ECON 382, Econometrics I Class Notes*

Pierangelo De Pace [†1]
Augusto Gonzalez-Bonorino [†2]
Pomona College
Department of Economics

January 27, 2025

Contents

- The present notes are based on a combination of various sources: (i) my class notes from when I was an undergraduate and then a graduate student at Bocconi University in Milan, (ii) the class notes from my Ph.D. years at the Johns Hopkins University in Baltimore, (iii) the teaching material I prepared for the sections of undergraduate econometrics I used to teach when I was a teaching assistant at the Johns Hopkins University, (iv) the textbook I adopted for this course of econometrics, the third edition of "A Guide to Modern Econometrics" by Marno Verbeek (John Wiley and Sons, 2008), and (v) some other statistics and econometrics textbooks that I have been reading in all this time. These notes are not intended to be substitutes for the in-class lectures or the textbook itself, which may contain many more details, examples, explanations, and practical implications than these pages. Rather, they should be thought of as supplemental material, to be complemented with the computer lab handouts and the solutions to the homework assignments I will make available at the due time. These notes are still in their draft form. They are currently under heavy revision and may contain errors and imprecisions.

[†1] Pomona College, Department of Economics; Carnegie Building, Room 205 - 425, N College Avenue - Claremont, CA 91711, USA. E-mail: pierangelo.depace@pomona.edu. Telephone: +1 9096218744.
[†2] The current version of Pierangelo's notes have been updated, edited, and complemented with exercises, definitions, and R code simulations produced by Augusto G.B. Any changes have been based on my personal notes from PhD-level econometric courses and data-driven econometric projects, as well as from my code examples from ECON 57: Economic Statistics. E-mail: agxa2024@pomona.edu

1 What is Econometrics?

Economic theory typically makes statements and hypotheses that are primarily qualitative in nature. In its most quantitative version, economics expresses theory in mathematical form, with little regard to empirical verification. Economic statistics provides methods to collect, process, and present economic data, but is not concerned with using the data to test theory. Econometrics develops tools and special methods to analyze observational (as opposed to experimental) data - i.e., data coming from uncontrolled environments, which is often the case in the social sciences. This lack of control often creates special problems when the researcher tries to establish causal relationships between variables.

More specifically, economics cannot be a proper experimental science (contrary to physics and natural sciences, for example, economists cannot and will not conduct large-scale experiments on economies). If we have any hope for it ever to become a proper "science" rather than a set of opinions, we need to be able to refute and reject wrong theories. This is the purpose of econometrics and econometricians: we develop tools to judge economic theories by their empirical relevance. The lack of experimentation implies that we have to resort to historical data and see what laws and principles are permanent and hidden. This is in fact a form of data sciences developed specifically with social sciences in mind.

1.1 Basic Methodology

Two are the primary purposes of an econometric model: (i) to empirically verify qualitative economic theory, using observed data, and (ii) to discover and develop new economic theory, by exploring the characteristics of the data.

To achieve these goals, an econometrician generally follows these lines:

i States the economic theory to be tested;

ii Specifies the mathematical model of the theory;

iii Specifies the econometric model of the theory;

iv Obtains the data;

v Estimates the parameters of the econometric model;

vi Tests hypotheses suggested by economic theory and concerning the econometric model parameters;

vii Forecasts and predicts;

viii Uses the econometric model for control and/or policy purposes.

An important tool of econometrics is regression analysis, the study of the dependence of one variable (the dependent variable, or regressand) on one or more other variables (the explanatory variables, or regressors). **The ultimate goal is to estimate or predict the population mean of the dependent variable using a combination of explanatory variables**.

1.2 Important Concepts

The aspiring econometrician must pay attention to several key distinctions:

- ***Statistical vs. Deterministic Relationships***: In regression analysis, we deal with random variables with probability distributions and are interested in the statistical relationships among them, as opposed to the deterministic relationships among non-random (non-stochastic) variables.

- ***Correlation vs. Causation***: Correlation is a pure mathematical relationship between variables that is inferable from the data. Causation refers to a behavior mechanism between two variables, a more stringent condition according to which the behavior of one variable is caused by another one. We must not mistake correlation for causation. Although variables with a causal relationship are bound to be correlated, the correlation between two variables does not imply that one of the two causes the other.

- ***Regression vs. Causation***: Regression analysis tries to describe the dependence of one variable on other variables but will not imply causation. To seek evidence of causation, we must look at outside statistics and formal theory.

- ***Regression vs. Correlation***: Whereas correlation analysis is concerned primarily with a linear relationship between two variables (the correlation coefficient), regression analysis attempts to predict the average value of one variable from an array of other exogenous factors. In regression analysis, we treat the explanatory and dependent variables as fundamentally different. The dependent variables are assumed to be stochastic, the explanatory variables are assumed to be nonrandom - i.e., fixed in repeated sampling. In correlation analysis, any two variables are treated symmetrically as random variables.

1.3 Types of Data

Econometrics uses three fundamentally different types of data:

i **Time Series Data**: A list of observations on values that a variable, concerning the same individual, takes at different points in time. For example, GDP in the USA from 1990 to 2010. The use of such variables in econometrics introduces problems of autocorrelation and non-stationarity.

ii **Cross-Section Data**: Data on variables collected repeatedly at the same point in time for a number of individuals. For example, GDP in all European countries in 2010. The use of such variables in econometrics can introduce problems of heterogeneity and heteroskedasticity.

iii **Pooled (panel) Data**: Data which are elements of both time series and cross-sectional data. A specific subset of pooled data is represented by panel data, in which the same cross-sectional units (families, individuals, firms) are surveyed over time. For example, GDP in all European countries from 1990 to 2010.

2 Basic Review of Statistics - Part I

Basic Statistics I Rnotebook

In this section we will be reviewing some elementary properties of statistics that you should already know and master from previous courses. You are also required to review (or study) Appendices A and B in the textbook before we even start talking about serious econometrics.

2.1 Estimators

Definition 2.1 (Estimator)**.** An estimator, $\widehat{\theta}$, is a statistic, a function of the random sample of data that we use to estimate the value of some unknown population parameter, θ. Since it is a function of random variables, $\widehat{\theta}$ is a random variable. Note that both θ and $\widehat{\theta}$ may be vectors.

Example. We may be interested in the average GPA, μ, of the students at Pomona College and may want to use the sample mean of the GPA's, X_i, of N randomly selected students to estimate it. The formula for the estimator of μ, the population mean, would simply be

$$\widehat{\mu} = \bar{X} = \frac{\sum_{i=1}^{N} X_i}{N}$$

In this case, the sample mean, a statistic, is employed to estimate the population mean, the population parameter of interest.
Since an estimator is a random variable, it must have a distribution, which we can usually use to compute its moments.

2.2 Expected Value, Variance, and Covariance

Let X be a univariate random variable.

Definition 2.2 (Continuous Expected Value)**.** The expected value of a continuous random variable, X, distributed according to some probability density function, $f_\theta(x)$, depending on a vector of parameters, θ - i.e., $X \sim f_\theta(x)$ - is defined as

$$E(X) = \int_{-\infty}^{\infty} x f_\theta(x) dx$$

with $\int_{-\infty}^{\infty} f_\theta(x)dx = 1$.

Click to show/hide example

Definition 2.3 (Discrete Expected Value)**.** The expected value of a discrete random variable, X, with range $\mathcal{X}$ is defined as

$$E(X) = \sum_{x \in \mathcal{X}} x p(x)$$

where $p(x)$ is the probability attached to the outcome x-i.e., $p(x) = \text{Prob}(X = x)$, with $\sum_{x \in \mathcal{X}} p(x) = 1$.

Click to hide/show example

Definition 2.4 (Continuous Variance)**.** The variance of a continuous random variable, X, distributed according to some probability density function, $f_\theta(x)$, depending on a vector of parameters, θ - i.e., $X \sim f_\theta(x)$ - is defined as

$$\text{Var}(X) = \int_{-\infty}^{\infty} [x - E(X)]^2 f_\theta(x) dx$$

with $\int_{-\infty}^{\infty} f_\theta(x) dx = 1$.

Click to hide/show example

Definition 2.5 (Discrete Variance)**.** The variance of a discrete random variable, X, with range $\mathcal{X}$ is defined as

$$\text{Var}(X) = \sum_{x \in \mathcal{X}} [x - E(X)]^2 p(x),$$

where $p(x)$ is the probability attached to the outcome x-i.e., $p(x) = \text{Prob}(X = x)$, with $\sum_{x \in \mathcal{X}} p(x) = 1$.

Click to hide/show example

Definition 2.6 (Variance with $\mu = E(X)$)**.** The variance of a continuous or discrete random variable, X, with expected value $\mu = E[X]$, is defined as

$$\begin{aligned}\operatorname{Var}(X) &= E\left[(X-\mu)^2\right] \\ &= E\left(X^2\right) - [E(X)]^2 \\ &= E\left(X^2\right) - \mu^2\end{aligned}$$

Definition 2.7 (Covariance)**.** The covariance of two continuous or discrete random variables, X and Y, with expected values $\mu_x = E(X)$ and $\mu_y = E(Y)$, is defined as

$$\begin{aligned}\operatorname{Cov}(X,Y) &= E\left[(X-\mu_x)(Y-\mu_y)\right] \\ &= E(XY) - E(X)E(Y) \\ &= E(XY) - \mu_x\mu_y.\end{aligned}$$

While the expectation is a linear operator, the variance and covariance are not. Let X, Y, W, and Z be four generic random variables and let a, b, c, and d be four constant real numbers. Then we have the following basic properties for the expectation, the variance, and the covariance operators:

- $E(aX + bY + c) = aE(X) + bE(Y) + c$;
- $\operatorname{Var}(a + bX) = b^2 \operatorname{Var}(X)$;

- $\operatorname{Var}(aX \pm bY) = a^2 \operatorname{Var}(X) + b^2 \operatorname{Var}(Y) \pm 2ab \operatorname{Cov}(X, Y)$;
- $\operatorname{Cov}(aX + bY, cW + dZ) = ac \operatorname{Cov}(X, W) + ad \operatorname{Cov}(X, Z) + bc \operatorname{Cov}(Y, W) + bd \operatorname{Cov}(Y, Z)$.

Proof. Using the definition of variance and linearity of expectation:

$$\begin{aligned}
\operatorname{Var}(a + bX) &= E[(a + bX - E(a + bX))^2] \\
&= E[(a + bX - (a + bE(X)))^2] \\
&= E[(bX - bE(X))^2] \\
&= E[b^2(X - E(X))^2] \\
&= b^2 E[(X - E(X))^2] \\
&= b^2 \operatorname{Var}(X)
\end{aligned}$$

□

Proof. Let's derive this step by step:

$$\begin{aligned}
\operatorname{Var}(aX \pm bY) &= E[(aX \pm bY - E(aX \pm bY))^2] \\
&= E[(aX \pm bY - (aE(X) \pm bE(Y)))^2] \\
&= E[(a(X - E(X)) \pm b(Y - E(Y)))^2] \\
&= E[a^2(X - E(X))^2 + b^2(Y - E(Y))^2 \pm 2ab(X - E(X))(Y - E(Y))] \\
&= a^2 E[(X - E(X))^2] + b^2 E[(Y - E(Y))^2] \pm 2ab E[(X - E(X))(Y - E(Y))] \\
&= a^2 \operatorname{Var}(X) + b^2 \operatorname{Var}(Y) \pm 2ab \operatorname{Cov}(X, Y)
\end{aligned}$$

□

Proof. Using the definition of covariance and linearity of expectation:

$$\begin{aligned}
&\operatorname{Cov}(aX + bY, cW + dZ) \\
&= E[(aX + bY - E(aX + bY))(cW + dZ - E(cW + dZ))] \\
&= E[(a(X - E(X)) + b(Y - E(Y)))(c(W - E(W)) + d(Z - E(Z)))] \\
&= E[ac(X - E(X))(W - E(W)) + ad(X - E(X))(Z - E(Z)) \\
&\quad + bc(Y - E(Y))(W - E(W)) + bd(Y - E(Y))(Z - E(Z))] \\
&= acE[(X - E(X))(W - E(W))] + adE[(X - E(X))(Z - E(Z))] \\
&\quad + bcE[(Y - E(Y))(W - E(W))] + bdE[(Y - E(Y))(Z - E(Z))] \\
&= ac \operatorname{Cov}(X, W) + ad \operatorname{Cov}(X, Z) + bc \operatorname{Cov}(Y, W) + bd \operatorname{Cov}(Y, Z)
\end{aligned}$$

□

2.3 Small-Sample and Large-Sample Properties

Definition 2.8 (Unbiasedness)**.** An estimator $\widehat{\theta}$ is unbiased for a population parameter θ if $E(\widehat{\theta}) = \theta, \forall \theta$.

If two estimators are unbiased for the same population parameter, the one with the lower variance is said to be more efficient. In general, within the class of unbiased estimators, we look for the most efficient one.

Hereafter, except cases of confusion, we will denote the generic estimator for the population parameter, θ, as $\widehat{\theta}_N$, where N is the size of the sample of data. The subscript in this piece of notation indicates that the estimator, as in the example on the sample mean described above, depends on (is a function of) the sample size.

Definition 2.9 (Convergence in Probability)**.** $\widehat{\theta}_N$ converges in probability to θ if, for any arbitrary $\varepsilon > 0$,

$$\lim_{N\to\infty} \text{Prob}\left(\left|\widehat{\theta}_N - \theta\right| > \varepsilon\right) = 0$$

or

$$\lim_{N\to\infty} \text{Prob}\left(\left|\widehat{\theta}_N - \theta\right| < \varepsilon\right) = 1$$

In other words, $\widehat{\theta}_N$ converges in probability to θ if, for any arbitrary $\varepsilon > 0$ and $\eta > 0$, there exists a natural number, N_0, such that, for any $N \geq N_0$, $\text{Prob}\left(\left|\widehat{\theta}_N - \theta\right| > \varepsilon\right) < \eta$. Notation: plim $\widehat{\theta}_N = \theta$, or $\widehat{\theta}_N \xrightarrow{p} \theta$, or $\widehat{\theta}_N = \theta + o_p(1)$.[1]

Definition 2.10 (Consistency)**.** An estimator $\widehat{\theta}_N$ is consistent for θ if it converges in probability to θ.

Unbiasedness is a small sample property of an estimator. Consistency is an asymptotic property. Unbiasedness means that, in repeated sampling from the population, the average value of an estimator equals the true unknown parameter we would like to estimate. Consistency means that, as the sample size increases, the value of the estimator will converge to the true population parameter. In other words, as the sample size increases, the distribution of the estimator becomes more and more concentrated around the true value of the parameter being estimated, so that the probability of the estimator being arbitrarily close to zero converges to one. Unbiasedness does not imply consistency and consistency does not imply unbiasedness. A sufficient condition for an unbiased estimator to be consistent is that its variance shrinks to zero as the sample size grows to infinity. Also note that a biased estimator may be consistent.

1

Theorem 1 (Weak Law of Large Numbers)**.** Let $X_1, \cdots, X_N$ be independent and identically distributed (i.i.d.) random variables with finite mean, $E(X) = \mu$, and finite variance, $\text{Var}(X) = \sigma^2 < \infty$. Then $\bar{X} \xrightarrow{p} E(X)$, or $\bar{X} = E(X) + o_p(1)$.

[1] A random variable is said to be $o_p(1)$ if it converges in probability to 0 . In general, an $o_p(k)$ random variable is a variable, X, such that plim $\frac{X}{k} = 0$.

Proof. Assume $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$, then $\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{N}\right)$.[2] Then [2]

$$\begin{aligned}
\text{Prob}(|\bar{X} - \mu| > \varepsilon) &= \text{Prob}(\bar{X} - \mu < -\varepsilon \vee \bar{X} - \mu > \varepsilon) \\
&= \text{Prob}(\bar{X} - \mu < -\varepsilon) + 1 - \text{Prob}(\bar{X} - \mu < \varepsilon) \\
&= \text{Prob}\left(\frac{\sqrt{N}(\bar{X} - \mu)}{\sigma} < -\frac{\sqrt{N}\varepsilon}{\sigma}\right) + 1 + \\
&- \text{Prob}\left(\frac{\sqrt{N}(\bar{X} - \mu)}{\sigma} < \frac{\sqrt{N}\varepsilon}{\sigma}\right) \\
&= 1 + \Phi\left(-\frac{\sqrt{N}\varepsilon}{\sigma}\right) - \Phi\left(\frac{\sqrt{N}\varepsilon}{\sigma}\right) \longrightarrow 0 \text{ as } N \longrightarrow \infty
\end{aligned}$$

where $\Phi(\cdot)$ is the cumulative probability density function (CDF) of a standard normal random variable [3].

Note that $\frac{\sqrt{N}(\bar{X}-\mu)}{\sigma} \sim \mathcal{N}(0,1), \Phi\left(\frac{\sqrt{N}\varepsilon}{\sigma}\right) \longrightarrow 1$, and $\Phi\left(-\frac{\sqrt{N}\varepsilon}{\sigma}\right) \longrightarrow 0$ as $N \longrightarrow \infty$. This implies that $\text{Prob}(|\bar{X} - \mu| > \varepsilon) \longrightarrow 0$ as $N \longrightarrow \infty$. □

Visualization

Example. In order to show that unbiasedness does not imply consistency, consider the estimator $\widehat{\mu} = \bar{X} + \varepsilon$, where $X_1, \cdots, X_N$ are independent and identically distributed (i.i.d.) random variables with finite mean, μ, and finite variance, σ^2. Epsilon, ε, is a random variable such that $E(\varepsilon) = 0$. The estimator of the population mean $\widehat{\mu}$ is unbiased. In fact, $E(\widehat{\mu}) = E(\bar{X} + \varepsilon) = E(\bar{X}) + E(\varepsilon) = \mu$. However, μ is not consistent: $\text{plim}(\widehat{\mu}) = \text{plim}(\bar{X} + \varepsilon) = \text{plim}(\bar{X}) + \text{plim}(\varepsilon) = \mu + \text{plim}(\varepsilon)$. Unfortunately, $\text{plim}(\varepsilon)$ does not exist under our assumptions.

Theorem 2 (Central Limit Theorem)**.** Let $X_1, \cdots, X_N$ be independent and identically distributed (i.i.d.) random variables with finite variance, $\text{Var}(X) = \sigma^2 < \infty$, and finite mean, $E(X) = \mu$. Then $\sqrt{N}(\bar{X} - \mu) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right)$.

In other words, $\frac{\sqrt{N}(\bar{X}-\mu)}{\sigma} = z + o_p(1)$ and $\bar{X} = \mu + \frac{\sigma}{\sqrt{N}} z + o_p\left(\frac{1}{\sqrt{N}}\right)$, where z is the realization of a random variable with a standard normal distribution. To see this, let $W = \frac{\sqrt{N}(\bar{X}-\mu)}{\sigma}$. Since $\sigma W \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right)$ by the CLT, then

[2] In step 3, we normalize the mean differences $\bar{X} - \mu$ using z-score normalization to ensure the random variable is distributed standard normally. That is, dividing by the standard deviation of, in this case, the sample mean. Since the variance is $\frac{\sigma^2}{N}$ we know the std. dev is $\sqrt{\frac{\sigma^2}{N}} = \frac{\sigma}{\sqrt{N}}$. Hence it results in $\frac{\sqrt{N}(\bar{X}-\mu)}{\sigma}$

[3] The standard normal CDF, $\Phi(x)$, is defined as $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt$.

[3] [2] The assumption of normality is not actually needed, but simplifies the steps in the proof. This proof will be revisited later, without the normality assumption, after the explanation of the Chebyshev inequality.

$W \xrightarrow{d} \mathcal{N}(0,1)$. So $W = z + o_p(1)$, and, therefore, $\bar{X} = \mu + \frac{\sigma}{\sqrt{N}} z + o_p\left(\frac{1}{\sqrt{N}}\right)$.

The central limit theorem intuitively states that, as the sample size increases, the distribution of a particular transformation of i.i.d. random variables with finite mean and finite variance converges to the distribution of a standard normal random variable.

3 Mathematical Properties of the Linear Regression Model

In this section we will describe the linear regression model and its mathematical properties. We will start from the simplest case of a model with one explanatory variable (or regressor) and an intercept term, then we will continue with the more general case of a model with multiple explanatory variables and an intercept term.

3.1 Linear Model with One Regressor and One Intercept Term

In our first incarnation of the linear regression model, we will study the case of a linear model with one regressor and one intercept term.

Let y and x be two variables. Consider a sample of N observations for each of them, $\{y_i\}_{i=1}^N$ and $\{x_i\}_{i=1}^N$ - i.e., our data. To make the framework more realistic, we can think about the following scenario. We have 1000 individuals, randomly drawn from a reference population (for instance, from the entire population in the USA). For each of these individuals we have data on income and consumption. Let the variable y be consumption, the variable x personal income, and $N = 100$. Looking at the scatter plot of these two variables may give us a picture of the approximate relationship existing between them. For example, this relationship may be roughly linear[4]. Figure 1 provides a clearer graphical and intuitive representation. Income and consumption appear to be linked by a positive linear relationship.

If we want to define the characteristics of this relationship, we need to draw the straight line through the clouds of points in Figure 1 that best approximates the cloud of points. In principle, we could draw many lines, but we are actually looking for the best one. Assume that the one already plotted in Figure 1 is the line we are looking for, the one which best represents the relationship between personal income and consumption, here thought to be linear. **In plain English, this would imply that, in the sample of individuals surveyed, when income goes up then consumption, on average, goes**

[4]In this case, I have simulated the values of y based on the equation $y = 2x + \epsilon$ to ensure a linear trend. We would denote this the "true" equation for the dependent variable, one we can never know in the real world. Simulations are great for testing properties of a model because we engineer the ground truth, and we will use them throughout the course.

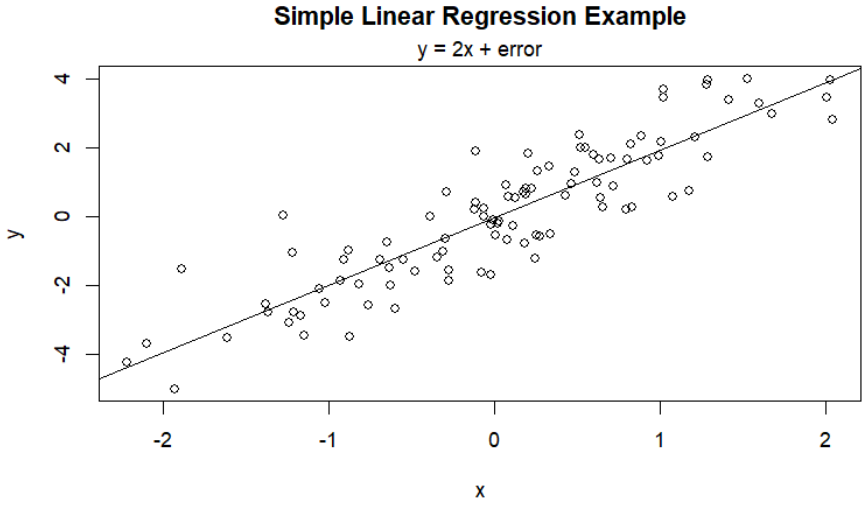

Figure 1: Simple Linear Regression example

up too. From Figure 1 and from the plotted straight line, we can also notice that increases in personal income are associated, on average, with smaller (and **constant**) increases in consumption - i.e., the marginal propensity to consume out of income is constant, positive, and less than one.

To draw that line, the word ***"best" should be defined in some way*** - i.e., according to some criterion that could be quantitatively established. For example, among the many, we may want to choose the line that minimizes a suitable **objective (or loss) function**. To be even more specific, the objective function may depend on the approximation errors. Very loosely speaking, by drawing that line, our purpose is to minimize the approximation errors, that is the errors that we make when we try to approximate the cloud of points using a straight line. Why not minimize a function of those errors, then?

If this is our ultimate goal, we may want to approximate the relationship between y and x using a model of the form $y_i = \alpha + \beta x_i + \varepsilon_i$, where ε_i is an error term - i.e., the error we make when we approximate the i-th observation of y, y_i, with a linear function of the i-th observation of x, x_i. Given that the sample size is equal to N, we have N approximation errors, one for each observation. At this point, We need to establish a criterion that will allow us to minimize, in some way, the approximation errors and to attach values to α and β in the linear relationship that we think describes the link between the two variables under investigation.

3.1.1 Unconstrained Optimization

[5]

The problem of unconstrained optimization involves finding the values of variables that minimize or maximize some objective function when there are no restrictions on the values these variables can take. Consider a scalar-valued objective function $f : \mathbb{R}^n \to \mathbb{R}$ that is at least twice continuously differentiable. Note that while the function can take multiple inputs (n-dimensional vector), it must return a single real number (scalar).

