

ECONOMETRICS

Bruce E. Hansen©2000, 2001, 2002¹

University of Wisconsin
www.ssc.wisc.edu/~bhansen

Revised: January 2002

¹This manuscript may be printed and reproduced for individual or instructional use, but may not be printed for commercial purposes.

Contents

1	Matrix Algebra	2
1.1	Terminology	2
1.2	Matrix Multiplication	3
1.3	Identity Matrix	4
1.4	Trace of a Matrix	4
1.5	Matrix Inversion	5
1.6	Determinants	5
1.7	Eigenvalues	6
1.8	Positive Definite Matrices	7
1.9	Idempotent Matrices	7
1.10	Projection Matrices	7
2	Probability Theory	9
2.1	Random Vectors	9
2.2	Inequalities	9
2.3	Random Sampling	13
2.4	Weak Law of Large Numbers	13
2.5	Central Limit Theorem	15
2.6	Continuous Mapping Theorem	16
3	The Random Sampling Model, Regression, and Projection	17
3.1	Random Sampling	17
3.2	Regression	17
3.3	Linear Models	19
3.4	Linear Projection	20
3.5	Assumptions on the Regression Error	20
4	Multivariate Linear Regression	23
4.1	Model in Matrix Notation	23
4.2	Method of Moments Estimation	23
4.3	MME is OLS	25
4.4	Gaussian Quasi-MLE	25
4.5	Semiparametric Efficiency	26
4.6	Frisch-Waugh-Lovell (FWL) Theorem	27
4.7	Analysis of Variance	29
4.8	Bias	30
4.9	Variance-Covariance Matrix of Regression Error	31

4.10	Covariance Matrix of OLS Estimator	32
4.11	GLS and the Gauss-Markov Theorem	33
4.12	Estimation of Error Variance	34
4.13	Covariance Matrix Estimation	35
4.14	Standard Errors	37
5	Asymptotic Distribution Theory	38
5.1	Consistency of the OLS Estimator	38
5.2	Asymptotic Distribution of OLS Estimator	40
5.3	Consistent Covariance Matrix Estimation	41
5.4	Studentized Statistic	44
5.5	Asymptotic Confidence Interval	44
5.6	Functions of Parameters	45
5.7	t tests	47
5.8	Wald Tests	48
5.9	F Tests	50
5.10	Quasi-LR Tests	52
6	Exact Distribution Theory	54
6.1	Normal Regression Model	54
6.2	Monte Carlo Simulation	55
6.3	An Example	57
7	The Bootstrap	60
7.1	Definition of the Bootstrap	60
7.2	The Empirical Distribution Function	60
7.3	Computation	62
7.4	Bootstrap Estimation of Bias	62
7.5	Bootstrap Estimation of Variance	63
7.6	Efron's Percentile Interval	64
7.7	Alternative Percentile Interval	65
7.8	One-Sided Hypothesis Tests	65
7.9	Percentile-t Equal-Tailed Interval	66
7.10	Two-Sided Hypothesis Tests	66
7.11	Symmetric Percentile-t Intervals	67
7.12	Vector Tests	67
7.13	Asymptotic Expansions	67
7.14	One-Sided Tests	68
7.15	Symmetric Two-Sided Tests	69
7.16	Percentile Confidence Intervals	70
7.17	Bootstrap Methods for Independent Errors	71
8	Functional Form	72
8.1	Dummy Variables	72
8.2	NonLinearity in Regressors	74
8.3	Testing for Omitted NonLinearity	75
8.4	$\log(Y)$ versus Y as Dependent Variable	75
8.5	Multicollinearity	76

