Econometric Theory

Giuseppe Ragusa

Class Notes

Econometric Theory

Date: September 25, 2014

Contents

Chapter 1

Tools and Foundations

1.1 Preliminary notation and definitions

1.1.1 Random variables, distributions, expectations

This is a notes on the margin

A *random variable* is a numerical translation of the outcomes of a statistical experiment. In broad terms, it is a mathematical model for something we do not know but which has a range of possible values, possibly some more likely than others. Examples include the sum of the dots on rolled dice, the value of a share of Apple stock 3 months from now, and the name of the Italian prime minister 1 year from now.

A *random vector* in $\mathbb{R}^k$ is a k-tuple $X = (X_1, X_2, \ldots, X_k)$ of real random variables. In what follows without loss of generality, we focus on random vectors, since random variable are random vector of dimension $k = 1$.

Associated with a random vector is a right-continuous *distribution function* defied on $\mathbb{R}^k$ by

$$F_{X_1,\ldots,X_k}(x_1, \ldots, x_k) = \mathbb{P}(X_1 \leqslant x_1, \ldots, X_k \leqslant x_k)$$

for all $(x_1, \ldots, x_k) \in \mathbb{R}^k$. This is also know as the probability law of X. Two random vectors X and Y, defined on possibly different spaces, "have the same law": if their distribution functions are the same, and this is denoted by $F_X = F_Y$.

A random vector is said discrete is all its elements take a countable number values. It is continuous if all its elements take an uncountable numbers of values. A random value is mixed if some elements take countable number of values and some take an uncountable number of values.

By expectation of a random vector X is meant the (Lebesgue-Stieltjes) integral of X with respect to F_X, so that

$$\mathbb{E}(X) := \int x dF_X.$$

The expectation $\mathbb{E}(X)$ is also called the mean of the random vector X. Some important characteristics of random vectors can be expressed conveniently in terms of expectations. For instance, the covariance of two random vectors is

$$\mathrm{Cov}\,[Y,X] = \mathbb{E}[(Y - \mathbb{E}(Y))(X - \mathbb{E}(X))'].$$

The variance a random vector, $\mathrm{Var}\,[X] := \mathrm{Cov}\,[X,X]$, is the $(k \times k)$ matrix with element (i,j)

$$\mathbb{E}[(Y_i - \mathbb{E}(Y_i))(X_j - \mathbb{E}(X_j))'].$$

Higher order moments are the skewness

$$\sigma^3 = \mathbb{E}[(X - \mu_X)^3],$$

and the kurtosis

$$\sigma^4 = \mathbb{E}[(X - \mu_X)^4],$$

For any function g of a random vector we can form the expectation of $g(X)$

$$\mathbb{E}[g(x)] = \int g(x) dF_X.$$

It follows that the probability of a set S can be written as the expectation of the indicator function

$$\mathbb{P}(X \in S) = \int_S dF_X = \int 1_S(x) dF_X = \mathbb{E}(1_S(X)),$$

where

$$I_S(x) = \begin{cases} 1 & x \in S, \\ 0 & x \notin S. \end{cases}$$

For a univariate distribution function F, and for $0 \leqslant t \leqslant 1$, the quantity

$$F^{-1}(p) = \inf\{x : F(x) \geqslant p\}$$

is called the pth quantile or fractile of F. In particular, $F^{-1}(1/2)$ is the median of F. The function $F^{-1}(t)$, $0 \leqslant t \leqslant 1$, is called the inverse function of F.

1.1.2 Normal distribution

The normal distribution with mean μ and variance $\sigma^2 > 0$ corresponds to the distribution function

$$F(x) = \frac{1}{(2\pi)^{1/2}\sigma} \int_{-\infty}^{x} \exp\left[-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right] dt, \quad -\infty < x < \infty.$$

The notation $N(\mu, \sigma^2)$ will be used to denote either this distribution or a ranodm variable having this distribution. The special distribution function $N(0,1)$ is known as the standard normal distribution and it is often denoted by Φ.

A random vector $X = (X_1, \ldots, X_k)$ has the k-variate normal distribution with mean vector $\mu = (\mu_1, \ldots, \mu_k)$ and covariance matrix $\Sigma = (\sigma_{ij})_{k\times k}$ if, for every non-null vector $a = (a_1, \ldots, a_k)$, the random variable aX' is $N(a\mu, a\Sigma a')$, that is, $aX' = \sum_{j=1}^{k} a_j X_j$ has the normal distribution with mean $\sum_{j=1}^{k} a_j \mu_j$ and variance $a\Sigma a' = \sum_{j=1}^{k}\sum_{i=1}^{k} a_j a_k \sigma_{ji}$. The notation $N(\mu, \Sigma)$ will denote either this multivariate distribution or a random vector having this distribution.

1.1.3 Chi-squared distributions

Let Z be k-variate $N(\mu, I)$, where I denotes the identity matrix of order k. For the case $\mu = 0$, the distribution of $ZZ' = \sum_{j=1}^{k} Z_j^2$ is called the chi-squared with k degrees of freedom. For the case $\mu \neq 0$, the distribution is called noncentral chi-squared with k degrees of freedom and noncentrality parameter $\lambda = \mu\mu'$. The notation $\chi_k^2(\lambda)$ encompasses both cases and may denote either the random variableor the distribution. We also denote $\chi_k^2(0)$ simply by χ_k^2

1.1.4 I.I.D. and I.N.I.D.

With reference to a sequence X_i of random vectors, the abbreviation $I.I.D.$ will stand for "independent and identically distributed", while $I.N.I.D.$ will stand for "independent but not identically distributed".

1.1.5 Weak dependence

A sequence $\{X_i\}$ is (strictly) stationary if, for any given finite integer h and for any set of subscript $i_1, i_2, \ldots, i_h$, the joint distribution of $(X_i, X_{i_1}, X_{i_2}, \ldots, X_{i_h})$ depends only on $i_1 - i$, $i_2 - i$, ...,$i_h - i$ but not on i itself. A sequence of random vectors is covariance stationary (also referred to weakly stationary) if $\mathbb{E}(X_i)$ does not depends on i and $\mathrm{Cov}\,[X_i, X_{i-j}]$ depends (possibly) on j, but not on i.

A sequence $\{X_i\}$ is ergodic if for any two bounded functions f, g

$$\lim_{n\to\infty} |\mathbb{E}[f(X_i,\ldots,X_{i+k})g(X_{i+n},\ldots,X_{i+n+h})]| = |\mathbb{E}[f(X_i,\ldots,X_{i+k})]|\,|\mathbb{E}[g(X_{i+n},\ldots,X_{i+n+h})]|\,.$$

That is, a sequence is ergodic if it is asymptotically independent: random variables positioned far apart in the sequence are almost independently distributed. A stationary process that is also ergodic is an ergodic stationary process.

The sequence $\{X_i\}$ is called a martingale difference sequence if the expectation conditional on its past values, too, is zero:

$$\mathbb{E}(X_i|X_{i-1},X_{i-2},\ldots,X_1) = 0.$$

Importantly, a martingale difference sequence has no serial correlation, that is,

$$\mathbb{C}\text{ov}\left(X_{i-j},X_{i-k}\right) = 0, \quad \text{all } j \neq k.$$

1.1.6 Conditional expectations

Let Y and X be random vectors in $\mathbb{R}^\ell$ and $\mathbb{R}^k$, respectively. If $\mathbb{E}\,\|y\| < \infty$, then there is a function, say $\mu(X_1,\ldots,X_k)$ such that

$$\mathbb{E}[Y|X_1,\ldots,X_k] = \mu(X_1,\ldots,X_k),$$

or $\mathbb{E}[Y|X] = \mu(X)$. Technically, we should distinguish $\mathbb{E}[Y|X]$ from $\mathbb{E}[Y|X = x]$, some $x \in \mathbb{R}^k$: the former is a random variable, because X is a random vector; the latter is a real quantity.

The so called Law of Iterated Expectation (L.I.E.) is a very important result.

Proposition 1 ((Law of Iterated Expectation))**.** *Let W be a r.v. in $\mathbb{R}^m$ and $f : \mathbb{R}^m \to \mathbb{R}^k$. Then:*

$$\mathbb{E}[Y|f(W)] = \mathbb{E}[\mathbb{E}(Y|W)|f(W)].$$

The most important special case is when $W = (X,Z)$ and $f(W) = f(X,Z) = X$.

$$\mathbb{E}[y|X] = \mathbb{E}[\mathbb{E}(y|X,Z)|X].$$

1.2 The Conditional Expectation Function

The Conditional Expectation Function (CEF) for a random variable y, given a random vector x, is the expectation of y with x held fixed. The CEF is written $\mathbb{E}[y|x]$ and is a **function** of x. Because x is random, the CEF is random, though sometimes

we work with a particular value of the CEF, say $\mathbb{E}[y|\mathrm{x} = 10]$. For continuous y with conditional densities $f_y(u|\mathrm{x})$ at $y = u$, the CEF is

$$\mathbb{E}[y|\mathrm{x} = x] = \int u f_y(u|\mathrm{x} = x)du.$$

If y is discrete, the CEF equals $\sum_u u\,\mathbb{P}(y = u|\mathrm{x} = x)$, where $\mathbb{P}_y(y = u|\mathrm{x} = x)$ is the conditional probability mass function of y given $\mathrm{x} = x$.

Theorem 1. *The CEF decomposition property Given the random variable y and the random vector x, such that $\mathbb{E}\,\|\mathrm{x}\| < \infty$, we can always write*

$$y = \mathbb{E}[y|\mathrm{x}] + u,$$

where $\mathbb{E}[u|\mathrm{x}] = 0$

Theorem **??** says that any random variable y can be decomposed into a piece that is "explained" by x—that is, the CEF—and a piece left over that is mean independent of x.

1.3 Linear Projection

Let Y and $X_1, X_2, \ldots, X_k$ random variables with $\mathbb{E}(Y)^2 < \infty$ and $\mathbb{E}(X_j)^2 < \infty$, $j = 1, \ldots, k$. The linear projection of Y on a constant and on $X = (1, X_1, \ldots, X_K)$ is

$$\mathbb{E}^*(Y|X) = X\delta$$

where

$$\delta = \mathbb{E}[X'X]^{-1}\,\mathbb{E}[X'Y].$$

The linear projection of Y on X can be equivalently written in the error term form as

$$Y = \delta_0 + \delta_1 X_1 + \ldots + \delta_k X_k + \eta,$$

where the error term η is orthogonal to the vector X:

$$\mathbb{E}[X'\eta] = 0. \tag{1.3.1}$$

This orthogonality can be shown to hold by noting that $Y = X\delta + \eta$ implies

$$X'Y = X'X\delta + X'\eta.$$

Taking expectation of both sides and substituting (**??**) yields

$$\mathbb{E}[X'Y] = \mathbb{E}[X'X]\left(\mathbb{E}[X'X]^{-1}\,\mathbb{E}[X'Y]\right) + \mathbb{E}[X'\eta] = \mathbb{E}[X'Y] + \mathbb{E}[X'\eta],$$

which in turns gives (1.3.1).

The linear projection can also be defined directly from the orthogonality, by choosing the value δ that set

$$\mathbb{E}[X'(Y - X\delta)] = 0.$$

Such value is easily seen being (**??**).

The linear projection can also be defined in the case in which the projecting variable is a random vector. The linear projection of the random vector $Y = (Y_1, Y_2, \ldots, Y_k)$ on $X = 1, X_1, \ldots, X_k$ is

$$\mathbb{E}(Y|X) = \Pi X + \eta,$$

where

$$\Pi = \mathbb{E}[XX'] \, \mathbb{E}[YX'],$$

is a $(k \times k)$ matrix, and $\eta = (\eta_1, \eta_2, \ldots, \eta_k)$ is a $k \times 1$ vector.

Chapter 2

A modicum of Asymptotic Theory

2.1 Introduction

Asymptotic theory is concerned with the study of the limit behavior of sequences of random variables. In all of the following n is an index that tends to infinity, and asymptotic means taking limits as $n \to \infty$. In most situations n is the number of observations, so that usually asymptotic is equivalent to "large-sample theory". However, in many interesting situation in econometrics limit theorems often have nothing to do with observations. In that case n is just an index that goes to infinity.

2.2 Stochastic Convergences

In this section we study two basic modes of convergence: almost-sure convergence, convergence in probability, and convergence in distribution.

The sequence of random vectors in $\mathbb{R}^k$ $X_1, X_2, \ldots, X_n$ is denoted by $\{X_n\}$. Also, we let $F_n(x) := \mathbb{P}(X_n \leqslant x)$.

If h is a vector in $\mathbb{R}^k$, then $\|h\|$ denote the Euclidean distance

$$\|h\| = \left(\sum_{k=1}^{k} h_i^2 \right)^2 .$$

Definition 1. (Almost-sure convergence) We say that the sequence $\{X_n\}_n$ *converges with probability 1* or *strongly* or *almost surely* to X if for all $\epsilon > 0$

$$\mathbb{P}\left(\lim_{n \to \infty} \|X_n - X\| < \epsilon \right) = 1.$$

This is written $X_n \xrightarrow{wp1} X$, $X_n \xrightarrow{as} X$, or $X_n = X$ $w.p.1$ as $n \to \infty$.

When a sequence converge almost surely, the probability of observing a realization of X_n for which converges to X does not occur is zero.

Definition 2. (Convergence in Probability) We say that the sequence $\{X_n\}_n$ *converges in probability* to X if for all $\epsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}\left(\|X_n - X\| < \epsilon\right) = 1$$

This is written $X_n \xrightarrow{p} X$, $plim\, X_n = X$, or $X_n = X$ $w.p.a.1$ as $n \to \infty$.

When a sequence converges in probability, it becomes less and less likely that an element of the sequence lies beyond any specified distance ϵ from X, as n increases.

Converge in probability is also referred to as a weak consistency, and since this has been the most familiar stochastic converge concept in econometrics, the word weak is often simply dropped.

Almost-sure convergence is stronger than convergence in probability: with almost sure convergence, the probability measure P takes into account the joint distribution of the entire sequence X_n, but we convergence in probability, we only need concern ourselves sequentially with the joint probability of elements that appear in X_n.

Theorem 2. *Let $\{X_n\}$ be a sequence of random variables. If $X_n \xrightarrow{as} X$, then $X_n \xrightarrow{p} X$.*

Thus, almost sure convergence implies convergence in probability, but the converse does not in general hold. It holds however for the special case in which $X_n \xrightarrow{d} c$, where c is a constant.

Definition 3. (Convergence in Distribution) A sequence of random vectors X_n *converges in distribution* to a random vector X if

$$\mathbb{P}(X_n \leqslant x) \to \mathbb{P}(X \leqslant x),$$

for every x at which the limit distribution function F_X is continuous.

Convergence in distribution is denoted by $X_n \xrightarrow{d} X$, $X_n \Rightarrow X$, or $X_n \rightsquigarrow X$. If the distribution of X has a standard code, as $N(0,1)$ we usually write $X_n \xrightarrow{d} N(0,1)$ or $X_n \rightsquigarrow N(0,1)$.

The converge in distribution of random vectors in $\mathbb{R}^k$ is equivalent to convergence of linear combinations of elements of the random vector. In particular,

$$X_n \xrightarrow{d} X, \quad \text{if and only if } t'X_n \xrightarrow{d} t'X, \quad \text{for all } t \in \mathbb{R}^k.$$

This is known as the Cramér-Wold device and it allows to reduce higher dimensional problems to the one-dimensional case.

The following result relates convergence in distribution of convergence of certain moments of $\{X_n\}$.

Lamma 1. *(Portmanteau) For any random vectors X_n and X the following statements are equivalent:*

$\mathbb{P}(X_n \leqslant x) \to \mathbb{P}(X \leqslant x)$ *for all continuity points of* $P(X \leqslant x)$*;*

$\mathbb{E}\, f(X_n) \to \mathbb{E}\, f(X)$*, for all bounded continuous functions* f*.*

2.2.1 Relationship between Modes of Convergence

The following result explores the relationship between different modes of convergence.

Theorem 3. *Let $\{X_n\}$, $\{X\}$, $\{Y_n\}$, and $\{Y\}$ be random vectors. Then*

1. *$X_n \xrightarrow{p} X$ implies $X_n \xrightarrow{d} X$;*
2. *$X_n \xrightarrow{p} c$ for a constant c if and only if $X_n \xrightarrow{d} c$;*
3. *if $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} Y$, then $(X_n, Y_n) \xrightarrow{p} (X, Y)$;*
4. *if $X_n - Y_n \xrightarrow{p} 0$ and $Y_n \xrightarrow{d} Y$, then $X_n \xrightarrow{d} Y$.*

2.2.2 Continuous Mapping Theorem

The continuous-mapping theorem is a simple result, but it is extremely useful. If the sequence of random vectors X_n converges to X on g is continuous, then $g(X_n)$ converges to $g(X)$. This is true for each of the three modes of stochastic convergence.

Theorem 4. *(continuous mapping) Let $g : \mathbb{R}^k \mapsto \mathbb{R}^m$ be continuous at every point of a set C such that $\mathbb{P}(X \in C) = 1$.*

1. *If $X_n \xrightarrow{d} X$, then $g(X_n) \xrightarrow{d} g(X)$;*
2. *If $X_n \xrightarrow{p} X$, then $g(X_n) \xrightarrow{p} g(X)$;*
3. *If $X_n \xrightarrow{as} X$, then $g(X_n) \xrightarrow{as} g(X)$.*

Continuous functions of weakly convergent sequences converge to the functions evaluated at the probability limit. Theorem 4 is implicitly or explicitly used in the demonstrations of consistency and asymptotic normality of most econometric estimators.