The unconstrained optimization problem can be written as either:

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{or} \quad \max_{x \in \mathbb{R}^n} f(x)$$

First-Order Necessary Conditions At any local extremum (minimum or maximum) x^*, the first derivative (in the scalar case) or gradient (in the vector case) must equal zero. This gives us the first-order necessary condition:

Scalar Case ($n = 1$): For a function $f : \mathbb{R} \to \mathbb{R}$,

$$\frac{df(x^*)}{dx} = 0$$

Vector Case ($n > 1$): For a function $f : \mathbb{R}^n \to \mathbb{R}$,

$$\nabla f(x^*) = \begin{pmatrix} \frac{\partial f(x^*)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x^*)}{\partial x_n} \end{pmatrix} = \mathbf{0}$$

This condition is necessary but not sufficient for a local extremum. A point satisfying this condition is called a critical point or stationary point.

The Hessian Matrix The Hessian matrix, denoted as $H(x)$, is a square matrix of second-order partial derivatives. For a function $f : \mathbb{R}^n \to \mathbb{R}$, it is an $n \times n$ matrix constructed as follows:

$$H(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

The Hessian is symmetric when f is twice continuously differentiable (by Young's theorem), meaning $\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$.

[5] Notes adapted from Mathematics for Economists by Carl Simon and Lawrence Blume, Chapter 17.

Second-Order Conditions The second-order conditions help distinguish between minima, maxima, and saddle points:

Scalar Case ($n = 1$):

- For a local minimum:
  - Necessary: $\frac{d^2 f(x^*)}{dx^2} \geq 0$
  - Sufficient: $\frac{d^2 f(x^*)}{dx^2} > 0$
- For a local maximum:
  - Necessary: $\frac{d^2 f(x^*)}{dx^2} \leq 0$
  - Sufficient: $\frac{d^2 f(x^*)}{dx^2} < 0$

Vector Case ($n > 1$):

- For a local minimum:
  - Necessary: $H(x^*)$ is positive semidefinite
  - Sufficient: $H(x^*)$ is positive definite
- For a local maximum:
  - Necessary: $H(x^*)$ is negative semidefinite
  - Sufficient: $H(x^*)$ is negative definite

Local vs. Global Extrema A point x^* is a local minimum (maximum) if there exists some neighborhood N around x^* such that $f(x^*) \leq f(x)$ ($f(x^*) \geq f(x)$) for all $x \in N$. If this holds for all $x \in \mathbb{R}^n$, then x^* is a global minimum (maximum).

For strictly convex functions (where the Hessian is positive definite everywhere), any local minimum is also a global minimum. Conversely, for strictly concave functions (where the Hessian is negative definite everywhere), any local maximum is also a global maximum. These properties will be particularly relevant when we examine the least squares estimation problem in the next section.

3.1.2 Least Squares Minimization

Ordinary Least Squares (OLS) is the approach we will use in this course to determine the best values of α and β. Graphically, by using OLS, we will be able to draw the straight line passing through the cloud of points in the scatter plot. The OLS approach is based on the minimization of a particular objective function, the sum of the so-called squared residuals. The i-th residual is the difference between the true value of the dependent variable, y_i, which is associated with x_i, and the value of the linear relationship to be estimated

calculated at x_i. Note that there will be as many residuals as many observations we have in the dataset and that OLS is just one of the many approaches that can be adopted to approximate the relationship described by the points in the cloud. As mentioned earlier, "best" should be defined according to some arbitrary criterion. If the criterion - i.e., the loss function - changes, the straight line we draw will change as a consequence.

Formally, let us start from the model above, $y_i = \alpha + \beta x_i + \varepsilon_i$, which we can rewrite in terms of the error term as $\varepsilon_i = y_i - \alpha - \beta x_i$. Then, $\varepsilon_i^2 = (y_i - \alpha - \beta x_i)^2$. According to the OLS approach, we need to minimize $\sum_{i=1}^{N} \varepsilon_i^2$, the sum of the squared errors, with respect to α and β :

$$\min_{\alpha,\beta} \sum_{i=1}^{N} \varepsilon_i^2 = \min_{\alpha,\beta} \sum_{i=1}^{N} (y_i - \alpha - \beta x_i)^2 = \min_{\alpha,\beta} S(\alpha, \beta).$$

In practice, we need to choose the values of α and β that minimize a very specific loss function. Let us take the first-order conditions of this optimization problem by setting the partial derivatives of $S(\alpha, \beta)$ with respect to α and β equal to zero, $\frac{\partial}{\partial \alpha} S(\alpha, \beta) = 0$ and $\frac{\partial}{\partial \beta} S(\alpha, \beta) = 0$:

$$\begin{aligned}
\frac{\partial}{\partial \alpha} S(\alpha, \beta) = & -2 \sum_{i=1}^{N} (y_i - \alpha - \beta x_i) = 0 \\
& \Longrightarrow \sum_{i=1}^{N} (y_i - \alpha - \beta x_i) = 0 \\
& \Longrightarrow \sum_{i=1}^{N} y_i - N\alpha - \beta \sum_{i=1}^{N} x_i = 0 \\
& \Longrightarrow \widehat{\alpha} = \frac{\sum_{i=1}^{N} y_i}{N} - \beta \frac{\sum_{i=1}^{N} x_i}{N} \\
& = \bar{y} - \beta \bar{x}
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial}{\partial \beta} \beta S(\alpha, \beta) &= -2 \sum_{i=1}^{N} (y_i - \alpha - \beta x_i) x_i = 0 \\
\Longrightarrow \widehat{\beta} &= \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \\
&= \frac{\frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{N-1}}{\frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N-1}} \\
&= \frac{\widehat{\mathrm{Cov}(x, y)}}{\widehat{\mathrm{Var}(x)}}
\end{aligned}$$

For $\widehat{\beta}$ to exist, we need that $\widehat{\text{Var}(x)} \neq 0$. This is the equivalent, in the case of a linear regression model with one intercept term and one regressor, of what we will call later no-multicollinearity condition.

Note that

$$\begin{aligned}\widehat{\beta} &= \frac{\sum_{i=1}^{N}(x_i - \bar{x})\, y_i}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^{N}(y_i - \bar{y})\, x_i}{\sum_{i=1}^{N}(x_i - \bar{x})^2}.\end{aligned}$$

We can then replace β in the expression of $\widehat{\alpha}$ with $\widehat{\beta}$ to compute

$$\widehat{\alpha} = \bar{y} - \widehat{\beta}\bar{x}$$

The two values of α and $\beta, \widehat{\alpha}$ and $\widehat{\beta}$ respectively, optimize the loss function. To be sure that $\widehat{\alpha}$ and $\widehat{\beta}$ minimize $S(\alpha, \beta)$, we need to show that the Hessian matrix

$$H = \begin{bmatrix} \frac{\partial^2}{(\partial\alpha)^2} S(\widehat{\alpha}, \widehat{\beta}) & \frac{\partial^2}{\partial\alpha\partial\beta} S(\widehat{\alpha}, \widehat{\beta}) \\ \frac{\partial^2}{\partial\alpha\partial\beta} S(\widehat{\alpha}, \widehat{\beta}) & \frac{\partial^2}{(\partial\beta)^2} S(\widehat{\alpha}, \widehat{\beta}) \end{bmatrix}$$

is positive definite.

Definition 3.1 (Positive Definite matrix)**.** A $n \times n$ matrix, H, is positive definite if the quadratic form $v'Hv > 0, \forall v \in \mathbb{R}^n, v \neq 0$.

In our case, $\frac{\partial^2}{(\partial\alpha)^2} S(\alpha, \beta) = 2N > 0$, $\frac{\partial^2}{\partial\alpha\partial\beta} S(\alpha, \beta) = 2\sum_{i=1}^{N} x_i$, and $\frac{\partial^2}{(\partial\beta)^2} = 2\sum_{i=1}^{N} x_i^2 > 0$. We would need to prove that $v'Hv > 0$ for any $v \in \mathbb{R}^2$ and such that $v \neq 0$. That is, $2Nv_1^2 + 4v_1v_2\sum_{i=1}^{N} x_i + 2v_2^2\sum_{i=1}^{N} x_i^2 > 0$, where $v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \neq 0$. For a matrix to be positive definite, $\det(H) > 0$ is a necessary condition. However, it is not sufficient.

The two values of $\widehat{\alpha}$ and $\widehat{\beta}$ represent the intercept and the slope of the straight line approximating the cloud of points we started from.

Definition. The i-th residual, $\widehat{\varepsilon}_i = y_i - \widehat{y}_i = y_i - \widehat{\alpha} - \widehat{\beta}x_i$, is the error left from the approximation straight line for the i-th observation in the sample of data.

We can then write

$$y_i = \widehat{\alpha} + \widehat{\beta}x_i + \widehat{\varepsilon}_i = \widehat{y}_i + \widehat{\varepsilon}_i,$$

where $\widehat{y}_i = \widehat{\alpha} + \widehat{\beta}x_i$ is the so-called fitted value of y_i. So the difference between y_i and $\widehat{y}_i$ - i.e., the residual, which can be either positive or negative, or even equal to zero - is the approximation error that we make when we use the OLS approach to fit the cloud of points that we plotted based on the dataset.

3.2 Linear Model with Multiple Regressors and One Intercept Term

We can apply the same logic to the case of many regressors in the linear regression model. For example, we may think that the relationship of interest is not just between consumption and personal income, but that, to correctly describe individual consumption, we need to consider other variables, in addition to personal income, on the right-hand side of the linear model we want to use. In other words, this time, we have one dependent variable, y, and multiple, $K-1$, independent variables, $x_2, x_3, \ldots, x_K$. For each of them, a sample of size N. We would like to approximate the link between y and these regressors using the linear model

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots + \beta_K x_{Ki} + \varepsilon_i.$$

First of all, one should notice that, unlike the previous case with one regressor and one intercept term, this new framework cannot be represented graphically. Given the $K-1$ regressors, we would need a K-dimensional space to plot the corresponding cloud of points. This time, we could not approximate the K-dimensional cloud of points with a straight line, but we would need a K dimensional hyperplane. The algebra we employed above to find the values of the intercept term and the slope term that minimize the OLS objective function would become much more complicated. At the end of the day, we would find that the expressions to derive are so complicated and hard to intuitively understand, or even impossible to determine, that we would need to find a solution to this issue. This is where linear algebra and matrix notation become useful and step in. In this course we will make use of two different matrix notations to describe the same model. We will switch back and forth between one and the other depending on the situation and convenience. We should always remember, though, that the two are perfectly equivalent and interchangeable and whatever can be said using one notation can be said using the alternative.

3.2.1 Matrix Notation I

Define the vectors

$$x_i = \begin{pmatrix} 1 & x_{2i} & x_{3i} & \cdots & x_{Ki} \end{pmatrix}'$$

and

$$\beta = \begin{pmatrix} \beta_1 & \beta_2 & \cdots & \beta_K \end{pmatrix}'.$$

The first vector contains the i-th observations of each regressors. The second contains all the parameters (intercept term and slope coefficients) in the linear regression model. Note that $x_{1i} = 1$ for all i and that $x_i \in M(K,1)$ - i.e., x_i belongs to the class of matrices of dimensions K and 1 , that is, with K rows and 1 column - and $\beta \in M(K,1)$. The linear model can then be rewritten as

$$y_i = x_i'\beta + \varepsilon_i,$$

which implies that $\varepsilon_i = y_i - x_i'\beta$. It follows that the objective function to be minimized to implement the OLS approach is

$$S(\beta) = \sum_{i=1}^{N} \left(y_i - x_i'\beta\right)^2$$

The minimization problem to be solved to obtain the optimal $\widehat{\beta}$ is then $\min_\beta S(\beta)$. The first-order condition is

$$\begin{aligned}
\frac{\partial}{\partial\beta} S(\beta) = &-2\sum_{i=1}^{N} x_i \left(y_i - x_i'\beta\right) = 0 \\
&\Longrightarrow \left(\sum_{i=1}^{N} x_i x_i'\right)\beta = \sum_{i=1}^{N} x_i y_i \\
&\Longrightarrow \widehat{\beta} = \left(\sum_{i=1}^{N} x_i x_i'\right)^{-1} \sum_{i=1}^{N} x_i y_i
\end{aligned}$$

To calculate $\widehat{\beta}$, we had to invert $\sum_{i=1}^{N} x_i x_i'$. $\sum_{i=1}^{N} x_i x_i'$ is a $K \times K$ matrix. For $\left(\sum_{i=1}^{N} x_i x_i'\right)^{-1}$ to exist, $\sum_{i=1}^{N} x_i x_i'$ needs to be invertible - i.e., none of the regressors is a linear combination of the other regressors. This condition, that each regressor is linearly independent of the other regressors in the model, is known as the no-multicollinearity condition.

Once the minimization problem is solved and $\widehat{\beta}$ is found, the best linear approximation for y_i will be $\widehat{y}_i = x_i'\widehat{\beta}$, which implies that $y_i = \widehat{y}_i + \widehat{\varepsilon}_i$. The residuals are hence $\widehat{\varepsilon}_i = y_i - \widehat{y}_i$.

From the first-order condition, $\sum_{i=1}^{N} x_i \left(y_i - x_i'\beta\right) = 0$ when $\beta = \widehat{\beta}$. This means that

$$\sum_{i=1}^{N} x_i \left(y_i - x_i'\widehat{\beta}\right) = \sum_{i=1}^{N} x_i \widehat{\varepsilon}_i = 0$$

Thus, x_i and $\widehat{\varepsilon}_i$ are orthogonal, that is they are independent of each other, or perpendicular in the N-space. If an intercept term, say β_1, is included in the model, then $\sum_{i=1}^{N} \widehat{\varepsilon}_i = 0$, since $x_{1i} = 1, \forall i$.

3.2.2 Matrix Notation II

Define the $N \times K$ matrix of independent explanatory variables, X, and the $N \times 1$ matrix representing the dependent variable, y, as

$$X = \begin{bmatrix} 1 & x_{21} & \cdots & x_{K1} \\ 1 & x_{22} & \cdots & x_{K2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{2N} & \cdots & x_{KN} \end{bmatrix} = \begin{bmatrix} x_1' \\ x_2' \\ \cdots \\ x_N' \end{bmatrix} \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

The first object, X, is a matrix collecting all the N observations (rows) in the K regressors (columns). The second object, y, is the vector of N observations in the independent variable. We next define

$$\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_K \end{bmatrix}$$

ε is the $N \times 1$ vector of error terms, β is the $K \times 1$ vector of parameters in the linear model. We can hence rewrite the linear regression model as

$$y = X\beta + \varepsilon$$

from which, $\varepsilon = y - X\beta$.
Note that

$$S(\beta) = \sum_{i=1}^{N} \varepsilon_i^2 = \varepsilon'\varepsilon = (y - X\beta)'(y - X\beta) = y'y - 2y'X\beta + \beta'X'X\beta$$

The optimization problem for the implementation of the OLS approach is again $\min_\beta S(\beta)$. The first-order condition is

$$\begin{aligned} \frac{\partial}{\partial \beta} S(\beta) = 0 &\Longrightarrow -2\left(X'y - X'X\beta\right) = 0 \\ &\Longrightarrow \widehat{\beta} = \left(X'X\right)^{-1} X'y. \end{aligned}$$

This $\widehat{\beta}$ is identical to the $\widehat{\beta}$ derived in the previous section. That is, $\widehat{\beta} = \left(X'X\right)^{-1} X'y = \left(\sum_{i=1}^{N} x_i x_i'\right)^{-1} \sum_{i=1}^{N} x_i y_i$. The only difference is notation, since we initially defined the same model in two different ways.

To obtain $\widehat{\beta}$ we need again to assume no-multicollinearity. This time, the corresponding condition is that $(X'X)$, a $K \times K$ matrix, is invertible. A necessary and sufficient condition for $(X'X)$ to be invertible is that $\det\left(X'X\right) \neq 0$. Note that, for $(X'X)$ to be invertible, $\text{rank}\left(X'X\right) = K$, a necessary condition of which is that $N \geq K$ - i.e., the number of observations in the dataset is bigger than or equal to the number of regressors including the intercept term.

From all above, it follows that $\widehat{y} = X\widehat{\beta}$ and $y - \widehat{y} = \widehat{\varepsilon}$, which implies that $y = X\widehat{\beta} + \widehat{\varepsilon}$. Looking again at the first-order condition,

$$X'y - X'X\beta = 0 \Longrightarrow X'(y - X\beta) = 0$$

This condition should hold at $\widehat{\beta}$, implying that

$$X'(y - X\widehat{\beta}) = 0 \Longrightarrow X'\widehat{\varepsilon} = 0$$

which means that the set of regressors and the residuals are orthogonal to each other. This is the same property of orthogonality we saw in the previous section.

3.3 Matrix Algebra and Calculus Review

Before proceeding, let us review key concepts from matrix algebra that will be essential throughout this course.

Basic Definitions and Notation A matrix A of dimension $m \times n$ has m rows and n columns:

$$A = [a_{ij}]_{m \times n} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

Special matrices include:

- Vector: A matrix with either one row (row vector) or one column (column vector)
- Square matrix: $m = n$
- Diagonal matrix: Square matrix where $a_{ij} = 0$ for all $i \neq j$
- Identity matrix (I_n): Diagonal matrix with ones on the main diagonal

Matrix Operations

1. Transpose: A' or A^T flips matrix over its diagonal

$$(A')_{ij} = A_{ji}$$

2. Matrix Addition ($A + B$): Requires same dimensions

$$(A + B)_{ij} = A_{ij} + B_{ij}$$

3. Matrix Multiplication (AB): Requires columns of A match rows of B

$$(AB)_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$$

4. Scalar Multiplication (cA): $(cA)_{ij} = c \cdot a_{ij}$

Important Properties For matrices of compatible dimensions:

1. Transpose properties:
   - $(A')' = A$
   - $(AB)' = B'A'$
   - $(A + B)' = A' + B'$
2. Multiplication properties:
   - $AB \neq BA$ generally (not commutative)
   - $A(BC) = (AB)C$ (associative)
   - $A(B + C) = AB + AC$ (distributive)

Matrix Invertibility For a square matrix A:

- A is invertible if there exists A^{-1} such that $AA^{-1} = A^{-1}A = I$
- Properties of invertible matrices:
  - $(A^{-1})^{-1} = A$
  - $(AB)^{-1} = B^{-1}A^{-1}$
  - $(A')^{-1} = (A^{-1})'$
- A matrix is invertible if and only if:
  - Its determinant is non-zero: $\det(A) \neq 0$[6]
  - It has full rank: $\text{rank}(A) = n$
  - Its columns (rows) are linearly independent[7]

Quadratic Forms For a vector x and symmetric matrix A, the quadratic form $x'Ax$ is scalar-valued:

$$x'Ax = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i a_{ij} x_j$$

Properties of symmetric matrices:

[6] The determinant of a matrix can be computed in several ways. For a 2×2 matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, $\det(A) = ad - bc$. For larger matrices, we can use: (1) cofactor expansion along any row or column, (2) row reduction to triangular form, or (3) the product of eigenvalues. In econometrics, we often encounter the determinant when checking if $X'X$ is invertible, which is crucial for the existence of the OLS estimator.

[7] In the context of OLS, linear independence of the columns of X means no perfect multicollinearity among regressors. This has an important economic interpretation: each regressor must provide some unique information about the dependent variable that cannot be derived from a linear combination of other regressors.

- Positive definite: $x'Ax > 0$ for all $x \neq 0$
- Positive semidefinite: $x'Ax \geq 0$ for all x
- All eigenvalues of a positive definite matrix are positive[8]

Matrix Calculus Key derivatives for optimization:

- $\frac{\partial(x'a)}{\partial x} = a$ where a is a constant vector
- $\frac{\partial(a'x)}{\partial x} = a$ where a is a constant vector
- $\frac{\partial(x'Ax)}{\partial x} = (A + A')x$
- If A is symmetric, $\frac{\partial(x'Ax)}{\partial x} = 2Ax$ [9]
- $\frac{\partial(b'x)}{\partial x} = b$ where b is a constant vector
- $\frac{\partial(x'x)}{\partial x} = 2x$ [10]
- $\frac{\partial \text{tr}(AX)}{\partial X} = A'$ where tr denotes the trace [11]
- $\frac{\partial \text{tr}(XA)}{\partial X} = A$
- $\frac{\partial \text{tr}(AXB)}{\partial X} = A'B'$
- $\frac{\partial \ln|X|}{\partial X} = (X')^{-1}$ [12]

Some useful second derivatives:

- $\frac{\partial^2(x'Ax)}{\partial x \partial x'} = A + A'$
- If A is symmetric, $\frac{\partial^2(x'Ax)}{\partial x \partial x'} = 2A$ [13]

[8] Eigenvalues λ and eigenvectors v satisfy $Av = \lambda v$. For symmetric matrices like $X'X$, all eigenvalues are real, and positive definiteness means all eigenvalues are positive. The condition number (ratio of largest to smallest eigenvalue) of $X'X$ indicates the severity of multicollinearity in regression analysis. A large condition number suggests near multicollinearity and potential numerical instability in computing $(X'X)^{-1}$.

[9] This result is central to deriving the OLS estimator, where we minimize the quadratic form $(y - X\beta)'(y - X\beta)$.

[10] This derivative appears when minimizing sum of squared residuals and in ridge regression where we add a penalty term $\lambda\beta'\beta$ to the objective function.

[11] Trace derivatives are particularly useful in panel data models and when working with variance-covariance matrices.

[12] This derivative is essential in maximum likelihood estimation of multivariate models, such as Vector Autoregressions (VARs) and multivariate GARCH models. It also appears in the estimation of covariance matrices in Seemingly Unrelated Regression (SUR) models.

[13] Second derivatives are crucial for verifying the conditions for consistent estimation in nonlinear models and for computing standard errors using the Hessian matrix. In maximum likelihood estimation, the negative of the expected Hessian gives us the Information matrix, which is key for hypothesis testing and efficiency analysis.

Remember that in econometric applications, these derivatives often appear in combination. For example, the normal equations in OLS combine several of these rules:

$$\frac{\partial(y - X\beta)'(y - X\beta)}{\partial\beta} = -2X'y + 2X'X\beta = 0$$

These concepts will be particularly important when:

- Expressing the linear regression model in matrix form
- Deriving the OLS estimator through optimization
- Understanding the properties of variance-covariance matrices
- Analyzing the precision of our estimates

3.3.1 Notable Matrices

Recall that $\widehat{y} = X\widehat{\beta}$, from which $\widehat{y} = X\left(X'X\right)^{-1}X'y = P_x y$. The matrix $P_x = X\left(X'X\right)^{-1}X'$ known as the projection matrix. Now consider $\widehat{\varepsilon} = y - \widehat{y} = y - X\widehat{\beta} = y - P_x y = \left(I - P_x\right)y = M_x y . M_x = \left(I - P_x\right)$ is another important matrix, known as the annihilator matrix.

Both matrices are idempotent - i.e., $\left(P_x\right)^J = P_x$ and $\left(M_x\right)^J = M_x, \forall J$. Also note that, if $K = N$, then $\widehat{y} = P_x y = X\left(X'X\right)^{-1}X'y = XX^{-1}\left(X'\right)^{-1}X'y = y$ and $\widehat{\varepsilon} = 0$.

These two matrices will be useful later in the course.

4 Statistical Properties of the OLS Estimator

We will define statistical properties for the objects in the linear regression model, $y = X\beta + \varepsilon$, or its alternative (and equivalent) version, $y_i = x_i'\beta + \varepsilon_i$. From now on, y and ε will be random variables. We can look at X in two different ways: as a deterministic variable, or as another random variable. If deterministic, in repeated sampling, all the entries in X would not change and X itself could be taken as given. For simplicity, we will assume X to be deterministic. In this new framework, the optimal β for the linear approximation of the link between y and X is still $\widehat{\beta} = \left(X'X\right)^{-1}X'y$. This formula reflects the mathematical properties of the linear regression model within the OLS approach and does not depend on the statistical hypotheses we are going to make later. All this given, the so-called Gauss-Markov assumptions are to be imposed on the model to derive some nice statistical properties for $\widehat{\beta}$.

[13] All the results we will derive in this and the next sections would not change, if X was random. However, if random, the notation would need to be adjusted and the derivations slightly modified.

4.1 Gauss-Markov Assumptions

The Gauss-Markov assumptions (GMAs) can be stated in two ways, (a) and (b), depending on the notation we adopt to describe the linear regression model:

(a) $y_i = x_i'\beta + \varepsilon_i$;	(b) $y = X\beta + \varepsilon$;
(ia) $E(\varepsilon_i) = 0, \forall i$;	(ib) $E(\varepsilon) = 0$;
(iia) $\{\varepsilon_i\}_{i=1}^N \wedge \{x_i\}_{i=1}^N$	(iib) $\text{Var}(\varepsilon) = \sigma^2 I_N$;
(iiia) $\text{Var}(\varepsilon_i) = \sigma^2, \forall i$;	(iiib) $E(\varepsilon \mid X) = 0$;
(iva) $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0, \forall i \neq j$;	(ivb) $\text{Var}(\varepsilon \mid X) = \sigma^2 I_N$.

