)\xrightarrow{D}\mathcal{N}\left[0,\sigma_0^2 S^{-1}\right].$$

Thus,

$$\left(\widehat{\beta}-\beta_0\right)\overset{a}{\backsim}\mathcal{N}\left[0,\sigma_0^2\left(X'X\right)^{-1}\right],$$

which is the result that we derived earlier.

9.4 NLLS as an Extremum Estimator

Nonlinear least squares (NLLS) extends the principle of least squares to models where the systematic component is nonlinear in the parameters. Although the estimator cannot generally be expressed in closed form, its large-sample properties follow directly from the general extremum-estimation results presented in Section 9.2. This section illustrates how consistency and asymptotic normality arise as straightforward consequences of those general theorems.

The NLLS estimator can be represented as in extremum estimator form as

$$\widehat{\beta}_{NLLS}=\arg\max_{\theta\in\Theta}Q_T\left(Y,X,\theta\right),$$

where

$$Q_T\left(Y,X,\theta\right)=-\sum_{t=1}^{T}\left(y_t-m\left(x_t,\beta\right)\right)^2=-\left[Y-m\left(X,\beta\right)\right]'\left[Y-m\left(X,\beta\right)\right].$$

9.4.1 Consistency of NLLS

The conditions for consistency of the global minimum or the local minimum can be obtained from Theorems 93, 94, or 95.[5] We shall consider only the latter case because the FONC are needed to prove asymptotic normality. Next, we restate Theorem 95 for the special case of NLLS.

Theorem 98 (Consistency of NLLS) *There exists an open neighborhood N of β_0 such that:*

A) $\partial m_t/\partial\beta$ exists and is continuos on N,

[5] The global minimum is the value of β that globally minimizes the SSR function over the parameter space, while the local minimum is any root of the FONC that corresponds to a local minimum.

B) $m_t(\beta)$ is continuous in $\beta \in N$,
C) $T^{-1}\sum_{t=1}^{T} m_t(\beta_1)\, m_t(\beta_2)$ converges uniformly in $\beta_1, \beta_2 \in N$,
D) $\lim T^{-1}\sum_{t=1}^{T} [m_t(\beta_0) - m_t(\beta)]^2 \neq 0$ if $\beta \neq \beta_0$.
Then a root of the equations

$$\frac{\partial Q_T}{\partial \beta} = 0$$

is consistent in the sense of Theorem 95.

Proof. (Sketch) Note that

$$\begin{aligned}
\frac{1}{T}Q_T &= -\frac{1}{T}\sum_{t=1}^{T}[y_t - m_t(\beta)]^2 \\
&= -\frac{1}{T}\sum_{t=1}^{T}[y_t - m_t(\beta_0) + m_t(\beta_0) - m_t(\beta)]^2 \\
&= -\frac{1}{T}\sum_{t=1}^{T}[u_t + m_t(\beta_0) - m_t(\beta)]^2 \\
&= -\frac{1}{T}\sum_{t=1}^{T}u_t^2 - \frac{1}{T}\sum_{t=1}^{T}[m_t(\beta_0) - m_t(\beta)]^2 - \frac{2}{T}\sum_{t=1}^{T}[m_t(\beta_0) - m_t(\beta)]\, u_t \\
&= A_1 + A_2 + A_3.
\end{aligned}$$

The term A_1 converges to σ_0^2 in probability by WLLN2. The term A_2 converges to a function that has a local minimum at β_0 because of assumptions C and D. Finally, A_3 converges to 0 in probability uniformly (<, >)amemiya. ■

Of course, the consistency of the NLLS estimator can be proven "directly" if we assume that

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\left.\frac{\partial m_t}{\partial \beta}\right|_{\beta_0}\left.\frac{\partial m_t}{\partial \beta'}\right|_{\beta_0} = \lim_{T\to\infty}\frac{1}{T}Z(\beta_0)'\, Z(\beta_0) = S, \qquad (9.12)$$

and use the Gauss-Newton result, with the Taylor series approximation made around β_0:

$$\widehat{\beta} = \beta_0 + \left[T^{-1}Z(\beta_0)'\, Z(\beta_0)\right]^{-1}\left[T^{-1}Z(\beta_0)'\,(Y - m(X, \beta_0))\right].$$

If S is invertible we find that $\widehat{\beta}$ is consistent, because the last term converges in probability to 0. This last result cannot usually be applied because (9.12) is not a sufficient condition for consistency given that we need to impose further restrictions on m.

When $m_t(\beta)$ has a very simple form, consistency can be proved more simply and with fewer assumptions by using Theorem 93, Theorem 94, or Theorem 95 directly rather than Theorem 98, as we shall show in the following example.

Example 99 *Consider the nonlinear regression model with* $m_t(\beta_0) = (\beta_0 + x_t)^2$ *and assume that:*

i) $a \leq \beta_0 \leq b$ *where* a *and* b *are real numbers such that* $a < b$*,*

ii) $\lim_{T\to\infty} T^{-1} \sum_{t=1}^{T} x_t = q$*, and*

iii) $\lim_{T\to\infty} T^{-1} \sum_{t=1}^{T} x_t^2 = p > q^2$*.*

We will prove that $\widehat{\beta}$ *is consistent. We have that* $-\frac{1}{T}\sum_{t=1}^{T}\left[y_t - (\beta + x_t)^2\right]^2 \xrightarrow{p} Q(\beta)$ *uniformly, where*

$$Q(\beta) = -\sigma_0^2 - \left(\beta_0^2 - \beta^2\right)^2 - 4(\beta_0 - \beta)^2 p - 4\left(\beta_0^2 - \beta^2\right)(\beta_0 - \beta) q.$$

But, because

$$Q(\beta) \leq -\sigma_0^2 - (\beta_0 - \beta)^2 (\beta_0 + \beta + 2q)^2,$$

$Q(\beta)$ *is uniquely maximized at* $\beta = \beta_0$*. Therefore, the estimator in question is consistent.*

9.4.2 Asymptotic Normality of NLLS

We shall now prove the asymptotic normality of the NLLS estimator of β by making assumptions on m_t to satisfy the assumptions of Theorem 97.

First, consider assumption C of Theorem 97. We have

$$\frac{\partial Q_T}{\partial \beta} = 2\sum_{t=1}^{T} [y_t - m_t(\beta)] \frac{\partial m_t}{\partial \beta}.$$

Therefore, we have

$$\frac{1}{\sqrt{T}} \left.\frac{\partial Q_T}{\partial \beta}\right|_{\beta_0} = \frac{2}{\sqrt{T}} \sum_{t=1}^{T} u_t \left.\frac{\partial m_t}{\partial \beta}\right|_{\beta_0}.$$

If S in (9.12) is a nonsingular matrix, then using the same arguments we used for OLS, we have,

$$\frac{1}{\sqrt{T}}\left.\frac{\partial Q_T}{\partial \beta}\right|_{\beta_0}=\frac{2}{\sqrt{T}}\sum_{t=1}^{T}u_t\left.\frac{\partial m_t}{\partial \beta}\right|_{\beta_0}\xrightarrow{D}\mathcal{N}\left(0,4\sigma_0^2S\right).$$

Next, we have

$$\begin{aligned}\frac{1}{T}\frac{\partial^2 Q_T}{\partial\beta\partial\beta'} &= \frac{2}{T}\sum_{t=1}^{T}u_t\frac{\partial^2 m_t}{\partial\beta\partial\beta'}+\frac{2}{T}\sum_{t=1}^{T}\left[m_t\left(\beta_0\right)-m_t\left(\beta\right)\right]\frac{\partial^2 m_t}{\partial\beta\partial\beta'}-\frac{2}{T}\sum_{t=1}^{T}\frac{\partial m_t}{\partial\beta}\frac{\partial m_t}{\partial\beta'}\\ &= A_1+A_2+A_3.\end{aligned}$$

It can be shown that A_1 and A_2 converge to 0 in probability uniformly. Thus

$$\frac{1}{T}\frac{\partial^2 Q_T}{\partial\beta\partial\beta'}\xrightarrow{P}-2S.$$

Then,

$$\sqrt{T}\left(\widehat{\beta}-\beta_0\right)\xrightarrow{D}\mathcal{N}\left(0,\sigma_0^2S^{-1}\right).$$

This results can also be derived by using the analogy with OLS from the transformed model that is used to obtain the NLLS estimator. Recall that

$$\sqrt{T}\left(\widehat{\beta}-\beta_0\right)=\left[T^{-1}Z\left(\beta_0\right)'Z\left(\beta_0\right)\right]^{-1}\left[T^{-1/2}Z\left(\beta_0\right)'u\right].$$

The first term converges to S^{-1} and the second to $\mathcal{N}\left(0,\sigma_0^2S\right)$. Then $\sqrt{T}\left(\widehat{\beta}-\beta_0\right)\xrightarrow{D}\mathcal{N}\left(0,\sigma_0^2S^{-1}\right)$ by Slutsky.

Example 100 *Consider the same model of Example 99 with the additional assumption that* $\lim_{T\to\infty}T^{-1}\sum_{t=1}^{T}x_t^4=r$. *Next, we obtain the asymptotic distribution of the value of* β *that minimizes the SSR. From equation (9.8) we know that*

$$\sqrt{T}\left(\widehat{\beta}-\beta_0\right)=-\left[\frac{1}{T}\left.\frac{\partial^2 Q_T}{\partial\beta^2}\right|_{\widetilde{\beta}}\right]^{-1}\frac{1}{\sqrt{T}}\left.\frac{\partial Q_T}{\partial\beta}\right|_{\beta_0},$$

where $\widetilde{\beta}$ *lies between* $\widehat{\beta}$ *and* β_0. *We have*

$$\frac{1}{\sqrt{T}}\left.\frac{\partial Q_T}{\partial\beta}\right|_{\beta_0}=\frac{4}{\sqrt{T}}\sum_{t=1}^{T}u_t\left(\beta_0+x_t\right).$$

It can be shown that

$$\frac{1}{\sqrt{T}} \left. \frac{\partial Q_T}{\partial \beta} \right|_{\beta_0} \xrightarrow{D} \mathcal{N}\left[0, 16\sigma_0^2\left(\beta_0^2 + p + 2\beta_0 q\right)\right]. \tag{9.13}$$

We also have

$$\frac{1}{T} \left. \frac{\partial^2 Q_T}{\partial \beta^2} \right|_{\widetilde{\beta}} = \frac{4}{T}\sum_{t=1}^{T}\left[y_t - \left(\widetilde{\beta} + x_t\right)^2\right] - \frac{8}{T}\sum_{t=1}^{T}\left(\widetilde{\beta} + x_t\right)^2.$$

Then

$$\frac{1}{T} \left. \frac{\partial^2 Q_T}{\partial \beta^2} \right|_{\widetilde{\beta}} \xrightarrow{p} -8\left(\beta_0^2 + p + 2\beta_0 q\right), \tag{9.14}$$

$$\sqrt{T}\left(\widehat{\beta} - \beta_0\right) \xrightarrow{D} \mathcal{N}\left[0, \frac{\sigma_0^2}{4\left(\beta_0^2 + p + 2\beta_0 q\right)}\right].$$

The large-sample behavior of NLLS therefore mirrors that of OLS: both are consistent and asymptotically normal with variance proportional to the inverse of the expected information matrix.

The difference lies in computation. Because $Q_T(\theta)$ is nonlinear and generally nonconcave, numerical optimization methods are required to obtain $\widehat{\theta}$. Algorithms such as Gauss–Newton or Levenberg–Marquardt are commonly used in practice. These procedures rely precisely on the differentiability assumptions invoked in the proofs of consistency and asymptotic normality.

By establishing that NLLS satisfies the same regularity conditions as the linear case, the extremum-estimation framework provides a unified foundation for both.

The next section applies the same logic to the maximum-likelihood estimator, for which the criterion function corresponds to the log-likelihood rather than a sum of squared residuals.

9.5 MLE as an Extremum Estimator

Maximum likelihood estimation can also be viewed as a special case of extremum estimation in which the criterion function corresponds to the log-likelihood. The general theory developed in Section 9.2 applies directly once we verify the conditions for consistency and asymptotic normality. This perspective emphasizes that the well-known asymptotic efficiency of MLE arises not from its algebraic form but from the regularity properties of the likelihood function as an extremum criterion.

The Maximum Likelihood Estimator can be represented as in extremum estimator form as

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} Q_T(Y, X, \theta),$$

where

$$Q_T(Y, X, \theta) = \ell(\theta; Y | X).$$

9.5.1 Consistency of MLE

The conditions for consistency of the global MLE or the local MLE can be obtained from Theorems 93, 94, or 95.[6]

In order to demonstrate the consistency of the MLE, we will use the following Lemma:

Lemma 101 (Jensen's Inequality) *If $h(x)$ is a concave function and x is a random variable, then*

$$h(\mathcal{E}[x]) \geq \mathcal{E}[h(x)].$$

If $h(\cdot)$ is strictly concave, then the inequality is strict unless x equals a constant with probability one.

Next, we restate Theorem 93 for the special case of MLE.

Theorem 102 (Consistency of MLE) *Suppose that:*

A) Θ is a compact subset of $\mathbb{R}^k$,

B) $\ell_T(Y, \theta)$ is continuous in $\theta \in \Theta$ for all Y and is a measurable function of Y for all $\theta \in \Theta$,

C) $T^{-1}\ell_T(\theta)$ converges uniformly in probability to a nonstochastic function $Q(\theta) = \mathcal{E} \ln L(\theta; y)$,

i) (Identification) $Q(\theta)$ attains a unique global maximum at θ_0,

ii) $Q(\theta)$ is continuous in $\theta \in \Theta$.

Then $\widehat{\theta} \xrightarrow{p} \theta_0$.

[6] The global MLE is the value of θ that globally maximizes the likelihood function over the parameter space, while the local MLE is any root of the FONC that corresponds to a local maximum.

Proof. Note that

$$\frac{1}{T}\ell_T(\theta) = \frac{1}{T}\sum_{t=1}^{T}\ln L(\theta; y_t).$$

From assumption C) we have

$$\frac{1}{T}\sum_{t=1}^{T}\ln L(\theta; y_t) \xrightarrow{p} \mathcal{E}\ln L(\theta; y) \quad \text{(uniformly)},$$

But

$$\mathcal{E}\ln L(\theta; y) = \int \ln L(\theta; y)\ f(y; \theta_0)\ dy = Q(\theta).$$

Then, for any θ we have

$$\begin{aligned} Q(\theta) - Q(\theta_0) &= \int \ln\left[\frac{L(\theta; y)}{f(y; \theta_0)}\right] f(y; \theta_0)\ dy \\ &\leq \ln \int L(\theta; y)\, dy \quad \text{(Jensen's inequality)} \\ &= 0. \end{aligned}$$

Thus, $Q(\theta) \leq Q(\theta_0)$ for all $\theta \in \Theta$. This implies that $\widehat{\theta} \xrightarrow{p} \theta_0$. ■

One can always specialize the conditions of Theorems 94 and 95 for the likelihood function case, and a variety of different types of regularity conditions are available to achieve consistency of the MLE in addition to those stated above. The possibilities are considered broad enough that the property of consistency is assumed to hold quite generally in practice for the MLE.

However, there are cases where the global MLE is inconsistent, whereas a root of the ML FONC can be consistent, as in the following example.

Example 103 *Let y_t, $t = 1, 2, \cdots, T$, be independent with the common distribution defined by*

$$f(y_t) = \begin{cases} \mathcal{N}(\mu_1, \sigma_1^2) & \textit{with probability } p \\ \mathcal{N}(\mu_2, \sigma_2^2) & \textit{with probability } 1-p \end{cases}.$$

This distribution is called a mixture of normal distributions. The likelihood function is given by

$$L = \prod_{t=1}^{T} \left[\frac{p}{\sqrt{2\pi\sigma_1^2}} \exp\left[-0.5\left(y_t-\mu_1\right)^2/\sigma_1^2\right] \right.$$
$$\left. + \frac{1-p}{\sqrt{2\pi\sigma_2^2}} \exp\left[-0.5\left(y_t-\mu_2\right)^2/\sigma_2^2\right] \right].$$

If we put $\mu_1 = y_1$ and let $\sigma_1 = 0$, the term of the product that corresponds to $t = 1$ goes to infinity, and, consequently, L goes to infinity. Hence, the global MLE cannot be consistent. Note that this example violates postulate i) of assumption C) of Theorem 102 because $Q(\theta)$ does not attain a global maximum at θ_0. However, the conditions of Theorem 95 are generally satisfied by this model.

Example 104 *Let $\{y_t\}$, $t = 1, 2, \cdots, T$, be i.i.d. with probability distribution*

$$f(y_t) = \begin{cases} 1 & \text{with probability } \pi_0 \\ 0 & \text{with probability } 1-\pi_0 \end{cases}.$$

This distribution is called a Bernoulli distribution. The log-likelihood function is given by

$$\ell(\pi; Y) = \sum_{t=1}^{T} \left[y_t \ln \pi + (1-y_t)\ln(1-\pi)\right].$$

In this case, it is trivial to verify that

$$\widehat{\pi} = \frac{\sum_{t=1}^{T} y_t}{T}$$

is the MLE estimator and that it is unbiased (given that $\mathcal{E}[y_t] = \pi_0$) and consistent by applying WLLN2 or SLLN2 directly, but we will prove consistency using Theorem 102 instead.

Given that $0 \leq \pi \leq 1$, we have that Θ is compact, furthermore, ℓ is differentiable, thus also continuous. This means that assumptions A) and B) *of Theorem 102 are satisfied. Next, we verify assumption C) of this theorem. It is trivial to verify that*

$$\frac{\ell(\pi; Y)}{T} = \frac{1}{T}\sum_{t=1}^{T}\left[y_t \ln \pi + (1-y_t)\ln(1-\pi)\right]$$

converges in probability uniformly to

$$Q(\pi) = \pi_0 \ln \pi + (1-\pi_0)\ln(1-\pi) = \mathcal{E}\ln L(\theta; y).$$

Finally, note that $\pi = \pi_0$ maximizes $Q(\pi)$; thus $\widehat{\pi} \xrightarrow{p} \pi_0$.

9.5.2 Asymptotic Normality of MLE

The asymptotic normality of the MLE estimator can be derived by using Theorem 97. Nevertheless, remember that when we derived the Cramer-Rao Lower bound, we stated that if the likelihood function was correctly specified the following condition was satisfied:

$$\mathcal{E}\left.\frac{\partial^2 \ell}{\partial\theta\partial\theta'}\right|_{\theta_0} = -\mathcal{E}\left.\left[\frac{\partial \ell}{\partial\theta}\frac{\partial \ell}{\partial\theta'}\right]\right|_{\theta_0}. \tag{9.15}$$

In the extremum estimator contexts, this means that $A(\theta_0) = -B(\theta_0)$ in which case we have

Theorem 105 *Under the assumptions of Theorem 97 and assumption (9.15), the MLE estimator satisfies*

$$\sqrt{T}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{D} \mathcal{N}\left(0, -\left[\lim_{T\to\infty}\mathcal{E}\frac{1}{T}\left.\frac{\partial^2 \ell}{\partial\theta\partial\theta'}\right|_{\theta_0}\right]^{-1}\right). \tag{9.16}$$

Proof. Left as an exercise. ■

Recall that condition (9.15) is satisfied only when the model is correctly specified. In the QML context, the asymptotic distribution of the QMLE estimator is not (9.16) and has to be derived from Theorem 97 directly.

In our discussion of the Cramer-Rao Lower bound, we considered the case of unbiased estimators. We can extend this concept for the case of consistent estimators.

Definition 106 *A consistent estimator is said to be asymptotically efficient if it satisfies that* $A(\theta_0) = -B(\theta_0)$.

Thus, the MLE under the appropriate assumptions is asymptotically efficient by definition. An asymptotically efficient estimator is also referred to as best asymptotically normal (BAN for short).