2.2.3 Slutsky's lemma

Lamma 2. *(Slutsky) Let* $\{X_n\}$, $\{X\}$, *and* $\{Y_n\}$ *be random vectors or variables. If* $X_n \xrightarrow{d} X$ *and* $Y_n \xrightarrow{p} c$ *for a constant* c, *then*

1. $X_n + Y_n \xrightarrow{d} X + c$;
2. $Y_n X_n \xrightarrow{d} cX$;
3. $Y_n^{-1} X_n \xrightarrow{d} c^{-1}X$, *provided* $c \neq 0$;
4. $Y_n^{-1/2} X_n \xrightarrow{d} c^{-1/2}X$, *provided* $c \neq 0$.

The Slutsky's lemma is also valid for matrices Y_n and c, where the provision that $c \neq 0$ is changed to c being invertible.

Example 1. Let $\{X_n\}$ be a $k \times 1$ vector and $X_n \xrightarrow{d} N(0,\Sigma)$, where Σ is the $k \times k$ matrix of variance of X_n. Let $\{\Sigma_n\}$ a sequence of positive definite $k \times k$ matrices such that $\Sigma_n \xrightarrow{p} \Sigma$. Then, by (iii) of Lemma 2 we have that $\Sigma_n^{-1/2} X_n \xrightarrow{d} \Sigma^{-1/2} X \sim N(0,I)$.

2.2.4 Boundedness in Probability

If a sequence is bounded in probability means that there exists a compact set to which *all* X_n give probability almost one.

Definition 4. (Boundedness in Probability) The sequence $\{X_n\}$ is said to be *bounded in probability* (or uniformly tight) is for every $\epsilon > 0$ there exists a constant M such that

$$\sup_n P\left(\|X_n\| > M\right) < \epsilon.$$

This useful theorem gives a simple criterion to check whether a sequence is bounded in probability: a sequence that converges in distribution is bounded in probability.

Theorem 5. *(Prohorov's theorem) If* $X_n \xrightarrow{d} X$, *then* $\{X_n\}$ *is bounded in probability.*

Corollary 1. *If* $X_n \xrightarrow{p} c$ *for some constant* c, *then* $X_n = O_p(1)$.

2.2.5 Stachastic o and O Symbols

It is convenient to have short expressions for terms that converge in probability to zero or converge in distribution. The notation $o_p(1)$ (small oh-P-one) is short for a sequence of random vectors that converges to zero in probability. The expression $O_p(1)$ denotes a sequence that is bounded in probability. More generally, for a given sequence of random variables a_n,

$$\begin{aligned} X_n &= o_p(a_n), && \text{means} \quad X_n = Y_n a_n, \quad \text{and } Y_n \xrightarrow{p} 0 \\ X_n &= O_p(a_n), && \text{means} \quad X_n = Y_n a_n, \quad \text{and } Y_n = O_p(1), \end{aligned}$$

This expresses that the sequence X_n converges in probability to zero or is bounded in probability at the *rate* a_n. There are many rules of calculus with o_p and O_p symbols, which we will apply without comment throughout the lectures. In particular,

$$\begin{aligned} o_p(1) + o_p(1) &= o_p(1) \\ o_p(1) + O_p(1) &= O_p(1) \\ O_p(1)o_p(1) &= o_p(1) \\ (1 + o_p(1))^{-1} &= O_p(1) \\ o_p(a_n) &= a_n o_p(1) \\ o_p(O_p(1)) &= o_p(1) \end{aligned}$$

2.3 Weak Laws of Large Numbers

"Weak Laws of Large Numbers" (WLLN) refer to convergence in probability of averages of random variables. In particular, let $\overline{X}_n$ be the average of the first n of a sequence of random vectors $X_1, X_2, \ldots, .$, that is

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

The first results is a WLLN for iid sequence of random vectors for which $\overline{X}_n$ converges in probability to the expected value of X_i.

Theorem 6. *(Kolmogorov WLLN) Let $\{X_n\}$ be a sequence of iid random vectors. Then, if $E|X_i| < \infty$, then*

$$\overline{X} \xrightarrow{p} \mu_X := \mathbb{E}[X_i]$$

The following result provide a WLLN under relaxation of the iid assumptions, but at the expense of assuming existence of variances and restricting their growth with increasing n.

Theorem 7. *(Chebyshev WLLN) Let* $\{X_1, X_2 \ldots, \}$ *be uncorrelated with means* $\mu_1 := \mathbb{E}[X_1]$, $\mu_2 := \mathbb{E}[X_2], \ldots$ *and variances* $\sigma_1^2 : \mathrm{Var}\,[X_1]$, $\sigma_2^2 := \mathrm{Var}\,[X_2], \ldots$. *If*

$$\lim_{n\to\infty} \frac{1}{n^2} \sum_{1}^{n} \sigma_i^2 = 0,$$

then

$$\overline{X} - \frac{1}{n} \sum_{i=1}^{n} \mu_i \xrightarrow{p} 0.$$

2.4 Central Limit Theorem

The central limit theorem (CLT) pertains to the convergence in distribution of (normalized) sums of random variables. As for the WLLN the simplest form of CLT applies to iid summands.

Theorem 8. *(Lindberg-Lévy CLT) Let* $\{X_n\}$ *be a sequence of iid random variables with mean* μ *and variance* $\sigma^2 < \infty$. *Then*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu) = \sqrt{n}(\bar{X} - \mu) \xrightarrow{d} N(0, \sigma^2).$$

Compared with the WLLN for iid sequences, the Lindberg-Lévy CLT imposes a single additional requirement, i.e., that $\sigma^2 < \infty$. Note that this implies that $E|X_i| < \infty$.

CLT theorems are usually spelled out for random variables. The so-called Cramer-Wold device makes sufficient establishing CLT for random variables.

Theorem 9. *(Cramer-Wold device) Let* $\{X_n\}$ *be a sequence of random vectors in* $\mathbb{R}^k$ *and suppose that for any* $\lambda \in \mathbb{R}^k$ *such that* $\lambda'\lambda = 1$, $\lambda' Z_n \xrightarrow{d} \lambda' X$. *Then the limiting distribution function of* X_n *exists and equal to the distribution of* X.

Corollary 2. *Let* $\{X_n\}$ *a random vector in* $\mathbb{R}^k$ *where each element satisfies the assumption of Theorem 8. Let* $\mu := (\mu_1, \ldots, \mu_k)$ *and* $\Sigma := \mathrm{Var}\,[X_i]$. *Then,*

$$\sqrt{n}\Big(\overline{X} - \mu\Big) \xrightarrow{d} N(0, \Sigma), \tag{2.4.1}$$

Different CLT are available for other sampling situations. As for the WLLN, there exists a trade-off between the dependence or heterogeneity that can be allowed and the moment requirements that need to be imposed to obtain a CLT.

Theorem 10. *(Liapounov CLT) Let $\{X_n\}$ be a sequence of independent random scalars with $\mu_i := \mathbb{E}(X_i)$, $\sigma_i^2 := \mathrm{Var}[X_i]$, and $E|X_i - \mu_i|^{2+\delta} < \Delta < \infty$ for some $\delta > 0$ and all i, If $\bar{\sigma}_n^2 > \delta' > 0$ for all n sufficiently large, then*

$$\sqrt{n}\frac{\left(\overline{X} - \mu\right)}{\bar{\sigma}_n} \xrightarrow{d} N(0,1).$$

We conclude this section with the simplest CLT for time series sequences.

Theorem 11. *(CLT for mds) Let $\{X_n\}$ be a martingale difference sequence that is also stationary and ergodic. Then, for $\mu = \mathbb{E}(X_i)$ and $\sigma^2 = \mathrm{Var}[X_i]$, we have*

$$\sqrt{n}\left(\overline{X} - \mu\right) \xrightarrow{d} N(0,\sigma).$$

2.5 Taylor's theorem and the Delta Method

A central result of calculus is Taylor's theorem concerning the expansion of a sufficiently smooth function about a point.

Theorem 12. *Taylor's Theorem Suppose that $f(x)$ has r derivatives at the point a. Then*

$$f(a+\Delta) = f(a) + \Delta f'(a) + \ldots + \frac{\Delta^r}{r!} f^{(r)}(a) + o(\Delta^r)$$

where the last term can also be written as

$$\frac{\Delta^r}{r!}\left[f^{(r)}(a) + o(1)\right].$$

If, in addition, the $(r+1)$st derivative of f exists in a neighborhood of a, the remainder $o(\Delta^r)$ can be written as

$$\frac{\Delta^{r+1}}{(r+1)!}\left[f^{(r+1)}(\xi)\right],$$

where ξ is a point between a and $a + \Delta$.

An consequence is the following result which greatly extends the usefulness of the central limit theorem.

Theorem 13. *Delta's Rule If*

$$\sqrt{n}(X_n - \mu) \xrightarrow{d} N(0,\sigma^2),$$

then

$$\sqrt{n}(f(X_n) - f(\mu)) \xrightarrow{d} N(0,\sigma^2[f'(\mu)]^2),$$

provided $f'(\mu)$ exists and is not zero.

Proof. By Taylor's theorem, with $a = \mu$ and $\Delta = X_n - \mu$,

$$f(X_n) = f(\mu) + (X_n - \mu)f'(\mu) + o_p(X_n - \mu)$$

and hence

$$\sqrt{n}(f(X_n) - f(\mu)) = \sqrt{n}(X_n - \mu)f'(\mu) + o_p(\sqrt{n}(X_n - \mu)).$$

The first term on the right side tends in distribution to $N(0, \sigma^2[f'(\mu)]^2)$. On the other hand, $\sqrt{n}(X_n - \mu) = O_p(1)$ by the central limit theorem. Thus, $o_p(\sqrt{n}(X_n - \mu)) = o_p(O_p(1)) = o_p(1)$. The result follows from 3-(iv). □

2.6 Inequalities

2.6.1 Markov's Inequality

The Markov's inequality is very useful. Let X be a r.v. taking only non-negative value. Fix a constant M. Then,

$$\mathbb{P}(X > M) \leqslant \frac{\mathbb{E}(X)^p}{M^p}.$$

2.6.2 Chebyshev's inequality

If X has a finite mean μ, variance σ^2, and $M > 0$, then

$$\mathbb{P}(|X - \mu| > M) \leqslant \frac{\sigma^2}{M^2}.$$

2.6.3 Jensen's Inequality

Suppose f is a convex function, that is,

$$\lambda f(x) + (1 - \lambda)f(y) \geqslant f(\lambda x + (1 - \lambda)y)$$

for all $\lambda \in (0, 1)$ and $x, y \in \mathbb{R}$. Then,

$$\mathbb{E}[f(X)] \geqslant f(\mathbb{E}(X)),$$

provided both expectation exists, i.e., $\mathbb{E}\,|X|$ and $\mathbb{E}\,|f(X)| < \infty$.

2.6.4 Hölder's inequality

If $p, q \in [1, +\infty]$ with $1/p + 1/q = 1$ then

$$\mathbb{E}\,|XY| \leqslant \|X\|_p \|Y\|_q,$$

where $\|X\|_r = (\mathbb{E}\,|X|^r)^{1/r}$.

The special case $p = q = 2$ is called the Cauchy-Schwartz inequality:

$$\mathbb{E}\,|XY| \leqslant (\mathbb{E}\,X^2\,\mathbb{E}\,Y^2)^{1/2}.$$

Exercise 2.1. *The Cauchy-Schwartz inequality is useful to derive easy to understand conditions on existence of certain moments. For instance, let X, Y, Z be random variables and suppose one wants to know what are the minimal requirement for the existence of $\mathbb{E}[XYZ]$. We have that $\mathbb{E}[XYZ] \leqslant \mathbb{E}\,|XYZ| \leqslant (\mathbb{E}\,X^2\,\mathbb{E}(YZ)^2)^{1/2}$. But $\mathbb{E}(YZ)^2 = \mathbb{E}\,Y^2Z^2 \leqslant (\mathbb{E}\,Y^4\,\mathbb{E}\,Z^4)^{1/2}$. Thus,*

$$\mathbb{E}[XYZ] \leqslant \sqrt{\mathbb{E}\,X^2}(\mathbb{E}\,Y^4\,\mathbb{E}\,Z^4)^{1/4}.$$

Hence, $\mathbb{E}[XYZ]$ will be finite if X have finite second moment and Y, Z have finite fourth moments.

Chapter 3

The Linear Model

3.1 What is an econometric model?

An econometric analysis usually deals with relationship between variables. Typically a theory exhibits how one set of variables, say, y, will be determined by another x. Demand may be determined by price; expenditure by income; labor supply by wage rates and so on. Often this relationships are deterministic in the sense that for any x there will be a unique y chosen by the agent or determined by a competitive equilibrium. A deterministic relationship is represented by a function, say $g(y,x) = 0$ with a unique solution for y given a value for x. One of the question faced by the econometrician is to relate this economic model to data on x and y is such a way that the theory can be tested to see if it is consistent with the evidence, and can be used to make predictions and decisions.

Let be now more specific about the variables. Instead of simply writing y,x, we write y_i and x_i, where y_i is the value of y for unit i, and x_i is a k-dimensional vector variables $\mathrm{x}_i = (x_{i1},x_{i2},\ldots,x_{ik})'$ for the same individual. The relationship relating y_i and x_i is now expressed as $g(y_i,\mathrm{x}_i)$.

The main difficulty is that economic data typically do not exhibit relationship that are deterministic. Agents that face the same wage rate choose how many hours to work in different ways. We resolve this by broadening the theory to, say, $g(y_i,\mathrm{x}_i) = u_i$, where for any particular pair of y_i, x_i we expect u_i to be zero but allow that it might not be. This introduce a source of *uncertainty* into the model that can be either attributed to measurement error or to *unobserved* factors that our theory does not take into account, that is are not in x_i. For instance, although two female workers face the same wage rate, they choose to work different amount of hours depending on their particular socio-economic situation.

We now contemplate the possibility that y_i and x_i are observed on several occasions,

say for n individuals, n firms, n points in time or in space, etc depending on the nature of the relationship. For each of those "agents" the relationship is supposed to hold apart to the deviations u_i $i = 1,\ldots,n$ that admit the possibility that agents with the same x_i will be observed to have different y_i. But since the theory is "true" we also expect that $u_i = 0$, at least approximately. The problem is: what does it mean for u_i to be approximately zero? The "trick" is to assume that u_i has a probability distribution. In this case, close to zero may mean $E[u_i] = 0$, where the expectation is taken with respect to the (unknown, yet existent) distribution of u_i. Having expectation equal to zero means that if we where to observe the *same* agent over and over again then the different realization of u_i would have a certain distribution centered a zero.

But the restriction that $E[u_i] = 0$ is not nearly enough to specify a probability model for (y_i, x_i) if **both** y_i and x_i are random. If only y_i is random and x_i is fixed than $E[u_i] = 0$ would be sufficient. [1] Indeed, if x_i is fixed, the distribution of y_i is determined by the distribution of u_i. If both y_i and x_i are random, then their distributions—or moment of their distributions—cannot be determined by imposing conditions on the expected value of u_i; we will need conditions on the *joint* distribution of (x_i, u_i).

Now you can say: why do we need x_i to be random? Reliance on fixed regressors or, more generally, fixed "exogenous" variables, can have unintended consequences. If x_i are taken as nonrandom then u_i and x_i are independent of one other, but this rules out important situations of interest. Viewing the x_i random draws along y_i forces us to think about the relationship between the error and the explanatory variables.

3.2 Causal relationship

But before going further into the probabilistic model for (y_i, x_i) it is necessary to define the objectives of the econometric analysis. The goal of most empirical studies in economics and other social sciences is to determine whether change in one of the x_i causes a change in y_i. For example, does having another year of education cause an increase in monthly salary? Does reducing class size cause an improvement in student performance? Does lowering the business tax rate cause an in increase in economic activity.

If y_i and x_i are related in a deterministic fashion, $y_i = h(x_i)$, then we are often interested in how y_i changes when elements of x_i change. Since we have already ruled out the possibility that a deterministic relationship is likely to hold, there would be other factors, that we have called u_i, affecting y_i. Nevertheless, we can

[1] Fixed means that if we were to observe the same agent over and over she will have the same value of x_i, say $\mathrm{x}_i = \bar{\mathrm{x}}$.

define the partial effects of the x_{ij} on the on the conditional expectation $E[y_i|\mathrm{x}_i]$. Assuming that the conditional expectation is differentiable and x_{ij} is a continuous variable, the partial derivative allows us to approximate the effect of x_{ij} on y_i:

$$\Delta \mathbb{E}[y_i|\mathrm{x}_i] \approx \frac{\partial \mathbb{E}[y_i|\mathrm{x}_i]}{\partial x_{ij}} \cdot \Delta x_{ij}, \quad \text{holding } x_{i1},\ldots,x_{ij-1},x_{ij+1},\ldots,x_k \text{ fixed.}$$

The partial derivative of $E[y|\mathrm{x}_i]$ with respect to x_{ij} is usually called the **partial effect** of x_{ij} on $E[y_i|\mathrm{x}_i]$.