The expected value of a vector-valued random variable is a vector of the same length. In this case, a vector of N zeros. The variance of a vector-valued random variable of N entries is a $N \times N$ matrix, the so-called **variance-covariance matrix**. I_N is the identity matrix with N rows and N columns. In this case, the elements along the main diagonal of $\sigma^2 I_N$ are the variances (all equal to each other) of the corresponding terms in the ε vector. All the other elements, those off the main diagonal, are covariances between indexed elements. Under this assumption, the matrix will then look like.

$$\Sigma = \begin{bmatrix} \sigma^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma^2 & \ddots & \vdots \\ \vdots & \vdots & \cdots & \ddots & \vdots \\ 0 & 0 & \cdots & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 I_N$$

When, as in this case, the variance-covariance matrix of the random error term in the linear regression model is diagonal and all the elements along the main diagonal are equal to each other, we say that the model has **spherical disturbances**.[14]

Assumption (iiia) is known as the homoskedasticity assumption. It states that along the sample, the variance of the error term is constant. Assumption (iva) is known as the no-autocorrelation assumption. Assumption (iib) incorporates assumptions (iiia) and (iva), assumptions (iiib) and (ivb) combine assumptions (ia)-(iva) all together. Assumptions (iia), (iiib), and (ivb) establish the orthogonality between the error term and the set of regressors, or the exogeneity of the set of regressors - i.e., the regressors are not correlated with the error term. Under the simplifying assumption that the set of regressors is deterministic, (iia), (iiib), and (ivb) follow as a direct consequence. Under the

[14] Because all of the off-diagonal values are zero, we know there is no correlation between the error terms for different observations (i.e., errors are uncorrelated). Moreover, since all the elements in the diagonal are equal, the variance is constant across observations (we call this homoskedasticity). We call it "spherical" because, geometrically, the error terms are distributed uniformly in all directions.

assumption of stochastic regressors, the property of unbiasedness that we prove in the next section needs (iia) or (iiib) and (ivb) to be verified.

Given that the two models under the two notations are perfectly equivalent and that the Gauss-Markov assumptions can be indifferently stated in a way or another, depending on which notation we are using, whatever is proved under a given notation can be proved under the other. On a case-by-case basis, we will choose the most convenient notation to prove the statistical properties of $\widehat{\beta}$.

Under the Gauss-Markov assumptions, and under the assumption that y and ε are random variables, $\widehat{\beta}$ is a random variable, too. To see this, just look at its formula, $\widehat{\beta} = (X'X)^{-1} X'y$, according to which the optimal β is a linear function of y, a random variable. In other words, $\widehat{\beta}$ is the statistic - i.e., a function of the data, in this framework thought to be random - that, within the OLS approach, we use to estimate the true, unknown value of β. As such, it is an estimator of the population parameter of interest, β, the OLS estimator of β, for which we can analyze the statistical properties in small and large samples.

4.2 Small-Sample Properties

We will prove the property of unbiasedness of the OLS estimator of $\beta, \widehat{\beta}$, derive the formula of its variance, state the Gauss-Markov Theorem, and analyze its distributional properties in small samples under a small modification of the Gauss-Markov assumptions.

4.2.1 Unbiasedness

To prove that $\widehat{\beta}$ is unbiased for the true population parameter, β, we need to show that, under the Gauss-Markov assumptions, $E(\widehat{\beta}) = \beta$:

$$\begin{aligned}
E(\widehat{\beta}) &= E\left[(X'X)^{-1} X'y\right] \\
&= E\left[(X'X)^{-1} X'(X\beta + \varepsilon)\right] \\
&= E\left[\left((X'X)^{-1} X'X\beta + (X'X)^{-1} X'\varepsilon\right]\right. \\
&= E\left[\beta + (X'X)^{-1} X'\varepsilon\right] \\
&\qquad = \beta + E\left[(X'X)^{-1} X'\varepsilon\right] \\
&\qquad = \beta + (X'X)^{-1} X'E(\varepsilon) \\
&\qquad = \beta
\end{aligned}$$

keeping in mind that X is deterministic (so the term $(X'X)^{-1} X'$ can be pulled out of the expectation operator), and utilizing the Gauss-Markov assumption that the expected value of ε is zero. Hence, $\widehat{\beta}$ is unbiased for β. This is true with or without the assumptions of homoskedasticity or no-autocorrelation, which we did not use in the proof.

[14] If, for the sake of exposition, we assume that X is stochastic, the proof would not look

4.2.2 Variance

Stanford Proofs

To find the variance of the OLS estimator, $\widehat{\beta}$:

$$\begin{aligned}
\operatorname{Var}(\widehat{\beta}) &= E\left[(\widehat{\beta}-\beta)(\widehat{\beta}-\beta)'\right] \\
&= E\left\{\left[(X'X)^{-1}X'\varepsilon + \not{\beta} - \not{\beta}\right]\left[(X'X)^{-1}X'\varepsilon + \not{\beta} - \not{\beta}\right]'\right\} \\
&= E\left[(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}\right]
\end{aligned}$$

Note that $\left[(X'X)^{-1}\right]' = (X'X)^{-1}$. Since X is a non-stochastic, deterministic term, the only random variable in the expression above is $\varepsilon\varepsilon'$. The expression can then be simplified as

$$\begin{aligned}
\operatorname{Var}(\widehat{\beta}) &= E\left[(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}\right] \\
&= (X'X)^{-1}X'E(\varepsilon\varepsilon')X(X'X)^{-1}
\end{aligned}$$

In general, if A is a vector-valued random variable, then $\operatorname{Var}(A) = E(AA') - E(A)[E(A)]'$. So,

$$\operatorname{Var}(\varepsilon) = E(\varepsilon\varepsilon') - E(\varepsilon)[E(\varepsilon)]' = E(\varepsilon\varepsilon') = \sigma^2 I_N,$$

by the Gauss-Markov assumptions. It follows that
$$

$$\begin{aligned}
\operatorname{Var}(\widehat{\beta}) &= (X'X)^{-1}X'E(\varepsilon\varepsilon')X(X'X)^{-1} \\
&= (X'X)^{-1}X'\sigma^2 I_N X(X'X)^{-1} \\
&= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} \\
&= \sigma^2(X'X)^{-1}
\end{aligned}$$

σ^2 is the unknown variance of the error term, constant along the sample by assumption. As long as σ^2 remains unknown, $\operatorname{Var}(\widehat{\beta})$ remains unknown. How

---

like just described. The proof would require the application of the so-called law of iterated expectations and would look like

$$\begin{aligned}
E(\widehat{\beta} \mid X) &= \beta + E\left[(X'X)^{-1}X'\varepsilon \mid X\right] \\
&= \beta + (X'X)^{-1}X'E(\varepsilon \mid X) \\
&= \beta
\end{aligned}$$

and thus

$$E(\widehat{\beta}) = E[E(\widehat{\beta} \mid X)] = \beta$$

The same logic would apply in the derivation of the variance of the OLS estimator, which follows in the next section, under the assumption of stochastic regressors.

can we estimate σ^2, then? We can either use a biased, $\widehat{\sigma}_B^2$, or an unbiased, $\widehat{\sigma}_U^2$, estimator:

$$\widehat{\sigma}_B^2 = \frac{\sum_{i=1}^N \widehat{\varepsilon}_i^2}{N-1} \Longrightarrow E\left(\widehat{\sigma}_B^2\right) \neq \sigma^2,$$
$$\widehat{\sigma}_U^2 = \frac{\sum_{i=1}^N \widehat{\varepsilon}_i^2}{N-K} \Longrightarrow E\left(\widehat{\sigma}_U^2\right) = \sigma^2.$$

In large samples, we could use either one interchangeably (both would converge to the same value, as N grows to infinity), but in a small samples it would be better to use the unbiased estimator.

Given an unbiased estimate of $\sigma^2, \widehat{\sigma}_U^2$, can estimate the variance of the OLS estimator as

$$\widehat{\mathrm{Var}(\widehat{\beta})} = \widehat{\sigma}_U^2 \left(X'X\right)^{-1}$$

Note that $\widehat{\mathrm{Var}(\widehat{\beta})}$ is a $K \times K$ matrix. If we take the elements along its main diagonal, we can estimate the standard deviation of $\widehat{\beta}$ - i.e., the standard error of $\widehat{\beta}$:

$$SE(\widehat{\beta}) = \widehat{SD(\widehat{\beta})} = \sqrt{\mathrm{diag}\left[\widehat{\sigma}_U^2 \left(X'X\right)^{-1}\right]}$$

where $\mathrm{diag}(\cdot)$ is the diagonal operator, which takes the elements along the main diagonal of a square matrix and puts them into a column vector. $SE(\widehat{\beta})$ is a $K \times 1$ vector.

4.2.3 Gauss-Markov Theorem

Theorem (Gauss-Markov Theorem). Under the Gauss-Markov assumptions, the OLS estimator is the Best Linear Unbiased Estimator (BLUE).
$\widehat{\beta}$ is a linear estimator (as argued earlier, this follows from its form - i.e., $\widehat{\beta}$ is linear with respect to y) and is unbiased, as we proved above. According to this theorem, which we will not prove, it is best in the sense that it has the lowest variance within the class of linear unbiased estimators. That is, given any other linear and unbiased estimator of $\beta, \tilde{\beta}, \mathrm{Var}(\tilde{\beta}) - \mathrm{Var}(\widehat{\beta}) \geq 0$ - i.e., $\mathrm{Var}(\tilde{\beta}) - \mathrm{Var}(\widehat{\beta})$ is a positive semi-definite matrix.

4.2.4 Normality

In small samples, we do not know the distribution of $\widehat{\beta}$, unless we impose an additional assumption on the error term. If we assume that $\varepsilon \sim \mathcal{N}\left(0, \sigma^2 I_N\right)$, then $\widehat{\beta} \sim \mathcal{N}\left(\beta, \sigma^2 \left(X'X\right)^{-1}\right)$. In practice, we are extending the two Gauss-Markov assumptions (ib) and (iib) with an assumption of normality for the error term.

In small samples, if ε is not normal, then $\widehat{\beta}$ is not normal. Thus, if we do not make the normality assumption described above, the distribution of $\widehat{\beta}$ will remain unknown and statistical inference on the parameters of the linear regression model will not be possible.

4.3 Large-Sample (Asymptotic) Properties

If we relax the Gauss-Markov assumptions, the small-sample properties of the OLS estimator are typically unknown. If the error term in the linear regression model is not normal, then the OLS estimator is not normal in small samples. If $E(\varepsilon) \neq 0$, then $E(\widehat{\beta}) \neq \beta$, and $\widehat{\beta}$ is a biased estimator for β. When the small-sample properties of the OLS estimator are not "good", an alternative is to evaluate the quality of the OLS estimator based on asymptotic theory, that is, when the sample is large.

4.3.1 Chebyshev's Inequality

Theorem (Chebyshev's Inequality). Let X be a random variable, with finite mean, $E(X) = \mu$, and $\text{Var}(X) = \sigma^2 < \infty$. Then:

$$\text{Prob}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}, \quad \forall k > 0$$

If $k = \frac{\alpha}{\sigma}$, with $\alpha > 0$, then $\text{Prob}(|X - \mu| \geq \alpha) \leq \frac{\sigma^2}{\alpha^2}$.
Proof. Let $W \geq 0$ be a weakly positive random variable and let $c \in \mathbb{R}^+$- i.e., $c > 0$. The derivations that follow will be better understood if we consider an example probability density function for W, as in Figure 2.

Figure 2: Example probability density function of a weakly positive random variable, $W \geq 0$.

Since W is weakly positive, the domain of it density function is the entire positive semi-axis plus zero. The strictly positive real number, c lies somewhere in the positive semi-axis. Since W is a random variable, we can calculate its expectation as a weighted average:

$$\begin{aligned} E(W) &= \text{Prob}(W \leq c)E(W \mid W \leq c) + \text{Prob}(W \geq c)E(W \mid W \geq c) \\ &\geq \text{Prob}(W \geq c)E(W \mid W \geq c) \\ &\geq \text{Prob}(W \geq c)c \end{aligned}$$

The last inequality holds since $E(W \mid W \geq c) \geq c$, as one can easily figure out from Figure 2. It follows that

$$E(W) \geq \text{Prob}(W \geq c)c$$

which we can rearrange as

$$\text{Prob}(W \geq c) \leq \frac{E(W)}{c}$$

Let $W = (X-\mu)^2 \geq 0$ and $c = k^2\sigma^2$, with $k > 0$. The last expression can then be rewritten as

$$\text{Prob}\left[(X-\mu)^2 \geq k^2\sigma^2\right] \leq \frac{E\left[(X-\mu)^2\right]}{k^2\sigma^2}$$
$$\implies \text{Prob}(|X-\mu| \geq k\sigma) \leq \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}$$

4.3.2 Weak Law of Large Numbers, Revisited

In an earlier section we made the proof of the Weak Law of Large Numbers easier by adding an unnecessary assumption of normality on the random variable, X.

This time we will prove the law without the normality assumption, using instead Chebyshev's inequality.

Theorem (Weak Law of Large Numbers, WLLN). Let $X_1, \cdots, X_N$ be i.i.d. random variables with finite mean, $E(X) = \mu$, and finite variance, $\text{Var}(X) = \sigma^2 < \infty$. Then $\bar{X} \xrightarrow{p} E(X)$, or $\bar{X} = E(X) + o_p(1)$.

Proof. Consider $\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N\bar{X}}$. Then $E(\bar{X}) = \mu$ and $\text{Var}(\bar{X}) = \frac{\sigma^2}{N}$. Using Chebyshev's inequality for $\bar{X}$:

$$\text{Prob}\left(|\bar{X}-\mu| \geq k\frac{\sigma}{\sqrt{N}}\right) \leq \frac{1}{k^2}$$

Let $k = \frac{\varepsilon\sqrt{N}}{\sigma}$, where $\varepsilon > 0$. Then

$$\text{Prob}(|\bar{X}-\mu| \geq \varepsilon) \leq \frac{\sigma^2}{N\varepsilon^2}$$

As $N \longrightarrow \infty$, the right-hand side of the above inequality converges to 0 . By the squeeze theorem we have that $\text{Prob}(|\bar{X}-\mu| \geq \varepsilon) \longrightarrow 0$. That is, we derived the definition of convergence in probability for $\bar{X}$. So, $\text{plim}(\bar{X}) = \mu$, or $\bar{X} \xrightarrow{p} E(X)$.

4.3.3 Consistency I

Let $k \in \{1, \ldots, K\}$. Consider β_k - i.e, the k-th element in the vector of model parameters, $\beta-$ and its OLS estimator $\widehat{\beta}_k$. Since $\widehat{\beta}_k$ is a random variable, we can apply Chebyshev's inequality, provided that its mean and variance are finite. We know that $E\left(\widehat{\beta}_k\right) = \beta_k$ and $\text{Var}\left(\widehat{\beta}_k\right) = \sigma^2 c_{kk}$, where c_{kk} is the k-th element along the main diagonal of $(X'X)^{-1}$. Hence:

$$\text{Prob}\left(\left|\widehat{\beta}_k - \beta_k\right| \geq \alpha\right) \leq \frac{\text{Var}\left(\widehat{\beta}_k\right)}{\alpha^2} = \frac{\sigma^2 c_{kk}}{\alpha^2}$$

If we fix α, as N increases to infinity, all the elements along the main diagonal of the matrix $X'X$ will increase. As a result, the elements along the main diagonal of $(X'X)^{-1}$ will be shrinking to zero, including c_{kk}. It follows that

$\lim_{N\to\infty} \frac{\sigma^2 c_{kk}}{\alpha^2} = 0$. The implication is that $\lim_{N\to\infty} \text{Prob}\left(\left|\widehat{\beta}_k - \beta_k\right| > \alpha\right) = 0$, that is $\text{plim}\left(\widehat{\beta}_k\right) = \beta_k$. In other words, $\widehat{\beta}_k$ is a consistent estimator of β_k.

4.3.4 Consistency II

To show the consistency of $\widehat{\beta}$, recall that

$$\widehat{\beta} = (X'X)^{-1} X'y = \beta + (X'X)^{-1} X'\varepsilon \Longrightarrow \widehat{\beta} - \beta = (X'X)^{-1} X'\varepsilon$$

from which

$$\begin{aligned}
\text{plim}(\widehat{\beta} - \beta) &= \text{plim}\left[\beta + (X'X)^{-1} X'\varepsilon - \beta\right] \\
&= \text{plim}\left[(X'X)^{-1} X'\varepsilon\right] \\
&= p\,\text{plim}\left[\left(\frac{X'X}{N}\right)^{-1} \frac{X'\varepsilon}{N}\right].
\end{aligned}$$

Assume that $\frac{X'X}{N}$ converges to some non-singular matrix, $\sum_{XX}$. Then

$$\text{plim}(\widehat{\beta} - \beta) = \sum_{XX}^{-1} plim \frac{X'\varepsilon}{N}.$$

Note that

$$\text{plim}(\widehat{\beta} - \beta) = \sum_{XX}^{-1} p\lim \frac{X'\varepsilon}{N} = \sum_{XX}^{-1} E\left(X'\varepsilon\right) = 0 \Longleftrightarrow E\left(X'\varepsilon\right) = 0$$

From the Gauss-Markov assumptions, X and ε are statistically independent. So we have that $E\left(X'\varepsilon\right) = 0$ and that $\widehat{\beta}$ is a consistent estimator for β, as long as we assume $\frac{X'X}{N} \xrightarrow{p} \sum_{XX}$, where $\sum_{XX}$ is a non-singular matrix.

4.4 Statistical Inference

Under the Gauss-Markov assumptions, $E(\varepsilon) = E(\varepsilon \mid X) = 0$ and $\text{Var}(\varepsilon) = \text{Var}(\varepsilon \mid X) = \sigma^2 I_N$. If we also have that $\frac{X'X}{N} \xrightarrow{p} \sum_{XX}$, then, by the Central Limit Theorem,

$$\sqrt{N}(\widehat{\beta} - \beta) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2 \sum_{XX}^{-1}\right).$$

Whereas in small samples, exact inference is only possible under the additional assumption of normally-distributed error term, in large samples inference is made possible by the central limit theorem and the weak law of large numbers.

The way to check whether the Gauss-Markov assumptions and the normality assumption are satisfied in small samples is to analyze the residuals of the model, looking for normality and non-autocorrelation. If the model is good, the residuals, which are estimates of the error terms, will be non-autocorrelated and will exhibit a constant variance. In small samples, if the residuals are normally distributed, we can safely use the inference techniques we are going to see in the remaining part of this course. In large samples, we do not strictly need the residuals to be normal, since we can rely on the central limit theorem and the weak law of large numbers for inference purposes.

4.5 Application

We will derive the properties of the OLS estimator in the case of a linear regression model with one exogenous regressor and a constant term - i.e., $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

4.5.1 Unbiasedness

We already know that $\widehat{\beta}_1 = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^N (x_i - \bar{x})^2}$ and that $\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$. To show that $\widehat{\beta}_1$ is an unbiased estimator:

$$
\begin{aligned}
&E\left(\widehat{\beta}_1\right) = E\left[\frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^N (x_i - \bar{x})^2}\right] \\
&= E\left[\frac{\sum_{i=1}^N (x_i - \bar{x}) y_i}{\sum_{i=1}^N (x_i - \bar{x})^2}\right] \\
&= E\left[\frac{\sum_{i=1}^N (x_i - \bar{x})(\beta_0 + \beta_1 x_i + \varepsilon_i)}{\sum_{i=1}^N (x_i - \bar{x})^2}\right] \\
&= E\left[\frac{\sum_{i=1}^N (x_i - \bar{x})\beta_0}{\sum_{i=1}^N (x_i - \bar{x})^2} + \beta_1 \frac{\sum_{i=1}^N (x_i - \bar{x}) x_i}{\sum_{i=1}^N (x_i - \bar{x})^2} + \frac{\sum_{i=1}^N (x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^N (x_i - \bar{x})^2}\right] \\
&= E\left[\frac{\beta_0}{\sum_{i=1}^N (x_i - \bar{x})} \sum_{i=1}^N (x_i - \bar{x})^2 \right. \\
&\left. \beta_1 \frac{\sum_{i=1}^N (x_i - \bar{x})(x_i - \bar{x})}{\sum_{i=1}^N (x_i - \bar{x})^2} + \frac{\sum_{i=1}^N (x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^N (x_i - \bar{x})^2}\right] \\
&= E\left[\beta_1 + \frac{\sum_{i=1}^N (x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^N (x_i - \bar{x})^2}\right] \\
&= \beta_1 + \frac{\sum_{i=1}^N (x_i - \bar{x}) E(\varepsilon_i)}{\sum_{i=1}^N (x_i - \bar{x})^2} \\
&= \beta_1.
\end{aligned}
$$

We could bring all constant or deterministic terms - including the explanatory variable, x_i - outside of the expectation operator. Since the expectation of ε is zero, $\widehat{\beta_1}$ is an unbiased estimator for β_1.

To show that $\widehat{\beta}_0$ is an unbiased estimator:

$$\begin{aligned}
E\left(\widehat{\beta}_0\right) &= E\left(\bar{y} - \widehat{\beta}_1 \bar{x}\right) \\
&= E\left[\frac{\sum_{i=1}^{N} y_i}{N} - \frac{\sum_{i=1}^{N}(x_i - \bar{x})\, y_i}{\sum_{i=1}^{N}(x_i - \bar{x})^2}\frac{\sum_{i=1}^{N} x_i}{N}\right] \\
&= E\left\{\frac{\sum_{i=1}^{N}(\beta_0 + \beta_1 x_i + \varepsilon_i)}{N} - \left[\beta_1 + \frac{\sum_{i=1}^{N}(x_i - \bar{x})\,\varepsilon_i}{\sum_{i=1}^{N}(x_i - \bar{x})^2}\right]\frac{\sum_{i=1}^{N} x_i}{N}\right\}
\end{aligned}$$

$$\begin{aligned}
&= E\left[\beta_0 + \frac{\beta_1 \frac{\sum_{i=1}^{N} x_i}{N}}{N} + \frac{\sum_{i=1}^{N}\varepsilon_i}{N} - \frac{\beta_1 \frac{\sum_{j=1}^{N} x_i}{N}}{N} - \frac{\sum_{i=1}^{N} x_i}{N}\frac{\sum_{i=1}^{N}(x_i - \bar{x})\,\varepsilon_i}{\sum_{i=1}^{N}(x_i - \bar{x})^2}\right] \\
&= \beta_0 + \frac{\sum_{i=1}^{N} E(\varepsilon_i)}{N} - \frac{\sum_{i=1}^{N} x_i}{N}\frac{\sum_{i=1}^{N}(x_i - \bar{x})\, E(\varepsilon_i)}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \\
&= \beta_0.
\end{aligned}$$

4.5.2 Consistency

To show that $\widehat{\beta}_1$ is consistent:

$$\begin{aligned}
\operatorname{plim}\left(\widehat{\beta}_1\right) &= \operatorname{plim}\left[\frac{\sum_{i=1}^{N}(x_i - \bar{x})\, y_i}{\sum_{i=1}^{N}(x_i - \bar{x})^2}\right] \\
&= \operatorname{plim}\left[\beta_1 + \frac{\sum_{i=1}^{N}(x_i - \bar{x})\,\varepsilon_i}{\sum_{i=1}^{N}(x_i - \bar{x})^2}\right] \\
&= \beta_1 + \operatorname{plim}\left[\frac{\sum_{i=1}^{N}(x_i - \bar{x})\,\varepsilon_i}{\sum_{i=1}^{N}(x_i - \bar{x})^2}\right] \\
&= \beta_1 + \operatorname{plim}\left[\frac{\frac{\sum_{i=1}^{N}(x_i - \bar{x})\varepsilon_i}{N}}{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}}\right] \\
&= \beta_1 + \operatorname{plim}\left[\frac{\operatorname{Cov}(x_i, \varepsilon_i)}{\widehat{\operatorname{Var}(x_i)}}\right] \\
&= \beta_1 + \frac{\operatorname{Cov}(x_i, \varepsilon_i)}{\Lambda_{xx}} \\
&= \beta_1,
\end{aligned}$$

where we assumed that $\operatorname{plim}\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N} = \Lambda_{xx} \neq 0$, given that the elements x_i are treated as deterministic. If the regressor is treated as a random variable,

then we need to assume that plim $\frac{\sum_{i=1}^{N}(x_i-\bar{x})^2}{N} = \text{Var}(x_i) \neq 0$. Note that, by the usual Gauss-Markov assumptions, since x_i and ε_i are independent of each other, their covariance is zero.