Example 107 *Consider the same model as in Example 104. Next we derive its asymptotic distribution.*

Once again, it would be trivial to demonstrate the asymptotic normality of $\widehat{\pi}$ *by invoking CLT1, but here we will use Theorem 105 instead.*

For that we will need to following expressions:

$$\begin{aligned}\frac{\partial \ell}{\partial \pi} &= \frac{\sum y_t}{\pi} - \frac{T - \sum y_t}{1-\pi} \\ \frac{\partial^2 \ell}{\partial \pi^2} &= -\frac{\sum y_t}{\pi^2} - \frac{T - \sum y_t}{(1-\pi)^2}.\end{aligned}$$

The last expression shows that assumption A) of Theorem 97 is satisfied. Next we verify assumption B).

$$\left.\frac{1}{T}\frac{\partial^2 \ell}{\partial \pi^2}\right|_{\theta_0} = -\frac{\sum y_t}{T\pi_0^2} - \frac{T - \sum y_t}{T(1-\pi_0)^2}.$$

Thus

$$A(\pi_0) = \lim \mathcal{E}\left.\frac{1}{T}\frac{\partial^2 \ell}{\partial \pi^2}\right|_{\theta_0} = -\frac{1}{\pi_0(1-\pi_0)}.$$

It is trivial to verify that $\frac{1}{T}\frac{\partial^2 \ell}{\partial \pi^2} \xrightarrow{p} A(\pi_0)$. *Thus assumption B) is satisfied.*

Next, lets focus on assumption C); for that we first derive $B(\pi_0)$:

$$\begin{aligned}\left.\frac{1}{T}\left[\frac{\partial \ell}{\partial \pi}\frac{\partial \ell}{\partial \pi'}\right]\right|_{\theta_0} &= \frac{1}{T}\sum_{t=1}^{T}\left[\frac{y_t}{\pi_0} - \frac{1-y_t}{1-\pi_0}\right]^2 \\ &= \frac{1}{T}\sum_{t=1}^{T}\left[\frac{y_t^2}{\pi_0^2} + \frac{(1-y_t)^2}{(1-\pi_0)^2} - \frac{2y_t(1-y_t)}{\pi_0(1-\pi_0)}\right] \\ &= \frac{1}{T}\sum_{t=1}^{T}\left[\frac{y_t}{\pi_0^2} + \frac{1-y_t}{(1-\pi_0)^2}\right] \quad \textit{because } y_t^2 = y_t \textit{ and } y_t(1-y_t) = 0.\end{aligned}$$

Thus

$$B(\pi_0) = \lim \mathcal{E} \frac{1}{T} \left[\frac{\partial \ell}{\partial \pi} \frac{\partial \ell}{\partial \pi'} \right] \Bigg|_{\pi_0} = \frac{1}{\pi_0 (1 - \pi_0)},$$

and $B(\pi_0) = -A(\pi_0)$.

Finally, we need to prove that $T^{-1/2} \left(\partial \ell_T / \partial \pi \right)_{\widetilde{\pi}} \xrightarrow{D} \mathcal{N}[0, B(\theta_0)]$, *with* $\widetilde{\pi} \xrightarrow{p} \pi_0$. *But*

$$\frac{1}{\sqrt{T}} \frac{\partial \ell}{\partial \pi}\Bigg|_{\widetilde{\pi}} = \frac{1}{\sqrt{T}} \frac{\sum (y_t - \widetilde{\pi})}{\widetilde{\pi}(1 - \widetilde{\pi})}.$$

As $\mathcal{E}(y) = \pi_0$ *and* $\mathcal{V}(y) = \pi_0 (1 - \pi_0)$

$$\frac{1}{\sqrt{T}} \frac{\sum (y_t - \widetilde{\pi})}{\sqrt{\widetilde{\pi}(1 - \widetilde{\pi})}} \xrightarrow{D} \mathcal{N}(0, 1),$$

then

$$\frac{1}{\sqrt{T}} \frac{\sum (y_t - \widetilde{\pi})}{\widetilde{\pi}(1 - \widetilde{\pi})} \xrightarrow{D} \mathcal{N}[0, B(\pi_0)].$$

Finally, we have

$$\sqrt{T}(\widehat{\pi} - \pi_0) \xrightarrow{D} \mathcal{N}[0, \pi_0 (1 - \pi_0)].$$

Then, an estimator of the variance of $\widehat{\pi}$ *is*

$$\widehat{\mathcal{V}}(\widehat{\pi}) = \frac{\widehat{\pi}(1 - \widehat{\pi})}{T}.$$

9.5.3 Asymptotic Normality of QMLE

When the likelihood function is misspecified, the estimator obtained by maximizing the same criterion is the quasi–maximum likelihood estimator (QMLE). The general asymptotic theory for extremum estimators still applies, but the information-matrix equality no longer holds.

In that case:

$$\sqrt{T}\left(\widehat{\theta}_{QMLE} - \theta_0\right) \xrightarrow{D} \mathcal{N}\left(0, H_0^{-1} O_0 H_0^{-1}\right)$$

where

$$O_0 = \mathcal{E}\left[\frac{\partial \ell}{\partial \theta} \frac{\partial \ell}{\partial \theta'} \Bigg|_{\theta_0} \right] = \mathcal{E}\left[g_0 g_0' \right], \quad H_0 = \mathcal{E}\left[\frac{\partial^2 \ell}{\partial \theta \partial \theta'} \Bigg|_{\theta_0} \right].$$

The MLE and its QMLE version illustrate the full power of the extremum framework. Consistency follows from uniform convergence and identification, while asymptotic normality arises from differentiability and the central limit theorem. When the model is correctly specified, efficiency results from the information-matrix equality; when it is not, the same logic yields robust covariance formulas. The extremum framework therefore unifies least squares, nonlinear regression, and likelihood methods within a single asymptotic theory.

The next section extends the discussion to estimators whose criterion functions are not differentiable, such as the least absolute deviations estimator. These cases demonstrate how the extremum framework accommodates non-smooth optimization problems and motivates further generalizations to quantile and robust estimation.

9.6 LAD as a Extremum Estimator

The least absolute deviations estimator has certain robustness properties that makes it interesting.[7] In linear models, the least absolute deviations estimator is known to be asymptotically more efficient than least squares for thick-tailed distributions. Besides its practical importance, the LAD estimator poses an interesting theoretical problem because the general results of Section 9.2 can be used to prove consistency of the LAD estimator but not its asymptotic normality, even though the LAD estimator is an extremum estimator. Here, we shall sketch proofs of the asymptotic normality of the median, which is the LAD estimator in the i.i.d. sample case, and of the LAD estimator in a regression model. Finally, we generalize the concep to discuss quantile regressions.

9.6.1 The Sample Median

Suppose that we have a random sample $\{y_t\}$ of T observations on the random variable Y with common distribution function $\mathcal{F}$ and density function f. If we reorder the observations so that

$$y^1 \leq y^2 \leq \cdots \leq y^T,$$

[7] The LAD estimator is also known as the minimum absolute deviation (MAD) estimator, or the least absolute error (LAE) estimator.

then the sample median is y^r where $r = (T + 1)/2$. This is the solution to the LAD problem

$$\min_{\beta} \sum_{t=1}^{T} |y_t - \beta| .$$

To see this, note that for all $\beta \neq y_t$, $t = 1, 2, \cdots, T$, the "derivative" of the objective function,

$$\frac{\partial}{\partial \beta} \sum_{t=1}^{T} |y_t - \beta| = \sum_{t=1}^{T} \left(I \left\{ y_t - \beta < 0 \right\} - I \left\{ y_t - \beta > 0 \right\} \right)$$

is the number of observations below β minus the number of observations above β. At the sample median, increases and decreases in β increase the sum of absolute deviations (residuals). Furthermore, the sample median is the only value of β with this property so that it is also the unique LAD solution.

A comparison with OLS is instructive because it shows how OLS is relatively more sensitive to the largest and smallest observations in the sample. Consider first what would happen to the sample median and mean if any observation above the median had been larger. The sample median would be the same, whereas the sample mean would increase. Indeed, as we artificially increase such an observation more and more the sample mean increases proportionately while the sample median remains constant. On the other hand, decreasing an observation strictly above the sample median will decrease the sample mean. The sample median remains unaffected by such decreases until the observation falls below the median observation. At that point, the observation that we are varying becomes the median observation and further decreases lower the median one for one until a third observation becomes the median. Thus the median has a bounded response to changes in one observation. Inefficiency in the sample mean comes in part from its excessive sensitivity to the outlying observations that are more common in fat-tailed distributions than in the normal.

The population median is defined by

$$\mathcal{F}(\beta_0) = \frac{1}{2},$$

and the binary random variable $J_t(\beta)$ by

$$J_t(\beta) = \begin{cases} 1 \text{ if } & Y_t \geq \beta \\ 0 \text{ if } & Y_t < \beta \end{cases} \tag{9.17}$$

for every real number β. Using (9.17), we define the sample median $\widehat{\beta}_{LAD}$ by

$$\widehat{\beta}_{LAD} = \inf \left\{ \beta \left| \sum_{t=1}^{T} J_t(\beta) \leq \frac{T}{2} \right. \right\}, \tag{9.18}$$

The median as defined above is clearly unique.[8]

The asymptotic normality of $\widehat{\beta}_{LAD}$ can be proved in the following manner: Using (9.18), we have for any y

$$\Pr\left[\widehat{\beta}_{LAD} < \beta_0 + T^{-1/2}y\right] = \Pr\left[\sum_{t=1}^{T} J_t\left(\beta_0 + T^{-1/2}y\right) \leq \frac{T}{2}\right].$$

Define

$$P_t = 1 - \Pr\left[y_t < \beta_0 + T^{-1/2}y\right].$$

Then because of a Taylor expansion

$$P_t \simeq \frac{1}{2} - T^{-1/2} f(\beta_0) y,$$

we have

$$\Pr\left[\sum_{t=1}^{T} J_t\left(\beta_0 + T^{-1/2}y\right) \leq \frac{T}{2}\right] = \Pr\left[T^{-1/2}\sum_{t=1}^{T}\left(J_t\left(\beta_0 + T^{-1/2}y\right) - P_t\right) \leq f(\beta_0) y\right].$$

It can be proved that $T^{-1/2}\sum_{t=1}^{T}\left(J_t\left(\beta_0 + T^{-1/2}y\right) - P_t\right) \xrightarrow{D} \mathcal{N}\left(0, \frac{1}{4}\right)$ given that J_t is Bernoulli. Thus, it can be shown that (Amemiya, 1985):

$$\sqrt{T}\left(\widehat{\beta}_{LAD} - \beta_0\right) \xrightarrow{D} \mathcal{N}\left(0, \frac{1}{4} f(\beta_0)^{-2}\right). \tag{9.19}$$

[8] If the sample is (1,2,8), 2 is the unique median. If the sample is (1,2,3,8), any point in the closed interval [2,3] may be defined as a median. The definition (9.18) picks 2 as the median. This ambiguity vanishes as the sample size approaches infinity.

The consistency of $\widehat{\beta}_{LAD}$ follows from (9.19). However, it also can be proved by direct application of Theorem 93.

Consistency follows from arguments analogous to those used for NLLS, replacing differentiability with convexity. The population criterion function is $Q(\theta) = -E\left[|y-\theta|\right]$ which attains a unique maximum at the median of y. Since the absolute value function is convex and continuous, the sample criterion $Q_T(\theta)$ is continuous and measurable.

9.6.2 LAD Linear Regression

Consider the LRM

$$Y = X\beta_0 + u,$$

where X is a $T \times k$ matrix of bounded constants such that $\lim_{T\to\infty} T^{-1}X'X = M$ is a finite positive definite matrix and u is a $T-$vector of i.i.d. random variables with continuous density function $f(\cdot)$ such that $\int_0^\infty f(\lambda)\, d\lambda = \frac{1}{2}$ and $f(z) > 0$ for all z in a neighborhood of 0. It is assumed that the parameter space B is compact. The LAD estimator $\widehat{\beta}_{LAD}$ is defined to be a value of β that minimizes

$$S_T(\beta) = \sum_{t=1}^{T} |y_t - \beta' x_t|. \tag{9.20}$$

In contrast to the OLS estimator, the LAD estimator is selected by minimizing the sum of absolute (rather than squared) residuals. As already mentioned, this estimator is robust to outliers among the observations.

The LAD objective function (9.20) may be written as

$$\begin{aligned} S_T(\beta) &= \sum_{t=1}^{T} |y_t - \beta' x_t| \\ &= \sum_{t=1}^{T} (y_t - \beta' x_t)\operatorname{sgn}(y_t - \beta' x_t), \end{aligned}$$

where $\operatorname{sgn}(z) = -1$ if $z < 0$ and $\operatorname{sgn}(z) = 1$ if $z > 0$. By further noting that $\operatorname{sgn}(z) = z/\operatorname{abs}(z)$, we can rewrite (9.20) as a weighted least squares objective function

$$S_T(\beta) = \sum_{t=1}^{T} \frac{(y_t - \beta' x_t)^2}{|y_t - \beta' x_t|}. \tag{9.21}$$

The weights for each observation are simply the reciprocals of the absolute deviations. If $D(\beta)$ is a diagonal matrix with elements $D_{t,t}(\beta) = |y_t - \beta' x_t|^{-1}$, the weighted least-squares objective function can be stated in matrix form as

$$S_T(\beta) = (Y - X\beta)' D(\beta)(Y - X\beta).$$

Although a direct solution to the problem is difficult to achieve because β appears in the numerator and denominator of (9.21), we can solve the problem by a procedure known as iterative weighted least squares. For some starting value β_1 (perhaps the OLS estimate of β), we form a $T \times k$ matrix $W = D(\beta_1) X$ and compute the subsequent estimate as $\beta_2 = (W'X)^{-1} W'Y$. The resulting estimate is used to compute new weights $D(\beta_2)$, and the process is repeated until the sequence of estimates converge.

The asymptotic normality of the LAD estimator $\widehat{\beta}_{LAD}$ is based on the following three fundamental results:

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} x_t \operatorname{sgn}\left(y_t - \widehat{\beta}' x_t\right) \overset{a.s.}{\to} 0$$

$$\begin{aligned}&\frac{1}{\sqrt{T}} \sum_{t=1}^{T} x_t \operatorname{sgn}\left(y_t - \widehat{\beta}' x_t\right) - \frac{1}{\sqrt{T}} \sum_{t=1}^{T} x_t \operatorname{sgn}(u_t) \\ &- \left\{ \frac{1}{\sqrt{T}} \sum_{t=1}^{T} x_t \mathcal{E}\left[\operatorname{sgn}\left(y_t - \widehat{\beta}' x_t\right)\right] - \frac{1}{\sqrt{T}} \sum_{t=1}^{T} x_t \mathcal{E}\left[\operatorname{sgn}(y_t - \beta_0' x_t)\right] \right\} \overset{p}{\to} 0\end{aligned}$$

and

$$\begin{aligned}\frac{1}{\sqrt{T}} \sum_{t=1}^{T} x_t \mathcal{E}\left[\operatorname{sgn}\left(y_t - \widehat{\beta}' x_t\right)\right] &\overset{LD}{=} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} x_t \mathcal{E}\left[\operatorname{sgn}(y_t - \beta_0' x_t)\right] \\ &+ \operatorname{plim} \frac{1}{T} \sum_{t=1}^{T} x_t \left[\frac{\partial}{\partial \beta'} \mathcal{E}\left[\operatorname{sgn}(y_t - \beta' x_t)\right]\right]_{\beta_0} \sqrt{T}\left(\widehat{\beta}_{LAD} - \beta_0\right).\end{aligned}$$

These results imply

$$\sqrt{T}\left(\widehat{\beta}_{LAD} - \beta_0\right) \overset{LD}{=} \left\{ \operatorname{plim} \frac{1}{T} \sum_{t=1}^{T} x_t \left[\frac{\partial}{\partial \beta'} \mathcal{E}\left[\operatorname{sgn}(y_t - \beta' x_t)\right]\right]_{\beta_0} \right\}^{-1} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} x_t \operatorname{sgn}(u_t).$$

Noting that $\mathcal{E}\left[\operatorname{sgn}\left(y_t - \beta' x_t\right)\right] = 1 - 2\mathcal{F}\left[x_t'\left(\beta - \beta_0\right)\right]$, $\mathcal{E}\left[\operatorname{sgn}\left(u_t\right)\right] = 0$, and $\mathcal{V}\left[\operatorname{sgn}\left(u_t\right)\right] = 1$, we obtain

$$\sqrt{T}\left(\widehat{\beta}_{LAD} - \beta_0\right) \xrightarrow{D} \mathcal{N}\left(0, \frac{1}{4} f\left(0\right)^{-2} S^{-1}\right).$$

The asymptotic variance depends inversely on the squared density of the error at the median, reflecting the fact that a steeper density (a "sharper" median) implies more precise estimation.

Although the LAD estimator is less efficient than OLS under normal errors, it is more robust to outliers and remains consistent under much weaker conditions.

The asymptotic distribution of LAD resembles that of OLS, except that the error variance is replaced by a term involving the conditional density of the disturbance at its median.

Because it depends only on the sign of the residuals, the LAD estimator is less sensitive to large outliers and is particularly suitable for heavy-tailed data.

9.6.3 Quantile Regression

The LAD estimator is a particular case of a broader class of estimators known as quantile regressions (Koencker and Hallock, 2001). These estimators minimize a weighted sum of absolute residuals, where the weights depend on the desired quantile level $\tau \in (0, 1)$:

$$\widehat{\beta}_\tau = \arg\min \sum_{t=1}^{T} \rho_\tau\left(y_t - \beta' x_t\right),$$

with check function, often referred to as the tilted absolute value function, which is used to define the quantile of the conditional distribution of y, and it places different weights on residuals above and below the quantile being estimated:

$$\rho_\tau\left(u\right) = u\left(\tau - I\left(u < 0\right)\right) u,$$

For $\tau = 0.5$, the criterion reduces to the absolute value, and $\widehat{\beta}_{0.5}$ coincides with the LAD estimator ($\widehat{\beta}_{LAD}$).

The population objective function,

$$Q_\tau\left(\beta\right) = \mathcal{E}\left[\rho_\tau\left(y_t - \beta' x_t\right)\right]$$

is minimized at the conditional τ-th quantile of y_t given x_t.

Under suitable regularity conditions, the estimator satisfies

$$\sqrt{T}\left(\widehat{\beta}_{\tau}-\beta_0\right)\xrightarrow{D}\mathcal{N}\left(0,\tau\left(1-\tau\right)f\left(0\right)^{-2}S^{-1}\right),$$

where $f\left(0\right)$ denotes the conditional density of the error evaluated at zero.

The asymptotic variance depends on the quantile level τ, with the factor $\tau\left(1-\tau\right)$ reflecting the proportion of observations on either side of the fitted quantile.

Quantile regression extends the median regression framework to model the entire conditional distribution of the dependent variable. By estimating multiple quantiles, one can characterize heterogeneity in the effects of regressors across different parts of the distribution.

Because each quantile regression minimizes a convex but nonsmooth criterion, all results developed here for LAD apply directly.

This connection shows how the extremum framework naturally encompasses both smooth and nonsmooth estimators within a unified asymptotic theory.

9.7 Asymptotic Tests

If the confrontation of economic theories with observable phenomena is the objective of empirical research, then hypothesis testing is the primary tool of analysis. To receive empirical verification, all theories must eventually be reduced to a testable hypothesis. In the past several decades, least squares based tests have functioned admirably for this purpose. More recently, the use of increasingly complex statistical models has led to heavy reliance on maximum likelihood methods for both estimation and testing. In such a setting only asymptotic properties can be expected for estimators or tests. Often there are asymptotically equivalent procedures which differ substantially in computational difficulty and finite sample performance. Econometricians have responded enthusiastically to this research challenge by devising a wide variety of tests for these complex models.

Most of the tests used are based either on the Wald, Likelihood Ratio (LRT) or Lagrange Multiplier (LM) principle. These three general principles have a certain symmetry which has revolutionized the teaching of hypothesis tests and the development of new procedures. Essentially, the Lagrange

Multiplier approach starts at the null and asks whether movement toward the alternative would be an improvement, while the Wald approach starts at the alternative and considers movement toward the null. The Likelihood ratio method compares the two hypotheses directly on an equal basis. Next, we provide a unified development of the three principles beginning with the likelihood functions. The properties of the tests and the relations between them are developed and their forms in a variety of common testing situations are explained. Finally we will present techniques that are used to test nonnested hypothesis.

9.7.1 A General Formulation of the Tests

Let $f\left(Y;\theta_0\right)$ be the joint density of a $T-$vector of random variables characterized by a $k-$vector of parameters θ_0. We assume all the conditions used to prove the asymptotic normality of the MLE $\widehat{\theta}$. Here we shall discuss the asymptotic tests of the hypothesis

$$\mathrm{H}_0 : h\left(\theta\right) = 0, \tag{9.22}$$

where $h\left(\cdot\right)$ is a $q-$vector valued differentiable function with $q < k$. We assume that (9.22) can be equivalently written as

$$\theta = r\left(\alpha\right),$$

where α is a $p-$vector of parameters such that $p = k - q$. We denote the constrained maximum likelihood estimator subject to (9.22) as $\overline{\theta}$.

We define the Wald test, which is the asymptotic approximation to the very familiar t and F tests, as

$$\text{Wald} = -h\left(\widehat{\theta}\right)' \left\{ \left.\frac{\partial h}{\partial \theta'}\right|_{\widehat{\theta}} \left[\left.\frac{\partial^2 \ell}{\partial \theta \partial \theta'}\right|_{\widehat{\theta}} \right]^{-1} \left.\frac{\partial h}{\partial \theta}\right|_{\widehat{\theta}} \right\}^{-1} h\left(\widehat{\theta}\right). \tag{9.23}$$

The Lagrange Multiplier test (also known as Rao's score test) is derived from a constrained maximization principle. Maximizing the log-likelihood subject to the constraint $h\left(\theta\right) = 0$ yields a set of Lagrange multipliers which measure the shadow price of the constraint. If the price is high, the constraint should be rejected as inconsistent with the data. We define this test as

$$\text{LM} = -\left.\frac{\partial \ell}{\partial \theta'}\right|_{\overline{\theta}} \left[\left.\frac{\partial^2 \ell}{\partial \theta \partial \theta'}\right|_{\overline{\theta}} \right]^{-1} \left.\frac{\partial \ell}{\partial \theta}\right|_{\overline{\theta}}, \tag{9.24}$$

where $\overline{\theta}$ denotes the MLE of the constrained model.