If x_{ij} is a discrete variable (such as a binary variable), partial effects are computed by comparing $E[y|\mathrm{x}_i]$ at different settings of x_{ij}, holding other variables fixed.

3.3 The Linear Model

Although economic models often provide a mathematical form for g, the common and simplest approach is to assume that this function is linear. In this case, given a vector of parameters, $\beta_0,\beta_1,\ldots,\beta_k$, the econometric model becomes $g(y_i,\mathrm{x}_i) = y_i - \beta_0 - \beta_1 x_{1i} - \ldots - \beta_k x_{ik}$.[2] Thus, a linear $g(\cdot,\cdot)$ gives the **linear model**:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + u_i = \mathrm{x}_i\beta + u_i.$$

Although the theoretical relationship linking y_i and x_i is often mathematically much more complicated, the linear model amid its simplicity has become a workhorse of econometric modeling.

It will be often useful using the matrix notation, in which case, the linear model is

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

where

$$\underset{(n\times 1)}{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \underset{(n\times k)}{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & & x_{2k} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}, \quad \underset{(n\times 1)}{u} = \begin{pmatrix} u1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}.$$

We have at least three possibilities in order to specify a probabilistic model for (y_i,x_i) in the context of a linear model. The strongest assumption that u_i and x_i are independent and normally distributed:

[2]Notice that we could have expressed any of the x's as a linear function of the other x's and of y. We are here assuming that the theory is designed to explain the determination of y in by x.

Assumption (A1). $u_i|\mathrm{x}_i \sim N(0, \sigma^2)$.

We can weaken this considerably. First, we could relax normality and only assume independence:

Assumption (A2). $u_i \perp \mathrm{x}_i$

We can weaken this assumption further by requiring only mean independence

Assumption (A3). $E[u_i|\mathrm{x}_i] = 0$

Mean-independence means that the expected value of of u_i is equal to zero very **every** possible value of x_i. Under this belief, it follows that

$$\mathbb{E}[y_i|\mathrm{x}_i] = \mathrm{x}_i\beta + \mathbb{E}[u_i|\mathrm{x}_i] = \mathrm{x}_i\beta, \tag{3.3.1}$$

Under one of the Assumptions A1-A3, the conditional expectation of y_i given x_i is linear as in (3.3.1). Thus, it should be immediate to see that

$$\beta_j = \frac{\partial\, \mathbb{E}[y_i|\mathrm{x}_i]}{\partial x_{ij}}, \quad j = 1, \ldots, k.$$

In words, the coefficient β_j is the partial effect of x_{ij} on $E[y_i|\mathrm{x}_i]$.

Assumption A1 specifies a full **parametric** model for (y_i, x_i). By parametric we mean that once you now the parameters (β, σ), then you know everything about the conditional distribution of y_i, since

$$p(y_i|\mathrm{x}_i) = \sqrt{\frac{1}{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mathrm{x}_i\beta)}{2\sigma^2}\right). \tag{3.3.2}$$

It should be now clear that we are interested in estimating β in reason to the fact that under either one of Assumptions A1-A3, $\beta_1, \ldots, \beta_k$ "parameterize" the partial effect of x_{ij} on y_i.

3.3.1 Estimating β under Assumption A1

Under Assumption A1, β (and σ) can be estimated by Maximum Likelihood. But while the presentation of the previous section relied only on assumption on the joint distribution of u_i and x_i, to estimate β with MLE we need further assumptions.

First of all, the observations are drawn randomly:

Assumption (B1). *The pairs* (u_i, x_i) *are independent draws from the same distribution.*

We must also impose some structure on the variance of u_i conditional on x_i.

Assumption (B2)**.** *The variance of u_i conditionally on x_i is constant, that is,* $\mathrm{Var}[u_i|\mathrm{x}_i] = \mathrm{Var}[u_i] = \sigma^2$.

This assumption is usually referred to as homoskedasticity.

From (3.3.2) and from the assumption of independence between x_i and u_i, and independence of (u_i, x_i), we can write that

$$\begin{aligned} p(y_1,\ldots,y_n,\mathrm{x}_1,\ldots,\mathrm{x}_n;\beta,\sigma) &= p(y_1,\ldots,y_n|\mathrm{x}_1,\ldots,\mathrm{x}_n;\beta,\sigma)p(\mathrm{x}_1,\ldots,\mathrm{x}_n) \\ &\propto \prod_{i=1}^{n} p(y_i|\mathrm{x}_i;\beta,\sigma) \\ &= (2\pi\sigma^2)^{-n/2}\exp\left(-\frac{\sum_{i=1}^{n}(y_i-\mathrm{x}_i\beta)^2}{2\sigma^2}\right) \\ &= (2\pi\sigma^2)^{-n/2}\exp\left(-\frac{(Y-X\beta)'(Y-X\beta)}{2\sigma^2}\right). \end{aligned}$$

Thus, the log-likelihood is

$$\mathcal{L}(\beta,\sigma) := \ln p(y_1,\ldots,y_n,\mathrm{x}_1,\ldots,\mathrm{x}_n;\beta,\sigma) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{(Y-X\beta)'(Y-X\beta)}{2\sigma^2}.$$

The Maximum Likelihood estimator solves the First Order Conditions (FOC)

$$\begin{aligned} -X'(Y-X\beta) &= 0 \\ \frac{n}{2\sigma^2} - \frac{(Y-X\beta)'(Y-X\beta)}{2\sigma^2} &= 0. \end{aligned}$$

Solving these two equations gives the MLE estimator of β and σ^2

$$\hat{\beta}_{ml} = (X'X)^{-1}X'Y, \quad \hat{\sigma}^2_{ml} = \frac{1}{n}(Y-X\beta)'(Y-X\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i-\mathrm{x}_i'\beta)^2.$$

Under Assumption A1, $\hat{\beta}$ is unbiased for β:

$$\begin{aligned} E[\hat{\beta}] &= E[E[\hat{\beta}|\mathrm{x}_1,\ldots,\mathrm{x}_n]|\mathrm{x}_1,\ldots,\mathrm{x}_n] \\ &= \beta + E[E[u|\mathrm{x}_1,\ldots,\mathrm{x}_n]|\mathrm{x}_1,\ldots,\mathrm{x}_n] \\ &= \beta. \end{aligned}$$

The MLE for σ^2 is biased. It can be shown that

$$E[\hat{\sigma}^2_{ml}] = \frac{n-k-1}{n}\sigma^2.$$

This bias can be eliminated by premultiplying $\hat{\sigma}^2_{ml}$ by $(n-k-1)/n$ which leads to the following estimators:

$$\hat{\sigma}^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}(y_i - \mathrm{x}_i'\beta)^2.$$

The ordinary least squares estimator for β solves

$$\min_{\beta}\sum_{i=1}^{n}(y_i - \mathrm{x}_i\beta)^2 = \min_{\beta}(Y - X\beta)'(Y - X\beta).$$

This leads to

$$\hat{\beta}_{ml} = (X'X)^{-1}X'Y.$$

Under the normality assumption, the exact (conditional) distribution of β is given by

$$\hat{\beta}_{ml}|X \sim N(0, \sigma^2(X'X)^{-1}).$$

Often we are interested in a particular coefficient. Suppose for example we are interested in β_1. In that case we have

$$\hat{\beta}_1|X \sim N(\beta_1, V_{22}),$$

where V_{ij} is the (i,j) element of the matrix V.

The variance of the distribution of $\hat{\beta}$ can be estimated by

$$\hat{V} = \hat{\sigma}^2(X'X)^{-1}.$$

3.3.2 Estimating β under Assumption A2 (or A3)

If we believe that assumption Assumption A1 is not reasonable, we have to sort out two distinct problem. First, how we estimate β? A natural way to proceed is to use $\hat{\beta}_{ml}$ as an estimator of β. The problem is that if the conditional distribution of u_i given x_i is not normal, that the likelihood function maximized by $\hat{\beta}_{ml}$ is not the right likelihood. Second, without Assumption A1 we cannot use the properties of the normal to derive the distribution of $\hat{\beta}$.

Let's start with the problem of finding an estimator of β when normality is not assumed. under Assumption A2/A3, we have that

$$E[u_i|\mathrm{x}_i] = 0. \tag{3.3.3}$$

From (3.3.3) it follows (see Exercise **??**) that

$$E[\mathrm{x}_i'u_i] = E[\mathrm{x}_i'(y_i - \mathrm{x}_i\beta)] = 0. \tag{3.3.4}$$

Solving for β, we find that

$$\beta = E[\mathrm{x}_i'\mathrm{x}_i]^{-1}E[\mathrm{x}_i'y_i]. \tag{3.3.5}$$

So, the linear projection of y_i onto x_i solves (3.3.4), and thus this β is what we would like to know. Unfortunately, this quantity depends on moments of the joint distribution of (y_i, x_i). A way to proceed in this case is to apply the "analogy principle" and to substitute population moments, i.e. expectations, with sample moments (i.e. averages on the sample). Doing so, we obtain the following estimator

$$\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n}\mathrm{x}_i'\mathrm{x}_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathrm{x}_i y_i\right).$$

It should not be difficult to see that this estimator is equivalent to the MLE estimator of β.

3.4 Exercises

Exercise 3.1. *Suppose* (u_i, x_i), $i = 1, \ldots, n$ *are independently distributed. Assume also that* $y_i = E[y_i|\mathrm{x}_i] + u_i$. *Show that*

$$E[u_i|\mathrm{x}_1, \ldots, \mathrm{x}_n] = E[u_i|\mathrm{x}_i] = 0.$$

Exercise 3.2. *Suppose* y *is a random variable,* $E|y| < \infty$, *and* x *is a random vector of dimension* $(k \times 1)$ *are two random variables. Show that if* $E[y|x] = 0$, *then* $E[f(x)'y] = 0$, *where is a function* $f : \mathbb{R}^k \to \mathbb{R}^k$.

Chapter 4

Inference

4.1 Inference in the Linear model

From the previous section, two asymptotic distribution for the OLS estimator can be given depending on the assumption on the conditional variance of the error $\{e_i\}$:

$$\sqrt{N}\left(\hat{\beta}-\beta\right) \xrightarrow{d} N\left[0, E\left(x_i x_i'\right)^{-1} E\left(e_i^2 x_i x_i'\right) E\left(x_i x_i'\right)^{-1}\right], \tag{4.1.1}$$

$$\sqrt{N}\left(\hat{\beta}-\beta\right) \xrightarrow{d} N\left[0, \sigma^2 E\left(x_i x_i'\right)^{-1}\right]. \tag{4.1.2}$$

The approximation in (5.1.2) holds whenever

$$E\left(e_i^2 x_i x_i'\right) = \sigma^2 E\left(x_i x_i'\right),$$

which in turns hold if $E[e_i^2|x_i] = \sigma^2$, that is, when the distribution of the error is *conditionally* homoskedastic.

4.2 Variance Estimation

In order to make use of the asymptotic distributions in (**??**) and (5.1.2), we need to provide estimator of the asymptotic variance. To ease notation, we use the following notation:

$$Q = E[x_i x_i'], \quad V = E[e_i^2 x_i x_i'], \quad V = Q^{-1} V Q^{-1}.$$

Call an estimator of V, $\hat{V}$. By application of Slutsky's Lemma, as done in Example 1, if $\hat{V} \xrightarrow{p} V$, then

$$\sqrt{n}\hat{V}^{-1/2}(\hat{\beta}-\beta) \xrightarrow{d} N(0, I_k).$$

So, we will be interested in estimators of V that are consistent in probability.

The most widely used estimator of V is constructed by obtaining consistent estimator for Q and for V. Consider

$$\hat{Q} = \sum_{i=1}^{n} x_i x_i'/N, \qquad \hat{V} = \sum_{i=1}^{n} \hat{e}_i^2 x_i x_i'/N, \quad \hat{e}_i = y_i - x_i'\hat{\beta}.$$

By the LLN, $\hat{Q}$ is consistent if $E[x_i x_i'] < \infty$ which is the same assumption needed to derive consistency of the OLS estimator. Thus,

$$\hat{Q} \xrightarrow{p} E[x_i x_i'] = Q.$$

By assumption, $E[x_i x_i']$ has full rank. So, by the continuous mapping theorem

$$\hat{Q}^{-1} \xrightarrow{p} Q^{-1}.$$

To show that $\hat{V}$ is consistent is a little bit more involved since it is necessary to show that replacing the unobserved $\{e_i\}$ with the estimators $\{\hat{e}_i\}$ does not affect the consistency of $\hat{V}$. In practice, we can show—under additional assumptions on the moments of $\{x_i\}$ and $\{e_i\}$—that

$$\hat{V} = \sum_{i=1}^{n} e_i^2 x_i x_i' + o_p(1),$$

thus $\hat{V} \xrightarrow{p} E[e_i^2 x_i x_i'] = V$. Now, let

$$\hat{V} = \hat{Q}^{-1}\hat{V}\hat{Q}^{-1}.$$

Since matrix product is a continuous operation it follows—again by the continuous mapping theorem that

$$\hat{V} \xrightarrow{p} V.$$

If one is willing to assume conditional homoskedasticity, then

$$V = \sigma^2 E[x_i x_i']^{-1} = \sigma^2 Q^{-1}.$$

A consistent estimator is

$$\hat{\sigma}^2 \hat{Q}^{-1}$$

where

$$\hat{\sigma}^2 = \sum_{i=1}^{n} (y_i - x_i\hat{\beta})^2/N.$$

Given a consistent estimator of V, $\hat{V}$, we can define the standard error of $\hat{\beta}_j$, $j = 1, \ldots, k$.

Definition 5. The standard error of $\hat{\beta}_j$, denoted by $s(\hat{\beta}_j)$ is given by

$$s(\hat{\beta}_j) = \sqrt{\frac{\hat{V}_{(j,j)}}{N}},$$

where $A_{(j,j)}$ denotes the entry (j,j) of the matrix A.

4.3 t-statistic

The statistic

$$t_n(\beta) = \frac{\hat{\beta}_j - \beta_j}{\mathrm{s}(\hat{\beta}_j)}.$$

is labeled as t-statistic, z-statistic or a studentized statistic. The following two result hold:

$$t_n \xrightarrow{p} 0, \qquad t_n \xrightarrow{d} N(0,1).$$

We say that t_n is *asymptotically pivotal* since its asymptotic distribution does not depend on β os any other unknown parameters. For instance, the statistic $v_n(\beta) = \hat{\beta}_j - \beta$ is not asymptotically pivotal, since its asymptotic distribution, $N(0,\sigma^2_{\beta_j})$ depends on an unknown parameter (usually referred to as nuisance parameter).

4.4 Confidence Intervals

Abstractly, a confidence interval is a random interval that contains a parameter value with a fixed probability. Formally, for random vectors $\{w_i\}$ and a parameter θ, a confidence interval with coverage α is a set—function of $\{w_i\}$—$\mathcal{C}(w_1,\ldots,w_n)$ such that

$$P\left(\theta \in \mathcal{C}(w_1,\ldots,x_n)\right) = \alpha.$$

Since we rarely know the exact distribution of $\mathcal{C}(w_1,\ldots,x_n)$, we often relay on asymptotic approximation and we define the confidence interval accordingly:

$$\lim_{n\to\infty} P\left(\theta \in \mathcal{C}(w_1,\ldots,w_n)\right) = 1-\alpha.$$

Definition 6. Consider now, constructing a confidence interval for β_j, some $j = 1,\ldots,k$. Then, the interval will be function of $(y_1,\ldots,y_n,x_1,\ldots,x_n)$, say $\mathcal{C}(y,x)$. $\mathcal{C}(y,x)$ will be a valid confidence interval if

$$\lim_{n\to\infty} P\left(\beta_j \in \mathcal{C}(y,x)\right) = 1-\alpha.$$

Consider the following proposal for $\mathcal{C}(y,x)$

$$\mathcal{C}(y,x) = (\hat{\beta}_j - z_{\alpha/2}s(\hat{\beta}_j), \hat{\beta}_j + z_{\alpha/2}s(\hat{\beta}_j)), \tag{4.4.1}$$

where $z_{\alpha/2}$ is the $\alpha/2$-percentile of the standard normal distribution normal distribution, that is, if $z \sim N(0,1)$,

$$\Pr(z \geq z_{\alpha/2}) = \alpha/2.$$

To make sure (5.4.1) is a valid confidence interval we have to show that

$$\lim_{n\to\infty} P\left[\beta_j \in \left(\hat{\beta}_j - z_{\alpha/2}\cdot s(\hat{\beta}_j), \hat{\beta}_j + z_{\alpha/2}\cdot s(\hat{\beta}_j)\right)\right] = 1-\alpha,$$

or equivalently, that

$$\lim_{n\to\infty} P\left[\beta_j \notin \left(\hat{\beta}_j - z_{\alpha/2}\cdot s(\hat{\beta}_j), \hat{\beta}_j + z_{\alpha/2}\cdot s(\hat{\beta}_j)\right)\right] = \alpha,$$

Now the event $\beta_j \notin (\hat{\beta}_j - z_{\alpha/2}\cdot s(\hat{\beta}_j), \hat{\beta}_j + z_{\alpha/2}\cdot s(\hat{\beta}_j))$ is equivalent to the event

$$\left\{t_n \geqslant z_{\alpha/2}\right\} \cup \left\{t_n \leqslant -z_{\alpha/2}\right\}.$$

The probability of the event is then

$$P\left\{t_n \geqslant z_{\alpha/2}\right\} + P\left\{t_n \leqslant -z_{\alpha/2}\right\} = E1\left\{t_n \geqslant z_{\alpha/2}\right\} + E1\left\{t_n \leqslant -z_{\alpha/2}\right\}.$$

By boundedness of the indicator function $1(\cdot)$, the fact that $t_n \xrightarrow{d} N(0,1)$, Lemma 1 implies that, for $z \sim N(0,1)$,

$$\lim_{n\to\infty} E1\left\{t_n \geqslant z_{\alpha/2}\right\} = E1\left\{z \geqslant z_{\alpha/2}\right\} = \alpha/2,$$

and

$$\lim_{n\to\infty} E1\left\{t_n \leqslant -z_{\alpha/2}\right\} = E1\left\{z \leqslant -z_{\alpha/2}\right\} = \alpha/2.$$

Thus,

$$\lim_{n\to\infty} P\left[\beta_j \notin \left(\hat{\beta}_j - z_{\alpha/2}\cdot s(\hat{\beta}_j), \hat{\beta}_j + z_{\alpha/2}\cdot s(\hat{\beta}_j)\right)\right] = \alpha.$$

Definition 7. A confidence interval for the $j-th$ element of β with asymptotic coverage $1-\alpha$ is given by:

$$\left(\hat{\beta}_j - z_{\alpha/2}\cdot s(\hat{\beta}_j), \hat{\beta} + z_{\alpha/2}\cdot s(\hat{\beta}_j)\right).$$

For example, if a 95% confidence interval is sought, $\alpha = 0.05$ and $z_\alpha = 1.96$.