To prove the consistency property of $\widehat{\beta}_0$:

$$\begin{aligned}
\text{plim}\left(\widehat{\beta}_0\right) &= p\lim\left(\bar{y} - \widehat{\beta}_1\bar{x}\right) \\
&= p\lim\left[\frac{\sum_{i=1}^{N} y_i}{N} - \left(\beta_1 + \frac{\sum_{i=1}^{N}(x_i-\bar{x})\varepsilon_i}{\sum_{i=1}^{N}(x_i-\bar{x})^2}\right)\frac{\sum_{i=1}^{N} x_i}{N}\right] \\
&= p\lim\left[\frac{\sum_{i=1}^{N}(\beta_0+\beta_1 x_i+\varepsilon)}{N} - \beta_1\frac{\sum_{i=1}^{N} x_i}{N} - \frac{\sum_{i=1}^{N}(x_i-\bar{x})\varepsilon_i}{\sum_{i=1}^{N}(x_i-\bar{x})^2}\frac{\sum_{i=1}^{N} x_i}{N}\right]
\end{aligned}$$

$$\begin{aligned}
&= p\lim\left[\beta_0 + \frac{\beta_1\frac{\sum_{i=1}^{N} x_i}{N}}{N} + \frac{\sum_{i=1}^{N}\varepsilon_i}{N} - \beta_1\frac{\sum_{i=1}^{N} x_i}{N} - \frac{\sum_{i=1}^{N}(x_i-\bar{x})\varepsilon_i}{\sum_{i=1}^{N}(x_i-\bar{x})^2}\frac{\sum_{i=1}^{N} x_i}{N}\right] \\
&= \beta_0 + \text{plim}\left(\frac{\sum_{i=1}^{N}\varepsilon_i}{N}\right) - \text{plim}\left(\frac{\sum_{i=1}^{N}(x_i-\bar{x})\varepsilon_i}{\sum_{i=1}^{N}(x_i-\bar{x})^2}\frac{\sum_{i=1}^{N} x_i}{N}\right) \\
&= \beta_0 + E(\varepsilon_i) - \frac{\text{Cov}(x_i,\varepsilon_i)}{\Lambda_{xx}}\Upsilon_x \\
&= \beta_0
\end{aligned}$$

where we assumed that plim $\frac{\sum_{i=1}^{N}(x_i-\bar{x})^2}{N} = \Lambda_{xx} \neq 0$ and plim $\frac{\sum_{i=1}^{N} x_i}{N} = \Upsilon_x$, given that the elements x_i are treated as deterministic. If the regressor is treated as a random variable, then we need to assume that plim $\frac{\sum_{i=1}^{N}(x_i-\bar{x})^2}{N} = \text{Var}(x_i) \neq 0$ and plim $\frac{\sum_{i=1}^{N} x_i}{N} = E(x_i)$. Note that, by the Gauss-Markov assumptions, $\text{Cov}(x_i,\varepsilon_i) = 0$ and $E(\varepsilon_i) = 0$.

Also note that, in order to prove unbiasedness of the OLS estimator, one needs strict exogeneity (orthogonality) between the error term and the set of regressors. To prove the consistency of $\widehat{\beta}_1$ and $\widehat{\beta}_0$, we need a weaker condition than orthogonality. We just need the regressor and the error term to be contemporaneously uncorrelated. A condition that, of course, is implied by orthogonality.

5 Goodness of Fit

How can we tell how well a linear regression model is describing the data? How well does the estimated regression line fit the data? To what extent is the variance of y captured by the variance of the regressors in the linear model? In this section we will describe some of the many criteria that can be used to answer these questions.

5.1 Coefficient of Determination

A very common method to assess the goodness of fit of the linear regression model is the coefficient of determination, R^2.

Definition 5.1 (Coefficient of Determination)**.** The coefficient of determination represents the portion of the sample variance of y_i that is explained by the model. It is defined as

$$\begin{aligned} R^2 &= \frac{\widehat{\operatorname{Var}\left(\widehat{y}_i\right)}}{\widehat{\operatorname{Var}\left(y_i\right)}} \\ &= \frac{\sum_{i=1}^{N}\left(\widehat{y}_i - \overline{\widehat{y}}\right)^2}{\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2} \\ &= \frac{\sum_{i=1}^{N}\left(\widehat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2} \\ &= \frac{ESS}{TSS} \\ &= \frac{\text{Explained Sum of Squares}}{\text{Total Sum of Squares}} \end{aligned}$$

Note that $\overline{\widehat{y}} = \bar{y}$, when an intercept term is included in the model, a shortcut we took above. In fact,

$$y_i = x_i'\beta + \varepsilon_i \Longrightarrow y_i = \widehat{y}_i + \widehat{\varepsilon}_i \Longrightarrow \bar{y} = \overline{\widehat{y}} + \overline{\widehat{\varepsilon}}$$

Since $\overline{\widehat{\varepsilon}} = \frac{\sum_{i=1}^{N}\widehat{\varepsilon}_i}{N} = 0$ from the first-order condition of a linear model including an intercept term, it follows that $\bar{y} = \overline{\widehat{y}}$.

5.1.1 Properties of the Coefficient of Determination

Consider $y_i = \widehat{y}_i + \widehat{\varepsilon}_i$, from which it can be proved that $\widehat{\operatorname{Var}\left(y_i\right)} = \widehat{\operatorname{Var}\left(\widehat{y}_i\right)} + \widehat{\operatorname{Var}\left(\widehat{\varepsilon}_i\right)}$ if we have an intercept term in the model. So

$$R^2 = \frac{\widehat{\operatorname{Var}\left(\widehat{y}_i\right)}}{\widehat{\operatorname{Var}\left(y_i\right)}}$$

$$
\begin{aligned}
&= \frac{\widehat{\operatorname{Var}\left(y_i\right)} - \widehat{\operatorname{Var}\left(\widehat{\varepsilon}_i\right)}}{\widehat{\operatorname{Var}\left(y_i\right)}} \\
&= 1 - \frac{\widehat{\operatorname{Var}\left(\widehat{\varepsilon}_i\right)}}{\widehat{\operatorname{Var}\left(y_i\right)}} \\
&= 1 - \frac{\sum_{i=1}^{N} \widehat{\varepsilon}_i^2}{\sum_{i=1}^{N} \left(y_i - \bar{y}\right)^2} \\
&= 1 - \frac{RSS}{\text{TSS}} \\
&= 1 - \frac{\text{Residuals Sum of Squares}}{\text{Total Sum of Squares}}.
\end{aligned}
$$

Note that $R^2 \in [0,1]$. Assuming that the total sum of squares is strictly positive - i.e., the sample variance of the dependent variable is not equal to zero

$$R^2 = 1 \Longleftrightarrow RSS = 0 \Longleftrightarrow \widehat{\varepsilon}_i = 0, \forall i \Longrightarrow \widehat{\operatorname{Var}}(\widehat{\varepsilon}_i) = 0.$$

$$R^2 = 0 \Longleftrightarrow \widehat{\operatorname{Var}}(\widehat{\varepsilon}_i) = \widehat{\operatorname{Var}}(y_i) \Longleftrightarrow \widehat{\operatorname{Var}}(\widehat{y}_i) = 0 \Longleftrightarrow x_i'\widehat{\beta}.$$

is constant - i.e., the model contains only an intercept term.

If the model does not contain an intercept terms, we have that $\sum_{i=1}^{N} \widehat{\varepsilon}_i \neq 0$, which implies that R^2 can be negative, if computed as $R^2 = 1 - \frac{\sum_{i=1}^{N} \hat{\varepsilon}_i^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2}$. To avoid this anomaly, always include an intercept term in your model.

After a least squares regression with a constant plus a set of exogenous explanatory variables, R^2 equals the square of the correlation coefficient between the observed and modeled (fitted) data values - i.e., $R^2 = \left[\operatorname{Corr}\left(y_i, \widehat{y}_i\right)\right]^2$.

Now, let us consider the following two models and their corresponding R^2 's:

(i) $y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i \Longrightarrow R_A^2$;

(ii) $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \nu_i \Longrightarrow R_B^2$.

It can be proved that $R_B^2 \geq R_A^2$ and that adding other regressors to a model will weakly increase its coefficient of determination, even if those regressors are not necessary to correctly explain the dependent variable.

Finally, R^2 does not tell whether: the independent variables are a true cause of the changes in the dependent variable, the correct regression was used, the most appropriate set of independent variables has been chosen, the model might be improved by using transformed versions of the existing set of independent variables.

Definition 5.2 (Adjusted R^2)**.**

$$\tilde{R}^2 = 1 - \frac{\frac{1}{N-K} \sum_{i=1}^{N} \widehat{\varepsilon}_i^2}{\frac{1}{N-1} \sum_{i=1}^{N} \left(y_i - \bar{y}\right)^2},$$

where K is the number of regressors.

There is a negative relationship between the number of regressors and the adjusted R^2, which is just a modified version of the coefficient of determination. The adjusted R^2 introduces a penalty for the inclusion of additional explanatory variables. Also note that, generally, $\tilde{R}^2 < R^2$. The two are equal if and only if the model contains only a constant term, which implies that both the R^2 and the adjusted R^2 are equal to 0 .

6 Basic Review of Statistics - Part II

In this section we will cover the basic theory behind statistical inference. The general frameworks of confidence intervals and hypothesis testing will be described, examples based on the linear regression model discussed.

6.1 Confidence Intervals

When we use OLS to estimate the linear regression model, $y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots + \beta_K x_{Ki} + \varepsilon_i$, we obtain a set of point estimates, $\left\{\widehat{\beta}_i\right\}_{i=1}^{K}$, for the unknown model parameters. The generic $\widehat{\beta}_i$ is the estimator of the parameter β_i, a function of the sample of data, and, as such, a random variable. Sometimes, however, rather than just obtaining point estimates for the model parameters, we may want to construct interval estimates based on interval estimators.

6.1.1 Method of the Pivotal Quantity

Suppose X is a random variable such that $X \sim f_\theta(x)$, where $\theta \in \Theta$ is a vector of parameter which completely describes the probability density function, $f^{8\Theta}$ is the parameter space. Let us assume we have a random sample of N observations for the random variable, $X, \{X_i\}_{i=1}^{N}$. Let us define $\alpha \in (0,1)$. Popular choices of α are $0.01, 0.05$, and 0.10 . We would like to use the sample of data to estimate θ in the form of a confidence interval.

Definition (Confidence Interval). Let T_1 and T_2 be two statistics. They are random variables since they are functions of the random sample, that is, $T_1 = T_1(X_1, \ldots, X_N)$ and $T_2 = T_2(X_1, \ldots, X_N)$. If:
(i) $T_1 \leq T_2$ for any realization of the random sample, and
(ii) $\text{Prob}(T_1 \leq \theta \leq T_2) = 1 - \alpha$,
then the random interval (T_1, T_2) is a confidence interval for θ, with confidence level equal to $(1 - \alpha)$.

This means that, if we want to construct a confidence interval for θ, we have to find two statistics with such characteristics.

Definition (Pivotal Quantity). A pivotal quantity is a function $g(X_1, \ldots, X_N; \theta)$, which
(i) depends on the random sample and on the parameter(s) to be estimated (but not on other unknown parameters);

(ii) has a distribution that depends neither on θ nor on other unknown parameters, and

(iii) is continuous and invertible with respect to θ.

If we use this method to construct a confidence interval for θ, we must find two numbers $a, b \in \mathbb{R}$ such that

$$\text{Prob}\left[a \leq g\left(X_1, \ldots, X_N; \theta\right) \leq b\right] = 1 - \alpha, \quad \forall \theta$$

Since g is continuous and invertible with respect to θ, we can solve for θ to have

$$\text{Prob}\left[T_1\left(X_1, \ldots, X_N\right) \leq \theta \leq T_2\left(X_1, \ldots, X_N\right)\right] = 1 - \alpha, \quad \forall \theta$$

This is exactly the definition of confidence interval stated above. To sum up, if we can find a pivotal quantity for a given parameter, we can estimate a confidence interval for that parameter.

6.1.2 Application I

Consider a linear regression model with one regressor, $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. We know that $\widehat{\beta}_1 = \frac{\text{Cov}(x_i, y_i)}{\text{ar}(x_i)\sqrt{2}}$ and $\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$, which are the point estimates of the model parameters. This time we would like to estimate a confidence interval for β_1, i.e., find sample-based values for two suitable statistics, T_1 and T_2, such that $\text{Prob}\left(T_1\left(x_1, \ldots, x_N\right) \leq \beta_1 \leq T_2\left(x_1, \ldots, x_N\right)\right) = 1 - \alpha$, implying that (T_1, T_2) is a $(1-\alpha)$-level confidence interval for β_1.

Consider the expression

$$t = \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\frac{\widehat{\sigma}^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}}} = \frac{\widehat{\beta}_1 - \beta_1}{SE\left(\widehat{\beta}_1\right)}.$$

Is this a pivotal quantity for β_1 ? t is a function of the random sample of data and of the parameter of interest, β_1. It does not depend on any other unknown parameters. So the first condition for t to be a pivotal quantity is met.

This function is the ratio between a normal random variable with mean equal to zero and the square root of a χ^2-distributed random variable with $N - K$ degrees of freedom divided by $N - K$. It follows that

$$t \sim T_{N-K},$$

where $N - K$ is the number of degrees of freedom. A T distribution is very similar to a normal distribution with mean zero. The key difference is that the T has fatter tails than the standard normal. A nice property of the T distribution is that, as the degrees of freedom increase to infinity, the fatter tails get thinner and the T distribution approaches a standard normal distribution. As we have only one regressor and a constant term, $K = 2$. The probability density function

[148] For example, if $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$, then $\theta = (\mu, \sigma)$.

of t is fully defined by the number of its degrees of freedom, which, in this case, is a known parameter. So, the second condition is met.
Finally, t is continuous and invertible with respect to the parameter β_1. In fact, it is linear, so it is one-to-one and onto. Thus, the third condition is also satisfied and t is a pivotal quantity.

At this point, let us write the condition

$$\text{Prob}\left[a \leq \frac{\widehat{\beta}_1 - \beta}{SE\left(\widehat{\beta}_1\right)} \leq b\right] = 1 - \alpha \Longrightarrow \text{Prob}\left(a \leq t_{N-2} \leq b\right) = 1 - \alpha$$

where t_{N-2} is a generic random variable with a T_{N-2} distribution. We can use a plot of the T_{N-2} distribution, as in Figure 3, to find values for a and b such that $\text{Prob}\left(a \leq t_{N-2} \leq b\right) = 1 - \alpha$ is satisfied. If we want the confidence interval to be symmetric, $a + b = 0$ is an additional condition that must be satisfied. We then have

$$a = t_{N-2;\frac{\alpha}{2}} \text{ and } b = t_{N-2;1-\frac{\alpha}{2}},$$

where $t_{N-2;\frac{\alpha}{2}}$ is the value in the support of a T_{N-2} distribution, which leaves a probability mass equal to $\frac{\alpha}{2}$ on its left. Analogously, $t_{N-2;1-\frac{\alpha}{2}}$ is the value in the support of a T_{N-2} distribution, which leaves a probability mass equal to $1 - \frac{\alpha}{2}$ on its left. In other words, $\text{Prob}\left(t_{N-2} \leq t_{N-2;\frac{\alpha}{2}}\right) = \frac{\alpha}{2}$ and $\text{Prob}\left(t_{N-2} \leq t_{N-2;1-\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2}$. By symmetry, $t_{N-2;\frac{\alpha}{2}} = -t_{N-2;1-\frac{\alpha}{2}}$.

Figure 3: The T_{N-2} distribution used to calculate a and b. Each shaded area in the two tails contains a probability mass equal to $\frac{\alpha}{2}$.

So the probability above becomes

$$\text{Prob}\left[-t_{N-2;1-\frac{\alpha}{2}} \leq \frac{\widehat{\beta}_1 - \beta_1}{SE\left(\widehat{\beta}_1\right)} \leq t_{N-2;1-\frac{\alpha}{2}}\right] = 1 - \alpha$$

By inverting the pivotal quantity in the probability with respect to β_1, we have

$$\text{Prob}\left[\widehat{\beta}_1 - t_{N-2;1-\frac{\alpha}{2}} \cdot SE\left(\widehat{\beta}_1\right) \leq \beta_1 \leq \widehat{\beta}_1 + t_{1-\frac{\alpha}{2}} \cdot SE\left(\widehat{\beta}_1\right)\right] = 1 - \alpha$$

The $(1 - \alpha)$-level confidence interval for β_1 is hence

$$\left[\widehat{\beta}_1 - t_{N-2;1-\frac{\alpha}{2}} \cdot SE\left(\widehat{\beta}_1\right), \widehat{\beta}_1 + t_{N-2;1-\frac{\alpha}{2}} \cdot SE\left(\widehat{\beta}_1\right)\right].$$

If $N \longrightarrow \infty \Longrightarrow t \xrightarrow{d} \mathcal{N}(0,1)$ and $t_{N-2;1-\frac{\alpha}{2}} \longrightarrow z_{1-\frac{\alpha}{2}}$, where $z_{1-\frac{\alpha}{2}} = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$. The confidence interval for β_1 becomes

$$\left[\widehat{\beta}_1 - z_{1-\frac{\alpha}{2}} \cdot SE\left(\widehat{\beta}_1\right), \widehat{\beta}_1 + z_{1-\frac{\alpha}{2}} \cdot SE\left(\widehat{\beta}_1\right)\right].$$

What is the interpretation of this confidence interval? Once we run a regression and construct (for example) a 95% confidence interval for β_1, this does not mean that the true value of β_1 is contained in this interval with a 95% probability. It merely means that in repeated sampling from the reference population, if we estimate the same linear regression model and construct a confidence interval for β_1 at each replication, the true value of β_1 will be contained in those confidence intervals 95% percent of the times. In other words, when we estimate a confidence interval based on the unique sample of data which is available, we may be just 95% confident that the true value of β_1 lies in the estimated interval.

6.2 Hypothesis Testing

We will describe the general framework of hypothesis testing and show how to apply the method of the pivotal quantity onto statistical tests in the frame of the linear regression model. The general formulation of a hypothesis testing problem requires:
(i) a population with a known shape/form and an unknown parameter to be estimated, that is, a random variable, $X \sim f_\theta(x)$, where $\theta \in \Theta$ is the parameter of interest;
(ii) two statistical hypotheses,

$$\begin{cases} H_0: & \theta \in \Theta_0 \qquad\qquad \text{(null hypothesis)} \\ H_1: & \theta \in \Theta_1 \quad \text{(alternative hypothesis)} \end{cases}$$

where Θ_1 and Θ_0 represent a partition of the parameter space, Θ. In other words, $\Theta_0, \Theta_1 \subseteq \Theta$, such that $\Theta_0 \cap \Theta_1 = \emptyset$ and $\Theta_0 \cup \Theta_1 = \Theta$. A statistical hypothesis on θ is a claim on the value of θ.
(iii) a random sample, $\{X_i\}_{i=1}^N$.

A test, relative to a generic hypothesis, is a procedure through which we decide whether to reject the null hypothesis, H_0, or not based on the information within the random sample. The test is defined completely by a two-way partition of the sample space (Θ_s - i.e., the set of all possible values of $\widehat{\theta}\Big)$ into a rejection, or critical, region, R, and an acceptance region, A, such that $R \cup A = \Theta_s$, and $R \cap A = \emptyset$. If $\widehat{\theta} \in R$, then we reject the null hypothesis. If $\widehat{\theta} \in A$, then we do not reject the null hypothesis.

Definition (Type I and Type II Errors). A Type I Error is the error we make if we reject H_0 and H_0 is true - i.e., when the true value of the parameter, θ, lies in Θ_0 (false positive). A Type II Error, conversely, is the error we make if we fail to reject H_0 and H_0 is false - i.e., the parameter, θ, lies in Θ_1 (false negative).

Definition (Power Function). The power function of a given test with critical region, R, is a function $\pi_R : \Theta \longrightarrow [0, 1]$ such that

$$\underset{\theta \in \Theta}{\pi_R}(\theta) = \text{Prob}(\widehat{\theta} \in R) = \text{ Probability of rejecting } H_0 \text{ as } \theta \text{ changes.}$$

According to the definition above, the power function describes the probability of rejecting the null even when H_0 is true. In such a case, it provides the probability of Type I error:

$$\underset{\theta \in \Theta_0}{\pi_R}(\theta) = \text{Prob (Type I Error)}$$

When $\theta \in \Theta_1$ - i.e., the null hypothesis is false - then the power function provides the probability of correctly rejecting H_0 when H_0 is false. Then,

$$1 - \underset{\theta \in \Theta_1}{\pi_R}(\theta) = \text{Prob (Type II Error)}.$$

In general,

$$1 - \underset{\theta \in \Theta}{\pi_R}(\theta) = \text{Prob}(\widehat{\theta} \in A) = \text{ Probability of not rejecting } H_0 \text{ as } \theta \text{ changes.}$$

Finally, the power function changes as the test of interest changes.
Definition (Size of a Test). The size of a test is defined as $\sup_{\theta \in \Theta_0} \pi_R(\theta)$. It represents the "maximum" probability of rejecting the null when the null is actually true, or, otherwise, the highest probability of Type I error.

Definition (Power of a Test). The power of a test is defined as $\underset{\theta \in \Theta_1}{\pi_R}(\theta)$. It represents the probability of correctly rejecting the null. As the power of a test increases, the probability of Type II Error falls.

6.2.1 Neyman-Pearson Approach

According to the Neyman-Pearson approach to hypothesis testing, we first fix the size, $\alpha \in (0, 1)$, of the test and then construct a test such that $\sup_{\theta \in \Theta_0} \pi_R(\theta) \leq \alpha$. Thus, we can minimize the probability of a Type I error by controlling the size. While we have full control on this probability, we do not have any on the probability of Type II error. However within the class of tests satisfying the above condition, we can choose the one with the highest power. Such a test is known as uniformly most powerful test.

6.2.2 Application II

We will now use this approach to run statistical tests on the slope coefficient of the linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. We will execute the so-called two-sided t-test on β_1,

$$\begin{cases} H_0: & \beta_1 = 0 \\ H_1: & \beta_1 \neq 0 \end{cases}.$$

The first step is to construct a rejection region for this statistical test. We will make use of the method of the pivotal quantity. Intuitively, the rejection region, R, should be based on the solution of the following inequality, $\left|\widehat{\beta}_1\right| > h$.

That is, we want to fix a value for h, such that, when the estimator, $\widehat{\beta}_1$, is large enough in absolute value, we will reject the null hypothesis. So,

$$\begin{aligned} & \widehat{\beta}_1 < -h \quad \vee \widehat{\beta}_1 > h \\ \Longrightarrow & \frac{\widehat{\beta}_1 - \beta_1}{SE\left(\widehat{\beta}_1\right)} < \frac{-h - \beta_1}{SE\left(\widehat{\beta}_1\right)} \vee \frac{\widehat{\beta}_1 - \beta_1}{SE\left(\widehat{\beta}_1\right)} > \frac{h - \beta_1}{SE\left(\widehat{\beta}_1\right)}. \end{aligned}$$

Under the Neyman-Pearson approach, $\text{Prob}_{\theta \in \Theta_0}(\widehat{\theta} \in R) = \alpha$. In this case, $\theta = \beta_1$, and $\Theta_0 = \{0\}$. As such, under the null hypothesis, we have $\beta_1 = 0$, from which

$$\text{Prob}_{\beta_1 = 0}\left[\frac{\widehat{\beta}_1}{SE\left(\widehat{\beta}_1\right)} < \frac{-h}{SE\left(\widehat{\beta}_1\right)} \vee \frac{\widehat{\beta}_1}{SE\left(\widehat{\beta}_1\right)} > \frac{h}{SE\left(\widehat{\beta}_1\right)}\right] = \alpha$$

From the section on confidence intervals, we know that $\frac{\widehat{\beta}_1}{S.E.\left(\widehat{\beta}_1\right)} \sim T_{N-K}$, with $K = 2$ and N given, is a pivotal quantity. The rejection region can then be derived from

$$t_{N-2} < \frac{-h}{SE\left(\widehat{\beta}_1\right)} \quad \vee \quad t_{N-2} > \frac{h}{SE\left(\widehat{\beta}_1\right)}$$

where $t_{N-2} \sim T_{N-2}$ is a generic T-distributed random variable with $N-2$ degrees of freedom.

Since the distribution of t_{N-2} and all its characteristics are known, see Figure 4 , we can solve the equation

$$\text{Prob}\left[t_{N-2} < \frac{-h}{SE\left(\widehat{\beta}_1\right)} \quad \vee \quad t_{N-2} > \frac{h}{SE\left(\widehat{\beta}_1\right)}\right] = \alpha$$

for h, to see that

$$\begin{aligned} & t_{N-2;\frac{\alpha}{2}} = -t_{N-2;1-\frac{\alpha}{2}} = -\frac{h}{SE\left(\widehat{\beta}_1\right)} \\ \Longrightarrow & h = -t_{N-2;\frac{\alpha}{2}} \cdot SE\left(\widehat{\beta}_1\right) = t_{N-2;1-\frac{\alpha}{2}} \cdot SE\left(\widehat{\beta}_1\right). \end{aligned}$$

Figure 4: Construction of the rejection region in the example t-test, given a T_{N-K} distribution.

So, after estimating the linear regression model, we will reject the null of the statistical test above if $\left|\widehat{\beta}_1\right| > t_{N-2;1-\frac{\alpha}{2}} \cdot SE\left(\widehat{\beta}_1\right)$. If $N \longrightarrow \infty$, then we reject if $\left|\widehat{\beta}_1\right| > z_{1-\frac{\alpha}{2}} \cdot SE\left(\widehat{\beta}_1\right)$.