Finally, the Likelihood Ratio test is based upon the difference between the maximum of the likelihood under the null and under the alternative hypothesis and is defined as

$$\mathrm{LRT} = 2\left[\ell\left(\widehat{\theta}\right) - \ell\left(\overline{\theta}\right)\right]. \tag{9.25}$$

All three test statistics have the same limit distribution, χ^2_q, under the null hypothesis. Thus, the null hypothesis is rejected when the value of the test statistic is large.

Theorem 108 (Asymptotic distribution of the Wald test) *Let* $\sqrt{T}\left(\widehat{\theta} - \theta_0\right) \overset{D}{\to} \mathcal{N}\left(0, -\left[\lim_{T\to\infty} \mathcal{E}\frac{1}{T}\left.\frac{\partial^2 \ell}{\partial\theta\partial\theta'}\right|_{\theta_0}\right]^{-1} = -A\left(\theta_0\right)^{-1}\right)$. *Then, under* H_0, *Wald*$\overset{D}{\to}$ χ^2_q.

Proof. Recall that if $z \backsim \mathcal{N}(0, D)$ then $z'D^{-1}z \backsim \chi^2$. Under the null hypothesis we have that $h(\theta_0) = 0$; thus $h\left(\widehat{\theta}\right) - h(\theta_0) = h\left(\widehat{\theta}\right)$. Taking a first order Taylor series expansion we have

$$h\left(\widehat{\theta}\right) \simeq \left.\frac{\partial h}{\partial \theta'}\right|_{\theta_0}\left(\widehat{\theta} - \theta_0\right).$$

Thus

$$\sqrt{T}h\left(\widehat{\theta}\right) \overset{LD}{=} \sqrt{T}\left.\frac{\partial h}{\partial \theta'}\right|_{\theta_0}\left(\widehat{\theta} - \theta_0\right).$$

Then

$$\sqrt{T}h\left(\widehat{\theta}\right) \overset{D}{\to} \mathcal{N}\left(0, \left.\frac{\partial h}{\partial \theta'}\right|_{\theta_0}\left[-\lim_{T\to\infty}\mathcal{E}\frac{1}{T}\left.\frac{\partial^2\ell}{\partial\theta\partial\theta'}\right|_{\theta_0}\right]^{-1}\left.\frac{\partial h}{\partial\theta}\right|_{\theta_0}\right).$$

Finally, we have

$$\sqrt{T}h\left(\widehat{\theta}\right)'\left\{\left.\frac{\partial h}{\partial \theta'}\right|_{\theta_0}\left[-\lim_{T\to\infty}\mathcal{E}\frac{1}{T}\left.\frac{\partial^2\ell}{\partial\theta\partial\theta'}\right|_{\theta_0}\right]^{-1}\left.\frac{\partial h}{\partial\theta}\right|_{\theta_0}\right\}^{-1}\sqrt{T}h\left(\widehat{\theta}\right) \overset{D}{\to} \chi^2_q$$

or

$$h\left(\widehat{\theta}\right)'\left\{\left.\frac{\partial h}{\partial\theta'}\right|_{\theta_0}\left[-\lim_{T\to\infty}\mathcal{E}\left.\frac{\partial^2\ell}{\partial\theta\partial\theta'}\right|_{\theta_0}\right]^{-1}\left.\frac{\partial h}{\partial\theta}\right|_{\theta_0}\right\}^{-1}h\left(\widehat{\theta}\right)\xrightarrow{D}\chi_q^2.$$

The desired result follows from replacing θ_0 with $\widehat{\theta}$. ■

Theorem 109 (Asymptotic distribution of the LM test) *Assume that the conditions ensuring consistency and asymptotic normality of the MLE $\widehat{\theta}$ are satisfied. Let $\overline{\theta}$ and $\overline{\lambda}$ be the constrained MLE and the value of the Lagrange multiplier that satisfy*

$$\max_{\theta,\lambda}\ell\left(\theta;Y\right)-\lambda'h\left(\theta\right).$$

Then, under H_0 it follows that $LM\xrightarrow{D}\chi_q^2$

Proof. Expanding both $\partial\ell\left(\overline{\theta}\right)/\partial\theta$ and $h\left(\overline{\theta}\right)$ in a first-order Taylor series around the true θ_0 allows the FONC of the ML problem to be written as

$$\begin{aligned}\left.\frac{\partial\ell}{\partial\theta}\right|_{\theta_0}+\left.\frac{\partial^2\ell}{\partial\theta\partial\theta'}\right|_{\theta_0}\left(\overline{\theta}-\theta_0\right)-\left.\frac{\partial h}{\partial\theta}\right|_{\overline{\theta}}\overline{\lambda} &= 0\\ \left.\frac{\partial h}{\partial\theta'}\right|_{\theta_0}\left(\overline{\theta}-\theta_0\right) &= 0.\end{aligned}$$

Note that the second equation incorporates the fact that the first term in the Taylor series, $h\left(\theta_0\right)$, is zero under H_0, which we now assume is true. Premultiplying the first of the preceding equations by $T^{-1/2}$ and the second by $T^{1/2}$ leads to the matrix equation

$$\begin{bmatrix}-T^{-1}\left.\frac{\partial^2\ell}{\partial\theta\partial\theta'}\right|_{\theta_0} & \left.\frac{\partial h}{\partial\theta}\right|_{\overline{\theta}}\\ \left.\frac{\partial h}{\partial\theta'}\right|_{\theta_0} & 0\end{bmatrix}\begin{bmatrix}T^{1/2}\left(\overline{\theta}-\theta_0\right)\\ T^{-1/2}\overline{\lambda}\end{bmatrix}=\begin{bmatrix}T^{-1/2}\left.\frac{\partial\ell}{\partial\theta}\right|_{\theta_0}\\ 0\end{bmatrix}. \tag{9.26}$$

Observe that the right-hand side of the matrix equation converges in distribution to

$$\begin{bmatrix}T^{-1/2}\left.\frac{\partial\ell}{\partial\theta}\right|_{\theta_0}\\ 0\end{bmatrix}\xrightarrow{D}\begin{bmatrix}\mathcal{N}\left(0,-A\left(\theta_0\right)\right)\\ 0\end{bmatrix}.$$

In examining the asymptotic behavior of $T^{1/2}\left(\overline{\theta}-\theta_0\right)$ and $T^{-1/2}\overline{\lambda}$, we can write the first matrix of (9.26) as

$$\begin{bmatrix} -A\left(\theta_0\right) & \left.\frac{\partial h}{\partial \theta}\right|_{\overline{\theta}} \\ \left.\frac{\partial h}{\partial \theta'}\right|_{\theta_0} & 0 \end{bmatrix}.$$

Using partitioned inversion,[9] $T^{-1/2}\overline{\lambda}$ can be solved for, yielding

$$T^{-1/2}\overline{\lambda} = \left\{ \left.\frac{\partial h}{\partial \theta'}\right|_{\theta_0} \left[-A\left(\theta_0\right)\right]^{-1} \left.\frac{\partial h}{\partial \theta}\right|_{\overline{\theta}} \right\}^{-1} \left.\frac{\partial h}{\partial \theta'}\right|_{\theta_0} \left[-A\left(\theta_0\right)\right]^{-1} \left[T^{-1/2} \left.\frac{\partial \ell}{\partial \theta}\right|_{\theta_0} \right].$$

It follows by Slutsky's Theorem that

$$T^{-1/2}\overline{\lambda} \xrightarrow{D} \mathcal{N}\left(0, \left\{ \left.\frac{\partial h}{\partial \theta'}\right|_{\theta_0} \left[-A\left(\theta_0\right)\right]^{-1} \left.\frac{\partial h}{\partial \theta}\right|_{\overline{\theta}} \right\}^{-1}\right)$$

$$\left\{ \left.\frac{\partial h}{\partial \theta'}\right|_{\theta_0} \left[-A\left(\theta_0\right)\right]^{-1} \left.\frac{\partial h}{\partial \theta}\right|_{\overline{\theta}} \right\}^{1/2} T^{-1/2}\overline{\lambda} \xrightarrow{D} \mathcal{N}\left(0, I\right).$$

Thus

$$-\overline{\lambda}' \left.\frac{\partial h}{\partial \theta'}\right|_{\theta_0} \left[\left.\frac{\partial^2 \ell}{\partial \theta \partial \theta'}\right|_{\theta_0} \right]^{-1} \left.\frac{\partial h}{\partial \theta}\right|_{\overline{\theta}} \overline{\lambda} \xrightarrow{D} \chi^2_q.$$

But $\overline{\theta}$ and $\overline{\lambda}$ satisfy

$$\left.\frac{\partial \ell}{\partial \theta}\right|_{\theta_0} = \left.\frac{\partial h}{\partial \theta}\right|_{\overline{\theta}} \overline{\lambda}$$

Then

$$-\left.\frac{\partial \ell}{\partial \theta'}\right|_{\theta_0} \left[\left.\frac{\partial^2 \ell}{\partial \theta \partial \theta'}\right|_{\theta_0} \right]^{-1} \left.\frac{\partial \ell}{\partial \theta}\right|_{\theta_0} \xrightarrow{D} \chi^2_q.$$

The final result follows from replacing θ_0 with $\overline{\theta}$ which is consistent under H_0. ■

[9] The following result in the case of a symmetric matrix can be proved by direct multiplication:

$$\begin{bmatrix} A & C \\ C' & B \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} + A^{-1}C\left(B - C'A^{-1}C\right)^{-1}C'A^{-1} & -A^{-1}C\left(B - C'A^{-1}C\right)^{-1} \\ -\left(B - C'A^{-1}C\right)^{-1}C'A^{-1} & \left(B - C'A^{-1}C\right)^{-1} \end{bmatrix}.$$

Theorem 110 (Asymptotic distribution of the LRT test) *Assume that the conditions ensuring consistency and asymptotic normality of the MLE* $\widehat{\theta}$ *are satisfied. Let* $\overline{\theta}$ *be the constrained MLE. Then, under* H_0 *it follows that* $LRT\xrightarrow{D}\chi_q^2$

Proof. By a Taylor expansion we have

$$\ell\left(\theta_0\right)=\ell\left(\widehat{\theta}\right)+\left.\frac{\partial\ell}{\partial\theta'}\right|_{\widehat{\theta}}\left(\theta_0-\widehat{\theta}\right)+\frac{1}{2}\left(\theta_0-\widehat{\theta}\right)'\left.\frac{\partial^2\ell}{\partial\theta\partial\theta'}\right|_{\widehat{\theta}}\left(\theta_0-\widehat{\theta}\right).$$

Noting that the second term of the right hand side is 0 by the definition of $\widehat{\theta}$, we have

$$\ell\left(\widehat{\theta}\right)-\ell\left(\theta_0\right)\overset{LD}{=}\frac{T}{2}\left(\widehat{\theta}-\theta_0\right)'\left[-A\left(\theta_0\right)\right]\left(\widehat{\theta}-\theta_0\right).$$

Treating $\ell\left[r\left(\alpha\right)\right]=\ell\left(\alpha\right)$ as a function of α, we similarly obtain

$$\ell\left(\widehat{\alpha}\right)-\ell\left(\alpha_0\right)\overset{LD}{=}\frac{T}{2}\left(\widehat{\alpha}-\alpha_0\right)'\left[-A\left(\alpha_0\right)\right]\left(\widehat{\alpha}-\alpha_0\right).$$

Then

$$\text{LRT}=T\left(\widehat{\theta}-\theta_0\right)'\left[-A\left(\theta_0\right)\right]\left(\widehat{\theta}-\theta_0\right)-T\left(\widehat{\alpha}-\alpha_0\right)'\left[-A\left(\alpha_0\right)\right]\left(\widehat{\alpha}-\alpha_0\right).$$

But we known that

$$\sqrt{T}\left(\widehat{\theta}-\theta_0\right)\overset{LD}{=}\left[-A\left(\theta_0\right)\right]^{-1}\frac{1}{\sqrt{T}}\left.\frac{\partial\ell}{\partial\theta}\right|_{\theta_0}$$

$$\sqrt{T}\left(\widehat{\alpha}-\alpha_0\right)\overset{LD}{=}\left[-A\left(\alpha_0\right)\right]^{-1}\frac{1}{\sqrt{T}}\left.\frac{\partial\ell}{\partial\alpha}\right|_{\alpha_0}.$$

Since

$$\frac{1}{\sqrt{T}}\left.\frac{\partial\ell}{\partial\alpha}\right|_{\alpha_0}=\left.\frac{\partial r}{\partial\alpha}\right|_{\alpha_0}\frac{1}{\sqrt{T}}\left.\frac{\partial\ell}{\partial\theta}\right|_{\theta_0}$$

and

$$\frac{1}{\sqrt{T}}\left.\frac{\partial\ell}{\partial\theta}\right|_{\theta_0}\xrightarrow{D}\mathcal{N}\left(0,-A\left(\theta_0\right)\right).$$

Then

$$\text{LRT}=w'\left[\left(-A\left(\theta_0\right)\right)^{-1}-\left.\frac{\partial r}{\partial\alpha'}\right|_{\alpha_0}\left(-A\left(\alpha_0\right)\right)^{-1}\left.\frac{\partial r}{\partial\alpha}\right|_{\alpha_0}\right]w,$$

where $w \backsim \mathcal{N}(0, -A(\theta_0))$. Finally, defining

$$v = (-A(\theta_0))^{-1/2} w \backsim \mathcal{N}(0, I),$$

we obtain

$$\text{LRT} = v' \left[I - (-A(\theta_0))^{1/2} \left.\frac{\partial r}{\partial \alpha'}\right|_{\alpha_0} (-A(\alpha_0))^{-1} \left.\frac{\partial r}{\partial \alpha}\right|_{\alpha_0} (-A(\theta_0))^{1/2} \right] v.$$

But, because

$$-A(\alpha_0) = -\left.\frac{\partial r}{\partial \alpha}\right|_{\alpha_0} A(\theta_0) \left.\frac{\partial r}{\partial \alpha'}\right|_{\alpha_0},$$

it can be shown that $I - (-A(\theta_0))^{1/2} \left.\frac{\partial r}{\partial \alpha'}\right|_{\alpha_0} (-A(\alpha_0))^{-1} \left.\frac{\partial r}{\partial \alpha}\right|_{\alpha_0} (-A(\theta_0))^{1/2}$ is an idempotent matrix of rank q. Therefore, $\text{LRT} \xrightarrow{D} \chi^2_q$ ■

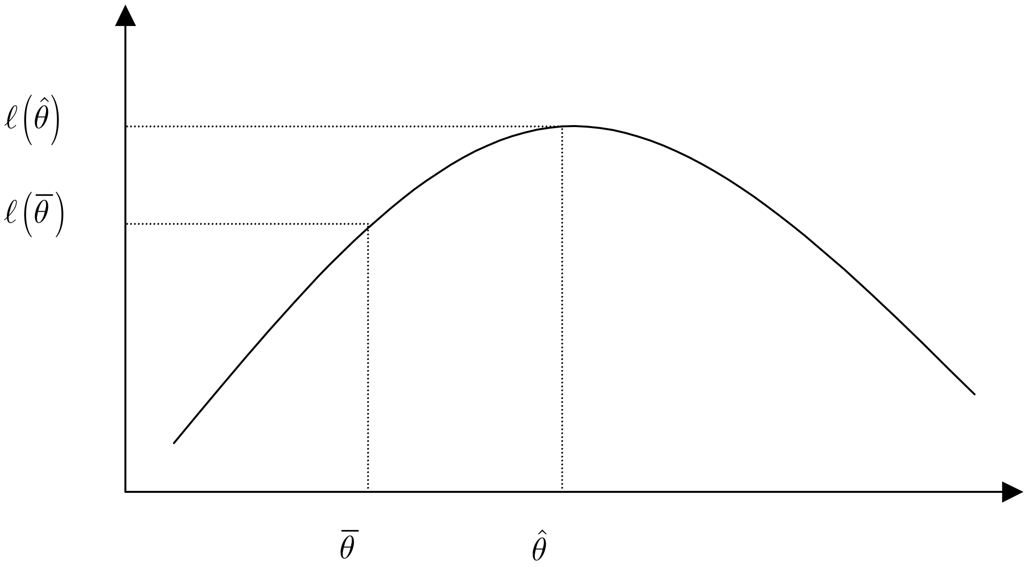

Figure 9.1: Geometric interpretation of the tests

The three principles are based on different statistics which measure the distance between H_0 (the null) and H_1 (the alternative). Figure 9.1 plots the log-likelihood function against θ for $q = 1$. The MLE under the alternative is $\widehat{\theta}$ and the hypothesized value is $\overline{\theta}$. The Wald test is based upon the horizontal difference between $\overline{\theta}$ and $\widehat{\theta}$, the LRT is based upon the vertical difference, and the LM test is based on the slope of the likelihood function at $\overline{\theta}$. Each is a reasonable measure of the distance between H_0 and H_1 and it is not surprising that when ℓ is a smooth curve well approximated by a quadratic, they all give the same test. This is established in the following Lemma.

Lemma 111 *If* $\ell = b - 1/2\left(\theta - \widehat{\theta}\right)' A \left(\theta - \widehat{\theta}\right)$ *where A is a symmetric positive definite matrix which may depend upon the data and upon known parameters, b is a scalar, $\widehat{\theta}$ is a function of the data, and $H_0 : \theta_0 = \overline{\theta}$. Then the Wald, LR, and LM tests are identical.*

Proof. Note that

$$\begin{aligned} \frac{\partial \ell}{\partial \theta} &= -\left(\theta - \widehat{\theta}\right)' A \\ \frac{\partial^2 \ell}{\partial \theta \partial \theta'} &= -A. \end{aligned}$$

Thus

$$\begin{aligned} \text{Wald} &= \left(\overline{\theta} - \widehat{\theta}\right)' A \left(\overline{\theta} - \widehat{\theta}\right) \\ \text{LM} &= \left.\frac{\partial \ell}{\partial \theta'}\right|_{\overline{\theta}} A^{-1} \left.\frac{\partial \ell}{\partial \theta}\right|_{\overline{\theta}} = \left(\overline{\theta} - \widehat{\theta}\right)' A \left(\overline{\theta} - \widehat{\theta}\right). \end{aligned}$$

Finally, by direct substitution we have: LRT= $\left(\overline{\theta} - \widehat{\theta}\right)' A \left(\overline{\theta} - \widehat{\theta}\right)$. ■

The results of Lemma 111 can not be generalized when the null hypothesis is nonlinear or the log-likelihood is not quadratic; thus despite sharing the same asymptotic distribution, these tests will generally differ in finite samples. Note that of the three tests, the easiest to derive is the Wald test given that it only requires to compute the unconstrained MLE. The LM test evaluates both the gradient and Hessian matrix in the constrained MLE, while LRT requires the computation of the constrained and unconstrained MLE. A major advantage of the LRT is that it is invariant to one-to-one transformations of the constraints. This property is shared by the LM test but not the Wald test.

The following example presents the results of applying the three tests.

Example 112 *Consider a set of T independent observations on a Bernoulli random variable as in example 104. We wish to test the null $H_0 : \pi = \pi_0$. We know that $\widehat{\pi} = \sum_{t=1}^{T} y_t/T$ and that*

$$\frac{\partial \ell}{\partial \pi} = \frac{\sum_{t=1}^{T}(y_t - \pi)}{\pi(1-\pi)}$$

and

$$-A(\pi) = \frac{1}{\pi(1-\pi)}.$$

The Wald test is given by

$$Wald = \frac{T(\pi_0 - \widehat{\pi})^2}{\widehat{\pi}(1-\widehat{\pi})}.$$

The LM test is

$$\begin{aligned} LM &= \left[\frac{\sum_{t=1}^{T}(y_t - \pi_0)}{\pi_0(1-\pi_0)}\right]^2 \frac{\pi_0(1-\pi_0)}{T} \\ &= \frac{T(\pi_0 - \widehat{\pi})^2}{\pi_0(1-\pi_0)}. \end{aligned}$$

Both clearly have a limiting χ_1^2 distribution. They differ in that the LM test uses an estimate of the variance under the null whereas the Wald uses an estimate under the alternative. When the null is true these will have the same probability limit and thus for large samples the tests will be equivalent. If the alternative is not close to the null, then presumably both tests would reject with very high probability for large samples.