8.6	Omitted Variables	77
8.7	Irrelevant Variables	78
8.8	Model Selection	79
9	NonLinear Regression	83
9.1	NonLinear Regression Models	83
9.2	NLLS Estimation	83
9.3	Concentration	84
9.4	Computation Using Linearization	84
9.5	Asymptotic Distribution	85
9.6	Identification	86
10	Feasible GLS	87
10.1	Skedastic Regression	87
10.2	Estimation of Skedastic Regression	88
10.3	Testing for Heteroskedasticity	88
10.4	Feasible GLS Estimation	89
10.5	Covariance Matrix Estimation	90
10.6	Commentary: FGLS versus OLS	90
11	Generalized Method of Moments	91
11.1	Endogeneity	91
11.2	Instrumental Variables	92
11.3	Reduced Form	93
11.4	Identification	94
11.5	Instrumental Variables Estimation	95
11.6	GMM Estimator	95
11.7	2SLS Estimator	96
11.8	Distribution of GMM Estimator	97
11.9	Optimal Weight Matrix	97
11.10	Estimation of the Efficient Weight Matrix	98
11.11	Over-Identification Test	99
11.12	GMM: The General Case	99
11.13	Hypothesis Testing: The Distance Statistic	100
11.14	GMM as Semiparametrically Efficient	101
11.15	Conditional Moment Restrictions	101
11.16	Continuously-Updated GMM	102
11.17	Empirical Likelihood	103
11.18	Bootstrap GMM Inference	104
11.19	Bekker Asymptotics	105
11.20	Identification Failure	107
12	Univariate Time Series	109
12.1	Stationarity and Ergodicity	109
12.2	Autoregressions	111
12.3	Stationarity of AR(1) Process	111
12.4	Lag Operator	112
12.5	Stationarity of AR(k)	112

12.6	Estimation	113
12.7	Asymptotic Distribution	114
12.8	Bootstrap for Autoregressions	114
12.9	Trend Stationarity	115
12.10	Testing for Omitted Serial Correlation	116
12.11	Model Selection	116
12.12	Autoregressive Unit Roots	117
13	Multivariate Time Series	119
13.1	Vector Autoregressions (VARs)	119
13.2	Estimation	120
13.3	Restricted VARs	120
13.4	Single Equation from a VAR	120
13.5	Testing for Omitted Serial Correlation	121
13.6	Selection of Lag Length in an VAR	122
13.7	Granger Causality	122
13.8	Cointegration	123
13.9	Cointegrated VARs	123
14	Limited Dependent Variables	125
14.1	Binary Choice	125
14.2	Count Data	127
14.3	Censored Data	127
14.4	Sample Selection	128
15	Panel Data	131
15.1	Individual-Effects Model	131
15.2	Fixed Effects	131
15.3	Dynamic Panel Regression	133

Chapter 1

Matrix Algebra

This chapter is presented for review and reference. The results are largely taken from the Appendix of Muirhead (1982).

1.1 Terminology

A **scalar** a is a single number.

A **vector** a is a $k \times 1$ list of numbers, typically arranged in a column. We write this as

$$a = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{pmatrix}$$

Equivalently, a vector a is an element of Euclidean k space, hence $a \in R^k$. If $k = 1$ then a is a scalar.

A **matrix** A is a $k \times r$ rectangular array of numbers, written as

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1r} \\ a_{21} & a_{22} & \cdots & a_{2r} \\ \vdots & \vdots & & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kr} \end{bmatrix} = [a_{ij}]$$

By convention a_{ij} refers to the i 'th row and j 'th column of A . If $r = 1$ or $k = 1$ then A is a vector. If $r = k = 1$, then A is a scalar.

A matrix can be written as a set of column vectors or as a set of row vectors. That is,

$$A = [a_1 \quad a_2 \quad \cdots \quad a_r] = \begin{bmatrix} \alpha'_1 \\ \alpha'_2 \\ \vdots \\ \alpha'_k \end{bmatrix}$$

where

$$a_i = \begin{bmatrix} a_{1i} \\ a_{2i} \\ \vdots \\ a_{ki} \end{bmatrix}$$

are column vectors and

$$\alpha'_j = [a_{j1} \quad a_{j2} \quad \cdots \quad a_{jr}]$$

are row vectors.

The **transpose** of a matrix, denoted $B = A'$, is obtained by flipping the matrix on its diagonal.

$$B = A' = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{k1} \\ a_{12} & a_{22} & \cdots & a_{k2} \\ \vdots & \vdots & & \vdots \\ a_{1r} & a_{2r} & \cdots & a_{kr} \end{bmatrix}$$

Thus $b_{ij} = a_{ji}$ for all i and j . Note that if A is $k \times r$, then A' is $r \times k$. If a is a $k \times 1$ vector, then a' is a $1 \times k$ row vector.