6.2.3 Application III

This time, after estimating the same linear regression model as in the previous example, we want to test whether β_1 is greater than a particular value γ or not (one-sided t-test):

$$\begin{cases} H_0: & \beta_1 > \gamma \\ H_1: & \beta_1 \leq \gamma \end{cases}$$

The construction of the rejection region for this test follows the same logic and intuition we saw in the previous example. The rejection region, R, will be defined over the portion of the sample space where $\widehat{\beta}_1 \leq h$, with h to be determined. According to the Neyman-Pearson approach,

$$\text{Prob}_{\beta_1>\gamma}\left(\widehat{\beta}_1 \leq h\right) = \alpha$$

$$\Longrightarrow \text{Prob}_{\beta_1>\gamma}\left[\frac{\widehat{\beta}_1 - \beta_1}{SE\left(\widehat{\beta}_1\right)} \leq \frac{h - \beta_1}{SE\left(\widehat{\beta}_1\right)}\right] = \alpha.$$

Under the null, we can replace β_1 with γ. We know that $\frac{\widehat{\beta}_1 - \gamma}{SE\left(\widehat{\beta}_1\right)} \sim T_{N-K}$, with $K = 2$, so we can impose the condition

$$\text{Prob}\left[t_{N-2} \leq \frac{h - \gamma}{SE\left(\widehat{\beta}_1\right)}\right] = \alpha$$

to be solved with respect to h. It follows that

$$\frac{h - \gamma}{SE\left(\widehat{\beta}_1\right)} = t_{N-2;\alpha} \Longrightarrow h = \gamma + t_{N-2;\alpha} \cdot SE\left(\widehat{\beta}_1\right).$$

We hence reject the null hypothesis if $\widehat{\beta}_1 \leq \gamma + t_{N-2;\alpha} \cdot SE\left(\widehat{\beta}_1\right)$. If $N \longrightarrow \infty$, then we reject if $\widehat{\beta}_1 \leq \gamma + z_\alpha \cdot SE\left(\widehat{\beta}_1\right)$.
Of course, if the form of the test is

$$\begin{cases} H_0: & \beta_1 < \gamma \\ H_1: & \beta_1 \geq \gamma \end{cases},$$

following the same line of reasoning as above, we reject the null hypothesis if $\widehat{\beta}_1 \geq \gamma + t_{N-2;1-\alpha} \cdot SE\left(\widehat{\beta}_1\right)$. If $N \longrightarrow \infty$, then we reject if $\widehat{\beta}_1 \geq \gamma + z_{1-\alpha} \cdot SE\left(\widehat{\beta}_1\right)$.

7 Statistical Inference and Prediction in the OLS Framework

The Gauss-Markov assumptions and the normality of the error term guarantee that

$$\widehat{\beta} \sim \mathcal{N}\left(\beta, \sigma^2 \left(X'X\right)^{-1}\right)$$

and

$$\widehat{\beta}_k \sim \mathcal{N}\left(\beta_k, \sigma^2 c_{kk}\right)$$

where c_{kk} is the element at the intersection of the k-th row and k-th column of the matrix $\left(X'X\right)^{-1}$. This normality result for the OLS estimator holds true in small samples under the normality assumption on the error term and is approximately true in large samples, even if we do not assume normality for the error term, thanks to the central limit theorem.

We can write

$$z_k = \frac{\widehat{\beta}_k - \beta_k}{\sigma\sqrt{c_{kk}}} \sim \mathcal{N}(0,1).$$

Since we do not know the value of σ, we have to estimate it. If we replace σ with its unbiased estimate, $\widehat{\sigma}$, then, in small samples,

$$t_k = \frac{\widehat{\beta}_k - \beta_k}{\widehat{\sigma}\sqrt{c_{kk}}} \sim T_{N-K},$$

with $\widehat{\sigma}^2 = \frac{\sum_{i=1}^{N} \widehat{\varepsilon}_i^2}{N-K}$. The numerator of t_k is a normally distributed random variable with mean equal to zero, its denominator is the square root of a χ^2 distributed random variable (because $\widehat{\varepsilon}_i^2$ is normal, from which $\sum_{i=1}^{N} \widehat{\varepsilon}_i^2$ has a χ^2 distribution) divided by $N-K$. It follows that $t_k \sim T_{N-K}$.

7.1 Simple Hypothesis on a Coefficient: t-Test

Consider the model $y = X\beta + \varepsilon$, where $\beta \in M(K,1)$. We estimate this vector of parameters with $\widehat{\beta} \in M(K,1)$ and then test the following simple hypothesis concerning the k-th element in β :

$$\begin{cases} H_0: & \beta_k = \beta_k^0 \\ H_1: & \beta_k \neq \beta_k^0 \end{cases}$$

Under the null, the statistic $t_k = \frac{\widehat{\beta}_k - \beta_k^0}{SE\left(\widehat{\beta}_k\right)} \sim T_{N-K}$ can be used to run the test. The idea is straightforward. Once β_k has been estimated using $\widehat{\beta}_k$, if the difference between $\widehat{\beta}_k$ and β_k^0 is large in absolute value, we will reject the null hypothesis. In other words, we will reject the null if the absolute value of t_k

is large. The definition of "large" in this case depends on the size, α, of the test. The null will be rejected if the probability of observing a value of $|t_k|$ or larger is smaller than a given significance level, α. Said in yet another way, we will reject H_0 if the value of t_k we observe is bigger, in absolute value, than a suitable critical value calculated using a T_{N-K} distribution.

According to this argument, we have two, perfectly equivalent, ways to run the test above and decide whether to reject or not the null hypothesis. We can compare the value of the test statistic with a critical value properly computed. The critical value, t^*, is chosen such that $\text{Prob}\,(|t| > t^*) = \alpha$, where $t \sim T_{N-K}$. It follows that $t^* = t_{N-K,1-\frac{\alpha}{2}}$. For example, for N large enough, H_0 is rejected at the 5% level if $|t_k| > 1.96$. Alternatively, we can base our decision on the socalled p-value, p, or probability level. The p-value is the probability of observing a more extreme value of the test statistic than the one we actually observe, assuming that the null hypothesis is true. In this case, $p = 2\,\text{Prob}\,(t > |t_k|)$, where $t \sim T_{N-K}$. If $p < \alpha$, we reject the null hypothesis.

In a one-sided test, the null and alternative hypotheses may be

$$\begin{cases} H_0: & \beta_k \leq \beta_k^0 \\ H_1: & \beta_k > \beta_k^0 \end{cases}$$

Intuitively, we will reject the null for large values of the t statistic. The critical value, t^*, that we should use for running the test satisfies the condition $\text{Prob}\,(t > t^*) = \alpha$, from which $t^* = t_{N-K,1-\alpha}$. The p-value associated with this test is $p = \text{Prob}\,(t > t_k)$.

7.2 Confidence Intervals

An interval estimate of a particular level of confidence, α, for β_k will be made of all values of β_k^0 for which the null that $\beta_k = \beta_k^0$ is not rejected by a t-test. That is, the following condition should be satisfied:

$$\begin{gathered} -t_{N-K;1-\frac{\alpha}{2}} < t_k < t_{N-K;1-\frac{\alpha}{2}} \\ \Longrightarrow \widehat{\beta}_k - t_{N-K;1-\frac{\alpha}{2}} \cdot SE\left(\widehat{\beta}_k\right) < \beta_k < \widehat{\beta}_k + t_{N-K;1-\frac{\alpha}{2}} \cdot SE\left(\widehat{\beta}_k\right), \end{gathered}$$

and

$$\left[\widehat{\beta}_k - t_{N-K;1-\frac{\alpha}{2}} \cdot SE\left(\widehat{\beta}_k\right); \widehat{\beta}_k + t_{N-K;1-\frac{\alpha}{2}} \cdot SE\left(\widehat{\beta}_k\right)\right]$$

is a $(1-\alpha)$-level confidence interval for β_k. If $\alpha = .05$, we have again that $t_{N-K;1-\frac{\alpha}{2}} \longrightarrow 1.96$, as $N \longrightarrow \infty$.

7.3 Linear Restriction of the Coefficients: t-Test

Consider the model

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots + \beta_K x_{Ki} + \varepsilon_i.$$

We would like to run a test on a linear combination of its parameters:

$$\begin{cases} H_0: & r_1\beta_1 + r_2\beta_2 + \ldots + r_K\beta_K = q \quad (\text{ or } r'\beta = q) \\ H_1: & H_0 \text{ is false} \end{cases}$$

where $r \in \mathbb{R}^K$ and $q \in \mathbb{R}$. Under the Gauss-Markov assumptions, $\widehat{\beta}$ is BLUE for β and $\text{Var}(\widehat{\beta}) = \sigma^2 (X'X)^{-1}$. It follows that
(i) the linear combination $r'\widehat{\beta}$ is BLUE for $r'\beta$, with variance $\text{Var}\left(r'\widehat{\beta}\right) = r' \text{Var}(\widehat{\beta}) r$;
(ii) $\widehat{\text{Var}\,(r'\,\widehat{\beta}}\,\Big)$ is the estimate of $\text{Var}\left(r'\widehat{\beta}\right)$;
(iii) $SE\left(r'\widehat{\beta}\right) = \sqrt{\text{Var}(\text{r}'\,\widehat{\beta}\,\Big)} = \sqrt{r' \text{Var}(\widehat{\beta}) r}$.

If the error is normally distributed, then under the null hypothesis $\widehat{\beta} \sim \mathcal{N}\left(\beta, \sigma^2 (X'X)^{-1}\right)$. This is approximately true in large samples even if the error term is not normal. Then

$$\begin{aligned} r'\widehat{\beta} &\sim \mathcal{N}\left(r'\beta, r' \text{Var}(\widehat{\beta}) r\right) \\ &\Longrightarrow \frac{r'\widehat{\beta} - r'\beta}{SE\left(r'\widehat{\beta}\right)} \sim T_{N-K} \\ &\Longrightarrow t = \frac{r'\widehat{\beta} - q}{SE\left(r'\widehat{\beta}\right)} \sim T_{N-K} \end{aligned}$$

We can then use this statistic to run the test above. In some cases, reparameterization is useful and more straightforward to run the same test.

7.3.1 A Reparameterization Trick

Instead of running a t-test, we might rearrange the terms in the model in a more convenient way so that running the test is more straightforward. For example, consider a model with two exogenous regressors:

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i.$$

We want to run the test

$$\begin{cases} H_0: & \beta_2 = \beta_3 \quad (\beta_2 - \beta_3 = 0) \\ H_1: & \beta_2 \neq \beta_3 \end{cases}.$$

In this case, $r = \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}$ and $q = 0$. The test statistic would be

$$t = \frac{\widehat{\beta}_2 - \widehat{\beta}_3}{SE\left(\widehat{\beta}_2 - \widehat{\beta}_3\right)}$$

However, we can just add and subtract $\beta_3 x_{2i}$ on the right-hand side of the equation to obtain the equivalent model:

$$\begin{aligned} y_i &= \beta_1 + (\beta_2 - \beta_3)\, x_{2i} + \beta_3\, (x_{3i} + x_{2i}) + \varepsilon_i \\ &= \beta_1 + \beta_2^* x_{2i} + \beta_3\, (x_{3i} + x_{2i}) + \varepsilon_i. \end{aligned}$$

At this point, to test the null hypothesis above, we can run the simple t-test

$$\begin{cases} H_0: & \beta_2^* = 0 \\ H_1: & \beta_2^* \neq 0 \end{cases}$$

whose test statistic is

$$t = \frac{\widehat{\beta}_2^*}{SE\left(\widehat{\beta}_2^*\right)}$$

7.4 Joint Test of Significance on Regression Coefficients: F-Test

Consider $J < K$ parameters in the linear regression model. This time we want to test the hypothesis that each of the J parameters are equal to zero. The alternative hypothesis is that at least one of them is not equal to zero:

$$\begin{cases} H_0: & \beta_{K-J+1} = \cdots = \beta_K = 0 \\ H_1: & H_0 \text{ is false} \end{cases}$$

In other words, we select J parameters in the linear model and test the joint hypothesis that they are all equal to zero. To run this test, let us estimate the unrestricted model first,

$$y_i = \beta_1 + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \varepsilon_i$$

i.e., the original model on which the restrictions we want to test have not been applied. We can thus calculate the residual sum of squares, $S_1 = RSS_U$, of the unrestricted model. Next, we can estimate the model with the restrictions in the null applied and the corresponding residual sum of squares, $S_0 = RSS_R$, for
the restricted equation.
Note that $S_0 > S_1$, because if we add zero-restrictions to the regression we omit some variables and the accuracy of the model deteriorates. [9] We can show that, under H_0,

$$\frac{S_0 - S_1}{\sigma^2} \sim \chi_J^2$$

where J is the number of restrictions and σ^2 is the variance of ε_i. In theory, we could use this statistic to run the test. The idea is that, if the restricted model is much worse than the unrestricted one at explaining y_i, S_0 would be much

bigger than S_1, the value of the test statistic would get big value and, based on the critical values of a χ^2_J distribution, we would reject the null hypothesis. The problem is that σ^2 is not known. The best we can do is to replace σ^2 with its unbiased estimate, $\widehat{\sigma}^2 = \frac{S_1}{N-K}$, and use

$$F_J = \frac{(S_0 - S_1)/J}{S_1/(N-K)} \sim F_{J,N-K}$$

as a test statistic. This test statistic is the ratio of two χ^2-distributed variables, an F-distributed random variable with degrees of freedom equal to the degrees of freedom of the two χ^2 variables at the numerator and denominator. Recall that $R^2 = 1 - \frac{RSS}{TSS}$. The test statistic is then equivalent to

$$F_J = \frac{\left(R_1^2 - R_0^2\right)/J}{\left(1 - R_1^2\right)/(N-K)}$$

where R_1^2 is the R^2 of the unrestricted model and R_0^2 is the R^2 of the restricted one. Intuitively, we will reject the null if S_0 is much bigger than S_1. That is, the larger the value of F_J (F_J is defined only on the positive real axis), the more likely we are to reject.

The actual test procedure is identical to the previous cases. we choose a size, α, to determine a critical value which defines the rejection region, R. If we reject - i.e., the value of the test statistic is larger than the critical value computed from an appropriate F distribution with given degrees of freedom - we will conclude that the restricted model does a poor job at explaining the variance of the dependent variable and should not be used. More specifically, the critical value, F^*, is chosen such that $\text{Prob}\left(F > F^*\right) = \alpha$, where $F \sim F_{J,N-K}$. It follows that $F^* = F_{J,N-K;1-\alpha}$. The p-value is then $p = \text{Prob}\left(F > F_J\right)$.

7.5 Multiple Linear Restrictions: Wald Test

Rather than testing a set of simple restrictions or just one linear restriction as in the previous paragraphs, sometimes we may be interested in running a test with multiple linear restrictions of the model parameters. That is, we may want to run the test:

$$\begin{cases} H_0: & R\beta = q \\ H_1: & R\beta \neq q \end{cases}$$

where $\beta \in \mathbb{R}^K$, $R \in \mathcal{M}(J,K)$ (the space of $J \times K$ matrices), and $q \in \mathbb{R}^J$. Here J denotes the number of linear restrictions in the null hypothesis. Under the normality assumption of the error term,

$$\widehat{\beta} \sim \mathcal{N}\left(\beta, \text{Var}(\widehat{\beta})\right)$$

and consequently,

[149] Remember the discussion on the R^2.

$$R\widehat{\beta} \sim \mathcal{N}\left(R\beta, \text{Var}(R\widehat{\beta})\right)$$

where $\text{Var}(R\widehat{\beta}) = R\,\text{Var}(\widehat{\beta})R'$. It can be shown that:

$$(R\widehat{\beta} - q)' \left[\text{Var}(R\widehat{\beta})\right]^{-1} (R\widehat{\beta} - q) \sim \chi^2_J$$

This follows because the quadratic form of a normal random vector with its covariance matrix inverse follows a chi-squared distribution. Note that the outer two terms are normal random variables and the center term is a constant unknown matrix. Replacing the unknown variance with its consistent estimator $\widehat{\text{Var}}(R\widehat{\beta})$, we obtain the asymptotic version:

$$(R\widehat{\beta} - q)' \left[R\widehat{\text{Var}}(\widehat{\beta})R'\right]^{-1} (R\widehat{\beta} - q) \overset{a}{\sim} \chi^2_J$$

This holds in large samples, if, as N grows to infinity, the denominator converges in probability to the true unknown population variance-covariance matrix of $R\hat{\beta}$. That is, if $\widehat{\text{Var}}(R\widehat{\beta}) \xrightarrow{p} \text{Var}(R\widehat{\beta})$ as $N \to \infty$.

To run this test, we calculate the sample value of the Wald test statistic,

$$W = (R\widehat{\beta} - q)' \left[R\widehat{\text{Var}}(\widehat{\beta})R'\right]^{-1} (R\widehat{\beta} - q)$$

using sample data. Intuitively, we will reject the null hypothesis if the test statistic is large ($R\hat{\beta}$ is far from q). We set a critical threshold (significance level) based on a size, α. Given the shape of a χ^2 distribution, we reject the null when the critical threshold is less than the realized value of the test statistic.

When the sample is not large, the statistic under consideration has an F distribution, which we will need to compute an appropriate critical value, given the size of the test. Under the Gauss-Markov assumptions and the normality of the error term,

$$R\widehat{\beta} \sim \mathcal{N}\left(R\beta, \sigma^2 R(X'X)^{-1}R'\right) \implies (R\widehat{\beta} - R\beta) \sim \mathcal{N}\left(0, \sigma^2 R(X'X)^{-1}R'\right)$$

So the Wald statistic can be rewritten as

$$\frac{(R\widehat{\beta} - R\beta)' \left[R(X'X)^{-1}R'\right]^{-1} (R\widehat{\beta} - R\beta)}{\sigma^2} \sim \chi^2_J$$

In this last expression, the only element we do not know is σ^2. Substituting σ^2 with its unbiased estimator $\widehat{\sigma}^2 = \frac{\sum_{i=1}^N \widehat{\varepsilon}_i^2}{N-K}$ yields:

$$W = \frac{\{(R\widehat{\beta} - R\beta)' \left[R(X'X)^{-1}R'\right]^{-1} (R\widehat{\beta} - R\beta)\}/J}{\widehat{\sigma}^2} \sim F_{J,N-K}$$

The intuition to use to run the test is the same as before: we will reject the null for large values of the test statistic.

The bottom line is that the statistic may have two distributions: in large samples it is χ^2-distributed, in small samples it is F-distributed under a set of conditions. Depending on the situation, the Wald test can be used either in its χ^2 or F forms to test multiple restrictions of the regression coefficients.

We need the Gauss-Markov assumptions - to assure the unbiasedness of $\widehat{\beta}$ - and the normality assumption of the error term for the test statistic to be distributed as an F in small samples. Using the χ^2 version of the test in small samples may lead to unreliable inference. If the errors are not independent and identically distributed, the Wald test in its F form may lead to unreliable inference as well, since the Wald statistic is F-distributed only if the Gauss-Markov assumptions hold.

If the errors are not i.i.d., the χ^2 version should be preferred, provided that the sample is large enough and that $\text{Var}(R\widehat{\beta})$ can be estimated consistently in presence of autocorrelation and/or heteroskedasticity.**If the errors are i.i.d., the two versions of the test are asymptotically equivalent**, as intuitively argued above.

Key considerations:

- Use χ^2 version for large samples or non-i.i.d. errors (with robust variance)
- Use F version for small samples under Gauss-Markov/normality
- As $N \to \infty$, $J \times F_{J,N-K} \xrightarrow{d} \chi^2_J$

7.6 Prediction

Consider the usual linear model in the form $y_i = x_i'\beta + \varepsilon_i$. Using this model, we want to predict the value of y for a given x_0, not in the original sample we are using for estimation. In other words, we want to predict y_0, where $y_0 = x_0'\beta + \varepsilon_0$. It follows that $\widehat{y}_0 = x_0'\widehat{\beta}$ is a predictor for y_0 and that $y_0 = \widehat{y}_0 + \widehat{\varepsilon}_0$.

We can show that the estimator, $\widehat{y}_0$, is unbiased for y_0 :

$$E(\widehat{\beta}) = \beta \Longrightarrow E\left(\widehat{y}_0\right) = E\left(y_0 - \widehat{\varepsilon}_0\right) = y_0$$

Since $\widehat{y}_0$ is a random variable, we can compute its variance,

$$\text{Var}\left(\widehat{y}_0\right) = \text{Var}\left(x_0'\widehat{\beta}\right) = x_0'\,\text{Var}(\widehat{\beta})x_0 = \sigma^2 x_0'\left(X'X\right)^{-1}x_0.$$

The prediction error is defined as $\widehat{y}_0 - y_0 = x_0'\widehat{\beta} - x_0'\beta - \varepsilon_0 = x_0'(\widehat{\beta} - \beta) - \varepsilon_0$, which, on average, should be equal to zero since $E\left(y_0 - \widehat{y}_0\right) = y_0 - y_0 = 0$.

Finally, the prediction error has a variance:

$$
\begin{aligned}
\operatorname{Var}(\widehat{y}_0 - y_0) &= \operatorname{Var}(\widehat{y}_0) + \operatorname{Var}(y_0) - 2\operatorname{Cov}(\widehat{y}_0, y_0) \\
&= \sigma^2 x_0' (X'X)^{-1} x_0 + \operatorname{Var}(x_0'\beta + \varepsilon_0) - 2\operatorname{Cov}(\widehat{y}_0, x_0'\beta + \varepsilon_0) \\
&= \sigma^2 x_0' (X'X)^{-1} x_0 + \operatorname{Var}(\varepsilon_0) - 2\widehat{\operatorname{Cov}(\widehat{y}_0, x_0'\beta)} - 2\operatorname{Cov}(\widehat{y}_0, \varepsilon_0) \\
&= \sigma^2 x_0' (X'X)^{-1} x_0 + \sigma^2 - 2\operatorname{Cov}\left(x_0'\widehat{\beta}, \varepsilon_0\right) \\
&= \sigma^2 x_0' (X'X)^{-1} x_0 + \sigma^2 \\
&= \left[1 + x_0' (X'X)^{-1} x_0\right] \sigma^2,
\end{aligned}
$$

under the assumption that $\widehat{\beta}$ and ε_0 are not correlated (in fact, ε_0 does not enter the expression of $\widehat{\beta}$). We can use the variance of the prediction error and the normality assumption in large samples to compute confidence intervals for y_0 or the prediction error itself.

8 Interpreting the Linear Regression Model

We have thus far looked at how to estimate the linear regression model and how to make inference on the model parameters. We will now learn how to interpret the coefficients of the model and, later, what criteria should be considered to determine whether or not a model is "good".

8.1 Marginal Effects

Consider the usual linear model, $y = X\beta + \varepsilon$, written in matrix notation. Under the assumption that $E(\varepsilon \mid X) = 0$, we can interpret the regression model as describing the expected value of y conditional on the values of the explanatory variables in X. That is, under the condition that ε and X are orthogonal, $E(y \mid X) = E(X\beta + \varepsilon \mid X) = E(X\beta \mid X) + E(\varepsilon \mid X) = X\beta$. If we apply OLS (or any other estimation method), how do we interpret the elements in $\widehat{\beta}$?

Consider the same linear regression model written in its alternative matrix notation, $y_i = x_i'\beta + \varepsilon_i$. It is easy to see that the k-th element in β, β_k, measures the expected change in y_i associated with a unit change in x_{ki}, when all the other variables in x_i are kept fixed (ceteris paribus condition). More precisely,

$$
\beta_k = \frac{\partial}{\partial x_{ki}} E(y_i \mid x_i) = \frac{\triangle E(y_i \mid x_i)}{\triangle x_{ki}}
$$

In other words, β_k represents the effect on the dependent variable of a marginal change in the k-th explanatory variable.

Example. The ceteris paribus condition is not always satisfied. Consider the model

$$
y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \varepsilon_i.
$$

The model is still linear in the parameters even if it contains a squared regressor. In fact,

$$X = \begin{bmatrix} 1 & x_{11} & x_{11}^2 \\ 1 & x_{12} & x_{12}^2 \\ \vdots & \vdots & \vdots \\ 1 & x_{1N} & x_{1N}^2 \end{bmatrix}$$

and $y = X\beta + \varepsilon$ can still be estimated by OLS. However,

$$\frac{\partial}{\partial x_{1i}} E\left(y_i \mid x_i\right) = \beta_1 + 2\beta_2 x_{1i}.$$

That is, the expected change in y_i associated with a change in x_{1i} also depends on the value of x_{1i} from which we start. Said in another way, the ceteris paribus condition does not hold in this case.

Example. Now consider the model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i} x_{2i} + \varepsilon_i.$$

By the same argument seen above, this model is linear in the parameters and can be estimated by OLS. The term $x_{1i}x_{2i}$ is an interaction term between x_{1i} and x_{2i}. To find the marginal effect of x_1 on y,

$$\frac{\partial}{\partial x_{1i}} E\left(y_i \mid x_i\right) = \beta_1 + \beta_2 x_{2i}$$

So, the marginal effect of x_{1i} on y_i does depend on another regressor and the ceteris paribus condition does not hold again.