The likelihood ratio test statistic is given by:

$$LRT = 2T\left[\widehat{\pi}\ln\left(\frac{\widehat{\pi}}{\pi_0}\right) + (1-\widehat{\pi})\ln\left(\frac{1-\widehat{\pi}}{1-\pi_0}\right)\right],$$

which has a less obvious limiting distribution and is slightly more awkward to calculate. A two-term Taylor series expansion of the statistic about $\widehat{\pi} = \pi_0$ establishes that under the null the three will have the same distribution.

9.7.2 The Tests in Special Cases

Next, we shall find explicit formulae for the three tests for the HLRM with the linear hypothesis $Q'\beta = 0$. In this case, (9.23)-(9.25) are reduced to

$$\begin{aligned}
\text{Wald} &= T\frac{S_T\left(\overline{\beta}\right) - S_T\left(\widehat{\beta}\right)}{S_T\left(\widehat{\beta}\right)} \\
\text{LM} &= T\frac{S_T\left(\overline{\beta}\right) - S_T\left(\widehat{\beta}\right)}{S_T\left(\overline{\beta}\right)} \\
\text{LRT} &= T\ln\left[\frac{S_T\left(\overline{\beta}\right)}{S_T\left(\widehat{\beta}\right)}\right].
\end{aligned}$$

Thus we can easily show that Wald≥LRT≥LM; that is, despite the fact that the three tests share the same asymptotic distribution, in this particular case, the values of these tests would make it easier to reject the null hypothesis with the Wald test. . This inequalities do not always hold for the nonlinear model, or for the linear model with nonlinear constraints.

In the NLLS model when u is normal, we have

$$\begin{aligned}
\text{Wald} &= \frac{Th\left(\widehat{\beta}\right)'\left[\left.\frac{\partial h}{\partial \beta'}\right|_{\widehat{\beta}}\left(\widehat{Z}'\widehat{Z}\right)^{-1}\left.\frac{\partial h}{\partial \beta'}\right|_{\widehat{\beta}}\right]^{-1}h\left(\widehat{\beta}\right)}{S_T\left(\widehat{\beta}\right)} \\
\text{LM} &= \frac{T\left[Y - m\left(\overline{\beta}\right)\right]'\overline{Z}\left(\overline{Z}'\overline{Z}\right)^{-1}\overline{Z}'\left[Y - m\left(\overline{\beta}\right)\right]}{S_T\left(\overline{\beta}\right)} \\
\text{LRT} &= T\left[\ln T^{-1}S_T\left(\overline{\beta}\right) - \ln T^{-1}S_T\left(\widehat{\beta}\right)\right].
\end{aligned}$$

These three tests hold asymptotically even if u is not normal.

9.7.3 Nonnested Hypotheses

The comparison of different hypotheses, i.e. of competing models, is the basis of model specification. It may be performed along two main lines. The first one consists in associating with each model a loss function and in retaining the specification implying the smallest (estimated) loss. In practice, the loss function is defined either by updating some a-priori knowledge on the models given the available observations (the Bayesian point of view), or by introducing some criterion taking into account the trade-off between the goodness of fit and the complexity of the model (for instance the AIC or BIC). This approach, called model choice or model selection, has already been discussed before. The second approach is hypothesis testing theory. For model selection we have to choose a decision rule explaining for which observations we prefer to retain each hypothesis. In the simplest case of two hypotheses H_0 and H_1, this is equivalent to the definition of the critical region giving the set of observations for which H_0 is rejected. However, the determination of the decision rule is not done on the same basis as model choice. The basis of hypothesis testing theory is to introduce the probability of errors: first error type (to reject H_0 when it is true), and second error type (to reject H_1 when it is true), then to choose a critical region for which the first error type probability is smaller than a given size (generally 5%) and the second error type probability is as small as possible. Hypothesis testing theory is usually advocated when H_0 may be considered as a "limit case" of the second hypothesis H_1. Broadly speaking the model $\mathrm{H}_0 \cup \mathrm{H}_1$ can be reduced to the submodel H_0 by imposing some restrictions on the parameters in which case H_0 is said to be nested in H_1.

Here we are interested in the opposite case, where none of the hypotheses is a particular case of another one. These hypotheses may be entirely distinct (globally nonnested hypotheses) or may have an intersection (partially nonnested hypotheses).

The hypotheses may concern either the whole conditional distribution of y_t given x_t, or simply some conditional moments, such as the conditional expectation. We successively consider these two situations.

When the hypotheses concern the whole conditional distribution and have a parametric form, they may be written as

$$\begin{aligned} H_g &: \left\{ g\left(y_t \left| x_t ; \alpha\right.\right), \alpha \in A \subset \mathbb{R}^G \right\} \\ H_h &: \left\{ h\left(y_t \left| x_t ; \beta\right.\right), \beta \in B \subset \mathbb{R}^H \right\}. \end{aligned}$$

The first hypothesis H_g (for instance) is valid if the true conditional distribution $f\left(y_t \left| x_t\right.\right)$ can be written as $g\left(y_t \left| x_t ; \alpha_0\right.\right)$ for some $\alpha_0 \in A$. A test that is commonly proposed is:

$$R_g = \ell_g\left(\widehat{\alpha}\right) - \mathcal{E}_\alpha\left[\ell_g\left(\alpha\right)\right]\big|_{\widehat{\alpha}} - \ell_h\left(\widehat{\beta}\right) + \mathcal{E}_\alpha\left[\ell_h\left(\beta_\alpha\right)\right]\big|_{\widehat{\alpha}},$$

where β_{α} =plim$_{\alpha}\widehat{\beta}$ (meaning the probability limit is taken assuming $g(\alpha)$ is the true model) and$\hat{}$indicates maximum likelihood estimates. We reject H_g if R_g is larger than a critical value determined by the asymptotic distribution of R_g which is asymptotically normal with zero mean and variance equal to $\mathcal{E}\left(v_g^2\right)-\mathcal{E}\left(v_g w_g'\right)\left(\mathcal{E}w_g w_g'\right)^{-1}\mathcal{E}\left(v_g w_g\right)$, where $v_g=\ell_g(\alpha)-\ell_h(\beta_{\alpha})-\mathcal{E}_{\alpha}\left[\ell_g(\alpha)-\ell_h(\beta_{\alpha})\right]$ and $w=\partial\ell_g(\alpha)/\partial\alpha$.

A weakness of this test is its inherent asymmetry; the test of H_g against H_h based on R_g may contradict the test of H_h against H_g based on R_h. For example, H_g may be rejected by R_g and at the same time H_h may be rejected by R_h.

Given this problem, some researchers have proposed to artificially nest both models and introduce a third hypothesis in both H_g and H_h are nested and characterized by some equality constraints. In this case, the idea is to introduce mixtures of distributions of H_g and H_h. This model is

$$\mathcal{M}=(1-\lambda)\,g\left(y_t\left|x_t;\alpha\right.\right)+\lambda h\left(y_t\left|x_t;\beta\right.\right)\quad \lambda\in[0,1]\,.$$

The basic hypotheses are defined by the constraints

$$H_g:\lambda=0\quad\text{and}\quad H_h:\lambda=1.$$

The procedure consists in testing $\lambda=0$ against $\lambda>0$ and $\lambda=1$ against $\lambda<1$; that is, in applying a one-sided t-ratio test to the parameters λ or $1-\lambda$. It is possible that both hypotheses are not satisfied. Even though this compound model is attractive, it has the usual drawback of mixtures, given that under the null hypothesis, either the parameter α or β are not identified. Therefore, the properties of the $t-$ratio are unknown except in special cases.

Davidson and MacKinnon (1993) propose the following procedure for the OLS models (which with a few minor modifications can be extended to the NLLS model):. Consider the models

$$\begin{aligned} H_g &: Y=X'\alpha+u_1\\ H_h &: Y=Z'\beta+u_2 \end{aligned}$$

and the compounded model

$$H_C:Y=(1-\lambda)\,X'\alpha+\lambda Z'\beta+u.$$

Let $\widehat{\beta}$ be the OLS estimator under H_h. Define the following auxiliary model

$$H_{C'} : Y = (1-\lambda)\, X'\alpha + \lambda Z'\widehat{\beta} + u.$$

If $\widehat{\lambda}$ is the estimate of λ of this auxiliary model, it can be shown that the $t-$test of the estimate converges in distribution to a standard normal. If the null $\lambda = 0$ is not rejected, we conclude that there is evidence in favor of H_g.

Although the Wald, LM, and LR tests apply to nested hypotheses, similar asymptotic reasoning extends to nonnested model comparisons. Vuong (1989) showed that differences in maximized log-likelihoods can be standardized to construct a test for nonnested models, where the null hypothesis states that both models are equally close to the true data-generating process in Kullback–Leibler distance. This approach provides the theoretical foundation for many modern model-comparison criteria, including information-theoretic and predictive tests discussed later.

9.8 Further Reading

This chapter follows closely the classical and elegant treatment of extremum estimators developed in Amemiya (1985), which remains one of the most comprehensive and rigorous expositions available. The modern formulation of the theory, including detailed proofs and extensions to models with stochastic regressors and dependent data, is presented in Newey and McFadden (1994). For contemporary treatments that emphasize intuition and applications, Hayashi (2000) and Hansen (2022) provide particularly clear discussions. Mittelhammer, Judge, and Miller (2000) offer an integrated view of extremum, likelihood, and generalized method-of-moments estimators within a common asymptotic framework, while Ruud (2000) and Greene (2012) contain useful complementary presentations. Readers interested in nonsmooth extremum estimators, such as the LAD and quantile estimators, may consult Koenker and Hallock (2001) and Huber (1981).

9.9 Workout Problems

1. Prove that if $|y - x| < \xi$, then $x > y - \xi$.

2. Derive (9.11).

3. Derive the asymptotic distribution of the GLS estimator using extremum estimators.

4. Derive equations (9.13) and (9.14).

5. Prove Theorem 105.

6. Derive the asymptotic distribution of the MLE of the NLRM.

7. Derive the asymptotic variance of the estimator of β obtained by minimizing $\sum_{t=1}^{T}(y_t - \beta x_t)^4$, where y_t is independent with the distribution $\mathcal{N}(\beta_0 x_t, \sigma_0^2)$ and $\lim_{T\to\infty} T^{-1}\sum_{t=1}^{T} x_t^2$ is a finite, positive constant. You may assume consistency and asymptotic normality. Indicate the additional assumptions on x_t one needs. Note that if $Z \backsim \mathcal{N}(0,\sigma^2)$, $\mathcal{E}Z^{2k} = \sigma^{2k}(2k)!/\left(2^k k!\right)$.

8. Consider a nonlinear regression model

$$y_t = (\beta_0 + x_t)^2 + u_t,$$

where we assume that $\{u_t\}$ are i.i.d. with $\mathcal{E}u_t = 0$ and $\mathcal{V}u_t = \sigma_0^2$; the parameter space $B = [-0.5, 0.5]$; $\{x_t\}$ are i.i.d. with the uniform distribution over $[1,2]$, distributed independently of $\{u_t\}$. $[\mathcal{E}X^r = (r+1)^{-1}(2^{r+1}-1)$ for every positive or negative integer r except $r = -1$. $\mathcal{E}X^{-1} = \ln 2]$.

Define two estimators of β_0:

(**a**) $\widehat{\beta}$ minimizes $\sum_{t=1}^{T}\left[y_t - (\beta + x_t)^2\right]^2$ over B,

(**b**) $\widetilde{\beta}$ minimizes $\sum_{t=1}^{T}\left[y_t/(\beta + x_t)^2 + \ln(\beta + x_t)^2\right]$ over B.

If $\beta_0 = 0$, which of the two estimators do you prefer? Explain your preference on the basis of asymptotic results.

9. In the HLRM, prove that if $\text{H}_0 : Q'\beta = 0$, then Wald$\geq$LRT$\geq$LM.

10. In the LRM, prove that LM$= TR^2$.

Chapter 10

Limited Dependent Variables

10.1 Introduction

This document is intended to be an account of certain salient themes of the Limited Dependent Variable (LDV) literature, building upon the maximum-likelihood framework introduced in Chapter 7. By LDV we will mean instances of (dependent) variables for which (a) their range is intrinsically a finite discrete set and any attempt to extend it to the real line not only does not lead to useful simplification, but renders standard continuous approximations analytically misleading; (b) even though their range may be the real (half) line their behavior is conditioned on another process(es).

Examples of the first type are models of occupational choice, entry into labor force, entry into college upon high school graduation, utilization of recreational facilities, utilization of modes of transport, childbearing, etc. Examples of the latter are models of housing prices and wages in terms of the relevant characteristics of the housing unit or the individual (what is commonly referred to as hedonic price determination). Under this category we will also consider the case of truncated dependent observations. Both classes share the feature that the dependent variable's support provides essential information about the underlying behavioral process.

In examining these issues we shall make an attempt to provide an economic rationalization for the model considered, but our main objective will be to show why common procedures such as LS fail to give acceptable results; how one approaches these problems by maximum likelihood procedures and how one can handle problems of inference (chiefly by determining the limiting distributions of the relevant estimators). An attempt will be made to handle all problems in a reasonably uniform manner and by relatively elementary means. Throughout, estimation will rely on the maximum-likelihood methods and asymptotic tools developed in the preceding chapters.

This document is organized as follows: Section 10.2 presents models in which the dependent variable is a binomial qualitative response (QR) variable. Section 10.3 generalizes the results of Section 10.2, for models in which the dependent variable is qualitative but may take more than two values. Section 10.4 introduces techniques to handle multivariate QR models. Section 10.5 describes models for count data (that is, models in which the dependent variable is quantitative but may only take discrete values). Section 10.6 discusses methods for estimating models with limited dependent variables that are truncated or censored. Finally, Section 10.7 introduces models for duration data.

10.2 Binary Response Models

Consider first the problem faced by a youth completing high school; or by a married female who has attained the desired size of her family. In the instance of the former the choice to be modelled is going to college or not; in the case of the latter we need to model the choice of entering the labor force or not.

Suppose that as a result of a properly conducted survey we have observations on N individuals, concerning their socioeconomic characteristics and the choices they have made.

In order to free ourselves from dependence on the terminology of a particular subject when discussing these problems, let us note that, in either case, we are dealing with a binary choice; let us denote this by

Alternative 1: Going to College or Entering Labor Force

Alternative 2: Not Going to College or Not Entering Labor Force

Since the two alternatives are exhaustive we may make alternative 1 correspond to an abstract event $\mathcal{W}$ and alternative 2 correspond to its complement $\overline{\mathcal{W}}$. In this context it will be correct to say that what we are interested in, is the set of factors affecting the occurrence or nonoccurrence of $\mathcal{W}$. What we have at our disposal is some information about the attributes of these alternatives and the (socioeconomic) attributes of the individual exercising choice.

Of course we also observe the choices of the individual agent in question. Let

$$y_i = \begin{cases} 1 & \text{if individual } i \text{ chooses in accordance with event } \mathcal{W} \\ 0 & \text{otherwise} \end{cases} .$$

Let w_i be a vector of characteristics relative to the alternatives corresponding to the events $\mathcal{W}$ and $\overline{\mathcal{W}}$; finally, let r_i be the vector describing the socioeconomic characteristics of the ith individual economic agent.

We may be tempted to model this phenomenon as

$$y_i = \beta' x_i + u_i, \tag{10.1}$$

where $x_i = (w_i, r_i)$, and β is a vector of unknown constants.

The formulation in (10.1) and subsequent estimation by LS procedures was a common occurrence in the empirical research of the sixties and is referred to as the linear probability model.

10.2.1 Linear Probability Model

Although the temptation to think of this problem in a LS context is enormous, a close examination will show that this is also fraught with considerable problems. First, notice that the dependent variable can only assume two possible values, while no comparable restrictions are placed on the first component of the right hand side of (10.1). Second, note that if we insist on putting this phenomenon in the LS mold, then for observations in which $y_i = 1$ we must have $u_i = 1 - \beta' x_i$ while for observations in which $y_i = 0$ we must have $u_i = -\beta' x_i$.

Thus, the error term can only assume two possible values, and we are immediately led to consider that what we need is not a linear model "explaining" the choices individuals make, but rather a model of the probabilities corresponding to the choices in question. Thus, if we ask ourselves: what is the expectation of u_i, we shall conclude that if we define $\mathcal{F}_i = \Pr(y_i = 1)$ and $1 - \mathcal{F}_i = \Pr(y_i = 0)$ we have

$$\begin{aligned} \mathcal{E}(u_i) &= \mathcal{F}_i (1 - \beta' x_i) - (1 - \mathcal{F}_i) \beta' x_i \\ &= \mathcal{F}_i - \beta' x_i. \end{aligned}$$

Thus, in order for $\mathcal{E}(u_i) = 0$, we conclude that $\mathcal{F}_i = \Pr(y_i = 1)$ must be set equal to $\mathcal{E}(y_i) = \beta' x_i$. If that were the case, we would have

$$\begin{aligned}\mathcal{V}(u_i) &= \beta' x_i (1 - \beta' x_i)^2 + (1 - \beta' x_i)(\beta' x_i)^2 \\ &= \beta' x_i (1 - \beta' x_i).\end{aligned}$$

This means that if we wanted to estimate this model using LS, by construction, our residuals would be heteroskedastic. Of course, absent any other problems, we could manage this with a FGLS estimator. For our purpose, though, a more serious flaw is that without some ad hoc tinkering with the disturbances, we cannot be assured that the predictions from this model will truly look like probabilities. Given that we cannot constrain $\beta' x_i$ to the zero-one interval, this procedure produces both nonsense probabilities and negative variances.

Our requirement, then, is a model that will produce predictions that satisfy

$$\lim_{\beta' x_i \to \infty} \Pr(y_i = 1) = 1$$

and

$$\lim_{\beta' x_i \to -\infty} \Pr(y_i = 1) = 0.$$

Then, what we really should be asking is: what determines the probability that the $i-$th economic agent chooses in accordance with event $\mathcal{W}$, and (10.1) should be viewed as a clumsy way of going about it. We see that putting

$$\begin{aligned}\Pr(y_i = 1) &= \mathcal{F}_i(\beta' x_i) = \int_{-\infty}^{\beta' x_i} f(v)\, dv, \\ \Pr(y_i = 0) &= 1 - \mathcal{F}_i(\beta' x_i) = \int_{\beta' x_i}^{\infty} f(v)\, dv,\end{aligned}$$

where $f(\cdot)$ is a suitable density function with known parameters that formalizes the dependence of the probabilities of choice on the observable characteristics of the individual and/or the alternatives.

10.2.2 A Utility Maximization Motivation

Discrete dependent variable models are often cast in the form of index function (or random utility) models. Consider the decision to make a large purchase. The theory states that the consumer makes a marginal benefit - marginal cost calculation. Since marginal benefit is obviously not observable, we model the difference between benefit and cost as an unobservable variable, y^*, such that

$$y_i^* = \beta' x_i + u_i.$$

Given that we do not observe the net benefit of the purchase and only observe if it is was made or not, we define $y_i = 1$ if the purchase was made and 0 otherwise.[1] Thus

$$y_i = \begin{cases} 1 \text{ if } y_i^* > 0 \\ 0 \text{ if } y_i^* \leq 0 \end{cases}. \tag{10.2}$$

In this formulation, $\beta' x_i$ is called the index function.[2]

If $\mathcal{F}(u\,|X)$ is the cumulative distribution function of the disturbances, then just as in the continuous case, the model is characterized by the conditional distribution of Y given X:

$$\begin{aligned} \Pr(y_i = 1\,|x_i) &= \Pr(y_i^* = \beta' x_i + u_i > 0) \\ &= \Pr(u_i > -\beta' x_i) \\ &= 1 - \mathcal{F}(-\beta' x_i). \end{aligned}$$

Given that $\mathcal{E}(u) = 0$, if $f(\cdot)$ is a symmetric density function, then $1 - \mathcal{F}(-\beta' x_i) = \mathcal{F}(\beta' x_i)$. This is called the response probability.

10.2.3 Functional Forms

Assuming that u is independent of x, in order to estimate the binomial model we need to impose a distributional assumption on u.

In principle, any proper, continuous probability distribution defined over the real line will suffice. The normal distribution has been used in many analysis, giving rise to the **probit** model,

$$\mathcal{F}(\beta' x_i) = \int_{-\infty}^{\beta' x_i} \phi(t)\; dt = \Phi(\beta' x_i), \tag{10.3}$$

where Φ is the standard cumulative normal; thus, assuming a unit variance on u. This is simply an innocent normalization because: a) this parameter

[1] Given that this is a qualitative variable, we could have defined y^* as the net cost, in which case $y = 1$ would have meant that the purchase was not made.

[2] Here we restrict our discussion to the case in which the index function is linear, but nothing prevents us from postulating a nonlinear index function of the form $m(\beta, x_i)$.

would not be identified if we wanted to estimate it separately from β and b) from our theoretical motivation, what we were interested in is the sign of y^*.