A matrix is **square** if $k = r$. A square matrix is **symmetric** if $A = A'$, which implies $a_{ij} = a_{ji}$. A square matrix is **diagonal** if the only non-zero elements appear on the diagonal, so that $a_{ij} = 0$ if $i \neq j$. A square matrix is **upper (lower) diagonal** if all elements below (above) the diagonal equal zero.

A **partitioned matrix** takes the form

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1r} \\ A_{21} & A_{22} & \cdots & A_{2r} \\ \vdots & \vdots & & \vdots \\ A_{k1} & A_{k2} & \cdots & A_{kr} \end{bmatrix}$$

where the A_{ij} denote matrices, vectors and/or scalars.

1.2 Matrix Multiplication

If a and b are both $k \times 1$, then their inner product is

$$a'b = a_1b_1 + a_2b_2 + \cdots + a_kb_k = \sum_{j=1}^k a_jb_j$$

Note that $a'b = b'a$.

If A is $k \times r$ and B is $r \times s$, then we define their product AB by writing A as a set of row vectors and B as a set of column vectors (each which of length r). Then

$$\begin{aligned} AB &= \begin{bmatrix} a'_1 \\ a'_2 \\ \vdots \\ a'_k \end{bmatrix} [b_1 \quad b_2 \quad \cdots \quad b_r] \\ &= \begin{bmatrix} a'_1b_1 & a'_1b_2 & \cdots & a'_1b_r \\ a'_2b_1 & a'_2b_2 & \cdots & a'_2b_r \\ \vdots & \vdots & & \vdots \\ a'_kb_1 & a'_kb_2 & \cdots & a'_kb_r \end{bmatrix} \end{aligned}$$

When the number of columns of A equals the number of rows of B , we say that A and B , or the product AB , is **conformable**, and this is the only case where this product is defined.

An alternative way to write the matrix product is to use matrix partitions. For example,

$$\begin{aligned} AB &= \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \\ &= \begin{bmatrix} A_{11}B_{11} + A_{12}B_{21} & A_{11}B_{12} + A_{12}B_{22} \\ A_{21}B_{11} + A_{22}B_{21} & A_{21}B_{12} + A_{22}B_{22} \end{bmatrix} \end{aligned}$$

and

$$\begin{aligned} AB &= \begin{bmatrix} A_1 & A_2 & \cdots & A_r \end{bmatrix} \begin{bmatrix} B_1 \\ B_2 \\ \vdots \\ B_r \end{bmatrix} \\ &= A_1B_1 + A_2B_2 + \cdots + A_rB_r \\ &= \sum_{j=1}^r A_jB_j \end{aligned}$$

1.3 Identity Matrix

An important diagonal matrix is the **identity matrix**, which has ones on the diagonal. A $k \times k$ identity matrix is denoted as

$$I_k = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

Important obvious properties are that if A is $k \times r$, then $AI_r = A$ and $I_kA = A$.

We say that two vectors a and b are **orthogonal** if $a'b = 0$. The columns of $k \times r$, $r \leq k$, are said to be orthogonal if $A'A = I_r$. A square matrix A is called orthogonal if $A'A = I_k$.

1.4 Trace of a Matrix

The trace of a $k \times k$ square matrix A is the sum of its diagonal elements

$$\text{tr}(A) = \sum_{i=1}^k a_{ii}$$

Some straightforward properties are

$$\begin{aligned} \text{tr}(cA) &= c \text{tr}(A) \\ \text{tr}(A') &= \text{tr}(A) \\ \text{tr}(A+B) &= \text{tr}(A) + \text{tr}(B) \\ \text{tr}(I_k) &= k \\ \text{tr}(AB) &= \text{tr}(BA) \end{aligned}$$

The last result follows since

$$\begin{aligned}
 \text{tr}(AB) &= \text{tr} \begin{bmatrix} a'_1 b_1 & a'_1 b_2 & \cdots & a'_1 b_k \\ a'_2 b_1 & a'_2 b_2 & \cdots & a'_2 b_k \\ \vdots & \vdots & \ddots & \vdots \\ a'_k b_1 & a'_k b_2 & \cdots & a'_k b_k \end{bmatrix} \\
 &= \sum_{i=1}^k a'_i b_i \\
 &= \sum_{i=1}^k b'_i a_i \\
 &= \text{tr}(BA).
 \end{aligned}$$