8.2 Elasticities

Definition (Elasticity). The expected relative change in the dependent variable, y_i, associated with a relative change in the k-th variable in x_i, $\frac{\triangle\% E(y_i|x_i)}{\Delta\% x_{ki}}$, is referred to as the elasticity of y_i with respect to that variable. Said in another way, the elasticity of y_i with respect x_{ki} represents the percent change in y_i associated with a percent change in x_{ki}.

Suppose we have two variables, y_i and x_{ki}. We would like to compute the elasticity of y_i with respect to x_{ki}. To do so, we need to design a model such that, once we get the estimates of its coefficients, we can attach to each of them the meaning of elasticity. Specifically, we will use log-linear models. Consider

$$\log y_i = \left[\log x_i\right]' \gamma + \nu_i,$$

where $\log x_i = \begin{pmatrix} 1 & \log x_{2i} & \log x_{3i} & \cdots & \log x_{Ki} \end{pmatrix}'$ and $E\left(\nu_i \mid \log x_i\right) = 0$. This model is linear in the parameters. The elasticity of y_i with respect to x_{ki} is of the form

$$\frac{\partial}{\partial x_{ki}} E\left(y_i \mid x_i\right) \frac{x_{ki}}{E\left(y_i \mid x_i\right)} \approx \frac{\partial}{\partial \log x_{ki}} E\left(\log y_i \mid \log x_i\right) = \gamma_k$$

In a log-linear model elasticities are constant, whereas in a linear model elasticities are not constant. In fact,

$$\frac{\partial}{\partial x_{ki}} E\left(y_i \mid x_i\right) \frac{x_{ki}}{E\left(y_i \mid x_i\right)} = \frac{x_{ki}}{x_i'\beta}\beta_k.$$

That is, the elasticities vary with x_i.
Example. Consider the model

$$\log y_i = \beta_1 + \beta_2 \log x_{2i} + \beta_3 \log x_{3i} + \varepsilon_i$$

What is the meaning that should be given to, say, β_2 ? Let us take the total derivative of the model:

$$\frac{1}{y_i} dy_i = \frac{\beta_2}{x_{2i}} dx_{2i} + \frac{\beta_3}{x_{3i}} dx_{3i} + d\varepsilon_i.$$

Under the ceteris paribus condition, if we want to calculate the marginal effect of x_{2i} on $y_i, dx_{3i} = d\varepsilon_i = 0$ and:

$$\frac{dy_i}{y_i} \frac{x_{2i}}{dx_{2i}} = \beta_2 \Longrightarrow \beta_2 = \frac{\triangle \% y_i}{\triangle \% x_{2i}},$$

which is clearly the elasticity of y_i with respect to x_{2i}.

8.3 When to Use the Log-Linear Model?

How should we choose between a linear model and a log-linear model? We should consider the following aspects:
(i) The economic interpretation we would like to give to the model. Are we interested in absolute changes or elasticities?
(ii) The log-linear model may help reduce heteroskedasticity problems by decreasing the variance of the dependent variable. [10]
(iii) It is not always possible to take logs of variables, as the logarithmic function is only defined on a positive real domain.
(iv) We can apply logs to some regressors and leave the others in levels.

8.4 Semi-Elasticities

Consider

$$\log y_i = x_i'\beta + \varepsilon_i.$$

Then we can compute

$$\beta_k = \frac{\Delta \% E\left(y_i \mid x_i\right)}{\Delta x_{ki}} = \frac{\partial}{\partial x_{ki}} E\left(\log y_i \mid x_i\right) = \frac{\partial}{\partial x_{ki}} E\left(y_i \mid x_i\right) \frac{1}{E\left(y_i \mid x_i\right)},$$

which is the semi-elasticity of y_i with respect to x_{ki}, representing the percent change in y_i associated with a unit change in x_{ki}.

9 Model Selection and Multicollinearity

In the process of determining what makes a "good" model, we are interested in how one goes about selecting the regressors to include and the potential problems with poor selection. We will see that, when a relevant variable is excluded from the model, we have bias in the OLS estimator; and that, when an irrelevant variable is included in the model, then the OLS estimator is unbiased but no longer efficient - i.e., it does not exhibit anymore the lowest variance. Both situations represent a violation of the Gauss-Markov assumptions. Generally speaking, if the Gauss-Markov assumptions are violated, then the OLS estimator is not BLUE.

Consider a scenario where we have two competing models, $(A) y = X\beta + Z\gamma + \varepsilon$ and $(B) y = X\beta + \nu$, where X is an $N \times K$ matrix and Z has dimension $N \times W$ (that is, in model A we have W additional regressors, each with N observations). β is $K \times 1, \gamma$ is $W \times 1. y$ and ε are both $N \times 1$. Finally, assume that $\text{Var}(\varepsilon_i) = \text{Var}(\nu_i) = \sigma^2$. We need to decide which model is best and should be used for estimation.

9.1 Under-Specified Models: Omitted Variable Bias

Let us assume that model A is the "correct" or "true" model. That is, if we want to correctly explain y, we need to include the regressors in both X and Z. However, rather than estimating model A, we estimate model B. The question is: what happens to the OLS estimator, $\widehat{\beta}$? We will show that if we make this kind of mistake in the estimation procedure - that is, we estimate an underspecified model - then $\widehat{\beta}$ is biased for β.

Under model B,

$$\widehat{\beta}_B = (X'X)^{-1} X'y,$$

in which formula we can plug the expression for y from the true model, A. Whence we see that

$$\begin{aligned}
\widehat{\beta}_B &= (X'X)^{-1} X'(X\beta + Z\gamma + \varepsilon) \\
&= (X'X)^{-1} X'X\beta + (X'X)^{-1} X'Z\gamma + (X'X)^{-1} X'\varepsilon \\
&= \beta + (X'X)^{-1} X'Z\gamma + (X'X)^{-1} X'\varepsilon
\end{aligned}$$

To show that $\widehat{\beta}_B$ is biased:

$$\begin{aligned}
E\left(\widehat{\beta}_B\right) &= E\left[\beta + (X'X)^{-1} X'Z\gamma + (X'X)^{-1} X'\varepsilon\right] \\
&= \beta + (X'X)^{-1} X'Z\gamma + (X'X)^{-1} X'E(\varepsilon)
\end{aligned}$$

[14][10]We will analyze the problem of heteroskedasticity later in this course.

$$= \beta + (X'X)^{-1} X'Z\gamma,$$

under the assumption that X and Z are deterministic. If $(X'X)^{-1} X'Z\gamma \neq 0$, $\widehat{\beta}$ is biased for β. It follows that

$$E(\widehat{\beta}) = \beta \iff (X'X)^{-1} X'Z\gamma = 0 \iff X'Z = 0 \text{ and/or } \gamma = 0$$

If $X'Z = 0$, then there is no covariance between X and Z. If $\gamma = 0$, there is no "true" relationship between Z and y and the set of regressors in Z are not necessary in the model. In empirical applications, $X'Z$ will almost always be different from zero. So, if we exclude (omit) a relevant variable from the model - i.e., $\gamma \neq 0$ - we will get a biased estimator for β.

The sign of the bias, $E(\widehat{\beta}) - \beta = (X'X)^{-1} X'Z\gamma$, is important. The bias is positive if and only if $X'Z > 0$ and $\gamma > 0$, or $X'Z < 0$ and $\gamma < 0$. Otherwise, the bias is negative.

The Gauss-Markov assumption we have violated in this case, which leads to bias in the OLS estimator, is the assumption that $E(\nu \mid X) = 0$. In fact, if A is the true model, then $\nu = Z\gamma + \varepsilon$, and $E(\nu \mid X) = E(Z_\gamma \mid X) + E(\varepsilon \mid X) = \gamma E(Z \mid X)$, which is different from zero if $\gamma \neq 0$ and $E(Z \mid X) \neq 0$.

Example. Suppose student k in this class comes to my office the day before the final exam and asks me to be excused from it. I tell her she can skip the test, but I also propose to her a way to give her a grade for the final test, so that I will be later able to compute a final grade for the course. I tell her that, using the grades of her classmates, I will run the regression:

$$\text{final}_{\ i} = \beta_0 + \beta_1 \text{ midterm}_{\ i} + \beta_2 \text{ midterm}_{\ i} + \beta_3 \text{ homework}_{\ i} + \nu_i,$$

and predict her grade on the final based on the coefficient estimates from the regression above and her actual grades on the midterms and the homework assignments.

However, the grade a student gets on the final will also depend on the effort she puts into it. That is, in the regression above I am omitting some term β_4 effort $_i$, which might helpful for the explanation of the dependent variable. A regression with such an explanatory variable cannot be run, though, since I have no way to reliably measure effort. So, the $\widehat{\beta}$ estimates in the regression I can run will be biased, since I am excluding at least one relevant regressor from the model. In the setup of this example, Z contains the omitted regressor, effort. β_4 is the γ coefficient we defined in the theoretical paragraph above.

Assume that $\beta_4 > 0$ - i.e., that effort and the grade on final are positively related. Also assume that the covariance between effort and scores in prior tests and assignments is positive as well. Usually a student that put a lot of effort in the previous tests will do the same for the preparation of the final. Since, under these assumptions, $X'Z > 0$ and $\gamma > 0$, then bias will be positive and the expected performance of student k is likely to be overestimated. So, if I apply this approach to give her a grade for the final, at least on average, I am really

doing her a favor! In reality, we have no way of knowing what the true sign of the covariance between effort on the final and previous performance in the course is. In fact, it may be the case that the effort for the final will be lower if students already got good grades in the previous tests and the homework assignments, and viceversa. If this is what happens in reality, then bias will be negative and the expected performance of student k is likely to be underestimated.

9.2 Over-Specified Models: Inefficient OLS Estimator

Assume that model B is the correct model, but we estimate model A. We will find that $\widehat{\beta}_A$ is unbiased for β, but also that $\widehat{\beta}_A$ is not BLUE, that is, it will not be an efficient estimator (it will exhibit higher variance).

9.2.1 Unbiasedness of the OLS Estimator in the Over-Specified Model

What follows is an informal explanation of the reason why including unnecessary regressors in a linear regression model does not lead to a biased OLS estimator. Consider model $A, y = X\beta + Z\gamma + \varepsilon$. If we estimate this model, $\widehat{\beta}_A$ and $\widehat{\gamma}_A$ are the estimators for β and γ, respectively. We can rewrite the model more compactly. Define the objects

$$\tilde{\beta} = \begin{bmatrix} \beta \\ \gamma \end{bmatrix} \quad \tilde{X} = \begin{bmatrix} X & Z \end{bmatrix}$$

of dimensions $(K+W) \times 1$ and $N \times (K+W)$, respectively. Under this new notation, model A is equivalent to $y = \tilde{X}\tilde{\beta} + \varepsilon$. The estimator $\widehat{\tilde{\beta}}$, which will include the two estimators of the coefficients associated with X and Z, is then given by

$$\widehat{\tilde{\beta}} = \left(\tilde{X}'\tilde{X}\right)^{-1} \tilde{X}'y$$

Since the correct model is B, we can replace y with it true expression in B :

$$\begin{aligned} \widehat{\tilde{\beta}} &= \left(\tilde{X}'\tilde{X}\right)^{-1} \tilde{X}'(X\beta + \nu) \\ &= \left(\tilde{X}'\tilde{X}\right)^{-1} \tilde{X}'X\beta + \left(\tilde{X}'\tilde{X}\right)^{-1} \tilde{X}'\nu \end{aligned}$$

What is $\left(\tilde{X}'\tilde{X}\right)^{-1} \tilde{X}'X$? It is the estimator matrix, $\widehat{\delta}$ (of dimensions $(K+W)\times K$), of a regression of X on $\tilde{X}$:

$$X = \tilde{X}\delta + \eta = \begin{bmatrix} X & Z \end{bmatrix} \delta + \eta = X\delta_1 + Z\delta_2 + \eta$$

where $\delta = \begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix}$, with $\delta_1 \in M(K,K)$ and $\delta_2 \in M(W,K)$. It follows that $\widehat{\delta} = \left(\tilde{X}'\tilde{X}\right)^{-1} \tilde{X}'X$.

If we run the regression above, since model B is correct and does not include Z to explain y, we should get

$$\widehat{\delta} = \begin{bmatrix} \widehat{\delta}_1 \\ \widehat{\delta}_2 \end{bmatrix} = \left(\tilde{X}'\tilde{X}\right)^{-1} \tilde{X}'X = \begin{bmatrix} I_K \\ 0 \end{bmatrix}$$

Then we have

$$\begin{aligned} \widehat{\tilde{\beta}} &= \left(\tilde{X}'\tilde{X}\right)^{-1} \tilde{X}'X\beta + \left(\tilde{X}'\tilde{X}\right)^{-1} \tilde{X}'\nu \\ &= \widehat{\delta}\beta + \left(\tilde{X}'\tilde{X}\right)^{-1} \tilde{X}'\nu \\ &= \begin{bmatrix} I_K \\ 0 \end{bmatrix} \beta + \left(\tilde{X}'\tilde{X}\right)^{-1} \tilde{X}'\nu \end{aligned}$$

If we compute the expected value of this estimator,

$$\begin{aligned} E(\widehat{\tilde{\beta}}) &= E\left\{ \begin{bmatrix} \widehat{\beta}_A \\ \widehat{\gamma}_A \end{bmatrix} \right\} \\ &= E\left\{ \begin{bmatrix} I_K \\ 0 \end{bmatrix} \beta + \left(\tilde{X}'\tilde{X}\right)^{-1} \tilde{X}'\nu \right\} \\ &= \begin{bmatrix} \beta \\ 0 \end{bmatrix} + 0 \\ &= \begin{bmatrix} \beta \\ 0 \end{bmatrix} \end{aligned}$$

where we have assumed that $\tilde{X}$ is exogenous and $E(\nu) = 0$. This shows that, if we use a set of irrelevant variables in a linear regression model, the OLS estimator is still unbiased.

For a more formal proof of the unbiasedness of the OLS estimator, consider again model A, $y = X\beta + Z\gamma + \varepsilon$, and its estimated version, $y = X\widehat{\beta}_A + Z\widehat{\gamma}_A + \widehat{\varepsilon}$. From the section on notable matrices in Chapter 3, define $P_Z = Z\left(Z'Z\right)^{-1} Z'$ and $M_Z = I_N - P_Z = I_N - Z\left(Z'Z\right)^{-1} Z'$. Then, pre-multiply both sides of the estimated version of model A by $X'M_Z$:

$$X'M_Z y = X'M_Z X\widehat{\beta}_A + X'M_Z Z\widehat{\gamma}_A + X'M_Z\widehat{\varepsilon} \tag{1}$$

$$= X'M_Z X\widehat{\beta}_A + X'\left[Z - Z\left(Z'Z\right)^{-1} Z'Z\right]\widehat{\gamma}_A + X'\widehat{\varepsilon} - X'Z\left(Z'Z\right)^{-1} Z'\widehat{\varepsilon} \tag{2}$$

Note that $X'\left[Z - Z\left(Z'Z\right)^{-1} Z'Z\right]\widehat{\gamma}_A = 0$, and that $X'\widehat{\varepsilon} = Z'\widehat{\varepsilon} = 0$ by the orthogonality property derived from the first-order conditions. Then we have

$$X'M_Z y = X'M_Z X\widehat{\beta}_A$$

$$
\begin{aligned}
\implies \widehat{\beta}_A &= (X'M_Z X)^{-1} X'M_Z y \\
&= (X'M_Z X)^{-1} X'M_Z (X\beta + \nu), \\
&= (X'M_Z X)^{-1} X'M_Z X\beta + (X'M_Z X)^{-1} X'M_Z \nu \\
&= \beta + (X'M_Z X)^{-1} X'M_Z \nu,
\end{aligned}
$$

since we are assuming model B to be true. The unbiasedness of $\widehat{\beta}_A$ can finally be proved:

$$
\begin{aligned}
E\left(\widehat{\beta}_A\right) &= E\left[\beta + (X'M_Z X)^{-1} X'M_Z \nu\right] \\
&= \beta + (X'M_Z X)^{-1} X'M_Z E(\nu) \\
&= \beta.
\end{aligned}
$$

9.2.2 Inefficiency of the OLS Estimator in the Over-Specified Model

To calculate the variance of $\widehat{\beta}_A$:

$$
\begin{aligned}
\operatorname{Var}\left(\widehat{\beta}_A\right) &= \operatorname{Var}\left[\beta + (X'M_Z X)^{-1} X'M_Z \nu\right] \\
&= \operatorname{Var}\left[(X'M_Z X)^{-1} X'M_Z \nu\right] \\
&= (X'M_Z X)^{-1} X'M_Z \operatorname{Var}(\nu) M_Z' X \left[(X'M_Z X)^{-1}\right]' \\
&= \sigma^2 (X'M_Z X)^{-1} X'M_Z M_Z' X \left[(X'M_Z X)^{-1}\right]' \\
&= \sigma^2 (X'M_Z X)^{-1} X'M_Z X \left[(X'M_Z X)^{-1}\right]^{+} \\
&= \sigma^2 (X'M_Z X)^{-1},
\end{aligned}
$$

since M_Z is idempotent. If we estimated the correct model, B, $\operatorname{Var}\left(\widehat{\beta}_B\right) = \sigma^2 (X'X)^{-1}$. It turns out that $\sigma^2 (X'M_Z X)^{-1} \geq \sigma^2 (X'X)^{-1}$, since the matrix $X'M_Z X - X'X$ is negative semi-definite, or, otherwise $X'M_Z X - X'X \leq 0$, which finally reduces to $-X'Z (Z'Z)^{-1} Z'X \leq 0$. This inequality holds true because $X'Z (Z'Z)^{-1} Z'X$ is positive definite. Also, $\operatorname{Var}\left(\widehat{\beta}_A\right) = \operatorname{Var}\left(\widehat{\beta}_B\right) \iff X'Z = 0$.

The bottom line is that, if we estimate a liner regression model with superfluous regressors, the OLS estimator will be inefficient - i.e., its variance will be higher than it should. It follows that the usual t-tests would provide us with lower t statistics, since the standard errors will be bigger. Thus we would be overaccepting the null.

9.3 Model Selection Criteria

We have a variety of metrics we can use to assess how good a selection of regressors is, typically by comparing two models with the same (or similar) dependent variable.

9.3.1 Coefficient of Determination

The R^2, or coefficient of determination, represents the proportion of the variance of the dependent variable which is explained by the model:

$$R^2 = 1 - \frac{\sum_{i=1}^{N} \widehat{\varepsilon}_i^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2}$$

In principle the model better fits the data, and explains the variance of the dependent variable, as the R^2 increases. However, we also know that whenever we add regressors to the model, regardless of whether they are relevant or not, the R^2 will go up. As such, it only makes sense to compare the coefficient of determination of alternative models when these models are non-nested models, that is, pairs of models where the set of regressors of one does not contain the full set of regressors of the other - i.e., you cannot obtain one of these two models by imposing zero-restrictions on a subset of the regressors of the other.

When we add more regressors, one way to test whether the change in the R^2 we observe is statistically significant is to F-test (or Wald test) the hypothesis that the coefficients associated with the newly-added variables are jointly equal to zero. Of course, if you just add one regressor, a simple t-test is sufficient.

9.3.2 Adjusted R^2

A criterion which can be look at both for nested and non-nested models is the adjusted R^2,

$$\tilde{R}^2 = 1 - \frac{\frac{1}{N-K} \sum_{i=1}^{N} \widehat{\varepsilon}_i^2}{\frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{y})^2},$$

which is a function of the number of regressors, K. As described earlier in this course, the use of the adjusted R^2 alleviates the problem associated with the
R^2, which is weakly monotonic in the number of regressors in nested models, by introducing a penalty for the addition of explanatory variables.

9.3.3 Information Criteria

Two additional ways to select regressors in nested or non-nested models (with the same dependent variable) are the Akaike Information Criterion,

$$AIC = \log\left(\frac{1}{N}\sum_{i=1}^{N} \widehat{\varepsilon}_i^2\right) + \frac{2K}{N}$$

and the Schwarz Criterion (or Bayesian Information Criterion),

$$BIC = \log\left(\frac{1}{N}\sum_{i=1}^{N} \widehat{\varepsilon}_i^2\right) + \frac{K}{N}\log N$$

In model selection, we should minimize one of these two indicators, as both involve the sum of squared residuals, $\sum_{i=1}^{N} \widehat{\varepsilon}_i^2$, and a penalty term for the number of regressors K. Noting that $\frac{K}{N} \log N > \frac{2K}{N}$ if $N \geq 8$, the BIC will usually favor more parsimonious models - i.e., with less regressors. In the end, it is our choice whether we look at the AIC or the BIC for model selection purposes, as long as we make it explicit and we are consistent in an empirical work.

9.3.4 Non-Nested Models

Consider two non-nested models, (A) $y_i = x_i'\beta + \varepsilon_i$ and (B) $y_i = z_i'\gamma + \nu_i$, where $z_i \nsubseteq x_i$ and $x_i \nsubseteq z_i$. We can use $R^2, \bar{R}^2, AIC$, or BIC to select the best model. An alternative approach would be to use an encompassing test, in the form of a non-nested F-test. The intuition is as follows: if model A is correct, it must be able to encompass model B, or, in other words, explain model B 's results. If it is unable to do so, model A should be rejected.

Let us rewrite the set of regressors in x_i as $x_i' = \begin{pmatrix} x_{1i}' & x_{2i}' \end{pmatrix}$, with $x_{1i} \subseteq z_i$ and $x_{2i} \cap z_i = \emptyset$. We can then run the auxiliary regression containing all regressors of models A and B,

$$y_i = x_{2i}'\delta_A + z_i'\gamma + \eta_i$$

Then we run an F-test on the hypotheses

$$\begin{cases} H_0: & \delta_A = 0 \\ H_1: & \delta_A \neq 0 \end{cases}.$$

If we reject the null, $\delta_A \neq 0$, and hence there is evidence that elements in model A are relevant for explaining the variance of y. So a rejection is equivalent to a rejection of model B. This is not enough evidence in favor of model A, though. We also need to run the same test the other way around.

We will now rewrite z_i as $z_i' = \begin{pmatrix} z_{1i}' & z_{2i}' \end{pmatrix}$, where $z_{1i} \subseteq x_i$ and $z_{2i} \cap x_i = \emptyset$. We then run the same auxiliary regression as before, this time written as

$$y_i = x_i'\beta + z_{2i}'\delta_B + \eta_i.$$

The F-test will be

$$\begin{cases} H_0: & \delta_B = 0 \\ H_1: & \delta_B \neq 0 \end{cases}.$$

Again, rejecting the null in this second test is equivalent to rejecting model A, but it is not evidence in favor of model B. It may very well be that we reject in both cases, in which case we should consider a different model specification. If we reject only one or the other, we have come closer to finding a better set of regressors. In some cases, we may be unable to reject neither of the two models.

9.3.5 Box-Cox Transformation

Sometimes we may want to compare a linear model with a log-linear model. [11] Suppose we are trying to decide between $(A) y_i = x_i'\beta + \varepsilon_i$ and $(B) \log y_i = (\log x_i)'\gamma + \nu_i$. In this situation, we cannot use R^2, BIC, or AIC for selection purposes, since these two models have different dependent variables. What we can do, instead, is to use the Box-Cox Transformation.

First estimate $\widehat{y}_i$ and $\log \tilde{y}_i$, which are the fitted values for the two models. Then test the linear model against its log-linear alternative by running the regression

$$y_i = x_i'\beta + \delta\left(\log \widehat{y}_i - \log \tilde{y}_i\right) + u_i$$

and the test

$$\begin{cases} H_0: & \delta = 0 \\ H_1: & \delta \neq 0 \end{cases}.$$

If we reject the null hypothesis, we have some evidence against model A, which only has x_i as a set of regressors. But exactly as in the case of encompassing tests, we need to test the converse transformation. This time, we run

$$\log y_i = (\log x_i)'\gamma + \delta\left(\widehat{y}_i - e^{\log \tilde{y}_i}\right) + u_i$$

and test

$$\begin{cases} H_0: & \delta = 0 \\ H_1: & \delta \neq 0 \end{cases}$$

to find evidence against model B. If we reject, we have evidence against model B. If we reject in both cases, then neither of the two models appear to be appropriate and we should consider a different, more general, model specification.

9.4 Misspecifying the Functional Form

Linearity in the parameters of the relationship between the dependent variable and the set of regressors is a very restrictive assumption. Misspecification may occur when the true relationship is not linear. In such a situation, we would have to use non-linear least-squares (NLS) as opposed to the usual OLS approach. There are two types of non-linearity:
(i) Non-linearity in the regressors. For example, $y_i = \beta_0 + \beta_1 \log x_i + \varepsilon_i$ or $y_i = \beta_0 + \beta_1 x_i^2 + \varepsilon_i$.
(ii) Non-linearity in the parameters. For example, $y_i = g\left(x_i'; \beta\right) + \varepsilon_i$. Suppose we would like to estimate a Cobb-Douglas production function, $y_i = AK_i^{\alpha} L_i^{\beta} + \varepsilon_i$. This expression can be linearized by logging its left-hand and right-hand sides,

[11] 11 The log-linear case involves a model which is still linear in the parameters, but non-linear in the regressors.

but is not linear in its current form. We would need to use non-linear least squares if we wanted to estimate the model as it is.