Partly because of its mathematical convenience, the logistic distribution

$$\mathcal{F}\left(\beta' x_i\right) = \frac{e^{\beta' x_i}}{1 + e^{\beta' x_i}} = \Lambda\left(\beta' x_i\right), \tag{10.4}$$

has also been used in many applications. We shall use the notation $\Lambda(\cdot)$ to indicate the logistic cumulative distribution function. This model is called the **logit** model.

The question of which distribution to use is a natural one. The logistic distribution is similar to the normal except in the tails, which are considerably heavier (it closely resembles a t distribution with seven degrees of freedom). Therefore, for intermediate values of $\beta' x_i$ (say, between -1.2 and 1.2), the two distributions tend to give similar probabilities. The logistic distribution tends to give larger probabilities to $y = 0$ when $\beta' x_i$ is extremely small (and smaller probabilities to $y = 0$ when $\beta' x_i$ is very large) than the normal distribution. It is difficult to provide practical generalities in this basis, however, since they would require knowledge of β. However, we should expect different predictions from the two models if the sample contains very few observations of $y = 1$ or $y = 0$ and very wide variation in an important independent variable.

In practice, the logit and probit models tend to yield extremely similar results. In most cases, the only real difference between them is in the way the elements of β are scaled. This difference in scaling occurs because the variance of the distribution for which the logistic distribution is the c.d.f. is $\pi^2/3$, while that of the standard normal is of course unity. The logit estimates therefore all tend to be larger than the probit estimates, although usually by a factor of somewhat less than $\pi/\sqrt{3}$.

In view of their similar properties, it is perhaps curious that both the logit and probit models continue to be widely used, while models that genuinely differ from them are rarely encountered. There are as many ways in which such models could be specified as there are plausible choices for $\mathcal{F}(\cdot)$. Alternative response probabilities that approach zero or one less rapidly (that have thicker tails) can be constructed from the Student$-t$ or Cauchy distribution; the latter yields the **arctan** model:

$$\mathcal{F}\left(\beta' x_i\right) = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}\left(\beta' x_i\right), \tag{10.5}$$

which has the following p.d.f.:

$$f(z) = \frac{1}{\pi(1+z^2)}.$$

Because the behavior of the Cauchy distribution function in the tails is very different from that of either $\Phi(\cdot)$ or $\Lambda(\cdot)$, there is at least the possibility that a binary response model based on (10.5) might perform substantially better or worse than a probit or logit model.

All three choices for $\mathcal{F}(\cdot)$ that we have discussed are skew-symmetric around zero. That is, they have the property that $1 - \mathcal{F}(z) = \mathcal{F}(-z)$, which implies that $f(z) = f(-z)$. This is sometimes a convenient property, but there is no a priori reason for it to hold. Choices for $\mathcal{F}(\cdot)$ that do not have this property will potentially yield quite different results from those produced by the logit and probit models. One way to obtain the same effect is to specify the model $\mathcal{F}(h(\beta' x_i))$; where $\mathcal{F}(\cdot)$ is $\Phi(\cdot)$ or $\Lambda(\cdot)$, and $h(\cdot)$ is a nonlinear transformation. This suggests a way to test the validity of the skew-symmetry assumption.

A different strategy is followed by the **maximum score** estimator, which can be viewed as a distribution-free estimator of the parameters in QR models. By distribution-free estimators we mean that no distributional assumption on u is imposed. This method generalizes the concept of the LAD estimator that we discussed before. The basic idea is to define the score function

$$S_N(\beta) = \sum_{i=1}^{N} \left[y_i \mathcal{I}(\beta' x_i \geq 0) + (1 - y_i) \mathcal{I}(\beta' x_i < 0) \right],$$

where $\mathcal{I}$ is an indicator function such that

$$\mathcal{I}(z) = \begin{cases} 1 & \text{if event } z \text{ occurs} \\ 0 & \text{otherwise} \end{cases}.$$

Note that the score is the number of correct predictions we would make if we predicted y_i to be 1 whenever $\beta' x_i \geq 0$. The maximum score estimator is defined by

$$S_N\left(\widehat{\beta}\right) = \sup_{\beta \in B} S_N(\beta),$$

where the parameter space B is taken as

$$B = \{\beta \,|\beta'\beta = 1\}. \tag{10.6}$$

Clearly, (10.6) implies no loss of generality because $S_N(c\beta) = S_N(\beta)$ for any positive scalar c. Given that this objective function is not differentiable, solving this maximization problem can be conducted by direct search methods, or by transformations such as the ones used for the LAD estimator. Initial values for the iterative procedure can be obtained by using the QMLE of the linear probability model (OLS), subject to a transformation that satisfies (10.6). This method can also be generalized for the multinomial models that we discuss on Section 3 (see Amemiya, 1985 for details). Even though this estimator is known to be consistent, its asymptotic distribution is yet to be derived.

10.2.4 Estimation

Estimation is done by maximum likelihood. To construct the conditional likelihood, recall that if y is (conditionally) Bernoulli, such that $\Pr(y=1) = \mathcal{F}(\beta' x_i)$ and $\Pr(y=0) = 1 - \mathcal{F}(\beta' x_i)$, then we can write the conditional log-likelihood as

$$\begin{aligned}
\ell(\beta; Y\,|X) &= \sum_{i=1}^{N} \ln\left[\mathcal{F}(\beta' x_i)^{y_i}(1-\mathcal{F}(\beta' x_i))^{1-y_i}\right] \\
&= \sum_{i=1}^{N} \left[y_i \ln \mathcal{F}(\beta' x_i) + (1-y_i)\ln(1-\mathcal{F}(\beta' x_i))\right] \\
&= \sum_{y_i=1} \ln \mathcal{F}(\beta' x_i) + \sum_{y_i=0} \ln(1-\mathcal{F}(\beta' x_i)).
\end{aligned}$$

The FONC are:

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{N} \left[y_i \frac{f_i}{\mathcal{F}_i} - (1-y_i)\frac{f_i}{1-\mathcal{F}_i}\right] x_i,$$

or, more compactly

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{N} \left[\frac{(y_i - \mathcal{F}_i) f_i}{\mathcal{F}_i(1-\mathcal{F}_i)}\right] x_i, \tag{10.7}$$

where $f_i = f(\beta' x_i)$ and $\mathcal{F}_i = \mathcal{F}(\beta' x_i)$. Note that (10.7) is a highly nonlinear function of the unknown parameter vector β and, hence, can only be solved by numerical methods such as Newton-Raphson.

It is instructive to derive the FONC for the case in which $\mathcal{F}$ corresponds to the logit model, where (10.7) reduces to

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{N} (y_i - \Lambda_i) x_i. \tag{10.8}$$

Note that if x contains a constant term, the FONC imply that the average of the predicted probabilities must equal the proportion of ones in the sample. This also bears some similarity to the least squares normal equations if we view the term $y_i - \Lambda_i$ as a residual.[3]

The SOSC for the logit model are quite simple based on (10.8):

$$\frac{\partial^2 \ell}{\partial \beta \partial \beta'} = -\sum_{i=1}^{N} \Lambda_i (1 - \Lambda_i) x_i x_i'.$$

Note that the Hessian is always negative definite, so the log-likelihood is globally concave; thus, Newton's method will usually converge to the maximum in just a few iterations irrespective of the initial conditions.

The computation is more involved for the probit model. A useful simplification is obtained by defining

$$\delta_i = \begin{cases} \frac{\phi_i}{\Phi_i} & \text{if } y_i = 1 \\ \frac{-\phi_i}{1-\Phi_i} & \text{if } y_i = 0 \end{cases}.$$

With this variable, the FONC are simply

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{N} \delta_i x_i.$$

The SOSC can be obtained by using the result

$$\frac{\partial \phi(z)}{\partial z} = -z \phi(z).$$

[3] Although regularly observed in practice, the result has not been verified for the probit model.

Then

$$\frac{\partial^2 \ell}{\partial \beta \partial \beta'} = -\sum_{i=1}^{N} \delta_i \left(\delta_i + \beta' x_i\right) x_i x_i', \qquad (10.9)$$

which is also negative definite for all values of β.

Global concavity is a characteristic of the logit and probit models (provided that the explanatory variables are linearly independent); however, this is not necessarily the case for other choices of $\mathcal{F}(\cdot)$ where there may be local maxima.

10.2.5 Inference

Under mild regularity conditions (Amemiya, 1985), the MLE is consistent and asymptotically normal. The covariance matrix for the MLE can be estimated by using the inverse of the information matrix (negative of the Hessian) evaluated at the MLE. Alternative, we could use the OPG to obtain

$$\frac{\partial \ell}{\partial \beta} \frac{\partial \ell}{\partial \beta'} = \sum_{i=1}^{N} g_i^2 x_i x_i'$$

where $g_i = (y_i - \Lambda_i)$ for the logit model and $g_i = \delta_i$ for the probit model. Recall that in both cases we are assuming that the model is correctly specified. If not, an estimator of the full asymptotic covariance matrix should be considered.

Once an estimator of this matrix is obtained, inference can be conducted as usual, by using Wald, LR or LM tests. Next, we describe some of the special features of these tests in the binary response model.

Using Wald Tests

Wald tests are useful not only for conducting inference, but can also be used to obtain estimates of the variance of nonlinear function. For example, note that the binary probability model implies that

$$\mathcal{E}(y_i) = 1\left[\mathcal{F}(\beta' x_i)\right] + 0\left[1 - \mathcal{F}(\beta' x_i)\right] = \mathcal{F}(\beta' x_i).$$

The estimated probability, $\mathcal{F}\left(\widehat{\beta}' x_i\right)$, is a nonlinear function of the parameter estimates. To compute standard errors, we can use the linear approximation approach discussed for nonlinear hypotheses. In this case, we have,

$$\begin{aligned}\widehat{\mathcal{V}}\left[\mathcal{F}\left(\widehat{\beta}' x_i\right)\right] &= \left(\frac{\partial \mathcal{F}\left(\widehat{\beta}' x_i\right)}{\partial \widehat{\beta}}\right)' \widehat{\mathcal{V}}\left(\widehat{\beta}\right)\left(\frac{\partial \mathcal{F}\left(\widehat{\beta}' x_i\right)}{\partial \widehat{\beta}}\right) \\ &= f\left(\widehat{\beta}' x_i\right)^2 x_i' \widehat{\mathcal{V}}\left(\widehat{\beta}\right) x_i.\end{aligned}$$

which depends, of course, on the particular value of x used.

Another important feature of these models is that whatever distribution is used, it is important to note that the parameters of the model, like those of any nonlinear function are not necessarily the marginal effects we are accustomed to analyzing. In general,

$$\begin{aligned}\frac{\partial \mathcal{E}(y)}{\partial x} &= \left[\frac{\partial \mathcal{F}(\beta' x_i)}{\partial (\beta' x_i)}\right]\beta \\ &= f(\beta' x_i)\beta.\end{aligned}$$

It is obvious that these effects will vary with the values of x. In interpreting the estimated model, it is useful to calculate these at, the means of the regressors and, where necessary, other pertinent values.

As the estimated marginal effects, $f\left(\widehat{\beta}' x_i\right)\widehat{\beta}$, are nonlinear functions of the parameter estimates, we can also use the linear approximation approach to compute standard errors. For the marginal effects, let $\widehat{\gamma}_i = f\left(\widehat{\beta}' x_i\right)\widehat{\beta}$. Then

$$\widehat{\mathcal{V}}(\widehat{\gamma}_i) = \left[\frac{\partial \widehat{\gamma}_i}{\partial \widehat{\beta}}\right]' \widehat{\mathcal{V}}\left(\widehat{\beta}\right)\left[\frac{\partial \widehat{\gamma}_i}{\partial \widehat{\beta}}\right], \tag{10.10}$$

which also depend on the particular value of x used.

Using LR Tests

We demonstrated that the LRT, defined as

$$\text{LRT} = 2\left[\ell\left(\widehat{\beta}\right) - \ell\left(\overline{\beta}\right)\right]$$

has a χ^2_q asymptotic distribution (where $\overline{\beta}$ is the MLE under the null hypothesis and q is the number restrictions implied by the null hypothesis). Thus, we can use this test to conduct inference regarding any linear or nonlinear hypothesis.

The LRT has another important use in the case of the binary response model. In the context of OLS, the coefficient of determination of the multiple regression (R^2) has at least three useful interpretations.

a) It stands in a one-to-one relation to the $F-$statistic for testing the hypothesis that all the coefficients (with exception of the constant) are zero;

b) it is a measure of the reduction of variability of the dependent variable through the inclusion of independent variables;

c) it is the square of the simple correlation coefficient between predicted and actual values of the dependent variable within the sample.

Unfortunately, in the case of the discrete choice models under consideration we do not have a statistic that fits all three characterizations above. We can, on the other hand, define one that essentially performs the first two functions.

Consider the hypothesis

$$\mathrm{H}_0 : \beta_0 = 0$$

as against

$$\mathrm{H}_1 : \beta_0 \neq 0.$$

Hence[4]

$$\ell(0) = \sum_{i=1}^{N} \left[y_i \ln \mathcal{F}(0) + (1 - y_i) \ln (1 - \mathcal{F}(0)) \right] = N \ln \left(\frac{1}{2} \right)$$

and

$$\begin{aligned} \mathrm{LRT} &= 2 \left[\ell \left(\widehat{\beta} \right) - N \ln \left(\frac{1}{2} \right) \right] \\ &= 2 \left[\ell \left(\widehat{\beta} \right) + N \ln (2) \right] \backsim \chi^2_k, \end{aligned}$$

is a test statistic for testing H_0. This is not a useful basis for defining an analog of the R^2 statistic, for it implicitly juxtaposes the economically motivated model that defines the probability of choice as a function of $\beta' x_i$, and

[4] Here we used the fact that if $\mathcal{F}(\cdot)$ is symmetric, $\mathcal{F}(0) = 0.5$.

the model based on the principle of insufficient reason which states that the probability to be assigned to choice corresponding to the event $\mathcal{W}$ and that corresponding to its complement $\overline{\mathcal{W}}$ are both $\frac{1}{2}$.

It would be far more meaningful to consider the null hypothesis to be

$$\mathrm{H}_0 : \beta_0 = \begin{pmatrix} \beta_{1,0} \\ 0 \end{pmatrix}, \tag{10.11}$$

i.e., that all the coefficients with exception of the constant are zero. The null hypothesis as above would correspond to the unconditional Bernoulli model that we already discussed, thus assigning a probability to choice corresponding to event $\mathcal{W}$ by

$$\overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i = \mathcal{F}\left(\overline{\beta}\right) \quad \text{or} \quad \overline{\beta} = \mathcal{F}^{-1}\left(\overline{y}\right)$$

In this case, we can define an analog of the R^2, denoted by LRI (for likelihood ratio index) by

$$\mathrm{LRI} = 1 - \frac{\ell\left(\overline{\beta}\right)}{\ell\left(\widehat{\beta}\right)}. \tag{10.12}$$

This quantity has the property

a) $\mathrm{LRI} \in [0,1)$

b) the larger the contribution of the independent variables to the maximum of ℓ, the closer is LRI to 1

c) LRI stands in a one-to-one relation to the chi-square statistic for testing the hypothesis (10.11). In fact, under H_0

$$-2\ell\left(\overline{\beta}\right) \cdot \mathrm{LRI} \backsim \chi^2_{k-1}.$$

It is desirable, in empirical practice, that a statistic like LRI be reported and that a constant term be routinely included in the specification of the linear functional $\beta' x_i$.

Finally, we should also stress that LRI as in (10.12) does not have the interpretation as the square of the correlation coefficient between predicted and actual observations.

Using LM Test

The Lagrange Multiplier test statistic is

$$\mathrm{LM} = -\left.\frac{\partial \ell}{\partial \beta'}\right|_{\overline{\beta}} \left[\left.\frac{\partial^2 \ell}{\partial \beta \partial \beta'}\right|_{\overline{\beta}} \right]^{-1} \left.\frac{\partial \ell}{\partial \beta}\right|_{\overline{\beta}} \backsim \chi^2_q,$$

where the derivatives of the unrestricted model are evaluated at the restricted parameter vector.

An interesting application of this test for the logit model when the null hypothesis (10.11) is being tested is that

$$\mathrm{LM} = N\mathrm{R}^2. \tag{10.13}$$

where R^2 is the uncentered coefficient of determination in the regression of $(y_i - \overline{y})$ on x_i. In fact, this result can be extended for the case of any OLS regression, where, when the null hypothesis takes the form of (10.11), the LM test is equivalent to product of the number of observations and the uncentered R^2 of the dependent variable on a constant and the x variables.

10.2.6 Predictions

A useful summary of the predictive ability of the model is a 2 × 2 table of the hits and missed of the prediction rule:

$$\widehat{y_i} = \begin{cases} 1 \text{ if } \mathcal{F}\left(\widehat{\beta}' x_i\right) > \mathcal{F}^* \\ 0 \text{ otherwise} \end{cases}, \tag{10.14}$$

where $\mathcal{F}^*$ is a threshold. The usual threshold value is 0.5, under the logic that we predict a 1 if the model says that a 1 is more likely than a 0. However, it is important not to place too much emphasis on this measure of predictive ability. Consider for example, the naive predictor

$$\widehat{y_i} = \begin{cases} 1 \text{ if } \overline{y} > 0.5 \\ 0 \text{ otherwise} \end{cases},$$

where, again, $\overline{y}$ is the proportion of ones in the sample. This rule will always predict correctly $100\overline{y}$ percent of the observations, which means that the naive model does not have a zero LRI. In fact, if the proportion of ones in the sample is very high, it is possible to construct examples in which the second model will generate more correct predictions than the first!

A second consideration is that 0.5, although the usual choice, may not be a very good value to use for the threshold. If the sample is relatively unbalanced, that is, has many more ones than zeros, or vice versa, then by this prediction rule, it might never predict a 1 (or a 0). To consider an example, suppose that in a sample of 1,000 observations, only 100 have $y = 1$. We know that the average predicted probability in the sample will be 0.1. As such, it may require an extreme configuration of regressors even to produce an $\mathcal{F}$ of 0.2, to say nothing of 0.5. In such a setting, the prediction rule may fail every time to predict when $y = 1$. The obvious adjustment is to reduce $\mathcal{F}^*$. Of course, this comes at a cost. Table 10.1 illustrates this point; given that, in general, any prediction rule of the form in (10.14) will make two types of errors. It will incorrectly assign zeros as ones (C) and ones as zeros (B). In practice, these errors need not be symmetric in the cost that results. For example, this practice has been routinely used for predicting currency crises. But incorrectly classifying a country as a bad risk represents a missed opportunity, while incorrectly classifying a bad risk as good could lead to substantial costs. Changing $\mathcal{F}^*$ will always reduce the probability of one type of error while increasing the probability of the other. There is no correct answer as to the best value to choose. It depends on the setting and on the criterion (risk) function to be minimized.

		Predicted		
		$\widehat{y} = 1$	$\widehat{y} = 0$	Total
Actual	$y = 1$	A	B	A+B
	$y = 0$	C	D	C+D
	Total	A+C	B+D	A+B+C+D

Table 10.1: Hits and misses

10.2.7 Analysis of Proportions Data

Data for the analysis of binary response will be in one of two forms. The data we have considered thus far are individual; each observation consists of (y_i, x_i), the actual response of the individual and associated regressor vector. Grouped data usually consist of counts or proportions. Grouped data are obtained by observing the response of N_i individuals, all of whom have the same x_i. The observed dependent variable will consist of the proportions P_i, of the N_i individuals, who respond with $y = 1$. An observation is thus (N_i, P_i, x_i).[5]

In the grouped data setting, it is possible to use regression methods as well as ML procedures to analyze the relationship between P_i and x_i. The observation P_i is an estimate of the population $\pi_i = \mathcal{F}(\beta' x_i)$. If we treat this as a simple problem in sampling from a Bernoulli population, then,

$$P_i = \mathcal{F}(\beta' x_i) + u_i = \pi_i + u_i,$$

where

$$\mathcal{E}(u_i) = 0, \quad \mathcal{V}(u_i) = \frac{\pi_i (1 - \pi_i)}{N_i}. \tag{10.15}$$

This heteroskedastic regression format suggests that the parameters could be estimated by a nonlinear weighted least squares regression. But there is a simpler way to proceed. Since the function $\mathcal{F}(\beta' x_i)$ is strictly monotonic, it has an inverse. Consider, then, a Taylor series approximation to this function around the point $u_i = 0$:

$$\mathcal{F}^{-1}(P_i) = \mathcal{F}^{-1}(\pi_i + u_i) \approx \mathcal{F}^{-1}(\pi_i) + \left[\frac{\partial \mathcal{F}^{-1}(\pi_i)}{\partial \pi_i}\right] u_i.$$

But

$$\mathcal{F}^{-1}(\pi_i) = \beta' x_i$$

and

$$\frac{\partial \mathcal{F}^{-1}(\pi_i)}{\partial \pi_i} = \frac{1}{\frac{\partial \mathcal{F}(\pi_i)}{\partial \pi_i}} = \frac{1}{f(\pi_i)},$$

so

$$\mathcal{F}^{-1}(P_i) \approx \beta' x_i + \frac{u_i}{f(\pi_i)}.$$

[5] A typical example may be a study on the determinants of the unemployment rate for a sample of countries.