Remark 1. In general, the *joint* confidence for β and not simply the j-th element of β will be a more complicate region. These regions can be obtained by inverting test statistics.

4.5 Hypothesis Testing

4.5.1 Preliminaries

With hypothesis testing, as the name clearly puts it, we are concerned with testing whether some hypothesis concerning parameters is supported.

The hypothesis being tested is said null hypothesis and usually denoted by H_0, the alternative hypothesis is denoted by H_1. Null and alternative hypotheses can be divided into two main classes: simple and composite. A simple null and a simple alternative take the form

$$H_0 : \beta = \beta_0; \qquad H_1 : \beta_j = \beta_1.$$

A simple null and a composite alternative take the form

$$H_0 : \beta = \beta_0; \qquad H_1 : \beta \neq \beta_0.$$

If we are interested in carrying out a test on a particular element of the parameter vector β, say $\hat{\beta}_j$, then composite alternative are either dual-sided:

$$H_1 : \beta_j \neq \beta_{j0},$$

or one-sided:

$$H_1 : \beta_j \geqslant \beta_{j0}, \qquad \textit{or} \qquad H_1 : \beta_j \leqslant \beta_{j0}.$$

4.5.2 Testing hypothesis on β_j

Many concept of this section will be common to all testing situation, but it is much easier to explain them in the contest of testing hypothesis on β_j. In particular, we consider the simple null $H_0 : \beta_j = \beta_{j,0}$ and the dual sided alternative $H_1 : \beta_j \neq \beta_{j,0}$.

Under the null hypothesis, that is, when it is actually the case that $\beta_j = \beta_{j,0}$, $t_n(\beta_{j,0}) \xrightarrow{p} 0$. Then an intuitive testing procedure is to reject the null hypothesis when $t_n(\beta_{j,0})$ is large, say $|t_n(\beta_{j,0})| > k$ and to accept the null hypothesis when $t_n(\beta_{j,0})$ is small, say $|t_n(\beta_{j,0})| \leqslant k$.

When performing a test one may arrive at the correct decision, or one may commit one of two errors: rejecting the hypothesis when it is true (error of the first kind) or accepting it when it is false (error of the second kind). The consequences of these are often quite different. For example, if one tests for the presence of some disease, incorrectly deciding on the necessity of treatment may cause the patient discomfort and financial loss. On the other hand, failure to diagnose the presence of the ailment may lead to the patient's death.

Definition 8. The asymptotic type I error of a test procedure is the limit probability that the procedure incorrectly reject the null hypothesis. In our case,

$$\lim_{n\to\infty} P(|t_n(\beta_{j,0})| > k|H_0).$$

Definition 9. The asymptotic type II error of a test is the limit probability of accepting the null hypothesis when H_1 is indeed true. In our case,

$$\gamma(\beta_j) \stackrel{def}{=} \lim_{n\to\infty} P(|t_n(\beta_{j,0})| \leqslant k|H_1).$$

Definition 10. The asymptotic power of a test is the limit probability of correctly rejecting the null hypothesis. In our case,

$$\pi(\beta_j) \stackrel{def}{=} \lim_{n\to\infty} P(|t_n(\beta_{j,0})| > k|H_1) = 1 - \gamma(\beta_j).$$

Notice that both the type II error and the power of the test will depend on the alternative value of β_j. It is desirable to carry out the test in a manner which keeps the probabilities of the two types of error to a minimum. Unfortunately, both probabilities cannot be controlled simultaneously. We proceed by choosing the test—that in our case amount to choosing k—in such a way to control the error of the first kind.

We proceed by choosing a value of $\alpha \in (0,1)$ and we choose k in such a way to set the error of the first type equal to α. In practice, we need to find k such that

$$\alpha = \lim_{n\to\infty} P(|t_n(\beta_{j,0})| > k|H_0).$$

While this problem seems hard, it is actually fairly easy. Since $t_n(\beta_{j,0}) \xrightarrow{d} N(0,1)$ under H_0, we have that for $z_{\alpha/2}$ defined in the usual manner,

$$\alpha = \lim_{n\to\infty} P(|t_n(\beta_{j,0})| > z_{\alpha/2}|H_0).$$

So, the testing procedures defined as

$$\begin{cases} reject & \text{if } |t_n| > z_{\alpha/2} \\ accept & \text{if } |t_n| \leqslant z_{\alpha/2} \end{cases}$$

has asymptotic type I error equal to α. It is said a test of level α. While we know what the *size* of the error of the first kind is, we do not anything about the power of the test. For almost all the test we study it is the case that

$$\pi(\beta) \to 1, \qquad n \to \infty,$$

that is the power of the test goes to 1 as n increases to ∞.

In econometric parlance, when $|t_n| > 1.96$ it is common to say that the t-statistic is statistically significant, or that β_j is statistically significant (or statistically different from 0).

4.5.3 p-value

Definition 11. The $p - value$ is defined as

$$p_n \overset{def}{=} p(t_n) = P(|Z| > |t_n|) = 2(1 - \Phi(|t_n|)),$$

where $Z \sim N(0,1)$.

Significance tests can be deducted from $p-$value, since $p_n < \alpha$ iff $|t_n| > z_{a/2}$. So if $p_n > \alpha$, we cannot reject the null hypothesis at $\alpha\%$ significance level.

The magnitude of the p-value is not indication of the plausibility of the null hypothesis. The fact that p_n is very large, say $p_n = 0.98$, does not mean that H_1 is very likely (or the null very likely). In facts, by construction, the p-value is asymptotically uniformly distributed:

$$P(1 - p_n \leqslant u) = P(F(t_n) \leqslant u) = P(|t_n| \leqslant F^{-1}(u)) \to F(F^{-1}(u)) = u.$$

Testing multiple hypothesis

Suppose we wish to test multiple hypotheses on β. For instance

$$\mathrm{H}_0 : \beta_1 = \beta_{1,0} \text{ and } \beta_2 = \beta_{2,0}$$

or

$$\mathrm{H}_0 : \beta_1 + \beta_3 = \beta_{13,0}$$

A general formulation is the following:

$$\mathrm{H}_0 : R\beta = r$$

where R is a $Q \times K$ matrix with rank $Q \leq K$, and r is $Q \times 1$ vector. The matrix R allows to write compactly *linear* hypothesis on the parameter vector β. For instance, if $K = 3$

$$R = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \quad r = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

the null hypothesis

$$\mathrm{H}_0 : R\beta = r$$

means

$$\mathrm{H}_0 : \beta_1 = 0 \text{ and } \beta_2 = 0$$

Similarly, if

$$R = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}, \quad r = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

the null hypothesis

$$\mathrm{H}_0 : R\beta = r$$

means

$$\mathrm{H}_0 : \beta_1 + \beta_2 = 0 \text{ and } \beta_2 + \beta_3 = 0$$

An important special case is given when R is $(K-1) \times K$ and r is 0_K

$$R = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & & & \ddots & \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}, \quad r = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

In this case the null hypothesis being tested is that all the coefficients with the exception of the intercept are equal to zero in the regression. Under the classical assumption of the linear model, this case is dealt by the F-statistics. Recall that the F statistic is given by

$$F[K-1, N-K] = \frac{R^2/(K-1)}{(1-R^2)/(N-K)}$$

Under the classical assumptions

$$F[K-1, N-K] \sim F_{(K-1),(N-K)}$$

However, if the assumption $E\left(x_i x_i'\right) \neq \sigma^2 E(x_i x_i')$ then the usual F statistics is not valid for testing linear restrictions, not even asymptotically.

The $Wald$ statistics is a generalization of the F statistics

$$W = \left(R\hat{\beta} - r\right)' \left(R\hat{V}R'\right)^{-1} \left(R\hat{\beta} - r\right)$$

where we let

$$\hat{V} = \left(\sum_i^n x_i x_i'\right)^{-1} \sum_i^n \hat{e}_i^2 x_i x_i' \left(\sum_i^n x_i x_i'\right)^{-1}$$

denote the robust variance estimator.

The intuition of the test statistics is simple: if the null hypothesis is true, then we should expect that $R\hat{\beta} - r$ is close to zero. However, even if the null hypothesis is true, from sample to sample we could observe large values of $(R\hat{\beta} - r)$ that are due to the sampling variability of $\hat{\beta}$. To take into account this variability we rescale

the quantity by the estimated variance of $R\hat{\beta}$, $\left(R\hat{V}R'\right)^{-1}$. It would a nice exercise to show that under the null hypothesis

$$W \xrightarrow{d} \chi^2_Q$$

So, we will reject the null hypothesis if

$$W > c_\alpha, \quad z \sim \chi^2_Q,\ c_\alpha = \Pr(z \geq \alpha)$$

Chapter 5

Linear Model Inference

5.1 Inference in the linear model

From the previous section, two asymptotic distribution for the OLS estimator can be given depending on the assumption on the conditional variance of the error $\{e_i\}$:

$$\sqrt{N}\left(\hat{\beta}-\beta\right) \xrightarrow{d} N\left[0, E\left(x_i x_i'\right)^{-1} E\left(e_i^2 x_i x_i'\right) E\left(x_i x_i'\right)^{-1}\right], \tag{5.1.1}$$

$$\sqrt{N}\left(\hat{\beta}-\beta\right) \xrightarrow{d} N\left[0, \sigma^2 E\left(x_i x_i'\right)^{-1}\right]. \tag{5.1.2}$$

The approximation in (5.1.2) holds whenever

$$E\left(e_i^2 x_i x_i'\right) = \sigma^2 E\left(x_i x_i'\right),$$

which in turns hold if $E[e_i^2|x_i] = \sigma^2$, that is, when the distribution of the error is *conditionally* homoskedastic.

5.2 Variance Estimation

In order to make use of the asymptotic distributions in (**??**) and (5.1.2), we need to provide estimator of the asymptotic variance. To ease notation, we use the following notation:

$$Q = E[x_i x_i'], \quad V = E[e_i^2 x_i x_i'], \quad V = Q^{-1} V Q^{-1}.$$

Call an estimator of V, $\hat{V}$. By application of Slutsky's Lemma, as done in Example 1, if $\hat{V} \xrightarrow{p} V$, then

$$\sqrt{n}\hat{V}^{-1/2}(\hat{\beta}-\beta) \xrightarrow{d} N(0, I_k).$$

So, we will be interested in estimators of V that are consistent in probability.

The most widely used estimator of V is constructed by obtaining consistent estimator for Q and for V. Consider

$$\hat{Q} = \sum_{i=1}^{n} x_i x_i'/N, \qquad \hat{V} = \sum_{i=1}^{n} \hat{e}_i^2 x_i x_i'/N, \quad \hat{e}_i = y_i - x_i'\hat{\beta}.$$

By the LLN, $\hat{Q}$ is consistent if $E[x_i x_i'] < \infty$ which is the same assumption needed to derive consistency of the OLS estimator. Thus,

$$\hat{Q} \xrightarrow{p} E[x_i x_i'] = Q.$$

By assumption, $E[x_i x_i']$ has full rank. So, by the continuous mapping theorem

$$\hat{Q}^{-1} \xrightarrow{p} Q^{-1}.$$

To show that $\hat{V}$ is consistent is a little bit more involved since it is necessary to show that replacing the unobserved $\{e_i\}$ with the estimators $\{\hat{e}_i\}$ does not affect the consistency of $\hat{V}$. In practice, we can show—under additional assumptions on the moments of $\{x_i\}$ and $\{e_i\}$—that

$$\hat{V} = \sum_{i=1}^{n} e_i^2 x_i x_i' + o_p(1),$$

thus $\hat{V} \xrightarrow{p} E[e_i^2 x_i x_i'] = V$. Now, let

$$\hat{V} = \hat{Q}^{-1}\hat{V}\hat{Q}^{-1}.$$

Since matrix product is a continuous operation it follows—again by the continuous mapping theorem that

$$\hat{V} \xrightarrow{p} V.$$

If one is willing to assume conditional homoskedasticity, then

$$V = \sigma^2 E[x_i x_i']^{-1} = \sigma^2 Q^{-1}.$$

A consistent estimator is

$$\hat{\sigma}^2 \hat{Q}^{-1}$$

where

$$\hat{\sigma}^2 = \sum_{i=1}^{n} (y_i - x_i\hat{\beta})^2/N.$$

Given a consistent estimator of V, $\hat{V}$, we can define the standard error of $\hat{\beta}_j$, $j = 1, \ldots, k$.

Definition 12. The standard error of $\hat{\beta}_j$, denoted by $s(\hat{\beta}_j)$ is given by

$$s(\hat{\beta}_j) = \sqrt{\frac{\hat{V}_{(j,j)}}{N}},$$

where $A_{(j,j)}$ denotes the entry (j,j) of the matrix A.

5.3 t-statistic

The statistic

$$t_n(\beta) = \frac{\hat{\beta}_j - \beta_j}{\mathrm{s}(\hat{\beta}_j)}.$$

is labeled as t-statistic, z-statistic or a studentized statistic. The following two result hold:

$$t_n \xrightarrow{p} 0, \qquad t_n \xrightarrow{d} N(0,1).$$

We say that t_n is *asymptotically pivotal* since its asymptotic distribution does not depend on β os any other unknown parameters. For instance, the statistic $v_n(\beta) = \hat{\beta}_j - \beta$ is not asymptotically pivotal, since its asymptotic distribution, $N(0,\sigma^2_{\beta_j})$ depends on an unknown parameter (usually referred to as nuisance parameter).