9.4.1 Non-Linear Least Squares

When we have a linear model, $y_i = x_i'\beta + \varepsilon_i$, we minimize $\sum_{i=1}^{N} \varepsilon_i^2$ with respect to β to derive the OLS estimator, $\widehat{\beta}$. when we have a non-linear model, the concept is exactly the same. If we want to estimate the vector of parameters, β, in the model

$$y_i = g\left(x_i'; \beta\right) + \varepsilon_i,$$

the NLS estimator would be

$$\widehat{\beta} = \arg\min_{\beta} \sum_{i=1}^{N} \varepsilon_i^2 = \arg\min_{\beta} \sum_{i=1}^{N} \left[y_i - g\left(x_i'; \beta\right)\right]^2 = \arg\min_{\beta} S(\beta)$$

This time the problem is that we will not always be able to find a nice closedform expression for $\widehat{\beta}$, since it will depend on the shape of the function g. Most of the times, when we are not able to solve for β analytically, we need to adopt numerical procedures. However, even then, it might happen that a unique solution for the optimization problem cannot be found, or is hard to find. [12] A necessary condition for the consistency of $\widehat{\beta}$, if it exists, is that the objective function $\sum_{i=1}^{N} \left[y_i - g\left(x_i'; \beta\right)\right]^2$ has a unique global minimum.

9.4.2 Testing the Functional Form

The way to test whether a linear model is appropriate or we should consider non-linearities (of any kind) in the model specification is the RESET test, Regression Equation Specification Error Test. The RESET test tells us whether a linear model is appropriate or not, but will not tell us what alternative non-linear

model would be better. After estimating the usual linear model, we run the auxiliary regression

$$y_i = x_i'\beta + \alpha_2 \widehat{y}_i^2 + \alpha_3 \widehat{y}_i^3 + \cdots + \alpha_Q \widehat{y}_i^Q + \omega_i$$

where $Q \geq 2$. We then run the test
$$

$$\begin{cases} H_0: & \alpha_2 = \alpha_3 = \cdots = \alpha_Q = 0 \\ H_1: & H_0 \text{ not true} \end{cases}$$

$$
such that the number of restrictions is $Q-1$. To test the restrictions, we can use

[1412] For example, think of a situation in which the objective function has multiple local minima, or multiple global minima.

an F-test based on an $F_{Q-1;N-K-Q+1}$ distribution or we can run a Wald test based on a χ^2_{Q-1} distribution. If we reject the null, we should look at non-linear models, but if we do not reject the null, that doesn't mean that non-linearities should be ruled out completely.

9.5 Multicollinearity

Multicollinearity is a statistical phenomenon in which two or more regressors in a multiple regression model are highly correlated. With multicollinearity the coefficient estimates may change erratically in response to small changes in the model or the data. Multicollinearity does not reduce the reliability of the model as a whole. That is, a multiple regression model with correlated explanatory variables can still indicate how well the entire set of regressors explains the variance of the dependent variable, but it may not give valid results about any individual regressor, or about which regressors are redundant with others.

Two variables are collinear if there exists an exact linear relationship between them. Multicollinearity refers to a situation in which two or more explanatory variables in a multiple regression model are highly correlated. We have perfect multicollinearity if the correlation between two independent variables is equal to 1 or -1 , that is, if a regressor is a linear combination of other explanatory variables. In practice, we rarely face perfect multicollinearity in a data set. More commonly, multicollinearity arises when there is a high degree of correlation (either positive or negative) between two or more independent variables.

Multicollinearity may be present in a model if (i) large changes in the estimated regression coefficients are observed when a variable is added or deleted, or (ii) if there are statistically insignificant coefficients for the affected variables in a multiple regression, but an F-Test is able to reject the hypothesis that those coefficients are jointly insignificant.

In the presence of multicollinearity, the estimate of one variable's impact on the dependent variable while controlling for the other regressors tends to be less precise than if regressors were uncorrelated with one another. Under multicollinearity the standard errors of the affected coefficients tend to be large. As such, the test of the hypothesis that the coefficient is equal to zero against the alternative that it is not equal to zero leads to a failure to reject the null hypothesis. However, if a simple linear regression of the dependent variable on this explanatory variable is estimated, the coefficient will be found to be significant. Specifically, the analyst will reject the hypothesis that the coefficient is not significant and might falsely conclude that there is no linear relationship between that independent variable and the dependent one.

Multicollinearity does not actually bias results, it just produces large standard errors in the related independent variables. With enough data, these errors will be reduced. However, if other problems could cause bias in the regression estimates, multicollinearity could amplify that bias. Some common remedies to multicollinearity are: (i) Drop one of the variables. An explanatory variable may be dropped to produce a model with significant coefficients. In this way, however, we may lose information. Furthermore, omission of a relevant variable

results in biased coefficient estimates for the remaining explanatory variables. (ii) Obtain more data. This is the preferred solution. More data can produce more precise parameter estimates, that is, with lower standard errors.

Example. In order to understand what happens under perfect multicollinearity, consider the following situation. We have a model with multiple regressors. For simplicity, assume that the model contains two exogenous regressors. One of the two is a linear combination of the other regressor - that is, it can be written as a linear transformation of the other independent variable (from which it follows that the absolute value of the correlation between the two regressors is 1). Formally,

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$$

s.t. $x_{3i} = \delta + \gamma x_{2i}$. Assume that the two parameters δ and γ are known. Our objective is to estimate the model coefficients β_1, β_2, and β_3 using sample information. The consequence of perfect multicollinearity can be described in the following terms:

$$\begin{aligned} y_i &= \beta_1 + \beta_2 x_{2i} + \beta_3 (\delta + \gamma x_{2i}) + \varepsilon_i \\ &= (\beta_1 + \beta_3 \delta) + (\beta_2 + \beta_3 \gamma) x_{2i} + \varepsilon_i \\ &= \omega_1 + \omega_2 x_{2i} + \varepsilon_i \end{aligned}$$

where $\omega_1 = \beta_1 + \beta_3 \delta$ and $\omega_2 = \beta_2 + \beta_3 \gamma$.
By OLS we can obtain $\widehat{\omega}_1$ and $\widehat{\omega}_2$,

$$\widehat{\omega}_2 = \frac{\sum_{i=1}^{N} (x_{2i} - \bar{x}_2)(y_i - \bar{y})}{\sum_{i=1}^{N} (x_{2i} - \bar{x}_2)^2}$$

$$\widehat{\omega}_1 = \bar{y} - \widehat{\omega}_2 \bar{x}_2.$$

To determine β_1, β_2, and β_3, we need to solve the following system of linear equations,

$$\begin{cases} \widehat{\omega}_1 = \beta_1 + \beta_3 \delta \\ \widehat{\omega}_2 = \beta_2 + \beta_3 \gamma \end{cases}.$$

However, the system has two equations and three unknowns. Thus, β_1, β_2, and β_3 cannot be unambiguously identified. The practical implication of this example is that one should always make sure that no regressor is a linear combination of the other regressors before OLS can be safely applied on a linear regression model.

Example. Suppose that we decide to build a model of the profits of tire stores in a given city and we include annual sales of tires (in dollars) at each store and the annual sales tax paid by each store as independent variables. We would expect to estimate a positive relationship between profits and annual sales and a negative relationship between profits and taxes. However, since the tire stores are all in the same city, they all pay the same percentage sales tax. It follows that the sales tax paid will be a constant percentage of their total sales

(in dollars). Thus sales tax will be a perfect linear function of sales, and we will have perfect multicollinearity.

10 Heteroskedasticity and Autocorrelation

Consider again a linear model, $y = X\beta + \varepsilon$, under the Gauss-Markov assumptions, $E(\varepsilon) = E(\varepsilon \mid X) = 0$ and $\text{Var}(\varepsilon) = \text{Var}(\varepsilon \mid X) = \sigma^2 I_N$. The variance of the error term is constant along the sample and the error term is not autocorrelated, as it shows from its diagonal variance-covariance matrix, $\sigma^2 I_N$. Written in another way, $\text{Var}(\varepsilon_i) = \sigma^2, \forall i$ (homoskedasticity assumption), and $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0, \forall i \neq j$ (no-autocorrelation assumption). If the error term satisfies these two conditions, we fall into the case of spherical disturbances.

10.1 Non-Spherical Disturbances

With non-spherical disturbances one or both of the assumptions $\text{Var}(\varepsilon_i) = \sigma^2$, $\forall i$, and $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0, \forall i \neq j$, are violated. That is $\text{Var}(\varepsilon_i) = \sigma_i^2$ (heteroskedasticity) and/or $\text{Cov}(\varepsilon_i, \varepsilon_j) \neq 0$, for some i, j (serial correlation). As such, the second Gauss-Markov assumption above no longer holds and we would have that $\text{Var}(\varepsilon) = \text{Var}(\varepsilon \mid C) \neq \sigma^2 I_N$.

When the error term is heteroskedastic, the diagonal elements of the variancecovariance matrix of ε are not be constant. When the error term is autocorrelated, the variance-covariance matrix of ε is no longer diagonal. In the most general case of non-spherical error terms,

$$\text{Var}(\varepsilon \mid X) = \text{Var}(\varepsilon) = \sigma^2 \Psi$$

where Ψ is some positive-definite matrix, which may also depend on X.
If the assumption of spherical disturbances is violated, do we still have the property of unbiasedness for the OLS estimator, $\widehat{\beta}$? The answer is yes, since, when we proved unbiasedness, we never used the assumptions about the variance of the error term. However, we have another kind of problems that should be taken into account.

Under the assumption of homoskedasticity and no autocorrelation,

$$\begin{aligned}
\text{Var}(\widehat{\beta}) &= \text{Var}\left[(X'X)^{-1} X'y\right] \\
&= \text{Var}\left[\beta + (X'X)^{-1} X'\varepsilon\right] \\
&= (X'X)^{-1} X' \,\text{Var}(\varepsilon) X \,(X'X)^{-1} \\
&= \sigma^2 (X'X)^{-1}
\end{aligned}$$

Under autocorrelation and/or heteroskedasticity,

$$\text{Var}(\widehat{\beta}) = \text{Var}\left[(X'X)^{-1} X'y\right]$$

$$
\begin{aligned}
&= \text{Var}\left[\beta + (X'X)^{-1} X'\varepsilon\right] \\
&= \text{Var}\left[(X'X)^{-1} X'\varepsilon\right] \\
&= (X'X)^{-1} X' \,\text{Var}(\varepsilon) X \,(X'X)^{-1} \\
&= \sigma^2 (X'X)^{-1} X'\Psi X \,(X'X)^{-1}
\end{aligned}
$$

In general, under non-spherical disturbances, the variance of $\widehat{\beta}$ is different, implying that standard t - and F-tests, when they are based (as in many statistical packages) on the usual expression for the variance-covariance matrix of the OLS estimator, will no longer be valid and inference will be wrong. Moreover, since the Gauss-Markov assumptions have been violated, the OLS estimator will no longer be "best" - i.e., efficient.

10.2 Dealing with Heteroskedasticity or Autocorrelation

Several approaches can be adopted to deal with non-spherical disturbances.

- Active strategies consist of a transformation of the original model so that we can derive an alternative estimator which again satisfies the Gauss-Markov assumptions and is BLUE.
- Passive strategies use a consistent estimator of the standard errors. If we adopt such an approach, we generally stick to the OLS estimator and then consistently estimate the standard errors under non-spherical disturbances so that the inference will be reliable again. With heteroskedasticity and/or autocorrelation, the OLS estimator is reliable point-wise, because still unbiased (and consistent).
- Re-specification of the model, such that the new model no longer exhibits non-spherical disturbances.

10.2.1 Theoretical Foundation of the Active Strategy

Consider a model, $y = X\beta + \varepsilon$, with non-spherical disturbances. Instead of computing the usual OLS estimator, which we know would not be BLUE, since some Gauss-Markov assumptions are violated, we would like to modify the model to estimate an alternative BLUE estimator for β.

Theorem. If Ψ is a positive definite matrix, then there exists a matrix $P \in M(N \times N)$, which is square and non-singular, such that $\Psi^{-1} = P'P$.

Note that

$$
\begin{aligned}
\Psi^{-1} = P'P \Longrightarrow \Psi &= (P'P)^{-1} = P^{-1} (P')^{-1} \\
\Longrightarrow P\Psi P' &= PP^{-1} (P')^{-1} P' = I_N
\end{aligned}
$$

Under the Gauss-Markov assumptions, $E(\varepsilon \mid X) = 0$. Likewise, since P is assumed to be non-random, $E(P\varepsilon \mid X) = PE(\varepsilon \mid X) = 0$, from which

$$\operatorname{Var}(P\varepsilon \mid X) = P \operatorname{Var}(\varepsilon \mid X) P' = \sigma^2 P \Psi P' = \sigma^2 I_N$$

While ε does not satisfy the Gauss-Markov assumptions, $P\varepsilon$ does. Consider the transformed model $Py = PX\beta + P\varepsilon$, or $y^* = X^*\beta + \varepsilon^*$, where $y^* = Py$, $X^* = PX$, and $\varepsilon^* = P\varepsilon$. Under this model, the estimator $\widehat{\beta}^*$ is BLUE and has the exact same interpretation and meaning as in the original untransformed model, as we never modified the β term:

$$\begin{aligned}
\widehat{\beta}^* &= \left(X^{*\prime} X^*\right)^{-1} X^{*\prime} y^* \\
&= \left(X' P' P X\right)^{-1} X' P' P y \\
&= \left(X' \Psi^{-1} X\right)^{-1} X' \Psi^{-1} y.
\end{aligned}$$

$\widehat{\beta}^*$ is known as the Generalized Least Squares (GLS) estimator. Of course, if $\Psi = I_N$, then $\widehat{\beta}^* = \widehat{\beta}$. However, since Ψ is generally unknown, this estimator is not feasible. We will then need to replace Ψ with an estimator $\widehat{\Psi}$, where $\widehat{\Psi} \xrightarrow{p} \Psi$. Then we would have $\widehat{\beta}^* = \left(X' \widehat{\Psi}^{-1} X\right)^{-1} X' \widehat{\Psi}^{-1} y$, known as the Feasible Generalized Least Squares (FGLS) estimator.

It follows that

$$\begin{aligned}
\operatorname{Var}\left(\widehat{\beta}^*\right) &= \sigma^2 \left(X^{*\prime} X^*\right)^{-1} \\
&= \sigma^2 \left(X' P' P X\right)^{-1} \\
&= \sigma^2 \left(X' \Psi^{-1} X\right)^{-1} \\
\implies \widehat{\operatorname{Var}\left(\widehat{\beta}^*\right)} &= \widehat{\sigma}^2 \left(X' \widehat{\Psi}^{-1} X\right)^{-1},
\end{aligned}$$

with

$$\begin{aligned}
\widehat{\sigma}^2 &= \frac{\sum_{i=1}^N \widehat{\varepsilon}_i^{*2}}{N - K} \\
&= \frac{\left(y^* - X^* \widehat{\beta}^*\right)' \left(y^* - X^* \widehat{\beta}^*\right)}{N - K} \\
&= \frac{\left(y - X \widehat{\beta}^*\right)' \widehat{\Psi}^{-1} \left(y - X \widehat{\beta}^*\right)}{N - K}
\end{aligned}$$

$\operatorname{Var}\left(\widehat{\beta}^*\right) \leq \operatorname{Var}(\widehat{\beta})$, where $\widehat{\beta}$ is the OLS estimator of β in the untransformed model, $y = X\beta + \varepsilon$, with non-spherical disturbances. In fact,

$$\sigma^2 \left(X' \Psi^{-1} X\right)^{-1} \leq \sigma^2 \left(X' X\right)^{-1} X' \Psi X \left(X' X\right)^{-1}$$

by the Gauss-Markov theorem.

10.3 Heteroskedasticity

Let us just relax the requirement of homoskedasticity and keep the assumption of no autocorrelation. This is the case when $\text{Var}(\varepsilon \mid X) \neq \sigma^2 I_N$, but is still diagonal. In other words, the error terms are mutually uncorrelated, but the variance of ε_i varies over the observations in the sample. Assume that

$$\text{Var}\left(\varepsilon_i \mid X\right) = \text{Var}\left(\varepsilon_i \mid x_i\right) = \sigma^2 h_i^2$$

Usually, h_i, the heteroskedasticity function, is unknown, needs to be estimated somehow, and may depend on X. For now we will assume it is known and non-constant (otherwise, we would fall into the homoskedasticity case again). In matrix form,

$$\text{Var}(\varepsilon \mid X) = \sigma^2 \,\text{diag}\left(h_i^2\right) = \sigma^2 \Psi$$

where Ψ is a diagonal matrix, such that

$$\sigma^2 \Psi = \sigma^2 \begin{bmatrix} h_1^2 & 0 & \cdots & 0 \\ 0 & h_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & h_N^2 \end{bmatrix}$$

10.3.1 Active Strategy (1)

If the linear regression model exhibits heteroskedastic error terms, the OLS estimator is unbiased but inefficient. If we want to adopt an active strategy and solve the problem, we need to find a matrix P such that $\Psi^{-1} = P'P$. One can show that, if we have heteroskedasticity and no autocorrelation, the matrix P we are looking for is equal to $\text{diag}\left(h_{i_1}^{-1}\right)$, a diagonal matrix with values along the diagonal of the form $\frac{1}{h_1}, \frac{1}{h_2}, \ldots, \frac{1}{h_N}$. If we pre-multiply both sides of the linear model by P,

$$Py = PX\beta + P\varepsilon,$$

that is,

$$y^* = X^*\beta + \varepsilon^*$$

where

$$Py = y^* = \begin{bmatrix} \frac{y_1}{h_1} \\ \frac{y_2}{h_2} \\ \vdots \\ \frac{y_N}{h_N} \end{bmatrix} \Longrightarrow y_i^* = \frac{y_i}{h_i}, \forall i.$$

In this case, the GLS estimator for β in the model $y_i = x_i'\beta + \varepsilon_i$ is obtained by running OLS on $y_i^* = x_i^{*\prime}\beta + \varepsilon_i^*$, or otherwise the model

$$\frac{y_i}{h_i} = \left(\frac{x_i}{h_i}\right)' \beta + \frac{\varepsilon_i}{h_i}.$$

Using this model, we have that

$$\operatorname{Var}(\varepsilon_i^*) = \operatorname{Var}\left(\frac{\varepsilon_i}{h_i}\right) = \frac{\sigma^2 h_i^2}{h_i^2} = \sigma^2,$$

implying that the error term, ε_i^*, is homoskedastic and the OLS estimator BLUE:

$$\widehat{\beta}^* = \left(X'\Psi^{-1}X\right)^{-1} X'\Psi^{-1}y = \left(\sum_{i=1}^{N} h_i^{-2} x_i x_i'\right)^{-1} \sum_{i=1}^{N} h_i^{-2} x_i y_i$$

The GLS estimator for this model is a Weighted Least Squares (WLS) estimator, a special case of $\widehat{\beta}^* = \left(X'\Psi^{-1}X\right)^{-1} X'\Psi^{-1}y$, where Ψ has a specific form. The use of weights, h_i, in the formula implies that observations with a higher variance get a smaller weight in estimation, or, in other words, that the greatest weights are given to the observations that provide the most accurate information about the model parameters, and the smallest weights to those that provide relatively little information about β.

The variance of the OLS estimator applied on the transformed, homoskedastic model is

$$\begin{aligned}
\operatorname{Var}\left(\widehat{\beta}^*\right) &= \operatorname{Var}\left[\left(X'\Psi^{-1}X\right)^{-1} X'\Psi^{-1}y\right] \\
&= \operatorname{Var}\left[\left(X'\Psi^{-1}X\right)^{-1} X'\Psi^{-1}X\beta + \left(X'\Psi^{-1}X\right)^{-1} X'\Psi^{-1}\varepsilon\right] \\
&= \operatorname{Var}\left[\left(X'\Psi^{-1}X\right)^{-1} X'\Psi^{-1}\varepsilon\right] \\
&= \left(X'\Psi^{-1}X\right)^{-1} X'\Psi^{-1} \operatorname{Var}(\varepsilon)\Psi^{-1}X \left(X'\Psi^{-1}X\right)^{-1} \\
&= \sigma^2 \left(X'\Psi^{-1}X\right)^{-1} X'\Psi'\Psi^{-1}X \left(X'\Psi^{-1}X\right)^{-1} \\
&= \sigma^2 \left(X'\Psi^{-1}X\right)^{-1} X'\Psi^{-1}\bar{X} \left(X'\Psi^{-1}X\right)^{-1} \\
&= \sigma^2 \left(X'\Psi^{-1}X\right)^{-1} \\
&= \sigma^2 \left(\sum_{i=1}^{N} h_i^{-2} x_i x_i'\right)^{-1}.
\end{aligned}$$

However, since the form of heteroskedasticity is still unknown, in practice we cannot compute this estimator yet. We would eventually like to find an estimator of $h_i, \widehat{h}_i \xrightarrow{p} h_i$, to use in the GLS estimator formula and get the statistic

$$\widehat{\beta}^* = \left(\sum_{i=1}^{N} \widehat{h}_i^{-2} x_i x_i'\right)^{-1} \sum_{i=1}^{N} \widehat{h}_i^{-2} x_i y_i$$

which is known as the Feasible Generalized Least Squares (FGLS) estimator. The estimated variance of the FGLS estimator is

$$\widehat{\text{Var}\left(\widehat{\beta}^*\right)} = \widehat{\sigma}^2 \left(\sum_{i=1}^{N} \widehat{h}_i^{-2} x_i x_i'\right)^{-1}$$

where $\widehat{\sigma}^2 = \frac{1}{N-K}\sum_{i=1}^{N} h_i^{-2}\left(y_i - x_i'\widehat{\beta}\right)^2$. At this point we can use this variance-covariance matrix to run t-tests for simple linear restrictions, or run F-tests for multiple restrictions on the β coefficients. Statistical tests can be run in small samples if the error term is normal, so that $\widehat{\beta}^* \sim \mathcal{N}\left(\beta, \text{Var}\left(\widehat{\beta}^*\right)\right)$; or in large samples, when the central limit theorem implies that $\left(\widehat{\beta}^* - \beta\right) \xrightarrow{d} \mathcal{N}\left(0, \text{Var}\left(\widehat{\beta}^*\right)\right)$.

10.3.2 Passive Strategy

Suppose we have problems of heteroskedasticity, but we do not want to respecify the model and/or we are not able to use the active solution because we cannot estimate h_i. The second-best is a passive strategy. If we want to estimate $y = X\beta + \varepsilon$, or $y_i = x_i'\beta + \varepsilon_i$, we can use an OLS estimator, which we know is still unbiased and consistent under heteroskedasticity. However, it will not have the minimum variance. Under heteroskedasticity, the variance of the OLS estimator is

$$\begin{aligned}
\text{Var}(\widehat{\beta}) &= \text{Var}\left[\left(\sum_{i=1}^{N} x_i x_i'\right)^{-1} \sum_{i=1}^{N} x_i y_i\right] \\
&= \text{Var}\left[\left(\sum_{i=1}^{N} x_i x_i'\right)^{-1} \sum_{i=1}^{N} x_i x_i'\beta + \left(\sum_{i=1}^{N} x_i x_i'\right)^{-1} \sum_{i=1}^{N} x_i \varepsilon_i\right] \\
&= \left(\sum_{i=1}^{N} x_i x_i'\right)^{-1} \left(\sum_{i=1}^{N} x_i x_i' \sigma_i^2\right) \left(\sum_{i=1}^{N} x_i x_i'\right)^{-1}.
\end{aligned}$$

To consistently estimate this variance, we need to consistently estimate σ_i^2. In a famous 1980 Econometrica paper, White proved that

$$\frac{1}{N}\sum_{i=1}^{N} \widehat{\varepsilon}_i^2 x_i x_i' \xrightarrow{p} \frac{1}{N}\sum_{i=1}^{N} \sigma_i^2 x_i x_i'$$

So, we simply need to run the original untransformed regression, take the error terms, square them, and use the estimator above for $\widehat{\text{Var}(\widehat{\beta})}$. Then,

$$\widehat{\text{Var}(\widehat{\beta})} = \left(\sum_{i=1}^{N} x_i x_i'\right)^{-1} \left(\sum_{i=1}^{N} \widehat{\varepsilon}_i^2 x_i x_i'\right) \left(\sum_{i=1}^{N} x_i x_i'\right)^{-1}$$
$$\xrightarrow{p} \left(\sum_{i=1}^{N} x_i x_i'\right)^{-1} \left(\sum_{i=1}^{N} \sigma_i^2 x_i x_i'\right) \left(\sum_{i=1}^{N} x_i x_i'\right)^{-1} = \text{Var}(\widehat{\beta}).$$

When the sample is not large enough, using this formula could be dangerous. The standard errors constructed in this way are known as White standard errors, or heteroskedasticity-robust standard errors. If we use robust standard errors, we can safely make inference on the parameters of the linear regression model without estimating the heteroskedasticity function. However, one should keep in mind that robust standard errors are often larger than usual standard errors. If we have some idea of the form of heteroskedasticity affecting the error term of the model, a FGLS approach may provide a more efficient estimator.