This produces a heteroskedastic linear regression,

$$\mathcal{F}^{-1}(P_i) = z_i \approx \beta' x_i + v_i,$$

where

$$\mathcal{E}(v_i) = 0, \quad \mathcal{V}(v_i) = \frac{\mathcal{F}_i(1-\mathcal{F}_i)}{N_i f_i^2}, \tag{10.16}$$

because of (10.15).

The inverse function of the logistic model is particularly easy to obtain; if

$$P_i = \frac{e^{\beta' x_i}}{1+e^{\beta' x_i}},$$

then

$$\ln\left(\frac{P_i}{1-P_i}\right) = \beta' x_i + v_i.$$

This function is called the **logit** of P_i, hence the name "logit" model. For the normal distribution, the inverse function, $\Phi^{-1}(P_i)$, called the **normit** of P_i, must be approximated.[6] The usual approach is a ratio of polynomials.[7]

Weighted least squares regression based on (10.16) has to be conducted in a two-step procedure, given that the weights are functions of the unknown parameters. As always, OLS at the first step produces consistent but inefficient estimates. Then the weights,

$$\omega_i = \sqrt{\frac{N_i \phi_i^2}{\Phi_i(1-\Phi_i)}},$$

for the probit model and

$$\omega_i = \sqrt{N_i \Lambda_i (1-\Lambda_i)},$$

for the logit model can be evaluated on the first step estimates and used for WLS. This estimator has the same asymptotic properties of the MLE (<, >)amemiya.

Two complications arise in practice. The familiar result in (10.15) suggests that when the proportion is based on a large population, the variance

[6] The function normit + 5 is called the **probit** of P_i.

[7] GAUSS tip: the command `cdfni(p)` computes the inverse of the c.d.f. of the normal distribution.

of the estimator can be exceedingly low. This will resurface in implausibly low standard errors and high t ratios in the regression. Unfortunately, this is a consequence of the model. The same result will emerge in ML estimates based on proportions data. For grouped data, the log-likelihood is

$$\ell(\beta) = \sum_{i=1}^{N} n_i \left[P_i \ln \mathcal{F}(\beta' x_i) + (1 - P_i) \ln (1 - \mathcal{F}(\beta' x_i)) \right].$$

Second, both the MLE and the FGLS estimators break down if any of the proportions are 0 or 1. A number of ad hoc patches have been suggested; the one that seems to be most widely used is just to add or substract a small constant, say 0.001, from the observed value when it is 0 or 1.

10.3 Multinomial Response Models

Although many discrete dependent variables are binary, discrete variables that can take on three or more different values are by no means uncommon in economics. A variety of qualitative response models has been devised to deal with such cases. These fall into two types: models designed to deal with unordered responses and models designed to deal with ordered responses. An example of unordered response data would be the results from a survey of how people choose to commute to work. The possible responses might be: walk, bicycle, take the bus, drive, and take the subway. Although one could probably make the case for ordering these responses in certain ways, there is clearly no one natural way to do so. An example of ordered response data would be results from a survey where respondents are asked to say whether they strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree with some statement. Here there are five possible responses, which evidently can be ordered in a natural way.

10.3.1 Unordered Responses

The latent variable model (10.2) can be generalized for unordered response models. For the ith individual faced with J choices, suppose that the utility of choice j is

$$y_{i,j}^* = \beta' x_{i,j} + u_{i,j}.$$

If the consumer makes choice j in particular, we assume that $y^*_{i,j}$ is the maximum among the J utilities. Hence, the statistical model is driven by the probability that choice j is made, which is

$$\Pr\left(y^*_{i,j} > y^*_{i,k}\right) \quad \text{for all } k \neq j.$$

The model is made operational by a particular choice of distribution of the disturbances. As before, two models have been considered, logit and probit. Because of the need to evaluate multiple integrals of the normal distribution, the probit model has found rather limited use in this setting. Let y_i be a random variable indicating the choice made. If the J disturbances are i.i.d. with Weibull distribution

$$\mathcal{F}\left(u_{i,j}\right) = \exp\left(\exp\left(-u_{i,j}\right)\right),$$

then

$$\Pr\left(y_i = j\right) = \frac{e^{\beta' x_{i,j}}}{\sum_j e^{\beta' x_{i,j}}}.$$

There are two approaches that are used in order to estimate these models: the multinomial logit model and the conditional logit model. The first applies when the choices depend on individual specific attributes and the latter, when the choices depend on choice specific attributes.

An example of each can help us to clarify the concepts. Suppose that we are interested in evaluating the determinants of career choices for a sample of individuals; after conducting the survey, we decided to code the outcomes in some (arbitrary) fashion, like 0 for politicians, 1 for lawyers, 2 for painters, and so on. The set of regressors include the level of schooling, experience, gender, and other variables. As we assume that the choice of a profession depends solely in the characteristics of the individuals, we would use the multinomial logit model to tackle this problem. On the other hand, consider a model for shopping center choice by individuals. In this case, the choice may depend not only on the characteristics of the shopping centers (such as the number of stores) but also on the distance between the shopping center and the individual's home. In this case, each individual has different values of x for each choice. Given that this model depends on choice specific attributes, we model it using the conditional logit model.

Multinomial Logit

The multinomial logit model is designed to handle $J+1$ responses. According to this model, the probability that any one of them is observed is

$$\begin{aligned}\Pr(y_i = 0) &= \frac{1}{1+\sum_{j=1}^{J} e^{x_i'\beta^j}} \\ \Pr(y_i = l) &= \frac{e^{x_i'\beta^l}}{1+\sum_{j=1}^{J} e^{x_i'\beta^j}} \quad \text{for } l = 1, \ldots, J.\end{aligned}$$

Here β^1 trough β^J are $k-$vectors of parameters. When $J = 1$, it is easy to see that this model reduces to the ordinary logit model with a single index function $e^{x_i'\beta^1}$. For every additional alternative, another index function and k more parameters are added to the model.

Some authors prefer to write the multinomial logit model as

$$\Pr(y_i = l) = \frac{e^{x_i'\beta^l}}{\sum_{j=1}^{J} e^{x_i'\beta^j}} \quad \text{for } l = 0, \ldots, J,$$

by defining an extra parameter vector β^0, all elements of which are identically zero. This way of writing the model is more compact but does not make it as clear that the ordinary logit model is a special case of the multinomial one.

Estimation of the multinomial logit model is reasonably straightforward, since the log-likelihood function is globally concave. This log-likelihood function can be written as

$$\ell\left(\beta^1, \ldots, \beta^J\right) = \sum_{i=1}^{N}\sum_{j=0}^{J} d_{i,j} \ln\left[\Pr(y_i = j)\right],$$

where

$$d_{i,j} = \begin{cases} 1 \text{ if } j \text{ is chosen by individual } i \\ 0 \text{ otherwise} \end{cases}.$$

Inference can also be conducted in the usual fashion. As was the case with the ordinary logit model, always report $\ell\left(\widehat{\beta}^1, \ldots, \widehat{\beta}^j\right)$. If we wanted to test the hypothesis that all the slope coefficients are zero, we would obtain

$$\ell\left(\overline{\beta}^1, \ldots, \overline{\beta}^J\right) = \sum_{j=0}^{J} N_j \ln P_j,$$

where N_j is the number of individuals that made the choice j, and P_j is the proportion of observations that made the choice j ($P_j = N_j/N$).

Conditional Logit

The conditional logit model is designed to handle individual choice among J (not $J+1$) discrete alternatives, where one and only one of the alternatives can be chosen. According to this model, the probability that any one of them is observed is

$$\Pr\left(y_i = l\right) = \frac{e^{x'_{i,l}\beta}}{\sum_{j=1}^{J} e^{x'_{i,j}\beta}} \quad \text{for } l = 1, \ldots, J,$$

The log-likelihood of the conditional logit model is

$$\ell\left(\beta\right) = \sum_{i=1}^{N} \sum_{j=0}^{J} d_{i,j} \ln\left[\Pr\left(y_i = j\right)\right],$$

with $d_{i,j}$ being defined in the same fashion as in the case of the multinomial model. This function is also globally concave, and inference can be conducted as usual.

There are two key differences between the multinomial logit and conditional logit models. In the former, there is a single vector of independent variables for each observation, and there are J different vector parameters. In the latter, the values of the independent variables vary across alternatives, but there is just a single parameter vector β. The multinomial logit model is a straightforward generalization of the logit model that can be used to deal with any situation involving three or more unordered qualitative responses. In contrast, the conditional logit model is specifically designed to handle individual choices among discrete alternatives based on the characteristics of those alternatives.

Independence of Irrelevant Alternatives

One important property of the multinomial logit model is that

$$\frac{\Pr\left(y_i = k\right)}{\Pr\left(y_i = j\right)} = \frac{e^{x'_i\beta^k}}{e^{x'_i\beta^j}} = e^{x'_i\left(\beta^k - \beta^j\right)}, \tag{10.17}$$

for any two responses k and j (including response zero if we interpret β^0 as a vector of zeros). Thus the odds between any two responses depend solely on x_i and on the parameters associated with those two responses. They do not depend on the parameter vectors associated with any of the other responses. If fact, we see from (10.17) that the log of the odds between responses k and j is simply $x_i'\beta^*$, where $\beta^* = \beta^k - \beta^j$. Thus, conditional on either k or j being chosen, the choice between them is determined by an ordinary logit model with parameter vector β^*.

In the case of the conditional logit we have

$$\frac{\Pr(y_i = k)}{\Pr(y_i = j)} = \frac{e^{x'_{i,k}\beta}}{e^{x'_{i,j}\beta}} = e^{\beta'\left(x_{i,k} - x_{i,j}\right)}.$$

This property, which is analogous to (10.17), is called the independence of irrelevant alternatives (IIA) property. It implies that adding another alternative to the model, or changing the characteristics of another alternative that is already included, will not change the odds between alternative k and j.

The IIA property can be extremely implausible in certain circumstances. Suppose that there are initially two alternatives for travelling between two cities: flying Monopoly Airways and driving. Suppose further that half of all travelers fly and the other half drive. Then Upstart Airways enters the market and creates a third alternative. If Upstart offers a service identical to that of Monopoly, it must gain the same market share. Thus according to the IIA property, one third of the travelers must take each of the airlines and one third must drive. So the automobile has lost just as much market share from the entry of Upstart Airways as Monopoly Airways has! This seems very implausible. As a result, efforts have been devoted to the problem of testing the IIA property and finding tractable models that do not embody it. These models relax the assumption of independent disturbances, which is the source of the IIA property.

10.3.2 Ordered Responses

Previously, it did not matter how we coded each alternative because there was no ordinal representation between them. In this case, we cannot use the multinomial logit or probit, because the alternatives have an ordinal representation. Of course, OLS cannot be used because the codes for the alternatives represent rankings, but have no cardinal interpretation.

The most common way to deal with ordered response data is to use an ordered QR model, usually either the ordered probit or the ordered logit model. The model is motivated around the same structure as before; that is, consider the latent variable model

$$y_i^* = \beta' x_i + u_i,$$

where u is i.i.d. with c.d.f. $\mathcal{F}(\cdot)$, and, for reasons that will soon become evident x_i does not include a constant term. What we actually observe is a discrete variable that can take only $J + 1$ values:

$$y_i = \begin{cases} 0 \ \text{ if } y_i^* < a_0 \\ 1 \ \text{ if } a_0 \leq y_i^* < a_1 \\ \vdots \qquad \vdots \\ J \ \text{ if } a_{J-1} \leq y_i^* \end{cases}.$$

The parameters of this model are β and $a = (a_0, \cdots, a_{J-1})$. The a_j's are thresholds that determine what value of y a given value of y^* will map into. This is illustrated in Figure 10.1 for the case of three choices. The number of elements in a is always one fewer than the number of choices. When there are only two choices, this model becomes indistinguishable from an ordinary binary response model, with the single element of a playing the role of the constant term.

The probability that $y_i = 0$ is

$$\begin{aligned} \Pr(y_i = 0) &= \Pr(y_i^* < a_0) = \Pr(\beta' x_i + u_i < a_0) \\ &= \Pr(u_i < a_0 - \beta' x_i) \\ &= \mathcal{F}(a_0 - \beta' x_i). \end{aligned}$$

Similarly, for any $j > 0$ but different from J, we have

$$\begin{aligned} \Pr(y_i = j) &= \Pr(a_{j-1} \leq y_i^* < a_j) = \Pr(a_{j-1} \leq \beta' x_i + u_i < a_j) \\ &= \Pr(u_i < a_j - \beta' x_i) - \Pr(u_i \leq a_{j-1} - \beta' x_i) \\ &= \mathcal{F}(a_j - \beta' x_i) - \mathcal{F}(a_{j-1} - \beta' x_i). \end{aligned}$$

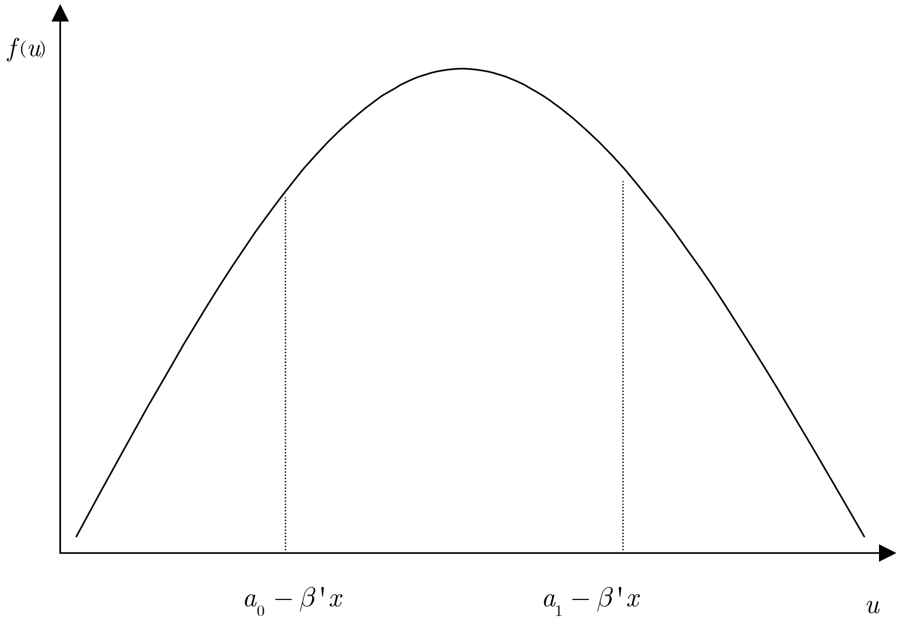

Figure 10.1: Probabilities with ordered observations

Finally, the probability that $y_i = J$ is

$$\begin{aligned}\Pr(y_i = J) &= \Pr(y_i^* \geq a_{J-1}) = \Pr(\beta' x_i + u_i \geq a_{J-1}) \\ &= \Pr(u_i \geq a_{J-1} - \beta' x_i) \\ &= 1 - \mathcal{F}(a_{J-1} - \beta' x_i).\end{aligned}$$

If $\mathcal{F}(\cdot)$ is symmetric, this last expression reduces to $\mathcal{F}(\beta' x_i - a_{J-1})$. Once we impose a distributional assumption for u, we replace $\mathcal{F}(\cdot)$ with the respective c.d.f. For example, the ordered probit model uses $\Phi(\cdot)$ and the ordered logit uses $\Lambda(\cdot)$.

The log-likelihood function is

$$\ell(\beta, a) = \sum_{i=1}^{N} \sum_{j=0}^{J} d_{i,j} \ln[\Pr(y_i = j)],$$

with $d_{i,j}$ being defined as in the case of the multinomial model. This function is also globally concave, and inference can be conducted in the usual fashion. However, remember that maximization of this objective function has to be done with respect to β and a. In order for all of the probabilities to be positive, we must have

$$a_0 < a_1 < \cdots < a_{J-1}.$$

10.4 Multivariate Probit Models

A natural extension of the probit model would be to allow more than one equation with correlated disturbances.[8] The general specification for a two-equation model would be

$$\begin{aligned} y_i^* &= \beta' x_i + u_i \qquad y_i = 1 \text{ if } y_i^* > 0, \ 0 \text{ otherwise} \\ z_i^* &= \delta' w_i + v_i \qquad z_i = 1 \text{ if } z_i^* > 0, \ 0 \text{ otherwise} \\ \mathcal{E}(u_i) &= \mathcal{E}(v_i) = 0 \\ \mathcal{V}(u_i) &= \mathcal{V}(v_i) = 1 \\ \operatorname{Cov}(u_i, v_i) &= \rho. \end{aligned}$$

To construct the log-likelihood, we will use a useful shorthand. Let

$$\begin{aligned} q_i &= 2y_i - 1, \\ r_i &= 2z_i - 1. \end{aligned}$$

Thus $q_i = 1$ if $y_i = 1$ and -1 if $y_i = 0$ (an analogous expression can be defined for r as a function of z). The bivariate normal c.d.f. is

$$\Pr(Y < y, Z < z) = \int_{-\infty}^{z} \int_{-\infty}^{y} \phi_2(t_1, t_2, \rho)\, dt_1\, dt_2,$$

which we denote $\Phi_2(y, z, \rho)$. The density is

$$\phi_2(y, z, \rho) = \frac{e^{-0.5\left(y^2+z^2-2\rho yz\right)/\left(1-\rho^2\right)}}{2\pi\left(1-\rho^2\right)^{1/2}}.$$

Now let

$$\begin{aligned} s_i &= q_i \beta' x_i, \\ t_i &= r_i \delta' w_i, \\ \rho_i^* &= q_i r_i \rho. \end{aligned}$$

[8] Consider for example the choice of having children and working.

Note the notational convention. The subscript 2 is used to indicate the bivariate normal density, ϕ_2 and c.d.f., Φ_2.

The probabilities that enter the likelihood function are

$$\Pr\left(Y = y_i, Z = z_i\right) = \Phi_2\left(s_i, t_i, \rho_i^*\right).$$

This accounts for all the necessary sign changes needed to compute probabilities for y and z equal to zero and one. Thus

$$\ell\left(\beta, \delta, \rho\right) = \sum_{i=1}^{N} \ln \Phi_2\left(s_i, t_i, \rho_i^*\right).$$

Maximization of this objective function can be performed numerically and inference is conducted in the usual fashion. An interesting hypothesis to check is $\rho = 0$, that is if u and v are uncorrelated. If this hypothesis is not rejected, we do not need to estimate the bivariate model, but two independent probit models, given that the choice variables are not correlated.

Extensions of this model to handle more than 2 outcome variables can be done by adding more equations. The practical obstacle to such an extension is the evaluation of higher order multivariate normal integrals. Some progress has been made on trivariate integration, but existing results are not sufficient to allow accurate and efficient evaluation for more than two variables in a sample of even moderate size. This is a reason why simulation methods for estimation of this type of models are increasingly popular.

10.5 Count Data

Assume that you are interested in modelling the determinants of the number of accidents of different models of cars or the number of orders placed by different customers. Even though these variables have a cardinal interpretation (they are quantitative random variables), their nature is discrete. Thus, even tough LS is an option, we would like to consider a model that accounts for these characteristics. The Poisson regression model has been widely used to study such data.

The model stipulates that each y_i is drawn from a Poisson distribution with parameter λ_i, which is related to the regressors x_i. The primary equation of the model is

$$\Pr\left(Y_i = y_i\right) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}, \qquad y_i = 0, 1, 2, \ldots$$

The most common formulation for λ_i is

$$\ln \lambda_i = \beta' x_i.$$

It is easily shown that

$$\mathcal{E}\left[y_i \left| x_i\right.\right] = \mathcal{V}\left[y_i \left| x_i\right.\right] = \lambda_i = e^{\beta' x_i},$$

so

$$\frac{\partial \mathcal{E}\left[y_i \left| x_i\right.\right]}{\partial x_i} = \lambda_i \beta.$$

With the parameter estimates in hand, this can be computed using any data vector desired.

In principle, the Poisson model is simply a nonlinear regression. But it is easier to estimate the parameters with ML techniques. The log-likelihood function is

$$\ell\left(\beta\right) = \sum_{i=1}^{N} \left[-\lambda_i + y_i \ln \lambda_i - \ln y_i!\right].$$

The FONC are

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{N} \left[y_i - \lambda_i\right] x_i.$$

The Hessian is

$$\frac{\partial^2 \ell}{\partial \beta \partial \beta'} = -\sum_{i=1}^{N} \lambda_i x_i x_i',$$

which is negative definite for all x and β, implying that the log-likelihood is globally concave. Estimation and inference can be conducted in the usual fashion.

The Poisson model has been criticized because of its implicit assumption that the variance of y equals its mean, and a number of extensions that relax this assumption have been proposed.