5.4 Confidence Intervals

Abstractly, a confidence interval is a random interval that contains a parameter value with a fixed probability. Formally, for random vectors $\{w_i\}$ and a parameter θ, a confidence interval with coverage α is a set—function of $\{w_i\}$—$\mathcal{C}(w_1,\ldots,w_n)$ such that

$$P\left(\theta \in \mathcal{C}(w_1,\ldots,x_n)\right) = \alpha.$$

Since we rarely know the exact distribution of $\mathcal{C}(w_1,\ldots,x_n)$, we often relay on asymptotic approximation and we define the confidence interval accordingly:

$$\lim_{n\to\infty} P\left(\theta \in \mathcal{C}(w_1,\ldots,w_n)\right) = 1-\alpha.$$

Definition 13. Consider now, constructing a confidence interval for β_j, some $j = 1,\ldots,k$. Then, the interval will be function of $(y_1,\ldots,y_n,x_1,\ldots,x_n)$, say $\mathcal{C}(y,x)$. $\mathcal{C}(y,x)$ will be a valid confidence interval if

$$\lim_{n\to\infty} P\left(\beta_j \in \mathcal{C}(y,x)\right) = 1-\alpha.$$

Consider the following proposal for $\mathcal{C}(y,x)$

$$\mathcal{C}(y,x) = (\hat{\beta}_j - z_{\alpha/2}s(\hat{\beta}_j), \hat{\beta}_j + z_{\alpha/2}s(\hat{\beta}_j)), \tag{5.4.1}$$

where $z_{\alpha/2}$ is the $\alpha/2$-percentile of the standard normal distribution normal distribution, that is, if $z \sim N(0,1)$,

$$\Pr(z \geq z_{\alpha/2}) = \alpha/2.$$

To make sure (5.4.1) is a valid confidence interval we have to show that

$$\lim_{n\to\infty} P\left[\beta_j \in \left(\hat{\beta}_j - z_{\alpha/2}\cdot s(\hat{\beta}_j), \hat{\beta}_j + z_{\alpha/2}\cdot s(\hat{\beta}_j)\right)\right] = 1-\alpha,$$

or equivalently, that

$$\lim_{n\to\infty} P\left[\beta_j \notin \left(\hat{\beta}_j - z_{\alpha/2}\cdot s(\hat{\beta}_j), \hat{\beta}_j + z_{\alpha/2}\cdot s(\hat{\beta}_j)\right)\right] = \alpha,$$

Now the event $\beta_j \notin (\hat{\beta}_j - z_{\alpha/2}\cdot s(\hat{\beta}_j), \hat{\beta}_j + z_{\alpha/2}\cdot s(\hat{\beta}_j))$ is equivalent to the event

$$\left\{t_n \geqslant z_{\alpha/2}\right\} \cup \left\{t_n \leqslant -z_{\alpha/2}\right\}.$$

The probability of the event is then

$$P\left\{t_n \geqslant z_{\alpha/2}\right\} + P\left\{t_n \leqslant -z_{\alpha/2}\right\} = E1\left\{t_n \geqslant z_{\alpha/2}\right\} + E1\left\{t_n \leqslant -z_{\alpha/2}\right\}.$$

By boundedness of the indicator function $1(\cdot)$, the fact that $t_n \xrightarrow{d} N(0,1)$, Lemma 1 implies that, for $z \sim N(0,1)$,

$$\lim_{n\to\infty} E1\left\{t_n \geqslant z_{\alpha/2}\right\} = E1\left\{z \geqslant z_{\alpha/2}\right\} = \alpha/2,$$

and

$$\lim_{n\to\infty} E1\left\{t_n \leqslant -z_{\alpha/2}\right\} = E1\left\{z \leqslant -z_{\alpha/2}\right\} = \alpha/2.$$

Thus,

$$\lim_{n\to\infty} P\left[\beta_j \notin \left(\hat{\beta}_j - z_{\alpha/2}\cdot s(\hat{\beta}_j), \hat{\beta}_j + z_{\alpha/2}\cdot s(\hat{\beta}_j)\right)\right] = \alpha.$$

Definition 14. A confidence interval for the $j-th$ element of β with asymptotic coverage $1-\alpha$ is given by:

$$\left(\hat{\beta}_j - z_{\alpha/2}\cdot s(\hat{\beta}_j), \hat{\beta} + z_{\alpha/2}\cdot s(\hat{\beta}_j)\right).$$

For example, if a 95% confidence interval is sought, $\alpha = 0.05$ and $z_\alpha = 1.96$.

Remark 2. In general, the *joint* confidence for β and not simply the j-th element of β will be a more complicate region. These regions can be obtained by inverting test statistics.

5.5 Hypothesis Testing

5.5.1 Preliminaries

With hypothesis testing, as the name clearly puts it, we are concerned with testing whether some hypothesis concerning parameters is supported.

The hypothesis being tested is said null hypothesis and usually denoted by H_0, the alternative hypothesis is denoted by H_1. Null and alternative hypotheses can be divided into two main classes: simple and composite. A simple null and a simple alternative take the form

$$H_0 : \beta = \beta_0; \qquad H_1 : \beta_j = \beta_1.$$

A simple null and a composite alternative take the form

$$H_0 : \beta = \beta_0; \qquad H_1 : \beta \neq \beta_0.$$

If we are interested in carrying out a test on a particular element of the parameter vector β, say $\hat{\beta}_j$, then composite alternative are either dual-sided:

$$H_1 : \beta_j \neq \beta_{j0},$$

or one-sided:

$$H_1 : \beta_j \geqslant \beta_{j0}, \qquad \textit{or} \qquad H_1 : \beta_j \leqslant \beta_{j0}.$$

5.5.2 Testing hypothesis on β_j

Many concept of this section will be common to all testing situation, but it is much easier to explain them in the contest of testing hypothesis on β_j. In particular, we consider the simple null $H_0 : \beta_j = \beta_{j,0}$ and the dual sided alternative $H_1 : \beta_j \neq \beta_{j,0}$.

Under the null hypothesis, that is, when it is actually the case that $\beta_j = \beta_{j,0}$, $t_n(\beta_{j,0}) \xrightarrow{p} 0$. Then an intuitive testing procedure is to reject the null hypothesis when $t_n(\beta_{j,0})$ is large, say $|t_n(\beta_{j,0})| > k$ and to accept the null hypothesis when $t_n(\beta_{j,0})$ is small, say $|t_n(\beta_{j,0})| \leqslant k$.

When performing a test one may arrive at the correct decision, or one may commit one of two errors: rejecting the hypothesis when it is true (error of the first kind) or accepting it when it is false (error of the second kind). The consequences of these are often quite different. For example, if one tests for the presence of some disease, incorrectly deciding on the necessity of treatment may cause the patient discomfort and financial loss. On the other hand, failure to diagnose the presence of the ailment may lead to the patient's death.

Definition 15. The asymptotic type I error of a test procedure is the limit probability that the procedure incorrectly reject the null hypothesis. In our case,

$$\lim_{n\to\infty} P(|t_n(\beta_{j,0})| > k|H_0).$$

Definition 16. The asymptotic type II error of a test is the limit probability of accepting the null hypothesis when H_1 is indeed true. In our case,

$$\gamma(\beta_j) \overset{def}{=} \lim_{n\to\infty} P(|t_n(\beta_{j,0})| \leqslant k|H_1).$$

Definition 17. The asymptotic power of a test is the limit probability of correctly rejecting the null hypothesis. In our case,

$$\pi(\beta_j) \overset{def}{=} \lim_{n\to\infty} P(|t_n(\beta_{j,0})| > k|H_1) = 1 - \gamma(\beta_j).$$

Notice that both the type II error and the power of the test will depend on the alternative value of β_j. It is desirable to carry out the test in a manner which keeps the probabilities of the two types of error to a minimum. Unfortunately, both probabilities cannot be controlled simultaneously. We proceed by choosing the test—that in our case amount to choosing k—in such a way to control the error of the first kind.

We proceed by choosing a value of $\alpha \in (0,1)$ and we choose k in such a way to set the error of the first type equal to α. In practice, we need to find k such that

$$\alpha = \lim_{n\to\infty} P(|t_n(\beta_{j,0})| > k|H_0).$$

While this problem seems hard, it is actually fairly easy. Since $t_n(\beta_{j,0}) \xrightarrow{d} N(0,1)$ under H_0, we have that for $z_{\alpha/2}$ defined in the usual manner,

$$\alpha = \lim_{n\to\infty} P(|t_n(\beta_{j,0})| > z_{\alpha/2}|H_0).$$

So, the testing procedures defined as

$$\begin{cases} reject & \text{if } |t_n| > z_{\alpha/2} \\ accept & \text{if } |t_n| \leqslant z_{\alpha/2} \end{cases}$$

has asymptotic type I error equal to α. It is said a test of level α. While we know what the *size* of the error of the first kind is, we do not anything about the power of the test. For almost all the test we study it is the case that

$$\pi(\beta) \to 1, \qquad n \to \infty,$$

that is the power of the test goes to 1 as n increases to ∞.

In econometric parlance, when $|t_n| > 1.96$ it is common to say that the t-statistic is statistically significant, or that β_j is statistically significant (or statistically different from 0).

5.5.3 p-value

Definition 18. The $p-value$ is defined as

$$p_n \overset{def}{=} p(t_n) = P(|Z| > |t_n|) = 2(1 - \Phi(|t_n|)),$$

where $Z \sim N(0,1)$.

Significance tests can be deducted from $p-$value, since $p_n < \alpha$ iff $|t_n| > z_{a/2}$. So if $p_n > \alpha$, we cannot reject the null hypothesis at $\alpha\%$ significance level.

The magnitude of the p-value is not indication of the plausibility of the null hypothesis. The fact that p_n is very large, say $p_n = 0.98$, does not mean that H_1 is very likely (or the null very likely). In facts, by construction, the p-value is asymptotically uniformly distributed:

$$P(1 - p_n \leqslant u) = P(F(t_n) \leqslant u) = P(|t_n| \leqslant F^{-1}(u)) \to F(F^{-1}(u)) = u.$$

Testing multiple hypothesis

Suppose we wish to test multiple hypotheses on β. For instance

$$\mathrm{H}_0 : \beta_1 = \beta_{1,0} \text{ and } \beta_2 = \beta_{2,0}$$

or

$$\mathrm{H}_0 : \beta_1 + \beta_3 = \beta_{13,0}$$

A general formulation is the following:

$$\mathrm{H}_0 : R\beta = r$$

where R is a $Q \times K$ matrix with rank $Q \leq K$, and r is $Q \times 1$ vector. The matrix R allows to write compactly *linear* hypothesis on the parameter vector β. For instance, if $K = 3$

$$R = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \quad r = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

the null hypothesis

$$\mathrm{H}_0 : R\beta = r$$

means

$$\mathrm{H}_0 : \beta_1 = 0 \text{ and } \beta_2 = 0$$

Similarly, if

$$R = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}, \quad r = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

the null hypothesis

$$\mathrm{H}_0 : R\beta = r$$

means

$$\mathrm{H}_0 : \beta_1 + \beta_2 = 0 \text{ and } \beta_2 + \beta_3 = 0$$

An important special case is given when R is $(K-1) \times K$ and r is 0_K

$$R = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & & & \ddots & \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}, \quad r = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

In this case the null hypothesis being tested is that all the coefficients with the exception of the intercept are equal to zero in the regression. Under the classical assumption of the linear model, this case is dealt by the F-statistics. Recall that the F statistic is given by

$$F[K-1, N-K] = \frac{R^2/(K-1)}{(1-R^2)/(N-K)}$$

Under the classical assumptions

$$F[K-1, N-K] \sim F_{(K-1),(N-K)}$$

However, if the assumption $E\left(x_i x_i'\right) \neq \sigma^2 E(x_i x_i')$ then the usual F statistics is not valid for testing linear restrictions, not even asymptotically.

The $Wald$ statistics is a generalization of the F statistics

$$W = \left(R\hat{\beta} - r\right)' \left(R\hat{V}R'\right)^{-1} \left(R\hat{\beta} - r\right)$$

where we let

$$\hat{V} = \left(\sum_i^n x_i x_i'\right)^{-1} \sum_i^n \hat{e}_i^2 x_i x_i' \left(\sum_i^n x_i x_i'\right)^{-1}$$

denote the robust variance estimator.

The intuition of the test statistics is simple: if the null hypothesis is true, then we should expect that $R\hat{\beta} - r$ is close to zero. However, even if the null hypothesis is true, from sample to sample we could observe large values of $(R\hat{\beta} - r)$ that are due to the sampling variability of $\hat{\beta}$. To take into account this variability we rescale

the quantity by the estimated variance of $R\hat{\beta}$, $\left(R\hat{V}R'\right)^{-1}$. It would a nice exercise to show that under the null hypothesis

$$W \xrightarrow{d} \chi^2_Q$$

So, we will reject the null hypothesis if

$$W > c_\alpha, \quad z \sim \chi^2_Q,\ c_\alpha = \Pr(z \geq \alpha)$$

Chapter 6

Instrumental Variables estimation

6.1 Introduction

We have a linear model

$$y_i = x_i'\beta + u_i, \quad i = 1, \ldots, n$$

and we are interested in estimating β, but we cannot assume that $E[x_i u_i] = 0$.

So, in this section, we explore methods to estimate β under the following assumption

$$E[x_i u_i] \neq 0.$$

The instrumental variable approach attempts at providing a solution to the estimation problem when $E[x_i u_i] \neq 0$.

A *valid* instrumental variable is a random vector of dimension $(k \times 1)$ such that

$$E[z_i u_i] = 0. \tag{6.1.1}$$

The instrumental variable is *relevant* if, in addition to (6.1.1), it holds that

$$E[z_i x_i'] \text{ has full rank } k. \tag{6.1.2}$$

Validity and relevance are sufficient to identify the parameter of interest β:

$$0 = E[z_i(y_i - x_i\beta)] = E[z_i y_i] - E[z_i x_i']\beta \implies \beta = E[z_i x_i']^{-1} E[z_i y_i].$$

The analogy principle suggests the following estimator

$$\hat{\beta}^{IV} = \left(\sum_{i=1}^{n} z_i x_i' \right)^{-1} \sum_{i=1}^{n} z_i y_i,$$

which is often referred to as the Instrumental Variable (IV) estimator. The asymptotic properties of $\hat{\beta}^{IV}$ can be studied along the lines of the OLS estimator.

Find a set of minimal assumptions under which:

1. $\hat{\beta}^{IV} = \beta + o_p(1)$;
   (a) $\sqrt{n}(\hat{\beta}^{IV} - \beta) \xrightarrow{d} N(0, V)$;
   (b) V can be consistently estimated.

The assumption on the rank $E[z_i x_i']$ looks very similar to the assumption on the rank of $E[x_i x_i']$ needed for identification of β when $E[x_i u_i] = 0$. There is however an important difference. The assumption on the rank of $E[x_i x_i']$ is a technical assumption: it fails only if the regressors are not chosen carefully. The assumption of the rank of $E[z_i x_i']$ is a substantive assumption and it involves the specification of the model. To get a better grasp of this assumption we analyze few specialization of the instrumental variables model.

6.2 Univariate IV model

Consider the simple univariate model

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{1i} + u_i, \\ &= x_i'\beta + u_i, \quad i = 1, \ldots, n \end{aligned}$$

where $\beta = (\beta_0, \beta_1)$, and $x_i = (1, x_{1i})$. While we are willing to assume that $E[u_i] = 0$, we are not willing to assume that $E[x_{1i} u_i] = 0$. What we are willing to assume is that there exists a random variable $\{w_i\}$ such that

$$E[w_i u_i] = 0.$$

The random vector $z_i = (1, w_i)$ is a valid instrument, since by construction

$$E[z_i u_i] = \begin{pmatrix} E[u_i] \\ E[w_i u_i] \end{pmatrix} = 0.$$

Is it a relevant instrument? Notice that in this particular case,

$$E[x_i z_i'] = E \begin{pmatrix} 1 \\ x_{1i} \end{pmatrix} \begin{pmatrix} 1 \\ z_{1i} \end{pmatrix}' = \begin{pmatrix} 1 & E[z_i] \\ E[x_{1i}] & E[w_i x_{ii}] \end{pmatrix}.$$

Since $E[z_i x_i']$ has full rank if its determinant is different from 0, we study the its determinant to find that

$$|E[z_i x_i']| = E[w_i x_{1i}] - E[x_{1i}]E[w_i] = cov(w_i, x_{1i}) \neq 0.$$

The instrumental variable z_i will be relevant if the covariance between w_i and x_{1i} is different from zero. In other word, w_i must be uncorrelated with u_i ($E[w_i u_i] = 0$) and must be correlated with the endogenous variable x_{1i} ($cov(w_i, x_{1i}) \neq 0$).

6.3 IV with one endogenous variable

Consider the general multivariate linear model

$$\begin{aligned} y_i &= \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots + \beta_k x_{ki} + u_i \\ &= x_i'\beta + u_i, \qquad i = 1, \ldots, n \end{aligned} \tag{6.3.1}$$

where $x_i = (1, x_{2i}, \ldots, x_{ki})$, and $\beta = (\beta_1, \beta_2, \ldots, \beta_k)$. Suppose that the following moment conditions are satisfied:

$$E[x_{j1} u_i] = 0, \quad i = 1, \ldots, k-1$$

but,

$$E(x_{ki} u_i) \neq 0.$$

In the applied parlance, x_{ki} is often referred to as "endogenous", while x_{ji}, $j = 1, \ldots, k-1$, are referred to as "exogenous" variables.

Suppose there exists a random variable $\{w_i\}$ that satisfies

$$E(w_i u_i) = 0.$$

Then, it holds that

$$E \begin{pmatrix} u_i \\ x_{2i} u_i \\ \vdots \\ x_{k-1,i} u_i \\ w_i u_i \end{pmatrix} = 0. \tag{6.3.2}$$

Let $z_i = (1, x_{2i}, \ldots, x_{k-1,i}, w_i)$. The moment conditions in (6.3.2) identify the parameter of interest:

$$0 = E[z_i' u_i] = E[z_i'(y_i - x_i\beta)] = E[z_i' y_i] - E[z_i' x_i]\beta,$$

which implies that

$$\beta = E(z_i'x_i)^{-1}E(z_i'y_i).$$

Recall that by identification we mean that we are able to express that parameter in terms of population moments of random variables whose realizations are observable.

The analogy principles suggests using the instrumental variable estimator:

$$\hat{\beta}^{IV} = \left(\sum_i^n z_i'x_i\right)^{-1} \sum_i^n z_i'y_i.$$

Let's look at the assumption on $\text{Rank}\left[E(z_ix_i')\right] = k$ more closely. We first make the following technical assumption

$$E[z_iz_i'] \text{ has full column rank.}$$

Let consider the linear projection of x_i on z_i

$$x_i = \Pi z_i + \eta_i, \quad \Pi = E(z_iz_i')^{-1}E(z_ix_i').$$

The dimension of Π is $(k \times k)$. Notice that if x_k is the only endogenous variable, we will have

$$\Pi = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & & & \vdots & 0 \\ \pi_{k1} & \pi_{k2} & \pi_{k3} & \cdots & \pi_{kk} \end{pmatrix}$$

Pre-multiplying z_i by $x_i' = \Pi' z_i + \eta_i'$ and taking expectation gives

$$E(z_ix_i') = E(z_iz_i')\Pi + E(z_i'\eta_i) = E(z_i'z_i)\Pi,$$

where the last equality follows from the fact that the error of a linear projection is orthogonal. This implies that

$$\begin{aligned} \text{Rank}\left[E(z_ix_i')\right] &= \min\left\{\text{Rank}\left[E(z_iz_i')\right], \text{Rank}\left[\Pi\right]\right\} \\ &= \min\left\{k, \text{Rank}\left[\Pi\right]\right\}. \end{aligned}$$

Then $\text{Rank}\left[E(z_i'x_i)\right] = k$ if and only if $\text{Rank}\left[\Pi\right] = k$. But $\text{Rank}\left[\Pi\right] = k$ if and only if $\pi_{kk} \neq 0$, that is, when the coefficient relative to w_i in the linear projection of x_{ki} on z_i.