10.3.3 Active Strategy (2) - Multiplicative Heteroskedasticity

If we know the shape of the variance-covariance matrix Ψ, the GLS estimator will work well. However, we frequently do not know the exact shape of this matrix. A common approach is to use an FGLS estimator assuming multiplicative heteroskedasticity, that is

$$\begin{aligned}\text{Var}\left(\varepsilon_i \mid x_i\right) = \sigma_i^2 &= \sigma^2 \exp\left(\alpha_1 z_{1i} + \alpha_2 z_{2i} + \ldots + \alpha_J z_{Ji}\right) \\ &= \sigma^2 \exp\left(z_i'\alpha\right) = \sigma^2 h_i^2\end{aligned}$$

where z_i is a vector of observed variables, possibly a function of the set of regressors x_i (in such a case, the error variance would depend on the set of regressors). The active strategy for FGLS estimation would then be:
(i) Estimate the model $y_i = x_i'\beta + \varepsilon_i$ with OLS, to get $\widehat{\beta} \xrightarrow{p} \beta$.
(ii) Compute $\log \widehat{\varepsilon}_i^2 = \log\left(y_i - x_i'\widehat{\beta}\right)^2$.
(iii) Note that $\sigma^2 \exp\left(z_i'\alpha\right) = \sigma^2 h_i^2 \Longrightarrow \log \sigma_i^2 = z_i'\alpha + \log \sigma^2$.
(iv) Estimate $\log h_i^2$ by running the regression $\log \widehat{\varepsilon}_i^2 = c + z_i'\alpha + \nu_i$, for which we have that $\widehat{\alpha} \xrightarrow{p} \alpha$.
(v) Compute $\widehat{h}_i^2 = \exp\left(z_i'\widehat{\alpha}\right)$.
(vi) Transform the original model into $\frac{y_i}{\widehat{h}_i} = \left(\frac{x_i}{\widehat{h}_i}\right)'\beta + \left(\frac{\varepsilon_i}{\widehat{h}_i}\right) \Longrightarrow y_i^* = x_i^{*\prime}\beta + \varepsilon_i{}^*$, and run OLS on this model to find the FGLS estimator $\widehat{\beta}^* \xrightarrow{p} \beta$.
(vii) Compute $\widehat{\sigma}^2 = \frac{1}{N-K}\sum_{i=1}^{N} \widehat{\varepsilon}_i^{*2}$.
(viii) Compute $\widehat{\text{Var}\left(\widehat{\beta}^*\right)} = \widehat{\sigma}^2 \left(\frac{\sum_{i=1}^{N} x_i x_i'}{\widehat{h}_i^2}\right)^{-1}$.

10.4 Testing for Heteroskedasticity

We will analyze two formal tests for determining if heteroskedasticity could be a problem.

10.4.1 Breusch-Pagan Test

We will work under the assumption that, if heteroskedasticity is present, $\sigma_i^2 = \sigma^2 h\left(z_i'\alpha\right)$. The function h determines how the variables in z_i affect the variance of the error term and is assumed to be unknown but continuously differentiable, such that $h(\cdot) > 0$, with $h(0) = 1$.

The Breusch-Pagan test is based on the following hypotheses:

$$\left\{\begin{array}{ll} H_0: & \text{no heteroskedasticity} \\ H_1: & \text{heteroskedasticity} \end{array}\right. \Longrightarrow \left\{\begin{array}{ll} H_0: & \alpha = 0 \\ H_1: & \alpha \neq 0 \end{array}\right.$$

If $h(t) = \exp(t)$, then $\sigma_i^2 = \sigma^2 \exp\left(\alpha_1 z_{1i} + \cdots + \alpha_J z_{Ji}\right)$. The Breusch-Pagan test should be run using the procedure that follows. Starting from the usual model, $y_i = x_i'\beta + \varepsilon_i$, estimate it by OLS and take the squared residuals, $\widehat{\varepsilon}_i^2$. Then run the auxiliary regression

$$\widehat{\varepsilon}_i^2 = \gamma + \alpha_1 z_{1i} + \alpha_2 z_{2i} + \ldots + \alpha_J z_{Ji} + \nu_i$$

under the assumption that h is a linear function. Then test the joint hypothesis that $\alpha_i = 0$ for all i. If we do not reject the null, then it follows that $\widehat{\varepsilon}_i{}^2$ will be assumed to be constant, implying constant variance of the error term. In such a case, we would conclude that there is no heteroskedasticity. To test the relevant joint hypothesis, we can use an F-test, a Wald test, or a Lagrange Multiplier test, based on the test statistic $BP = NR^2 \sim \chi_J^2$. The application of the Lagrange Multiplier test does not require the model to be estimated under the alternative. If BP is high (in a statistical sense - i.e., if compared to an appropriate critical value), we will reject the null and conclude that the variance is non-constant.

The Breusch-Pagan test requires a specific assumption on the nature of heteroskedasticity - i.e., we need to assume a specific form of the heteroskedasticity function, h.

10.4.2 White Test

An alternative to the Breusch-Pagan test is the White test. This test does not need assumptions on the form of h and further exploits the idea of a heteroskedasticity-consistent variance-covariance matrix for the OLS estimator derived by White in his 1980 Econometrica paper. To run the test, the first step is again to estimate the model $y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_K x_{Ki} + \varepsilon_i$. Then compute the squared residuals, $\widehat{\varepsilon}_i^2$, and run the auxiliary regression

$$\widehat{\varepsilon}_i^2 = \alpha_0 + \alpha_1 x_{1i} + \cdots + \alpha_K x_{Ki} + \alpha_{K+1} x_{1i}^2 + \cdots + \alpha_{K+K} x_{Ki}^2 + \sum_{l=1}^{K-1} \sum_{m=l+1}^{K} \alpha_{K+K+l} x_{li} x_{mi} + \nu_i$$

In practice, the auxiliary regression includes all regressors of the original model, all squared regressors, and all cross products between regressors. The White test proceeds exactly in the same way as the Breusch-Pagan test. That is, we test the null that all α 's, except α_0, are zero. If at least one of these coefficients is not equal to zero, we have heteroskedasticity. This hypothesis can be tested using an F-test, a Wald test, or a Lagrange-Multiplier test based on the test statistic $W = NR^2 \sim \chi_P^2$, where P is the number of regressors in the test regression, excluding the intercept.

11 Instrumental Variables (IV) Estimation

Under the Gauss-Markov assumptions, the OLS estimator is BLUE. To prove unbiasedness, we require that $\{x_i\}_{i=1}^N$ and $\{\varepsilon_i\}_{i=1}^N$ be independent of each other, or that $E(\varepsilon \mid X) = 0$. If, rather than assuming orthogonality, we simply assume that the error term and the regressors are contemporaneously independent, that is $E(\varepsilon_i x_i) = 0 \Longrightarrow \text{Cov}(\varepsilon_i, x_i) = 0$, we cannot prove unbiasedness, but we can still prove consistency. If we assume orthogonality, contemporaneous independence follows, but the opposite is not true. Sometimes, for economic reasons, we need to relax even the assumption of contemporaneous independence. If $E(\varepsilon_i x_i) \neq 0$, or otherwise $\text{Cov}(x_i, \varepsilon_i) \neq 0$, then the OLS estimator is biased and inconsistent, and we need to use alternative estimators for the linear model. Within this framework, we necessarily need to treat the regressors as stochastic variables.

We may have three situations in which $\text{Cov}(x_i, \varepsilon_i) \neq 0$. In such cases, we say that the regressors that covary with the error term are endogenous:
i Omitted variables;
ii Measurement error in one or more regressors;
iii Simultaneity and/or reverse causality of one or more regressors.
In all of these cases, OLS cannot be used as it will produce biased and inconsistent estimators. The solution is to use instrumental variables.

11.1 Example of Endogeneity and IV Estimation

Suppose we want to estimate the demand function of a given good or service in a given market. We have a sample of data for the quantities and for the prices (P_i). From economic theory, demand is defined as

$$D_i = \gamma_0 + \gamma_1 P_i + u_i$$

where $\gamma_1 < 0$, and supply is defined as

$$S_i = \alpha_0 + \alpha_1 P_i + \alpha_2 W_i + \varepsilon_i$$

where W_i is some exogenous variable which affects supply and $\alpha_1 > 0$. Our first instinct would be to use the dataset and OLS to estimate the demand function by regressing quantities on prices. However, we cannot use OLS in this context, because one of the regressors is endogenous to the system - i.e., it is not uncorrelated with the error term. So, we have a problem of endogeneity that should be addressed using proper instrumental variables.

To see why we have endogeneity problems, consider the equilibrium in this market:

$$\begin{aligned} D_i = S_i &\implies \gamma_0 + \gamma_1 P_i + u_i = \alpha_0 + \alpha_1 P_i + \alpha_2 W_i + \varepsilon_i \\ &\implies P_i = \frac{\alpha_0 - \gamma_0}{\gamma_1 - \alpha_1} + \frac{\alpha_2}{\gamma_1 - \alpha_1} W_i + \frac{\varepsilon_i - u_i}{\gamma_1 - \alpha_1} \end{aligned}$$

That is, at equilibrium, P_i is determined by u_i. According to the equilibrium expression for the price level, we have that $\operatorname{Cov}(P_i, u_i) \neq 0$. As such, if we try to estimate the demand function by OLS, we will obtain inconsistent and biased estimators for γ_0 and γ_1, since one of the Gauss-Markov assumptions (that the set of regressors and the error term should be independent of each other) is violated. P_i is assumed to be a random variable.

Suppose we use OLS to estimate γ_1. Then,

$$\begin{aligned} \widehat{\gamma}_1 &= \frac{\widehat{\operatorname{Cov}(P_i, D_i)}}{\widehat{\operatorname{Var}(P_i)}} \\ &\xrightarrow{p} \frac{\operatorname{Cov}(P_i, D_i)}{\operatorname{Var}(P_i)} \\ &= \frac{\operatorname{Cov}(P_i, \gamma_0 + \gamma_1 P_i + u_i)}{\operatorname{Var}(P_i)} \\ &= \gamma_1 \frac{\operatorname{Cov}(P_i, P_i)}{\operatorname{Var}(P_i)} + \frac{\operatorname{Cov}(P_i, u_i)}{\operatorname{Var}(P_i)} \\ &= \gamma_1 + \frac{\operatorname{Cov}(P_i, u_i)}{\operatorname{Var}(P_i)} \\ &\neq \gamma_1 \end{aligned}$$

since $\frac{\operatorname{Cov}(P_i, u_i)}{\operatorname{Var}(P_i)} \neq 0$. So, we have an inconsistent estimator.
If we follow the same steps for the OLS estimator of γ_0,

$$\begin{aligned} \widehat{\gamma}_0 &= \bar{D} - \widehat{\gamma}_1 \bar{P} \\ &= \gamma_0 + \gamma_1 \bar{P} - \widehat{\gamma}_1 \bar{P} + \bar{u} \\ &= \gamma_0 + (\gamma_1 - \widehat{\gamma}_1) \bar{P} \\ &\xrightarrow{p} \gamma_0 + \left[\cancel{\gamma_1} - \cancel{\gamma_1} - \frac{\operatorname{Cov}(P_i, u_i)}{\operatorname{Var}(P_i)} \right] E(P_i) \\ &\neq \gamma_0 \end{aligned}$$

since, again, $\frac{\text{Cov}(P_i,u_i)}{\text{Var}(P_i)} \neq 0$.

A solution to this estimation problem is the Two-Stage Least Squares, or 2SLS, estimator. This estimator is just one of the many instrumental variables estimators that could potentially be used. It is called in this way because it requires two stages to estimate a given equation. First of all, we need to identify the regressors which suffer from endogeneity. In this case, it is only P_i. Then we should regress them on the exogenous variables in the system. We should require these exogenous variables to be correlated with the endogenous variable. In the example, P_i is endogenous, W_i is exogenous, and, as it shows from the equilibrium expression, W_i and P_i are correlated. An instrument is a variable which satisfies the following two properties: (i) it is exogenous to the system (exogeneity), and (ii) it is correlated with the endogenous variable (relevance). If the instrument is exogenous and relevant, we say that the instrument is valid.

Going back to the example, in the first stage of the procedure, we estimate the model

$$P_i = \beta_0 + \beta_1 W_i + v_i$$

to get $\widehat{\beta}_0, \widehat{\beta}_1$, and $\widehat{P}_i$. In the second stage, we consider the original model we would like to estimate and replace P_i with the fitted values from the first stage, $\widehat{P}_i$. That is, we estimate the model

$$D_i = \gamma_0 + \gamma_1 + \widehat{P}_i + u_i$$

One can show that $\widehat{\gamma}_0$ and $\widehat{\gamma}_1$ are now consistent. In fact,

$$\begin{aligned}\text{Cov}\left(D_i, W_i\right) &= \text{Cov}\left(W_i, \gamma_0 + \gamma_1 P_i + u_i\right) \\ &= \gamma_1 \,\text{Cov}\left(W_i, P_i\right) + \text{Cov}\left(W_i, u_i\right)\end{aligned}$$

Since W_i is a valid instrument, it must be exogenous and, hence, should not covary with error term in the system. As such,

$$\text{Cov}\left(D_i, W_i\right) = \gamma_1 \,\text{Cov}\left(W_i, P_i\right) \Longrightarrow \gamma_1 = \frac{\text{Cov}\left(D_i, W_i\right)}{\text{Cov}\left(W_i, P_i\right)}$$

Using some algebra,

$$\begin{aligned}\widehat{\gamma}_1 &= \frac{\text{Cov}\left(\widehat{D_i, W_i}\right)}{\text{Cov}\left(\widehat{W_i}, P_i\right)} \\ &= \frac{\text{Cov}\left(\widehat{D_i}, \widehat{P_i}\right)}{\widehat{\text{Var}\left(\widehat{P_i}\right)}}\end{aligned}$$

which represents the OLS estimator of the coefficients in the regression at the second stage. To prove consistency of $\widehat{\gamma}_1$:

$$\widehat{\gamma}_1 \xrightarrow{p} \frac{\text{Cov}\left(W_i, D_i\right)}{\text{Cov}\left(W_i, P_i\right)}$$

$$
\begin{aligned}
&= \frac{\text{Cov}\left(W_i, \gamma_0 + \gamma_1 P_i + u_i\right)}{\text{Cov}\left(W_i, P_i\right)} \\
&= \gamma_1 \frac{\text{Cov}\left(W_i, P_i\right)}{\text{Cov}\left(W_i, P_i\right)} + \frac{\text{Cov}\left(W_i, u_i\right)}{\text{Cov}\left(W_i, P_i\right)} \\
&= \gamma_1 + \frac{\text{Cov}\left(W_i, u_i\right)}{\text{Cov}\left(W_i, P_i\right)} \\
&= \gamma_1
\end{aligned}
$$

We can prove consistency if and only if $\text{Cov}\left(W_i, u_i\right) = 0$ - i.e., W_i is exogenous - and $\text{Cov}\left(W_i, P_i\right) \neq 0$ - i.e., W_i is relevant. So, we can conclude that $\widehat{\gamma}_1 \xrightarrow{p} \gamma_1$ - that is, $\widehat{\gamma}_1$ is consistent - if and only if W_i is a valid instrument.

11.2 Two-Stage Least Squares Estimation

We use instrumental variables when the assumptions of orthogonality (i.e., statistical independence) and/or contemporaneous correlation between the regressors and the error term are violated. We first need to identify which variable(s) are endogenous in the equation we want to estimate. Once we have identified these variables, we need to find valid instruments, that is, variables which are exogenous and relevant at the same time. We can then use 2SLS techniques to perform estimation. In the first stage of 2SLS estimation, we regress each endogenous variable on the set of valid instruments and on the set of exogenous regressors in the original equation. Note that the constant term should be included in the first-stage regressions to avoid biased estimates. The number of instruments should be at least as large as the number of endogenous variables for the estimation to be feasible. At the second stage, we estimate the original equation after replacing the endogenous variables with their fitted values from the first-stage estimated equations. Consistency of the estimator is thus restored.

11.2.1 Properties of the Two-Stage Least Squares Estimator

When some of the regressors are endogenous and instrumental variables estimation is implemented, simple expressions for the moments of the estimator cannot, generally, be so obtained. Generally, instrumental variables estimators only have desirable asymptotic, not finite sample, properties, and inference is based on asymptotic approximations of the sampling distribution of the estimator. Even when the instruments are uncorrelated with the error in the equation of interest and when the instruments are not weak (i.e., they are strongly correlated with the endogenous regressors), the finite sample properties of the instrumental variables estimator may be poor or not defined at all.

The two-stage least squares estimator is a biased estimator, but it is consistent. As such, one should only apply it when a large sample of data is available. In large samples, it is approximately normally distributed. This means that we can use this estimator and its sampling distribution to make inference on the

parameters of the model in large samples. In small samples, the expression of the variance-covariance matrix of the 2SLS estimator is unknown. However, in large samples we have expressions that we can use as approximations. Finally, if we run the two stages using any statistical software, we will find that the standard errors which are computed at the second stage are wrong for the $2SLS$ estimator. Consistent estimates for the parameter will be obtained, but the standard errors and the associated t statistics cannot be used. Modern statistical packages, however, do have procedures that allow us to compute the correct standard errors and run the two stages in just a click.

11.3 The General Case: Multiple Endogenous Regressors with an Arbitrary Number of Instruments

Suppose we are given the model $y_i = x_i'\beta + \varepsilon_i$. From the first-order conditions of OLS estimation, $\widehat{\beta}$ can be obtained by solving for β the orthogonality condition

$$\begin{aligned}
&\sum_{i=1}^{N} x_i \varepsilon_i = 0 \,(K \text{ equations, if } x_i \text{ is } K \times 1) \\
&\qquad \Longrightarrow \sum_{i=1}^{N} x_i \left(y_i - x_i'\beta\right) = 0 \\
&\qquad \Longrightarrow \sum_{i=1}^{N} x_i \left(y_i - x_i'\widehat{\beta}\right) = 0 \\
&\qquad \Longrightarrow \widehat{\beta} = \left(\sum_{i=1}^{N} x_i x_i'\right)^{-1} \sum_{i=1}^{N} x_i y_i
\end{aligned}$$

$\sum_{i=1}^{N} x_i \widehat{\varepsilon}_i = 0$, or $X'\widehat{\varepsilon} = 0$, is the sample moment condition which states orthogonality between the set of regressors and the residuals. The corresponding population moment condition is

$$E\left(x_i \varepsilon_i\right) = E\left[x_i \left(y_i - x_i'\beta\right)\right] = 0 \text{ (} K \text{ equations, if } x_i \text{ is } K \times 1)$$

IV estimation is based on a similar idea. That the 2SLS estimator should be derived from a sample moment condition which states orthogonality between the residuals and a set of exogenous variables. Some of these exogenous variables may be included in the original linear regression model we are interested in (included instruments or included exogenous variables), some may be excluded (excluded instruments or excluded exogenous variables).

Let z_i be a vector of R exogenous variables, which may overlap with x_i. The moment conditions are

$$E\left(z_i \varepsilon_i\right) = E\left[z_i \left(y_i - x_i'\beta\right)\right] = 0 \,(R \text{ equations, if } z_i \text{ is } R \times 1)$$

If $R = K$, we have R equations and $K = R$ unknowns, and the R sample moment conditions are

$$\frac{1}{N}\sum_{i=1}^{N} z_i\left(y_i - x_i'\widehat{\beta}_{IV}\right) = 0, \text{ or } Z'\widehat{\varepsilon} = 0$$

$$\Longrightarrow \widehat{\beta}_{IV} = \left(\sum_{i=1}^{N} z_i x_i'\right)^{-1} \sum_{i=1}^{N} z_i y_i$$

If the model is written as $y = X\beta + \varepsilon$ and Z is a $N \times R$ matrix of instruments (which may contain the included exogenous regressors in X and an intercept term), then

$$\widehat{\beta}_{IV} = (Z'X)^{-1} Z'y$$

If $R > K$, then there are more instruments than regressors and it is not possible to solve for β using the sample counterpart of the population moment condition. In this case, we have a bigger number of equations than unknowns. Possible solutions to this issue are

- to drop excluded instruments, but we would lose information and efficiency;
- to choose β such that the R sample moments, $\frac{1}{N}\sum_{i=1}^{N} z_i\left(y_i - x_i'\beta\right)$, are as close as possible to zero.

If the model is $y = X\beta + \varepsilon, Z$ is the matrix of instruments, and we want to follow the second proposed solution, we can minimize with respect to β the quadratic form

$$Q_N(\beta) = \left[\frac{1}{N}Z'(y - X\beta)\right]' W_N \left[\frac{1}{N}Z'(y - X\beta)\right]$$

where $W_N \in M(R, R)$ is a positive definite symmetric matrix. W_N is a weighting matrix that tells us how much weight to attach to which linear combinations of the sample moments.

If we solve for β,

$$\widehat{\beta}_{IV} = (X'ZW_NZ'X)^{-1} X'ZW_NZ'y$$

provided that $\text{rank}\,(X'Z) = K$, a necessary condition to invert $X'ZW_NZ'X$. If $R = K, \widehat{\beta}_{IV} = (Z'X)^{-1} Z'y$, since $Z'X$ would be square and invertible, because $\text{rank}\,(X'Z) = K$

In general, if $R = K$, then β is exactly identified. [13] If $R > K$, then β is overidentified, so $\widehat{\beta}_{IV}$ changes depending on the choice of the weighting matrix, W_N. If $R < K$, then β is under-identified and cannot be estimated, since there are more unknowns than equations.

In the case when $R > K$, as long as $W_N \xrightarrow{p} W$, where W is positive definite and symmetric, $\widehat{\beta}_{IV}$ is consistent. But $\widehat{\beta}_{IV}$ will have generally different asymptotic variance-covariance matrices, depending on the choice of W_N. What is the

optimal W_N, then - i.e., the weighting matrix that leads to the most efficient IV estimator? It can be proved that

$$W_N^{OPT} = \left(\frac{1}{N} Z'Z\right)^{-1}$$

if ε is homoskedastic. Then,

$$\widehat{\beta}_{IV} = \left[X'Z\left(Z'Z\right)^{-1}Z'X\right]^{-1} X'Z\left(Z'Z\right)^{-1}Z'y$$

This estimator is known as the Generalized Instrumental Variable Estimator, GIVE, for which we have

$$\widehat{\text{Var}\left(\widehat{\beta}_{IV}\right)} = \widehat{\sigma}^2\left[X'Z\left(Z'Z\right)^{-1}Z'X\right]^{-1}$$

with $\widehat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}\widehat{\varepsilon}_i^2$. Hence, in large samples,

$$\widehat{\beta}_{IV} \overset{a}{\sim} \mathcal{N}\left(\beta, \text{Var}\left(\widehat{\beta}_{IV}\right)\right).$$

11.4 Testing for Valid Instruments

The excluded instruments to be used for estimation should be, first of all, relevant. Running an F-test on the slope coefficients of all excluded exogenous variables in each first-stage regression may provide evidence of the relevance of the instruments. Practitioners usually use the rule of thumb that the F statistic should be bigger than 10 for the instruments to be relevant. Note that this rule does not have solid theoretical foundations. To be even more precise, this approach should be used only when the number of endogenous variables is one. With multiple endogenous variables, one should use an extension of the F-test to be applied on multiple first-stage regressions, based on a multi-dimensional analog of the first-stage F-test.

Exogeneity is much harder to test. Usually, in applied work, one should propose intuitive, logical, theoretical reasons of why a set of instruments is exogenous.

More formally, one could use tests of over-identifying restrictions (for example, the Sargan test). The steps to follow to implement the Sargan test are: (i) take the residuals of the second-stage regression and regress them on an intercept term and the set of included and excluded exogenous variables; (ii) obtain the R^2 of the auxiliary regression and then compute the Sargan test statistic, $S = NR^2$; (iii) under the null that all instruments are exogenous (i.e., the slope coefficients associated with the excluded exogenous variables are jointly equal to zero), $S \sim \chi^2_{R-K}$. Of course, the null hypothesis is rejected for large values of

[14][13] Even in the case of exactly identified models, instrumental variables approaches produce finite sample estimators with no moments, so the estimator can be said to be neither biased nor unbiased, the nominal size of test statistics may be substantially distorted, and the estimates may commonly be far away from the true value of the parameter.

S. Note that the Sargan test can be used only if the number of excluded instruments is bigger than the number of endogenous variables - i.e., if the equation is over-identified. If it is not, the Sargan test statistic will be equal to 0 and the corresponding χ^2 distribution will have 0 degrees of freedom (so, it cannot be used).