10.6 Truncated and Censored Data

Continuous limited dependent variables are designed to handle samples that have been **truncated** or **censored** in some way. These two terms are easily confused. A sample has been truncated if some observations that should have been there are systematically excluded from the sample. For example, a sample of households with incomes under \$200,000 necessarily excludes all households with income over that level. It is not a random sample of all households. Thus, truncation occurs before a sample is taken, meaning that we eliminate a subset of the support of the distribution so that no observations are sampled from this discarded subset.

On the other hand, a sample has been censored if no observations have been systematically excluded, but some of the information contained in them has been suppressed. To continue with the previous example, suppose that households with all income levels are included in the sample, but for those with income in excess of \$200,000, the amount reported is always \$200,000. In this case, the censored sample is still a random sample of all households, but the values reported for high-income households are not the true values.

Econometricians have devised a large number of models for dealing with truncated and censored data. Here we discuss some of them.

10.6.1 Models for Truncated Data

Here, we are concerned with inferring the characteristics of a population from a sample drawn from a restricted part of that population.

Truncated Distributions

For our purposes, a truncated distribution is part of an untruncated distribution that is above or below some specified value.

Theorem 113 (Density of a truncated random variable) *If a continuous random variable, x, has p.d.f. $f(x)$ and b is a constant,*

$$f(x\,|x>b) = \frac{f(x)}{\Pr(x>b)}.$$

Proof. This follows directly from the definition of conditional probability. ■

This amounts to scaling the density so that it integrates to one over the range above b.

Most applications use the truncated normal distribution. If x has a normal distribution with mean μ and standard deviation σ,

$$\Pr(x>b) = 1-\Phi\left(\frac{b-\mu}{\sigma}\right) = 1-\Phi(\alpha),$$

where

$$\alpha = \frac{b-\mu}{\sigma}.$$

The truncated normal distribution is, then,

$$\begin{aligned} f\left(x \left| x > b\right.\right) &= \frac{f\left(x\right)}{1-\Phi\left(\alpha\right)} \\ &= \frac{\left(1/\sigma\right)\phi\left(\left(x-\mu\right)/\sigma\right)}{1-\Phi\left(\alpha\right)}. \end{aligned}$$

The standard normal and the truncated standard normal distributions are compared in Figure 10.2 (for $b = -0.5$).

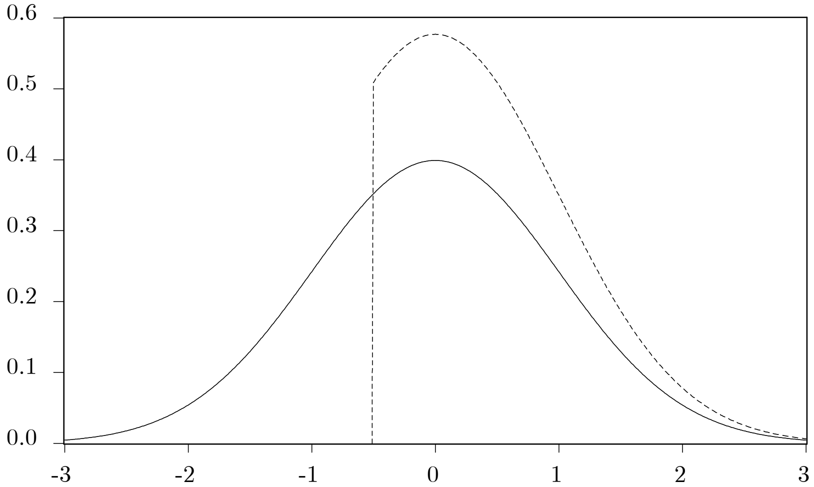

Figure 10.2: Truncated normal distribution

Moments of Truncated Distributions

We are usually interested in the mean and variance of the truncated random variable. These would be obtained in the usual fashion:

$$\mathcal{E}\left(x \left| x > b\right.\right) = \int_b^\infty x f\left(x \left| x > b\right.\right) dx,$$

for the mean and likewise for the variance.

If the truncation is from below, the mean of the truncated variable is greater than the mean of the original one. If the truncation is from above, the mean of the truncated variable is smaller than the mean of the original one. On the other hand, truncation reduces the variance compared to the variance of the untruncated distribution.

For the truncated normal distribution, we have:

Theorem 114 (Moments of the truncated normal distribution) *If* $x \backsim \mathcal{N}\left(0, \sigma^2\right)$ *and* b *is a constant,*

$$\begin{aligned} \mathcal{E}\left(x \left| truncation \right.\right) &= \mu + \sigma\lambda\left(\alpha\right), \\ \mathcal{V}\left(x \left| truncation \right.\right) &= \sigma^2\left(1 - \delta\left(\alpha\right)\right). \end{aligned}$$

where

$$\begin{aligned} \lambda\left(\alpha\right) &= \frac{\phi\left(\alpha\right)}{1 - \Phi\left(\alpha\right)} \quad \textit{if truncation is } x > b, \\ \lambda\left(\alpha\right) &= \frac{-\phi\left(\alpha\right)}{\Phi\left(\alpha\right)} \quad \textit{if truncation is } x < b, \\ \delta\left(\alpha\right) &= \lambda\left(\alpha\right)\left[\lambda\left(\alpha\right) - \alpha\right]. \end{aligned}$$

An important result is

$$0 < \delta\left(\alpha\right) < 1, \quad \text{for all values of } \alpha.$$

A result that we will use at several points below is

$$\frac{\partial\phi\left(\alpha\right)}{\partial\alpha} = -\alpha\phi\left(\alpha\right).$$

The function $\lambda\left(\alpha\right)$ is called the **inverse Mills ratio**.

The Truncated Regression Model

Consider the model

$$y_i = \beta' x_i + u_i, \quad u_i \backsim \mathcal{N}\left(0, \sigma^2\right),$$

so that

$$y_i \left| x_i \right. \backsim \mathcal{N}\left(\beta' x_i, \sigma^2\right).$$

We are interested in the distribution of y_i given that y_i is greater than the truncation point, b. This is precisely the model described before. It follows that

$$\mathcal{E}\left[y_i\left|y_i > b\right.\right] = \beta' x_i + \sigma\lambda\left(\alpha_i\right),$$

where now $\alpha_i = \left(b - \beta' x_i\right)/\sigma$. Then

$$\begin{aligned}
\frac{\partial\mathcal{E}\left[y_i\left|y_i > b\right.\right]}{\partial x} &= \beta + \sigma\left(\frac{\partial\lambda_i}{\partial\alpha_i}\right)\frac{\partial\alpha_i}{\partial x} \\
&= \beta + \sigma\left(\lambda_i^2 - \alpha_i\lambda_i\right)\left(-\frac{\beta}{\sigma}\right) \\
&= \beta\left(1 - \lambda_i^2 + \alpha_i\lambda_i\right) \\
&= \beta\left(1 - \delta\left(\alpha_i\right)\right).
\end{aligned}$$

Since the truncated variance is between zero and one, we conclude that for every element of x, the marginal effect is less than the corresponding coefficient. There is similar attenuation of the variance. In the subpopulation $y_i > b$, the regression variance is not σ^2 but

$$\mathcal{V}\left[y_i\left|y_i > b\right.\right] = \sigma^2\left(1 - \delta\left(\alpha_i\right)\right).$$

Whether the marginal effect or the coefficient β itself is of interest depends on the intended inferences of the study. If the analysis is to be confined to the subpopulation, the marginal effect is of interest. However, if the study is intended to extend to the entire population, it is the coefficients β that are actually of interest.

We now consider the estimation of the parameters of the truncated regression. One's first inclination might be to use OLS. For the subpopulation from which the data is drawn, we see that

$$y_i\left|y_i > b\right. = \beta' x_i + \sigma\lambda\left(\alpha_i\right) + v_i,$$

where v_i is y_i minus its conditional expectation. By construction, v_i has a zero mean, but it is heteroskedastic,

$$\mathcal{V}\left(v_i\right) = \sigma^2\left(1 - \delta\left(\alpha_i\right)\right),$$

which is a function of x_i. If we estimate a model regressing y on x, we have omitted a variable, the nonlinear term $\lambda\left(\alpha_i\right)$. All of the biases that arise because of an omitted variable can be expected. Furthermore, the OLS estimator will also be inconsistent.

Given that the density of y conditional on the value of $y > b$, is:

$$f\left(y_i \left| y_i > b\right.\right) = \frac{\frac{1}{\sigma}\phi\left(\frac{y_i - \beta' x_i}{\sigma}\right)}{1 - \Phi\left(\frac{b - \beta' x_i}{\sigma}\right)}.$$

Thus,

$$\begin{aligned} \ell\left(\beta, \sigma^2\right) &= -\frac{N}{2}\left(\ln\left(2\pi\right) + \ln\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i - \beta' x_i\right)^2 \\ &\quad -\sum_{i=1}^{N}\ln\left[1 - \Phi\left(\frac{b - \beta' x_i}{\sigma}\right)\right]. \end{aligned}$$

Maximization, while rather involved because of the extreme nonlinearity of the function, is straightforward. Inference is conducted in the usual fashion.

10.6.2 Models of Censored Data

A very common problem in microeconomic data is censoring of the dependent variable. When the dependent variable is censored, values in a certain range are all transformed to a single value. Consider for example the case of estimating the demand for tickets at a certain stadium. Our only measure is the number actually sold. However, whenever an event sells out, we know that the actual number demanded is larger than the number sold. The number of tickets demanded is censored when it is transformed to obtain the number sold.

Other examples are: household purchase of durable goods, number of extramarital affairs, number of hours worked, and number of arrests after release of prison. Each of these examples analyzes a dependent variable that is zero for a significant fraction of the observations. Conventional regression methods fail to account for the qualitative difference between limit (zero) observations and nonlimit (continuous) observations.

The Censored Normal Distribution

The relevant distribution theory for a censored variable is similar to that for a truncated one. Once again, we focus on the normal distribution. We also assume that the censoring point is zero, though this is only a convenient normalization.

In a truncated distribution, only the part of the distribution above $y = 0$ is relevant to our computations. To make the distribution integrate to one, we scale it up by the probability that an observation in the untruncated population falls in the range that interests us. When data are censored, the distribution that applies to the sample is a mixture of discrete and continuous distributions.

To analyze this distribution, we define a new random variable, y, transformed from the original one, y^*, by

$$y_i = \begin{cases} 0 & \text{if } y_i^* \leq 0 \\ y_i^* & \text{if } y_i^* > 0 \end{cases}.$$

The distribution that applies if $y^* \backsim \mathcal{N}(\mu, \sigma^2)$ is

$$\begin{aligned} \Pr(y = 0) &= \Pr(y^* \leq 0) \\ &= \Phi\left(-\frac{\mu}{\sigma}\right) \\ &= 1 - \Phi\left(\frac{\mu}{\sigma}\right), \end{aligned}$$

and if $y^* > 0$, y has the density of y^*.

This distribution is a mixture of discrete and continuous parts. The total probability is one, as required, but instead of scaling the second part, we simply assign the full probability in the censored region to the censoring point, in this case, zero.

Theorem 115 (Moments of the censored normal variable) *If $y^* \backsim \mathcal{N}(\mu, \sigma^2)$ and $y = b$ if $y^* \leq b$ else $y = y^*$ then*

$$\mathcal{E}[y] = \Phi b + (1 - \Phi)(\mu + \sigma\lambda)$$

and

$$\mathcal{V}[y] = \sigma^2 (1 - \Phi)\left[(1 - \delta) + (\alpha - \lambda)^2 \Phi\right],$$

where $\Phi[(b - \mu)/\sigma] = \Phi(\alpha) = \Pr(y^* \leq b) = \Phi$, $\lambda = \phi/(1 - \phi)$ and $\delta = \lambda^2 - \lambda\alpha$.

Proof. Note that

$$\begin{aligned}\mathcal{E}(y) &= \mathcal{E}(y\,|y=b)\Pr(y=b)+\mathcal{E}(y\,|y>b)\Pr(y>b)\\ &= b\Phi+(1-\Phi)(\mu+\sigma\lambda) \quad \text{(using Theorem 114).}\end{aligned}$$

For the variance, we use $\mathcal{V}(y)=\mathcal{E}$ [conditional variance]$+\mathcal{V}$ [conditional mean] and Theorem 114 to obtain $\mathcal{V}[y]=\sigma^2(1-\Phi)\left[(1-\delta)+(\alpha-\lambda)^2\Phi\right]$. ■

For censoring of the upper part of the distribution instead of the lower, it is only necessary to reverse the role of Φ and $1-\Phi$ and redefine λ as in Theorem 114.

The Censored Regression Model (Tobit)

The regression model based on the preceding discussion is referred to as the censored regression model or the Tobit model.[9] The regression is obtained by making the mean in the preceding correspond to a classical regression model. The general formulation is usually given in terms of an index function,

$$\begin{aligned}y_i^* &= \beta' x_i+u_i,\\ y_i &= \begin{cases}0 & \text{if } y_i^*\le 0\\ y_i^* & \text{if } y_i^*>0\end{cases}.\end{aligned}$$

Assuming normality on u, the log-likelihood function is given by

$$\ell(\beta,\sigma)=\sum_{y_i=0}\ln\left[1-\Phi\left(\frac{\beta' x_i}{\sigma}\right)\right]+\sum_{y_i>0}\ln\left[\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(y_i-\beta' x_i)^2}{2\sigma^2}}\right]. \qquad (10.18)$$

Several features of this function are worth mentioning: First, the second part of the log-likelihood resembles OLS in the observations that are not censored, while the first resembles the probit model. Second, unlike the probit model where the normalization $\sigma=1$ was harmless, in this case we must estimate σ. The reason is simple, in the probit model, β and σ cannot be identified separately, in the Tobit model they can. Third, if the Tobit structure is correct, OLS estimates will be inconsistent. Finally, (10.18) is globally concave, thus maximization and inference can be conducted in the usual fashion.

[9] This is in reference to Tobin, who first proposed the model.

10.6.3 Incidental Truncation

In Section 10.6.1, we discussed models in which the sample was truncated according to the value of the dependent variable. In many practical cases, however, truncation is based not on the value of the dependent variable but rather on the value of another variable that is correlated with it. For example, people may choose to enter the labor force only if their market wage exceeds their reservation wage. Then a sample of people who are in the labor force will exclude those whose reservation wage exceeds their market wage. If the dependent variable is anything that is correlated with either reservation or market wages, OLS will yield inconsistent estimates. In this case, the sample may be said to have been selected on the difference between market and reservation wages, and the problem that this type of selection causes is often referred to as **sample selectivity bias**.[10]

Selectivity refers to the presence of some characteristic of the treatment or control group that is associated with both the recipient of the treatment and its outcome, so as to lead to a false attribution of causality regarding treatment and causality. For example, suppose that we want to measure the quality of schools by post-school wages. Furthermore, assume that the econometrician does not have information regarding family background. What would happen to our results if some school administrators do not admit students randomly? For example, suppose that there is a selective admission process in some schools (only students from wealthy families are admitted). If these students are also more likely to be wealthy for reasons other than superior schooling, then this may lead the econometrician to attribute the effect of family background of a student to the actual effect that the training received on a particular school had on his post-school wages.

The best way to understand the key features of models involving sample selectivity is to examine a simple model in some detail. Suppose that y^* and z^* are two latent variables, generated by the bivariate process

$$\begin{bmatrix} y_i^* \\ z_i^* \end{bmatrix} = \begin{bmatrix} \beta' x_i \\ \gamma' w_i \end{bmatrix} + \begin{bmatrix} u_i \\ v_i \end{bmatrix}, \tag{10.19}$$

[10] It is curious that the problem is often times associated with biases, given that the only characteristic that we will hope that our estimators meets is consistency.

where

$$\begin{bmatrix} u_i \\ v_i \end{bmatrix} \backsim \mathcal{N}\left(0, \begin{bmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{bmatrix}\right).$$

The variables that are actually observed are y and z, and they are related to y^* and z^* as follows:

$$\begin{aligned} y_i &= y_i^* \text{ if } z_i^* > 0;\ y_i = 0 \text{ otherwise;} \\ z_i &= 1 \quad \text{if } z_i^* > 0;\ z_i = 0 \text{ otherwise.} \end{aligned}$$

There are two types of observations: ones for which both y and z are observed to be zero and ones for which $z = 1$ and $y = y^*$. The log-likelihood function for this model is thus

$$\sum_{z_i=0} \ln\left(\Pr\left[z_i = 0\right]\right) + \sum_{z_i=1} \ln\left(\Pr\left[z_i = 1\right] f\left(y_i^* \left| z_i = 1\right.\right)\right).$$

The first term is the summation over all observations for which $z = 0$ of the logarithms of the probability that $z = 0$. It is exactly the same as the corresponding term in a probit model for z by itself. The second term is the summation over all observations for which $z = 1$ of the probability that $z = 1$ times the density of y conditional on $z = 1$. Using the fact that we can factor a joint density any way we like, this second term can also be written as

$$\sum_{z_i=1} \ln\left(\Pr\left[z_i = 1 \left| y_i^* \right.\right] f\left(y_i^*\right)\right),$$

where $f\left(y_i^*\right)$ is the unconditional density of y^*, which is just a normal density with conditional mean $\beta' x$ and variance σ^2.

The only difficulty writing out the log-likelihood function explicitly is to calculate $\Pr\left[z_i = 1 \left| y_i^* \right.\right]$. Since u and v are bivariate normal, we can write

$$z_i^* = \gamma' w_i + \rho\left(\frac{y_i^* - \beta' x_i}{\sigma}\right) + h_i,$$

where

$$h_i \backsim \mathcal{N}\left(0, 1 - \rho^2\right).$$

It follows that

$$\Pr\left(z_i = 1\right) = \Phi\left(\frac{\gamma' w_i + \rho\left(\frac{y_i - \beta' x_i}{\sigma}\right)}{\sqrt{1 - \rho^2}}\right),$$

since $y = y^*$ when $z = 1$. Thus, the log-likelihood function becomes

$$\sum_{z_i=0} \ln\left(\Phi\left(\gamma' w_i\right)\right) + \sum_{z_i=1} \ln\left(\frac{1}{\sigma}\phi\left(\beta' x_i\right)\right) \qquad (10.20)$$
$$+ \sum_{z_i=0} \ln\left(\Phi\left(\frac{\gamma' w_i + \rho\left(\frac{y_i - \beta' x_i}{\sigma}\right)}{\sqrt{1-\rho^2}}\right)\right).$$

The first term looks like the corresponding term for a probit model. The second term looks like the log-likelihood function for a linear regression model with normal errors. The third term is one that we have not seen before.

ML estimates can be obtained in the usual way by maximizing (10.20). However, this maximization is relatively burdensome, and so instead of ML estimation a computationally simpler technique proposed by Heckman is often used. **Heckman's two-step method** is based on the fact that (10.19) can be rewritten as

$$y_i^* = \beta' x_i + \rho\sigma v_i + u_i.$$

The idea is to replace y^* by y and v by its mean conditional on $z = 1$ and on the realized value of $\gamma' w$. As we have seen, this conditional mean is the inverse Mills ratio. Hence, we have

$$y_i = \beta' x_i + \rho\sigma\lambda_i + \text{residual}_i. \qquad (10.21)$$

It is now easy to see how Heckman's two-step method works. In the first step, an ordinary probit model is used to obtain consistent estimates $\widehat{\gamma}$ of the parameters of the selection equation. In the second step, the selectivity regressor,

$$\lambda_i = -\frac{\phi\left(\gamma' w_i\right)}{\Phi\left(\gamma' w_i\right)},$$

is evaluated at $\widehat{\gamma}$ and regression (10.21) is estimated by OLS. This regression provides a test for sample selectivity as well as an estimation technique. The coefficient on the selectivity regressor is $\rho\sigma$. Since $\sigma \neq 0$, the ordinary t statistic for this coefficient to be zero can be used to test the hypothesis that $\rho = 0$; it will be asymptotically distributed as $\mathcal{N}(0,1)$ under the null hypothesis. Thus, if this coefficient is not significantly different from zero, the investigator may reasonably decide that selectivity is not a problem for this data set and proceed to use OLS as usual.

Even when the hypothesis $\rho = 0$ is rejected, OLS estimation of (10.21) yields consistent estimates of β. However, the OLS covariance matrix is valid only when $\rho = 0$. When selectivity is present, there are actually two problems. First, the residuals in (10.21) will be heteroskedastic. Second, the selectivity regressor is being treated like any other regressor, when it is in fact part of the error term. One could solve the first problem by using a heteroskedasticity-consistent covariance matrix estimator, but that would not solve the second problem. It is possible to obtain a valid covariance matrix estimate to go along with the two-step estimates of β from (10.21). However, the calculation is cumbersome, and the estimated covariance matrix is not always positive definite (see Greene, 1993 for details).

Although the two-step method for dealing with sample selectivity is widely used, it is preferable to use regression (10.21) only as a procedure for testing the null hypothesis that selectivity bias is not present. When the null is rejected, ML estimation based on (**??**) should be used in preference to the two-step method, unless it is computationally prohibitive.