6.3.1 Instrumental Variable Asymptotic

We discuss now the asymptotic behavior of the IV estimator. As usual, the first step is to notice that we can write

$$\hat{\beta}^{IV} = \left(\sum_{i}^{n} z_i' x_i\right)^{-1} \sum_{i}^{n} z_i' y_i = \beta + \left(\sum_{i}^{n} z_i' x_i\right)^{-1} \sum_{i=1}^{n} z_i' u_i.$$

Showing consistency of the IV estimator to β is tantamount to showing that

$$\left(\sum_{i}^{n} z_i' x_i\right)^{-1} \sum_{i=1}^{n} z_i' u_i$$

converges in probability to 0. If $E(z_i' u_i) = 0$, we have that

$$\left(\sum_{i}^{n} z_i' x_i\right)^{-1} \sum_{i=1}^{n} z_i' u_i = \left(\sum_{i}^{n} z_i' x_i / N\right)^{-1} \sum_{i=1}^{n} z_i' u_i / N,$$

and by the LLN and Assumption IV1,

$$\left(\sum_{i}^{n} z_i' x_i / N\right)^{-1} \xrightarrow{p} \left[E(z_i' x_i)\right]^{-1}$$

and

$$\sum_{i=1}^{n} z_i' u_i / N \xrightarrow{p} E(z_i' x_i) = 0.$$

Then using the continuous mapping theorem, it follows that

$$\left(\sum_{i}^{n} z_i' x_i / N\right)^{-1} \sum_{i=1}^{n} z_i' u_i / N \xrightarrow{p} \left[E(z_i' x_i)\right]^{-1} \times E(z_i' u_i) = 0$$

We summarize the consistency result below.

Suppose Assumption IV1 is satisfied and $E(z_i' u_i) = 0$. Then $\hat{\beta}^{IV} \xrightarrow{p} \beta$.

Notice that the consistency result is spelled out for the general IV estimator, regardless of the number of endogenous variables. We only requires that there are **exactly** as many instruments as variables to instruments. The same will be true for the normality result below.

The IV estimator is asymptotically normal if the following assumption is satisfied:

IV2

$$E(u_i^2 z_i' z_i) < \infty$$

This assumption is needed to apply the CLT to $\sum_{i=1}^n z_i' u_i / N$.

Suppose Assumption IV1 and IV2 are satisfied. If $E(z_i' u_i) = 0$, then

$$\sqrt{N}(\hat{\beta}^{IV} - \beta) \xrightarrow{p} N\left(0, \left[E(z_i' x_i)\right]^{-1} E(u_i^2 z_i' z_i) \left[E(x_i' z_i)\right]^{-1}\right).$$

The asymptotic variance of $\hat{\beta}^{IV}$ is thus given by

$$\text{aVar}(\hat{\beta}^{IV}) = \frac{\left[E(z_i' x_i)\right]^{-1} E(u_i^2 z_i' z_i) \left[E(x_i' z_i)\right]^{-1}}{N}.$$

This variance is consistently estimated by

$$\begin{aligned}\widehat{\text{aVar}(\hat{\beta}^{IV})} &= \frac{\left[\sum_i^n z_i' x_i / N\right]^{-1} \sum_{i=1}^n \hat{u}_i z_i' z_i / N \left(\sum_i^n x_i' z_i / N\right)^{-1}}{N} \\ &= \left[\sum_i^n z_i' x_i / N\right]^{-1} \sum_{i=1}^n \hat{u}_i z_i' z_i / N \left(\sum_i^n x_i' z_i / N\right)^{-1}\end{aligned}$$

where $\hat{u}_i = y_i - x_i \hat{\beta}^{IV}$. Incidentally, this is the variance matrix used to construct robust standard error in statistical packages.

An alternative variance is obtained if the following assumption is made:

$$E(u_i^2 z_i' z_i) = \sigma^2 E(z_i' z_i)$$

This is the IV equivalent of the *conditional homoskedastiticy* assumption. It holds if $E(u_i^2 | z_i) = \sigma^2$, that is the error is conditional (on z_i) homoskedastiticy. Under this assumption the asymptotic variance of the IV estimator simplifies to

$$\begin{aligned}\text{aVar}(\hat{\beta}^{IV}) &= \frac{\left[E(z_i' x_i)\right]^{-1} E(u_i^2 z_i' z_i) \left[E(x_i' z_i)\right]^{-1}}{N} \\ &= \frac{\left[E(z_i' x_i)\right]^{-1} \sigma^2 E(z_i' z_i) \left[E(x_i' z_i)\right]^{-1}}{N}\end{aligned}$$

6.4 Two Stage Least Squares

Consider again the basic linear model

$$y_i = x_i\beta + u_i, \quad x_i = (1, x_{2i}, x_{3i}, \ldots, x_{ki}). \tag{6.4.1}$$

As in the previous sections, we consider the case in which we are not comfortable with the key assumption for estimation of β by OLS. In particular, we believe that

$$E[x_i u_i] \neq 0. \tag{6.4.2}$$

Suppose now there is a random vector z_i of dimension $(\ell \times 1)$, $\ell \geq k$, such that

$$E[z_i u_i] = 0. \tag{6.4.3}$$

We are willing to make the following assumption:

TSLS1

$$\text{Rank}\left[E(z_i x_i')\right] = k, \quad \text{Rank}\left[E(z_i' z_i)\right] = \ell.$$

If $\ell = k$, we are back to the setup of the previous section. If $\ell > k$, then we are facing a new problem. In this last case, the conditions (6.4.3) is not enough to uniquely identify the parameter of interest β.

Indeed, if we proceed as usual we obtain that

$$0 = E[z_i(y_i - x_i'\beta)] = E[z_i y_i] - E[z_i x_i']\beta.$$

It might seem natural to obtain β by inverting $E[z_i x_i']$. But this is not possible! The dimension of $E[z_i x_i']$ is $\ell \times k$ and for $\ell > k$ this matrix is not a square matrix. In other word,

$$0 = E[z_i y_i] - E[z_i x_i']\beta \tag{6.4.4}$$

defines a overidentified system of liner equations: there are ℓ equations in k unknowns.

How do we proceed then? Let $\Lambda = Az$ denote a $(k \times 1)$ vector. Λ is a linear combination of instruments that effectively reduces the number of instruments, from ℓ to k. Multiplying both sides of (6.4.4), we obtain:

$$Az_i y_i = Az_i x_i'\beta + Az_i u_i.$$

Since $E[Az_i u_i] = AE[z_i u_i] = 0$, then if $AE[z_i x_i]$ is invertible

$$\beta = \left(AE[z_i x_i']\right)^{-1} AE[z_i y_i].$$

We have reduced the system to k equations in k unknowns by combining the $(\ell \times 1)$ of instruments into a new vector of dimension $(k \times 1)$. The analogy principle suggests the following estimator

$$\dot{\beta}_A = \left(A_n \sum_{i=1}^{n} z_i x_i' \right)^{-1} A_n \sum_{i=1}^{n} z_i y_i,$$

where $A_n \xrightarrow{p} A$. We have defined the estimator as $\dot{\beta}_A$ to stress the fact that it is a class of estimators, not a single estimator. In this class we want to choose the "best" member, that is, we want to choose a matrix A in such a way that the resulting estimator has an optimality property.

Consider the asymptotic distribution of $\ddot{\beta}_A$ given in the following proposition.

Suppose:

1. $\sqrt{n} \sum_{i=1}^{n} z_i x_i'/n \xrightarrow{d} N(0, \Omega := E[u_i^2 z_i z_i']) = 0$;

   (a) $\sum_{i=1}^{n} z_i x_i'/n \xrightarrow{p} S_{zx} := E[z_i x_i']$;

   (b) $A_n \xrightarrow{p} A$; has full column rank.

Then, for all $A \in \mathbb{R}^{k \times \ell}$ with full column rank, $\dot{\beta}_A = \beta + o_p(1)$ and $\sqrt{n}(\dot{\beta} - \beta) \xrightarrow{d} N(0, V_A)$, where

$$V_A = A S_{zx} \Omega S_{zx}' A'.$$

The "best" estimator is the one that corresponds with a matrix $A \in \mathbb{R}^{k \times \ell}$ that minimizes the asymptotic variance V_A. In turn, this matrix gives the best linear combination of instruments.

How do we find Az_i that gives the best linear combination? Recall what a linear projection of x onto z looks like:

$$x_i = \Pi z_i + \eta_i, \quad \Pi = E[x_i z_i'] E[z_i z_i']^{-1} = S_{xz} S_{zz}^{-1}.$$

The term Πz_i seems a good candidate as a linear combination of instruments. However, in order to construct the linear combination we need an estimator for Π. Again, by the analogy principle, we have

$$\Pi_n = \sum_{i=1}^{n} x_i z_i'/n \left(\sum_{i=1}^{n} z_i z_i'/n \right)^{-1} = \hat{S}_{xz} \hat{S}_{zz}^{-1}$$

Our estimator becomes

$$\dot{\beta}_{\Pi} = \left[\sum_{i=1}^{n} x_i z_i'/n \left(\sum_{i=1}^{n} z_i z_i'/n\right)^{-1} \sum_{i=1}^{n} z_i x_i'/n\right]^{-1} \left[\sum_{i=1}^{n} x_i z_i'/n \left(\sum_{i=1}^{n} z_i z_i'/n\right)^{-1} \sum_{i=1}^{n} z_i y_i/n\right]$$
$$= \left(\hat{S}_{xz}\hat{S}_{zz}^{-1}\hat{S}_{zx}\right)^{-1} \hat{S}_{xz}\hat{S}_{zz}^{-1}\hat{S}_{zy}. \tag{6.4.5}$$

The estimator in (**??**) is the so called two-stage least squares:

$$\hat{\beta}^{TSLS} = \left(\hat{S}_{xz}\hat{S}_{zz}^{-1}\hat{S}_{zx}\right)^{-1} \hat{S}_{xz}\hat{S}_{zz}^{-1}\hat{S}_{zy}.$$

At this point, you may have two questions: a) why is it called two-stage least squares; and b) is TSLS the best estimator in the sense of having the lowest symptotic variance among all the estimators of the type $\dot{\beta}_A$.

Let's answer part a) first. The TSLS estimator is **numerically equivalent** to a procedures that involves two least squares problems.

1. In the first step estimate the following linear model

   $$x_i = \Pi z_i + \eta_i,$$

   by least squares imposing $\sum_{i=1}^{n} z_i \eta_i'$. We obtain

   $$\hat{\Pi} = \sum_{i=1}^{n} x_i z_i \left(\sum_{i=1}^{n} z_i z_i\right)^{-1} = \hat{S}_{xz}\hat{S}_{zz}^{-1};$$

2. Form

   $$\hat{x}_i = \hat{\Pi} z_i,$$

   and estimate the linear model

   $$y_i = \hat{x}_i \beta + \epsilon_i,$$

   by imposing $\sum_{i=1}^{n} \hat{x}_i \epsilon_i = 0$, to obtain

   $$\hat{\beta} = \left(\sum_{i=1}^{n} \hat{x}_i \hat{x}_i'\right) \sum_{i=1}^{n} \hat{x}_i y_i.$$

Now,

$$\hat{\beta} = \left(\sum_{i=1}^{n} \hat{x}_i \hat{x}_i'\right) \sum_{i=1}^{n} \hat{x}_i y_i \tag{6.4.6}$$
$$= \left(\sum_{i=1}^{n} \hat{\Pi} z_i z_i' \hat{\Pi}'\right) \sum_{i=1}^{n} \hat{\Pi} z_i y_i \tag{6.4.7}$$
$$= \left(S_{xz}S_{zz}^{-1}S_{zz}S_{zz}^{-1}S_{zx}\right)^{-1} S_{xz}S_{zz}^{-1}S_{zy} \tag{6.4.8}$$
$$= \left(S_{xz}S_{zz}^{-1}S_{zx}\right)^{-1} S_{xz}S_{zz}^{-1}S_{zy} \tag{6.4.9}$$

that is the two stage least squares.

in the following special case model:

$$\begin{aligned} y_i &= \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots + \beta_k x_{ki} + u_i \\ &= x_i'\beta + u_i, \qquad i = 1, \ldots, n \end{aligned} \tag{6.4.10}$$

where $x_i = (1, x_{2i}, \ldots, x_{ki})$, and $\beta = (\beta_1, \beta_2, \ldots, \beta_k)$. Suppose that the following moment conditions are satisfied:

$$E[x_{j1} u_i] = 0, \quad i = 1, \ldots, k-1$$

but,

$$E(x_{ki} u_i) \neq 0.$$

If there are $(w_{1i}, w_{2i}, \ldots, w_{mi})$ variables such that

$$E[w_{ji} u_i] = 0, \quad j = 1, \ldots, m,$$

then

$$z_i = (1, x_{2i}, x_{3i}, \ldots, x_{k-1,i}, w_{1i}, w_{2i}, \ldots, w_{mi})$$

is a valid instrument of dimension $\ell = (k-1) + m$, with $\ell > k$ for $m > 1$.

In this special case the linear projection takes the form

$$x_i = \Pi z_i + \eta_i,$$

with

$$\Pi = \begin{pmatrix} I_{k-1} & 0_{1\times m} \\ \pi_{k1} \ldots \pi_{k,k-1} & \pi_{k,k} \ldots \pi_{k,m} \end{pmatrix}.$$

Thus,

$$\Pi z_i = \begin{pmatrix} 1 \\ x_{2i} \\ x_{3i} \\ \vdots \\ x_{k-1,i} \\ \pi_{k1} + \pi_{k2} x_{2i} + \ldots + \pi_{kk} w_{1i} + \ldots + \pi_{k\ell} w_{mi} \end{pmatrix}.$$

Also,

$$\hat{\Pi} = \begin{pmatrix} 1 \\ x_{2i} \\ x_{3i} \\ \vdots \\ x_{k-1,i} \\ \hat{\pi}_{k1} + \hat{\pi}_{k2} x_{2i} + \ldots + \hat{\pi}_{kk} w_{1i} + \ldots + \hat{\pi}_{k\ell} w_{mi} \end{pmatrix}.$$

and $\hat{\pi}_{kj}$, $j = 1, \ldots, \ell$ are obtained by running the following OLS regression:

$$\begin{aligned} x_{k,i} &= \pi z_i + u_i \\ &= \pi_{k1} + \pi_{k2} x_{2i} + \ldots + \pi_{kk} w_{1i} + \ldots + \pi_{k\ell} w_{mi} + r_i. \end{aligned}$$

In the second stage we regress

$$y \text{ on } 1, x_{2i}, x_{3i}, \ldots, x_{k-1,i}, \hat{\pi}_{k1} + \hat{\pi}_{k2} x_{2i} + \ldots + \hat{\pi}_{kk} w_{1i} + \ldots + \hat{\pi}_{k\ell} w_{mi}.$$

The first is called the first stage regression,

$$x_i = \Pi z_i + \eta_i,$$

the projection of $(1, x_2, x_3, \ldots, x_k)$ on $(1, x_1, x_2, \ldots, x_{k-1}, w_1, \ldots, w_s)$. We could be temped to say that the second stage consists in using $\hat{x}_i^*$ in place of x_i^*. But what we are looking for is a second stage that is also a regression.

Let

$$Z = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_N \end{pmatrix}, \quad X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix}, \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$$

Now, $\hat{x}_i^* = z_i \hat{\Pi} = z_i (Z'Z)^{-1} Z'X$ can be collapsed in matrix form as in

$$\hat{X}^* = \begin{pmatrix} \hat{x}_1^* \\ \hat{x}_2^* \\ \vdots \\ \hat{x}_N^* \end{pmatrix} = Z(Z'Z)^{-1} Z'X = P_Z X.$$

The TSLS estimator of equation (**??**) can be rewritten in matrix form as

$$\hat{\beta}^{TSLS} = (\hat{X}^{*'} X)^{-1} (\hat{X}^{*'} Y)$$

Notice that P_Z is symmetric and idempotent, $P_Z P_Z = P_Z$. Simple manipulations give that

$$\hat{\beta}^{TSLS} = (\hat{X}^{*'} X)^{-1} \hat{X}^{*'} Y = (X' P_z P_z X)^{-1} (X' P_z Y) = (\hat{X}^{*'} \hat{X}^*)^{-1} \hat{X}^{*'} Y,$$

But this expression is exactly the OLS estimator of the following regression

$$y_i = \hat{x}_i^* \beta + \epsilon_i.$$

Notice that the TSLS estimator can be written back in summation-form

$$\hat{\beta}^{TSLS} = \left(\sum_{i=1}^{n} \hat{x}_i^{*'} \hat{x}_i^* \right)^{-1} \sum_{i=1}^{n} \hat{x}_i^{*'} y_i.$$

Substituting $\hat{x}^* = z_i\hat{\Pi} = z_i\left(\sum_{i=1}^n z_i'z_i\right)^{-1} z_i'x_i$ we obtain an explicit expression for the TSLS estimator

$$\hat{\beta}^{TSLS} = \left[\sum_i^n x_i'z_i\left(\sum_i^n z_i'z_i\right)^{-1}\sum_i^n z_i'x_i\right]^{-1}\sum_i^n x_i'z_i\left(\sum_i^n z_i'z_i\right)^{-1}\sum_i^n z_i'y_i.$$

The assumption that only x_k is "endogenous" is for convenience only. It only buys as a simple way to check the assumption $E(z'x) = k+1$ is satisfied. Indeed, if x_k is the only endogenous regressor in (6.4.1) and $E[z_i'z_i]$ has full rank, $\text{Rank}\,[E(z'x)] = k+1$ **if and only if** at least on of $\theta_1,\ldots,\theta_s$ in the linear projection of x_{ik} on z_i, i.e.

$$x_{ik} = \delta + \delta_1 x_{i1} + \ldots \delta_{(k-1)}x_{i(k-1)} + \theta_1 w_{i1} + \ldots \theta_s w_{is} + \eta,$$

is different from 0. Of course we do not know $\theta_1,\ldots,\theta_s$, but we can estimate them consistently by using OLS. The conditions on the rank can be tested by using a F-test

$$H_0 : \theta_1 = \ldots = \theta_s = 0, \quad H_1 : \theta_j \neq 0 \text{ for at least one } j = 1,\ldots,s$$

When more endogenous variables are present, the rank condition can be tested using the test by Cragg and Donald (1997). However, the costume in applied economics is to run as many as first stages as there are endogenous variables and test that at least one of the instruments is different from zero in each regression.

6.5 TSLS Asymptotic

We can write the two-stage least square estimator as

$$\begin{aligned}\hat{\beta} &= \left[\sum_i^n x_i'z_i\left(\sum_i^n z_i'z_i\right)^{-1}\sum_i^n z_i'x_i\right]^{-1}\sum_i^n x_i'z_i\left(\sum_i^n z_i'z_i\right)^{-1}\sum_i^n z_i'y_i \\ &= \beta + \left[\sum_i^n x_i'z_i\left(\sum_i^n z_i'z_i\right)^{-1}\sum_i^n z_i'x_i\right]^{-1}\sum_i^n x_i'z_i\left(\sum_i^n z_i'z_i\right)^{-1}\sum_{i=1}^n z_iu_i\end{aligned}$$

By applying the Law of Large Numbers to each sums comprising the TSLS estima-

tor we easily find that

$$\sum_i^n x_i' z_i / N \xrightarrow{p} E(x_i' z_i)$$
$$\sum_i^n z_i' z_i / N \xrightarrow{p} E(z_i' z_i)$$
$$\sum_{i=1}^n z_i u_i / N \xrightarrow{p} E(z_i u_i) = 0.$$

By continuity, we obtain that

$$\left[\sum_i^n x_i' z_i / N \left(\sum_i^n z_i' z_i / N\right)^{-1} \sum_i^n z_i' x_i / N\right]^{-1} \sum_i^n x_i' z_i / N \left(\sum_i^n z_i' z_i / N\right)^{-1}$$
$$\xrightarrow{p} A \stackrel{def}{=} \left\{E(x_i' z_i)\left[E(z_i' z_i)\right]^{-1} E(z_i' x_i)\right\}^{-1} E(z_i' x_i)\left[E(z_i' z_i)\right]^{-1}. \quad (6.5.1)$$

Hence,

$$\left[\sum_i^n x_i' z_i \left(\sum_i^n z_i' z_i\right)^{-1} \sum_i^n z_i' x_i\right]^{-1} \sum_i^n x_i' z_i \left(\sum_i^n z_i' z_i\right)^{-1} \sum_{i=1}^n z_i u_i \xrightarrow{p} 0,$$

which implies that $\hat{\beta}^{TSLS}$ is consistent.

6.5.1 Asymptotic Normality

We make this assumption:

[TSLS2] $Var(z_i' u_i) = E\left(e_i^2 x_i x_i'\right) < \infty$

From the Central Limit Theorem it follows that

$$\sqrt{N} \sum_{i=1}^n z_i u_i / N \xrightarrow{d} N(0, E\left(u_i^2 z_i' z_i\right)).$$

Notice that we can write $\hat{\beta} - \beta_0 = A_N \sum_{i=1}^n z_i u_i / N$ where

$$A_N \stackrel{def}{=} \left[\sum_i^n x_i' z_i / N \left(\sum_i^n z_i' z_i / N\right)^{-1} \sum_i^n z_i' x_i / N\right]^{-1} \sum_i^n x_i' z_i / N \left(\sum_i^n z_i' z_i / N\right)^{-1}.$$

We know that $A_N \xrightarrow{p} A$, where

$$A = \left[E(x_i'z_i)E(z_i'z_i)^{-1}E(z_i'x_i)\right]^{-1} E(x_i'z_i)E(z_i'z_i)^{-1}.$$

But then

$$\sqrt{N}(\hat{\beta} - \beta_0) = A_N\sqrt{N}\sum_{i=1}^{n} z_i'u_i/N,$$

which implies that

$$\sqrt{N}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, AE(u_i^2 z_i'z_i)A').$$

Summarizing:

Suppose Assumptions TSLS1 and TSLS2 are satisfied. Then, if $E(z_i'u_i) = 0$,

$$\sqrt{N}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, AE(u_i^2 z_i'z_i)A'),$$

where $A = \left[E(x_i'z_i)E(z_i'z_i)^{-1}E(z_i'x_i)\right]^{-1} E(x_i'z_i)E(z_i'z_i)^{-1}$.

If we assume

$E\left(u_i^2 z_i'z_i\right) = \sigma^2 E[z_i'z_i]$

The variance matrix simplifies to

$$AE\left(u_i^2 z_i'z_i\right)A = \sigma^2\left[E(x_i'z_i)E(z_i'z_i)^{-1}E(z_i'x_i)\right]^{-1}$$

The variance covariance matrix can be estimated by the analogy principle: $AE\left(\widehat{u_i^2 z_i'z_i}\right)A' = \hat{A}\hat{E}\left(u_i^2 z_i'z_i\right)\hat{A}'$.

Chapter 7

Generalized Method of Moments

7.1 Introduction

The generalized method of moments (GMM) is a very flexible econometric technique to estimate parameter of interest when the model is expressed in terms of moment conditions. We have already studied econometric models defined in terms of moment conditions. The regression model

$$y = x\beta + u, \quad E(u|x) = 0$$

is estimated by OLS by using the following moment conditions

$$E(x'u) = E[x'(y - x\beta)] = 0$$

Given a sample of observations (y_i, x_i), $i = 1, \ldots, N$, the OLS estimator solves the following "empirical" moment conditions:

$$\sum_{i=1}^{n} x_i^{'}(y_i - x_i\beta) = 0$$

Note that $\sum_{i=1}^{n} x_i'(y_i - x_i\beta) = 0$ describes a K equations in K unknowns:

$$\sum_{i=1}^{n} x_{1i}^{'}(y_i - x_i\beta) = 0$$

$$\sum_{i=1}^{n} x_{2i}^{'}(y_i - x_i\beta) = 0$$

$$\vdots$$

$$\sum_{i=1}^{n} x_{Ki}^{'}(y_i - x_i\beta) = 0$$

The OLS is the value of β that solves the above K equations. If $\sum_{i=1}^{n} x_i' X_i$ is invertible,

$$\hat{\beta}^{OLS} = \left\{ \sum_{i}^{n} x_i x_i' \right\}^{-1} \sum_{i=1}^{n} x_i' y_i$$

If x_k is endogenous, and we have a random variable z such that $E(z_i u_i) = 0$, a consistent estimators of β can be obtained by the IV estimator. IV solves the following empirical moment conditions

$$\begin{aligned}
&\sum_{i=1}^{n} x_{1i}^{'}(y_i - x_i\beta) = 0 \\
&\sum_{i=1}^{n} x_{2i}^{'}(y_i - x_i\beta) = 0 \\
&\vdots \\
&\sum_{i=1}^{n} z_i^{'}(y_i - x_i\beta) = 0
\end{aligned}$$

Once again, there are K equations in K unknowns. If Rank$E(x_i' z_i) = k$, the IV estimator solves the empirical moment conditions

$$\hat{\beta}^{IV} = \left\{ \sum_{i}^{n} x_i' z_i \right\}^{-1} \sum_{i=1}^{n} z_i' y_i$$

Consider now the situation in which x_k is endogenous, but now there are many random variables that can act as instruments

$$E(z_j u) = 0, \quad j = 1, \ldots, L$$

The empirical moment conditions are in this case

$$\begin{aligned}
&\sum_{i=1}^{n} x_{1i}^{'}(y_i - x_i\beta) = 0 \\
&\sum_{i=1}^{n} x_{2i}^{'}(y_i - x_i\beta) = 0 \\
&\vdots \\
&\sum_{i=1}^{n} z_{1i}^{'}(y_i - x_i\beta) = 0 \qquad\qquad \vdots \\
&\sum_{i=1}^{n} z_{Ki}^{'}(y_i - x_i\beta) = 0
\end{aligned}$$

There $K-L+1$ equations and "only" K parameters to estimate. Indeed, the two stage least square does not solve the empirical moment conditions. It would be impossible, the moment conditions do not have a solutions. The TSLS (linear-)projects the endogenous variable onto the exogenous variables and the set of instruments. These procedure can be thought as dimension restrictions.

In general, the moment conditions take the following general form

$$E[g(w,\theta_0)] = 0, \quad \|E[g(w,\theta)]\| \neq 0, \text{ for any } \theta \neq \theta_0 \tag{7.1.1}$$

where $g(w,\theta)$ is a vector of functions of dimension M

$$g(w,\theta) = \begin{pmatrix} g_1(w,\theta) \\ g_2(w,\theta) \\ \vdots \\ g_M(w,\theta) \end{pmatrix}$$

w is a random vector and θ_0 is a $K \times 1$ parameter vector. Notice that $\|E[g(w,\theta)]\| \neq 0$, for any $\theta \neq \theta_0$ means that the moment conditions hold only at θ_0. For any other value of θ, at least one element of the function vector is different from 0. (For a vector x, $\|x\| = \sqrt{\sum_j x_j^2}$).

For example, for the OLS case, $w = (y,x)$, $\theta_0 = \beta$ and $g(w,\theta_0) = x'(y - x\beta)$. For the IV, $w = (y,x,z_1)$, $\theta_0 = \beta$ and $g(w,\theta_0) = z'(y - x\beta)$, $z = (1x_1 \ldots x_{K-1}z_1)$. For the TSLS, $w = (y,x,z_1,\ldots,z_L)$, $\theta_0 = \beta$ and $g(w,\theta_0) = z'(y - x\beta)$, $z = (1x_1 \ldots x_{K-1}z_1,\ldots,z_L)$.

If we have observations from w, say $(w_1,\ldots,w_N)$, then the empirical moment restrictions corresponding to (7.1.1) is

$$\sum_{i=1}^{n} g(w_i,\theta) = 0.$$

If $M > K$ there are more equations than there are parameters to estimate and a solutions does not generally exist.

Consider the following moment conditions

$$\begin{aligned} E[w - \theta_0] &= 0 \\ E[(w - \theta_0)^2 - 1] &= 0 \end{aligned}$$

Here, $g_1(w,\theta) = w - \theta$ and $g_2(w,\theta) = (w - \theta) - 1$, $g(w,\theta) = (g_1(w,\theta)g_2(w,\theta))$. Given a sample of size N, $(w_1,\ldots,w_N)$, the empirical moment conditions are

$$\sum_{i=1}^{n} w_i - \theta = 0, \quad \sum_{i=1}^{n} (w - \theta)^2 - 1 = 0$$

It is easy to see that in general there is not θ that solves both equations simultaneously. From the first equation we obtain that $\theta = \sum_{i=1}^{n} w_i$. But $\sum_{i=1}^{n} w_i$ does not solve the second equations (if not by luck). The problem is that there are 2 equations but only a parameter to estimate θ.

7.2 Asymptotic Theory

Hansen (1982) proposes to estimate θ by minimizing the distance between the empirical moment conditions and 0. The idea is that given a small sample the empirical moment conditions cannot be set equal to 0, but they can be made "small". The GMM estimator is the solution to the following optimization problem:

$$\min_{\theta \in \Theta} J_N(\theta; W), \quad J_N(\theta; W) = \left(\sum_{i=1}^{n} g(w_i, \theta) \right)' W \left(\sum_{i=1}^{n} g(w_i, \theta) \right)$$

W is a $M \times M$ positive definite matrix. Broadly speaking, $J_N(\theta; W)$ is a distance measuring the discrepancy between $\sum_{i=1}^{n} g(w_i, \theta)$ and its asymptotic value 0. $J_N(\theta; W)$ will be equal to zero when $\sum_{i=1}^{n} g(w_i, \theta) = 0$ and will be large when $\sum_{i=1}^{n} g(w_i, \theta)$ is large.

Proving consistency of $\hat{\beta}^{GMM}$ for the general nonlinear case is more difficult than proving consistency when $g(w, \theta)$ is linear in θ. But the intuition is rather simple. Suppose that $\sum_{i=1}^{n} g(w_i, \theta)/N$ obeys the uniform law of large numbers. That is a stronger requirement that $\sum_{i=1}^{n} g(w_i, \theta)/N$ obeys the WLLN for every $\theta \in \Theta$. Then the random function $J_N(\theta)$ converges uniformly to

$$J(\theta; W) = E[g(w, \theta)]' W E[g(w, \theta)]. \tag{7.2.1}$$

Because W is positive definite, θ_0 uniquely minimizes (7.2.1). Intuitively, the minimizer of

$$\arg \min_{\theta \in \Theta} J_N(\theta; W)$$

should converge to the minimizer of (7.2.1), θ_0.

Under regularity conditions:

$$\hat{\theta}(W) = \arg \min_{\theta \in \Theta} J_N(\theta; W) \xrightarrow{p} \theta_0$$

exists, and $\hat{\theta}(W) \xrightarrow{p} \theta_0$.

We can also derive the asymptotic distributions of $\hat{\theta}$.

Under regularity conditions:

$$\sqrt{N}(\hat{\theta}(W) - \theta_0) \xrightarrow{d} N(0, V)$$

where

$$\begin{aligned}
V(W) &= (\Gamma' W \Gamma)^{-1}(\Gamma' W S W \Gamma)(\Gamma' W \Gamma)^{-1} \\
\Gamma &= E\left[\frac{\partial g(w, \theta_0)}{\partial \theta}\right] \\
S &= \operatorname{Var}\left(\frac{1}{\sqrt{N}} \sum_{i=1}^{n} g(w_i, \theta_0)\right).
\end{aligned}$$

One of the theoretical advantages of GMM is that consistency and asymptotic normality hold for more general sampling scheme than iid. If the data are iid, then

$$\begin{aligned}
S = \operatorname{Var}\left(\frac{1}{\sqrt{N}} \sum_{i=1}^{n} g(w_i, \theta_0)\right) &= \frac{1}{N} \sum_{i=1}^{n} \operatorname{Var}\left(g(w_i, \theta_0)\right) \\
&= E\left[g(w_i, \theta_0) g(w_i, \theta_0)'\right]
\end{aligned}$$

For the time being we assume that $S = E\left[g(w_i, \theta_0) g(w_i, \theta_0)'\right]$, but keep in mind that with time series data the matrix S will not in general coincide with the expectation of the outer product of the moment functions.

Notice that we have indexed both the estimator and the asymptotic variance to stress the fact that the GMM procedures defines a class of consistent and asymptotically normal estimator.

The only requirement we have imposed on W is that it is a positive definite matrix. Two researchers with the same dataset, same model, but two different W will end up with different estimates of θ_0. Nevertheless, the two estimators are both consistent and asymptotically normal. The next questions is: is there an "optimal" way to pick W? Clearly the answer to this question depends on what we mean by "optimal". A standard approach (justified by theoretical considerations) is to pick a W in such a way to minimize the asymptotic variance of $\hat{\theta}$.