10.7 Models for Duration Data

Intuition might suggest that the longer a strike persists, the more likely it is that it will end within, say, next week. Or is it? It seems equally plausible that the longer a strike has lasted, the more difficult must be the problems that led to it in the first place, and hence the less likely it is that it will end in the next short time interval. A similar kind of reasoning could be applied to spells of unemployment, time until business failure, length of time between arrests, intervals between purchases, and so on. In each of these cases, it is not only the duration of the event, per se, which is interesting, but also the likelihood that the event will end in "the next period" given that it has lasted as long as it has.

This section will give a brief introduction to the econometric analysis of duration data (also known as longitudinal data). The variable of interest in the analysis of duration is the length of time from the beginning of some event either until it ends or until the measurement is taken, which may precede termination. Censoring is a pervasive and usually unavoidable problem in the analysis of duration data. The common cause is that the measurement is made while the process is ongoing. An obvious example can be drawn from medical research. Consider analyzing the survival times of heart transplant patients. Although the beginning times may be known with precision, at the time of the measurement, observations on any individual who are still alive are necessarily censored. Likewise, samples of spells of unemployment drawn from surveys will probably include some individuals who are unemployed at the time the survey is taken. For these individuals, duration, or survival, is at least the observed time but not equal to it. Estimation must account for the censored nature of the data. The consequences of ignoring censoring in duration data are not unlike those which arise in regression analysis.

10.7.1 Parametric Models of Duration

The variable of interest in the analysis of duration is the length of time that elapsed from the beginning of some event either until it ends or until the measurement is taken, which may precede termination. Observation will typically consist of a cross-section of durations, $t_1, t_2, \ldots, t_N$.

We'll use the term "spell" as a catchall for the different duration variables we might measure. Spell length is represented by the random variable T. A simple approach to duration analysis would be to apply regression analysis to the sample of observed spells. By this device, we could characterize the expected duration, perhaps conditioned on a set of covariates whose values were measured at the end of the period. We could also assume that conditioned on an x which has remained fixed from $T = 0$ to $T = t$, t has a normal distribution, as we commonly do in regression. We could then characterize the probability distribution of observed duration times. But normality is not particularly attractive for a number of reasons, not least of which is that duration is positive by construction, while a normally distributed variable can take negative values.

Theoretical Background

Suppose that the random variable T has a continuous p.d.f., $f(t)$, where t is a realization of T. The c.d.f. is

$$\mathcal{F}(t) = \int_0^t f(s)\ ds = \Pr[T \leq t].$$

We will usually be more interested in the probability that the spell is of length at least t, which is given by the survival function

$$\mathcal{S}(t) = 1 - \mathcal{F}(t) = \Pr[T \geq t].$$

Consider the question raised in the introduction of this section, which is, given that the spell has lasted until time t, what is the probability that it will end in the next short interval of time, say Δ? This is

$$l(t,\Delta) = \Pr\left[t \le T \le t+\Delta \,|T \ge t\right].$$

A useful function for characterizing this aspect of the distribution is the hazard rate,

$$\begin{aligned} \lambda(t) &= \lim_{\Delta\to 0} \frac{l(t,\Delta)}{\Delta} \\ &= \lim_{\Delta\to 0} \frac{\mathcal{F}(t+\Delta) - \mathcal{F}(t)}{\Delta \mathcal{S}(t)} \\ &= \frac{f(t)}{\mathcal{S}(t)}. \end{aligned}$$

Roughly, the hazard rate is the rate at which spells are completed after duration t, given that they last at least until t. As such, the hazard function gives an answer to our original question.

Given the way the question was posed at the outset, we might prefer to model the hazard function rather than the density, the c.d.f., or the survival function. Clearly, all four functions are related. The hazard function is

$$\lambda(t) = \frac{-\partial \ln \mathcal{S}(t)}{\partial t},$$

and

$$f(t) = \mathcal{S}(t)\,\lambda(t).$$

Another useful function is the integrated hazard function

$$\Lambda(t) = \int_0^t \lambda(s)\, ds,$$

for which

$$\mathcal{S}(t) = e^{-\Lambda(t)},$$

so

$$\Lambda(t) = -\ln \mathcal{S}(t).$$

Models of the Hazard Rate

Given the above relationships, the hazard function is more interesting than the survival rate or the density, because we can use the hazard function to obtain the density. For example, consider the case in which the hazard rate does not vary over time (that is, $\lambda(t) = \lambda$). This is characteristic of a process that has no memory; the conditional probability of "failure" in a given short interval is the same regardless of when the observation is made. From the earlier definition, we obtain the simple differential equation

$$\frac{-d \ln \mathcal{S}(t)}{dt} = \lambda$$

The solution is

$$\ln \mathcal{S}(t) = c - \lambda t$$

or

$$\mathcal{S}(t) = Ce^{-\lambda t}$$

where C is the constant of integration. The condition $\mathcal{S}(0) = 1$ implies that $C = 1$, and the solution is

$$\mathcal{S}(t) = e^{-\lambda t}$$

This is the exponential distribution, which has been used to model the time until failure of electronic components, precisely because of the memoryless property of the distribution. Estimation of λ is simple, since with the exponential distribution $\mathcal{E}(t) = 1/\lambda$. The MLE of λ would be $1/\overline{t}$.

A natural extension might be to model the hazard rate as a linear function, $\lambda(t) = \alpha + \beta t$. Then, $\Lambda(t) = \alpha t + 0.5\beta t^2$ and $f(t) = \lambda(t)\exp(-\Lambda(t))$. With an observed sample of durations, estimation of α and β is, at least in principle, a straightforward problem in ML.[11]

A distribution whose hazard function slopes upward (downward) is said to have positive (negative) duration dependence. For such distributions, the likelihood of failure at time t, conditional upon duration up to time t, is increasing (decreasing) in t. Our question in the beginning of this section about whether the strike is more or less likely to end at time t given that it lasted until time t can be framed in terms of positive or negative dependence. The assumed distribution has a considerable bearing on the answer. If one is

[11] To avoid a negative hazard function, one might depart from $\lambda(t) = \exp(g(t, \theta))$, where θ is the vector of parameters to be estimated.

unsure at the outset of the analysis whether the data can be characterized by positive or negative duration dependence, it is counterproductive to assume a distribution that displays one characteristic or the other over the entire range of t. Thus, the exponential distribution and our suggested extension could be problematic.

The literature contains several choices of duration models, including normal, inverse normal (inverse Gaussian), lognormal, F, gamma, Weibull (which is a popular choice) and many others. To illustrate the differences, we will examine a few of the simpler ones. Table 10.2 lists the hazard functions and survival functions for three commonly used distributions.

Distribution	**Hazard Function**	**Survival Function**	**Median Duration**
Exponential	λ	$e^{-\lambda t}$	$\ln 2/\lambda$
Weibull	$\lambda p\,(\lambda t)^{p-1}$	$e^{-(\lambda t)^{p}}$	$(1/\lambda)\,(\ln 2)^{1/p}$
Log-logistic	$\lambda p\,(\lambda t)^{p-1}/\left[1+(\lambda t)^{p}\right]$	$1/\left[1+(\lambda t)^{p}\right]$	$1/\lambda$

Table 10.2: Some survival distributions

All of these are distributions for a nonnegative random variable. Their hazard functions display very different behaviors. The hazard function of the exponential distribution is constant, that for the Weibull is monotonically increasing or decreasing depending on p,[12] and the hazard for the log-logistic distribution first increases and then decreases (see Figure 10.3).

The parameters of these models can be estimated by ML. The log-likelihood function is

$$\ell = \sum_{\text{uncensored observations}} \ln f\,(t) + \sum_{\text{censored observations}} \ln \mathcal{S}\,(t)\,.$$

For some distributions, it is convenient to formulate the log-likelihood function as

$$\ell = \sum_{\text{uncensored observations}} \ln \lambda\,(t) + \sum_{\text{all observations}} \ln \mathcal{S}\,(t)\,.$$

Inference about the parameters can be done in the usual way. The estimate of the median of the survival distribution depends on the model, and the deltha method can be used to estimate standard errors.

[12] The exponential distribution is a special case of the Weibull distribution, with $p = 1$.

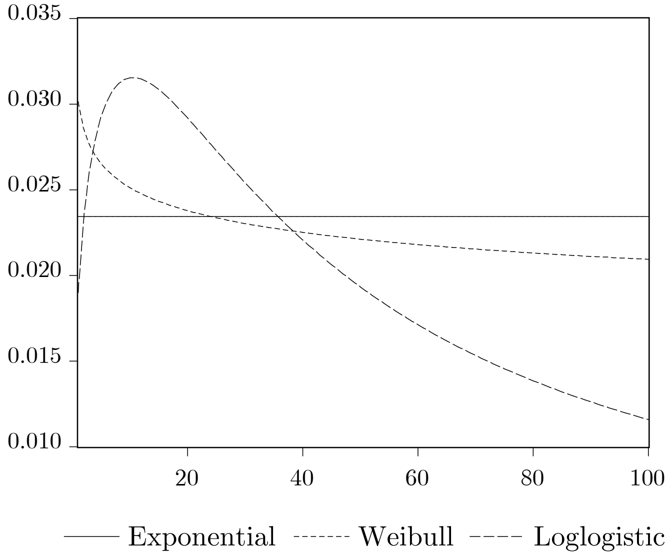

Figure 10.3: Hazard functions

10.7.2 Duration and Exogenous Variables

One limitation of the models given above is that external factors are not given a role in the survival function. The addition of covariates to duration models is fairly straightforward, although the interpretation of the coefficients in the models is less so. Consider for example the Weibull model (extensions to other distributions are direct). Let

$$\lambda_i = e^{-\beta' x_i},$$

making λ a function of a set of regressors is equivalent to changing the units of measurement on the time axis. For this reason, these models are sometimes called accelerated failure time models. Note, as well, that in all of the models listed (and generally), the regressors do not bear on the question of duration dependence, which is a function of p. Let

$$\begin{aligned}\sigma &= \frac{1}{p},\\ \delta_i &= \begin{cases}1 \text{ if the spell is completed}\\ 0 \text{ if it is censored}\end{cases},\\ w_i &= p\ln(\lambda_i t_i) = \frac{\ln t_i - \beta' x_i}{\sigma}.\end{aligned}$$

By making the change of variable, we find that

$$\begin{aligned}f(w_i) &= \left(\frac{1}{\sigma}\right)\exp\left(w_i - e^{w_i}\right),\\ \mathcal{S}(w_i) &= \exp\left(-e^{w_i}\right).\end{aligned}$$

The log-likelihood function is

$$\begin{aligned}\ell &= \sum_{i=1}^{N}\left[\delta_i \ln f(w_i) + (1-\delta_i)\ln\mathcal{S}(w_i)\right]\\ &= \sum_{i=1}^{N}\left[\delta_i(w_i - \ln\sigma) - e^{w_i}\right].\end{aligned}$$

Estimation and inference can be conducted as usual. Note that the hazard function generally depends on t, p, and x. The sign of the estimated coefficient suggests the direction of the effect of the variable on the hazard function when the hazard is monotonic. But in those cases, such as the log-logistic, in which the hazard is nonmonotonic, even this may be ambiguous. The magnitudes of the effects may also be difficult to interpret in terms of the hazard function. But in few cases, we do get a regression-like interpretation. In the Weibull and exponential models, $\mathcal{E}[t\,|x_i] = \exp(p\beta' x_i)$, while for the log-logistic model $\mathcal{E}[\ln t\,|x_i] = \beta' x_i$.

10.8 Further Reading

The discussion in this chapter draws primarily on Amemiya (1985) and Greene (2012), which provide complementary treatments of limited dependent variable models. Amemiya offers the formal econometric foundations and asymptotic properties of the estimators, while Greene emphasizes practical modeling and computational aspects.

Classic surveys by McFadden (1984) and Dhrymes (1986) provide the theoretical background for qualitative-response and limited dependent variable models, respectively. Maddala (1983) presents a comprehensive reference that includes the early literature on discrete choice, truncated, censored, and count data models.

Further complementary and modern perspectives can be found in Hayashi (2000), Hansen (2022), and Mittelhammer, Judge, and Miller (2000), all of which link limited dependent variable models to general maximum-likelihood and extremum estimation frameworks.

10.9 Workout Problems

1. Verify (10.8).

2. Verify (10.9).

3. In the case of the probit model, (10.10) reduces to $\widehat{\mathcal{V}}\left(\widehat{\gamma}_i\right) = \widehat{\phi}_i^2 \widehat{M}_i \widehat{\mathcal{V}}\left(\widehat{\beta}\right) \widehat{M}_i'$. Find an expression for $\widehat{M}_i$.

4. In the case of the logit model, (10.10) reduces to $\widehat{\mathcal{V}}\left(\widehat{\gamma}_i\right) = \left[\widehat{\Lambda}_i\left(1-\widehat{\Lambda}_i\right)\right]^2 \widehat{D}_i \widehat{\mathcal{V}}\left(\widehat{\beta}\right) \widehat{D}_i'$. Find an expression for $\widehat{D}_i$.

5. Verify (10.13).

6. Derive the density function of a continuous random variable that is truncated from below and above.

7. Derive the density function of a continuous random variable that is censored from below and above.

8. Complete the proof of Theorem 115.

Part II

The Second Part

Appendix

The First Appendix

To be included.

Afterword

To be included.

Bibliography

Abadie, A. and G. Imbens (2006). "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica* 74(1), 235-67.

Amemiya, T. (1985). *Advanced Econometrics.* Harvard University Press.

Angrist, J. and J. Pischke (2008). *Mostly Harmless Econometrics*, Princeton University Press.

Baltagi, B. (1999). *Econometrics.* Springer-Verlag.

Callaway, B. and P. Sant'Anna (2021). "Difference-in-Differences with Multiple Time Periods," *Journal of Econometrics* 225, 200–30.

Cattaneo, M., N. Idrobo, and R. Titiunik (2019). *A Practical Introduction to Regression Discontinuity Designs: Foundations*, Cambridge University Press.

Davidson, R. and J. MacKinnon (1993). *Estimation and Inference in Econometrics.* Oxford University Press.

Davidson, R. and J. MacKinnon (2004). *Econometric Theory and Methods.* Oxford University Press.

Dhrymes, P. (1986). "Limited Dependent Variables," in Z. Griliches and M. Intriligator (eds.) *Handbook of Econometrics* III. North-Holland.

Dhrymes, P. (2000). *Mathematics for Econometrics.* Springer-Verlag.

Diebold, F. and R. Mariano (1995). "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics* 13, 253-65.

Efron, B. and R. Tibshirani. (1998). *An Introduction to the Bootstrap.* Chapman & Hall.

Engle, R. (1984). “Wald, Likelihood Ratio, and Lagrange Multiplier Tests in Econometrics,” in Z. Griliches and M. Intriligator (eds.) *Handbook of Econometrics* II. North-Holland.

Franses, P. and D. van Dijk (2000). *Non-linear Time Series Models in Empirical Finance.* Cambridge University Press.

Gallant, R. (1987). *Nonlinear Statistical Models.* Wiley & Sons.

Gourieroux, C. and A. Monfort (1994). “Testing Nonnested Hypotheses,” in R. Engle and D. McFadden (eds.) *Handbook of Econometrics* IV. North-Holland.

Greene, W. (2012). *Econometric Analysis.* Macmillan.

Hansen, B. (2022). *Econometrics.* Princeton University Press.

Harvey, D., S. Leybourne, and P. Newbold (1997). “Testing the Equality of Prediction Mean Square Errors,” *International Journal of Forecasting* 13, 281-91.

Harville, D. (1997). *Matrix Algebra from a Statistician’s Perspective.* Springer.

Hayashi, F. (2000). *Econometrics.* Princeton University Press.

Heckman J. and B. Singer (1986). “Econometric Analysis of Longitudinal Data,” in Z. Griliches and M. Intriligator (eds.) *Handbook of Econometrics* III. North-Holland.

Hendry, D. (1984). “Monte Carlo Experimentation in Econometrics” in Z. Griliches and M. Intriligator (eds.) *Handbook of Econometrics* II. North-Holland.

Hill, C., W. Griffiths and G. Lim (2011). *Principles of Econometrics.* Wiley.

Horowitz, J. (2001). “The Bootstrap” in J. Heckman and E. Leamer (eds.) *Handbook of Econometrics* V. North-Holland.

Huber, P. (1981). *Robust Statistics.* Wiley.

Imbens, G. and D. Rubin (2015), *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge University Press.

Johnston, J. and J. DiNardo (1997). *Econometric Methods.* McGraw-Hill.

Judd, K. (1998). *Numerical Methods in Economics.* MIT Press.

Kleibergen, F. and R. Paap (2006). "Generalized Reduced Rank Tests Using the Singular Value Decomposition," *Journal of Econometrics* 133(1), 97–126.

Koenker, R. and K. Hallock (2001). "Quantile Regression," *Journal of Economic Perspectives* 15(4), 143–56.

Kuan, C. and H. White (1994). "Artificial Neural Networks: An Econometric Perspective," *Econometric Reviews* 13, 1-91.

Lam, J. and M. Veill (2002). "Bootstrap Prediction Intervals for Single Period Regression Forecasts," *International Journal of Forecasting* 18, 125-30.

Leamer, E. (1983a). "Let's Take the Con Out of Econometrics." *American Economic Review* 73, 31-43.

Leamer, E. (1983b). "Model Choice and Specification Analysis," in D. Belsley, Z. Griliches, M. Intriligator, and P. Schmidt (eds.) *Handbook of Econometrics* I. North-Holland.

Lovell, M. (1983). "Data Mining," *Review of Economics and Statistics* 65, 1-12.

Maddala, G. (1983). *Limited Dependent and Qualitative Variables in Econometrics.* Cambridge University Press.

McFadden, D. (1984). "Econometric Analysis of Qualitative Response Models," in Z. Griliches and M. Intriligator (eds.) *Handbook of Econometrics* II. North-Holland.

Mittelhammer, R. (1996). *Mathematical Statistics for Economics and Business.* Springer-Verlag.

Mittelhammer, R., G. Judge, and D. Miller (2000). *Econometric Foundations.* Cambridge University Press.

Newey, W. and K. West (1987). "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica* 55(3), 703–08.

Newey, W. and D. McFadden (1994). "Large Sample Estimation and Hypothesis Testing," in R. Engle and D. McFadden (eds.) *Handbook of Econometrics* IV. North-Holland.

Nychka, D., S. Ellner, D. McCaffrey, and R. Gallant (1990). "Statistics for Chaos," *Statistical Computing and Statistical Graphics Newsletter*, 4-11.

Peracchi, F. (2001). *Econometrics.* Wiley & Sons.

Pindyck, R. and D. Rubinfeld (1997). *Econometric Models and Economic Forecasts.* McGraw-Hill/Irwin.

Politis, D. and J. Romano (1994). "Stationary Bootstrap," *Journal of the American Statistical Association* 89, 1303-13.

Ramanathan, R. (1993). *Statistical Methods in Econometrics.* Academic Press.

Rao, R. (1973). *Linear Statistical Inference and Its Applications.* Wiley & Sons.

Ruud, P. (2000). *An Introduction to Classical Econometric Theory.* Oxford University Press.

Sheskin, D. (2000). *Parametric and Nonparametric Statistical Procedures.* Chapman & Hall.

Stock, J. and M. Watson (2003). *Introduction to Econometrics.* Addison-Wesley.

Stock, J. and M. Yogo (2005). "Testing for Weak Instruments in Linear IV Regression," in *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, D. Andrews and J. Stock (editors), Cambridge University Press.

Sullivan, R., A. Timmermann, and H. White (1998). "Dangers of Data-Driven Inference: The Case of Calendar Effects in Stock Returns," *Manuscript.* University of California, San Diego.

Sun, L. and S. Abraham (2021), "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects," *Journal of Econometrics* 225, 175–99.

Thisted, R. (1988). *Elements of Statistical Computing.* Chapman & Hall.

Ullah, A. (2004). *Finite Sample Econometrics.* Cambridge University Press.

Vigen, T. (2015). *Spurious Correlations.* Hachette Books.

Vuong, Q. (1989). "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses." *Econometrica* 57(2), 307–333.

White, H. (1980). "A Heteroskedastic-Consistent Covariance Matrix and a Direct Test for Heteroskedasticity," *Econometrica* 48, 421-48.

White, H. (1982). "Maximum Likelihood Estimation of Misspecified Models." *Econometrica*, 50, 1–25.

White, H. (1996). *Estimation, Inference and Specification Analysis.* Cambridge University Press.

White, H. (2000). "A Reality Check for Data Snooping," *Econometrica* 68, 1097-126.

Yule, G. (1926). "Why Do We Sometimes Get Nonsense-Correlations Between Time-Series? A Study in Sampling and the Nature of Time-Series," *Journal of the Royal Statistical Society* 89(1), 1–63.