It turns out the optimal W, is given by S^{-1}, the inverse of the variance of the empirical moment conditions. When $W = S^{-1}$, the variance of $\hat{\theta}(S^{-1}) \overset{def}{=} \hat{\theta}$, $V \overset{def}{=} V(S^{-1})$, reduces to

$$\begin{aligned}
V &= (\Gamma' W \Gamma)^{-1}(\Gamma' W S W \Gamma)(\Gamma' W \Gamma)^{-1} \\
&= (\Gamma' S^{-1} \Gamma)^{-1}(\Gamma' S^{-1} S S^{-1} \Gamma)(\Gamma' S^{-1} \Gamma)^{-1} \\
&= (\Gamma' S^{-1} \Gamma)^{-1}
\end{aligned}$$

To show that $W = S^{-1}$ is optimal we need to show that

$$(\Gamma' W \Gamma)^{-1}(\Gamma' W S W \Gamma)(\Gamma' W \Gamma)^{-1} - (\Gamma' S^{-1} \Gamma)^{-1}$$

is positive semi-definite. It is actually easier to show that

$$(\Gamma' S^{-1} \Gamma) - (\Gamma' W \Gamma)(\Gamma' W S W \Gamma)^{-1}(\Gamma' W \Gamma).$$

is positive semi-definite. We can rewrite the last expression as CPC', where

$$C = \Gamma' S^{-1/2}; \quad P = [I_M - S^{1/2} W \Gamma \left(\Gamma W S^{1/2} S^{1/2} W \Gamma\right)^{-1} \Gamma' W S^{1/2}]$$

It is easy to show that P is a symmetric, idempotent matrix. It follows that from every $x \in \mathbb{R}^K$, $x'CPC'x = x'CPPC'x = \sum^K d_i^2 \geq 0$.

The intuition of why S^{-1} is the optimal weighting matrix is rather simple. Broadly speaking, W weights the contribution of each moment conditions in pinning down the parameter estimates. Using S^{-1} guaranties that we are weighting the moments by the inverse of their precision measured by their variance.

The optimal GMM solves the following optimization problem

$$\min_{\theta \in \Theta} \left(\sum_{i=1}^n g(w_i, \theta) \right)' S^{-1} \left(\sum_{i=1}^n g(w_i, \theta) \right)$$

The variance of the empirical moment conditions is generally unknown and, in practice, the optimal GMM is infeasible. However, we can apply the analogy principle and replace S^{-1} with a consistent estimate:

$$\left[\sum_{i=1}^n g(w_i, \theta_0) g(w_i, \theta_0)' \right]^{-1}.$$

The above expression depends on the unknown θ_0. We can substitute θ_0 with *any* consistent estimator of θ_0. To obtain such consistent estimator, we can use a non-optimal GMM procedures. A common choice is:

$$\tilde{\theta} = \arg\min_{\theta \in \Theta} \left(\sum_{i=1}^n g(w_i, \theta) \right)' I_M \left(\sum_{i=1}^n g(w_i, \theta) \right)$$

where I_M denotes the $M \times M$ identity matrix. Other choices can be also considered (see the instrumental case below). Consider the following estimator

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \left(\sum_{i=1}^n g(w_i, \theta) \right)' \hat{S}^{-1} \left(\sum_{i=1}^n g(w_i, \theta) \right) \tag{7.2.2}$$

where

$$\hat{S} = \left(\sum_{i=1}^{n} g(w_i, \tilde{\theta}) g(w_i, \tilde{\theta})' \right).$$

The asymptotic variance of the GMM estimator in (7.2.2) is

$$(\Gamma' S^{-1} \Gamma)$$

This result says that replacing the optimal weighting matrix $\{E[g(w, \theta_0)g(w, \theta_0)]\}^{-1}$ with $\hat{S}^{-1}$ does not change the asymptotic variance of the estimator. The intuition is simple. It can be shown that

$$\{E[g(w, \theta_0)g(w, \theta_0)]\}^{-1} \xrightarrow{p} \hat{S}^{-1}$$

and that the difference between the truth and the estimator does not affect the distribution of the optimal GMM estimator.

7.2.1 Estimation of V

The variance of the optimal GMM estimator can be easily estimated. Let

$$\hat{\Gamma} = \sum_{i=1}^{n} \frac{\partial g(w_i, \hat{\theta})}{\partial \theta}$$

. Then the asymptotic variance of $\hat{\theta}$ is consistently estimated by

$$\hat{\Sigma} \overset{def}{=} \left(\hat{\Gamma}' \hat{S}^{-1} \hat{\Gamma} \right)^{-1} / N$$

Where $\hat{S}^{-1} = \sum_{i=1}^{n} g(w_i, \hat{\theta}) g(w_i, \hat{\theta})' / N$, that is, $\hat{S}^{-1}$ is now calculated with the efficient GMM estimator.

The asymptotic standard error of the j-th element of $\hat{\theta}$ are thus given by

$$\sigma(\hat{\theta}_j) \overset{def}{=} \sqrt{\hat{\Sigma}_{jj}}$$

Similarly, the variance of the non-optimal GMM can be estimated by

$$\hat{\Sigma}(W) \overset{def}{=} \left(\hat{\Gamma}' W \hat{\Gamma} \right)^{-1} \left(\hat{\Gamma}' W \hat{S} W \hat{\Gamma} \right) \left(\hat{\Gamma}' W \hat{\Gamma} \right)^{-1} / N$$

and, similarly, the standard error of the non-optimal GMM

$$\sigma_W(\hat{\theta}_j(W)) \overset{def}{=} \sqrt{\hat{\Sigma}_{jj}(W)}$$

7.3 GMM based testing

GMM provides with a very flexible for testing hypothesis on the parameter.

Suppose we want to test that the j-th element of θ_0 is equal to some value $\bar{\theta}_j$. Formally, the null hypothesis can be stated as

$$H_0 : \theta_{0,\ell} = \bar{\theta}_j.$$

Usually we are interested in testing $\theta_{0,\ell} = 0$, but this does not need to be the case and $\bar{\theta}_j$ can be any real number. Under the null hypothesis,

$$t_j \overset{def}{=} \frac{\hat{\theta}_j - \bar{\theta}_j}{\sqrt{\Sigma_{\hat{\beta},jj}}} \xrightarrow{d} N(0,1)$$

For the null hypothesis

$$H_0 : h(\theta_0) = 0$$

where $h : \mathbb{R}^K \to \mathbb{R}^R$, and $H(\theta_0)$ is the $R \times K$ matrix of first derivatives of h, is continuous and of full rank,

$$\mathcal{W} = h(\hat{\theta}(W))' \left(H(\hat{\theta}(W)) \hat{\Sigma}(W) H(\hat{\theta}(W))' \right)^{-1} h(\hat{\theta}(W)) \xrightarrow{d} \chi^2_R$$

An other approach is to directly use the GMM criterion function. This is sometimes called the GMM Distance statistic, and sometimes called a LR-like statistic (the LR is for likelihood-ratio). The idea was first put forward by Newey and West (1987). The GMM estimator is

$$\hat{\beta}(W) = \arg\min_{\theta \in \Theta} J_N(\theta; W)$$

Under the null hypothesis

$$\bar{\beta}(W) = \arg\min_{h(\theta)=0} J_N(\theta; W)$$

The two minimizing criterion functions are $J_N(\hat{\theta}; W)$ and $J_N(\bar{\theta}; W)$. The GMM distance statistics is

$$D = N \cdot J_N(\bar{\theta}) - J_N(\hat{\theta})$$

If the same weighting matrix is used for both the null and the alternative, then (i) $D \geq 0$; (ii) $D \xrightarrow{d} \chi^2_R$; and (iii) if h is linear, then D equals the Wald statistics.

If h is non-linear, the Wald statistic can work quite poorly. In contrast, current evidence suggests that the D statistic appears to have quite good sampling properties, and is the preferred test statistic. Newey and West (1987) suggested to use

the same weight matrix for both null and alternative, as this ensures that $D \geq 0$. This reasoning is not compelling, however, and some current research suggests that this restriction is not necessary for good performance of the test. The advantage of $\mathcal{D}$ over $\mathcal{W}$ is invariance; the numerical value of $\mathcal{D}$ does not depends on how the nonlinear restrictions are represented by h. On the other hand, you have to write a nonlinear optimization computer program to find the restricted efficient GMM when the hypothesis is non-linear.

7.3.1 Overidentified Restrictions

If the equation (7.1.1) is exactly identified, then it is possible to chose θ so that all the elements of the empirical moments are zero and the distance is zero, that is $J(\theta, W) = 0$. If the equation is overidentified, then the distance cannot be set to zero exactly. It turns out that, if the weighting matrix is chosen optimally, then the minimized distance is asymptotically chi-squared:

$$J(\hat{\theta}) \overset{def}{=} \min_{\theta \in \Theta} N\Big(\sum_{i=1}^{n} g(w_i, \theta)\Big)' \hat{S}^{-1} \Big(\sum_{i=1}^{n} g(w_i, \theta)\Big)' \xrightarrow{d} \chi^2_{M-K}$$

The so-called J statistics provides us with a specification test, testing whether all the restrictions of the model (that is, (7.1.1)) are satisfied. A large value of $J(\hat{\theta})$ will lead to reject the null hypothesis that $E[g(w, \theta_0)] = 0$.

7.4 The linear case

In this section we show that for an important special case—linear moment conditions with conditionally homoscedastic error—the GMM estimator is the *TSLS*. It is also shown that when the error are conditionally heteroscedastic, the *TSLS* is not efficient.

Let consider the model

$$y = x\beta + u$$

with the usual notation convention that x is a $(K+1) \times 1$ vector, $x = (1, x_1, \ldots, x_K)$. Suppose

$$E[z'(y - x\beta)] = 0$$

where $z = (1, x_1, \ldots, x_{K-1}, z_1, \ldots, z_L)$ is a $(K + L)$ vector. The GMM estimator solves

$$\min_{\beta} \sum_{i=1}^{n} z_i(y_i - x_i\beta) W \sum_{i=1}^{n} z_i'(y_i - x_i\beta)$$

The first order conditions are:

$$\left(\sum_i^n x_i' z_i\right) W \sum_{i=1}^n z_i'(y_i - x_i\beta) = 0$$

Rearranging,

$$\left(\sum_i^n x_i' z_i\right) W \sum_i^n z_i' y_i = \left(\sum_i^n x_i' z_i\right) W \sum_i^n z_i' x_i \beta$$

If $\left(\sum_i^n x_i' z_i\right) W \left(\sum_i^n z_i' x_i\right)$ is invertible

$$\hat{\beta} = \left\{\left(\sum_i^n x_i' z_i\right) W \left(\sum_i^n z_i' x_i\right)\right\}^{-1} \sum_i^n x_i' z_i W \sum_i^n z_i' y_i$$

The optimal GMM is obtained for $W = E[u^2 z' z]$.

7.4.1 Conditional homoscedasticity

If $E(u^2 z' z) = \sigma^2 E(z' z)$, then the optimal GMM is obtained by setting $W = \left(\sum_i^n z_i' z_i\right)^{-1}$:

$$\hat{\beta} = \left\{\left(\sum_i^n x_i' z_i\right)\left(\sum_i^n z_i' z_i\right)^{-1}\left(\sum_i^n z_i' x_i\right)\right\}^{-1} \times \sum_i^n x_i' z_i \left(\sum_i^n z_i' z_i\right)^{-1} \sum_i^n z_i' y_i$$

It is easy to see that in this case $\hat{\beta} = \hat{\beta}^{TSLS}$.

7.4.2 Heteroscedasticity

If $E(u^2 z' z) \neq \sigma^2 E(z' z)$, then the optimal GMM is different from the $TSLS$. The optimal GMM is obtained by setting W to a feasible estimator of $E(u^2 z' z)$. If $\tilde{\beta}$ is a consistent estimator of β,

$$W = \left(\sum_{i=1}^n \hat{u}_i^2 z_i' z_i\right)^{-1}, \quad \hat{u}_i = y_i - x_i\tilde{\beta}$$

$$\hat{\beta} = \left\{ \left(\sum_i^n x_i' z_i \right) \left(\sum_{i=1}^n \hat{u}_i^2 z_i' z_i \right)^{-1} \left(\sum_i^n z_i' x_i \right) \right\}^{-1} \times \sum_i^n x_i' z_i \left(\sum_{i=1}^n \hat{u}_i^2 z_i' z_i \right)^{-1} \sum_i^n z_i' y_i$$

The preliminary estimator of β, it is usually obtained in one of the following ways:

(1) Using the identity matrix:

$$\tilde{\beta} = \arg\min_\beta \left\{ \left(\sum_i^n x_i' z_i \right) \left(\sum_i^n z_i' x_i \right) \right\}^{-1} \sum_i^n x_i' z_i \sum_i^n z_i' y_i$$

(2) Using the homoscedastic optimal weighting matrix

$$\tilde{\beta} = \arg\min_\beta \left\{ \left(\sum_i^n x_i' z_i \right) \left(\sum_i^n z_i' z_i \right)^{-1} \left(\sum_i^n z_i' x_i \right) \right\}^{-1} \left(\sum_i^n x_i' z_i \right) \left(\sum_i^n z_i' z_i \right)^{-1} \sum_i^n z_i' y_i$$

So, while the TSLS is efficient under conditional homoscedasticity, the GMM estimator is efficient regardless of whether the variance of the empirical moment conditions is proportional to $E(z'z)$.

Chapter 8

System of equations

SUR The population model is a set of G linear equations

$$\begin{aligned} y_1 &= x_1\beta_1 + u \\ y_2 &= x_2\beta_2 + u \\ &\vdots \\ y_G &= x_G\beta_G + u \end{aligned} \tag{8.0.1}$$

where x_g is $1 \times K_g$,

$$x_g = (x_{g1}, \ldots, X_{gK_g}), \quad g = 1, \ldots, G$$

and β_g is a $K_g \times 1$, $g = 1, \ldots, G$. In many applications x_g is the same for all g, but the general statement allows for the dimension of x to vary across equations. The system (8.0.1) is often called Zellner's (1962) Seemingly unrelated regressions (SUR) model. The name comes from the fact that since each equation has its own β_g, it appears that the equations are unrelated. Correlation between errors in the equations provide links that can be exploited.

The system can be written as (**??**) by defining $y_i = (y_{i1}, y_{i2}, \ldots, y_{iG})$, $u_i = (u_{i1}, u_{i2}, \ldots, u_{iG})$, and

$$X_i = \begin{pmatrix} x_{i1} & 0 & 0 & \ldots & 0 \\ 0 & x_{i1} & 0 & \ldots & 0 \\ 0 & 0 & & & \vdots \\ \vdots & & & & 0 \\ 0 & 0 & 0 & \ldots & x_{iG} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_G \end{pmatrix}$$

When we need to write the equations for a particular random draw from the population, y_g, x_g, and u_g will also contain an i subscript: equation g becomes $y_{ig} = x_{ig}\beta_g + u_{ig}$.

Assume that we have independent, identically distributed cross section observations $\{(X_i, y_i) : i = 1, \ldots, N\}$, where X is a $G \times K$ matrix and y_i is a $G \times 1$ vector. The multivariate linear model for a random draw of the population can be expressed as

$$y_i = X_i\beta + u_i \tag{8.0.2}$$

where β is the $K \times 1$ parameter vector of interest and u_i is a $G \times 1$ vector of unobservables.

SOLS.1

1. $E(X_i'u_i) = 0$;

   (a) $A = E(X_i'X_i)$ is nonsingular (has rank K);

Under Assumption SOLS.1-SOLS.2,

$$\hat{\beta} \overset{def}{=} \left(\sum_{i=1}^{n} X_i'X_i\right)^{-1} \sum_{i=1}^{n} X_i'y_i \xrightarrow{p} \beta$$

Under Assumption SOLS.1-SOLS.2,

$$\sqrt{N}(\hat{\beta} - \beta) \xrightarrow{d} N(0, A^{-1}BA^{-1})$$

where $B = E(X_i'u_iu_i'X_i)$.

8.1 Generalized Least Squares

System of OLS is consistent under fairly weak assumptions.

If we strengthen Assumption SOLS.1 and add assumptions on the variance matrix of u_i, we can do better using a generalized least squares procedures (GLS). As we will see, GLS is not usually feasible because it requires knowing the variance matrix of the errors up to a multiplicative constant.

[SGLS.1]

1. $E(X_i \otimes u_i) = 0$

This is different from Assumption SOLS.1. Here each element of X_1 must be orthogonal (in expectation) to u_i.

$$E(X_i \otimes u_i) = E\left[\begin{pmatrix} x_{i1} & 0 & 0 & \dots & 0 \\ 0 & x_{i1} & 0 & \dots & 0 \\ 0 & 0 & & & \vdots \\ \vdots & & & & 0 \\ 0 & 0 & 0 & \dots & x_{iG} \end{pmatrix} \otimes u_i\right]$$

$$= E\left[\begin{pmatrix} x_{i1}u_1 & 0 & 0 & \dots & 0 \\ 0 & x_{i1}u_2 & 0 & \dots & 0 \\ 0 & 0 & & & \vdots \\ \vdots & & & & 0 \\ 0 & 0 & 0 & \dots & x_{iG}u_G \end{pmatrix}\right]$$

Let

$$V \stackrel{def}{=} E(u_i u_i')$$

[SGLS.2]

1. V is p.d. and $E(X_i' V^{-1} X_i)$ is nonsingular

Transform error to have

$$V^{-1/2} y_i = (V^{-1/2} X_i)\beta + V^{-1/2} u_i, y^* = X_i^* \beta + U_i^*$$

$$E(u_i^* u_i^{*'}) = I_G$$

$$\beta^* \stackrel{def}{=} \left(\sum_{i=1}^{n} X_i' V^{-1} X_i\right)^{-1} \left(\sum_{i=1}^{n} X_i' V^{-1} y_i\right)$$

In matrix notation

$$\beta^* \stackrel{def}{=} \left(X'(I_N \otimes V^{-1})X\right)^{-1} \left(X'(I_G \otimes V^{-1})Y\right)$$

$$X = (X_1' X_2' \dots X_N')', \quad NT \times G$$