1.5 Matrix Inversion

A $k \times k$ matrix A has **full rank**, or is **nonsingular**, if there is no $c \neq 0$ such that $Ac = 0$. In this case there exists a unique matrix B such that $AB = BA = I_k$. This matrix is called the **inverse** of A and is denoted by A^{-1} . Some properties include

$$\begin{aligned}
 AA^{-1} &= A^{-1}A = I_k \\
 (A^{-1})' &= (A')^{-1} \\
 (AC)^{-1} &= C^{-1}A^{-1} \\
 (A + C)^{-1} &= A^{-1}(A^{-1} + C^{-1})^{-1}C^{-1} \\
 A^{-1} - (A + C)^{-1} &= A^{-1}(A^{-1} + C^{-1})A^{-1}
 \end{aligned}$$

Also, if A is an orthogonal matrix, then $A^{-1} = A$.

The following fact about inverting partitioned matrices is sometimes useful

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} E^{-1} & -E^{-1}BD^{-1} \\ -D^{-1}CE^{-1} & F^{-1} \end{bmatrix}$$

where

$$\begin{aligned}
 E^{-1} &= A^{-1} + A^{-1}BF^{-1}CA^{-1} = (A - BD^{-1}C)^{-1} \\
 F^{-1} &= D^{-1} + D^{-1}CE^{-1}BD^{-1} = (D - CA^{-1}B)^{-1}
 \end{aligned}$$

1.6 Determinants

The notion of a determinant is defined for square matrices. While the determinant is seen often in econometrics, the actual definition is rarely used. I give the definition here for completeness.

If A is 2×2 , then its determinant is $\det A = a_{11}a_{22} - a_{12}a_{21}$.

For a general $k \times k$ matrix $A = [a_{ij}]$, we can define the determinant as follows. Let $\pi = (j_1, \dots, j_k)$ denote a permutation of $(1, \dots, k)$. There are $k!$ such permutations. There is a unique count of the number of inversions of the indices of such permutations (relative to the natural order $(1, \dots, k)$),

and let $\varepsilon_\pi = +1$ if this count is even and $\varepsilon_\pi = -1$ if the count is odd. Then the general definition of the determinant is

$$\det A = \sum_{\pi} \varepsilon_{\pi} a_{1j_1} a_{2j_2} \cdots a_{kj_k}$$

Some properties include

$$\begin{aligned} \det A &= \det A' \\ \det(\alpha A) &= \alpha^k \det A \\ \det(AB) &= (\det A)(\det B) \\ \det(A^{-1}) &= (\det A)^{-1} \\ \det \begin{bmatrix} A & B \\ C & D \end{bmatrix} &= (\det D) \det(A - BD^{-1}C) \text{ if } \det D \neq 0. \end{aligned}$$

Some additional properties include

- $\det A \neq 0$ if and only if A is nonsingular.
- If A is triangular (upper or lower), then $\det A = \prod_{i=1}^k a_{ii}$
- If A is orthogonal, then $\det A = \pm 1$

1.7 Eigenvalues

The characteristic equation of a square matrix A is

$$\det(A - \lambda I_k) = 0.$$

The left side is a polynomial of degree k in λ so has exactly k roots, which are not necessarily distinct and may be real or complex. They are called the **latent roots** or **characteristic roots** or **eigenvalues** of A . If λ_i is an eigenvalue of A , then $A - \lambda_i I_k$ is singular so there exists a non-zero vector h_i such that

$$(A - \lambda_i I_k) h_i = 0$$

The vector h_i is called a **latent vector** or **characteristic vector** or **eigenvector** of A corresponding to λ_i .

We now state some useful properties. Let λ_i and h_i , $i = 1, \dots, k$ denote the k eigenvalues and eigenvectors of a square matrix A . Let Λ be a diagonal matrix with the characteristic roots in the diagonal, and let $H = [h_1 \cdots h_k]$.

- $\det(A) = \prod_{i=1}^k \lambda_i$
- $\text{tr}(A) = \sum_{i=1}^k \lambda_i$
- A is non-singular if and only if all its characteristic roots are non-zero.
- If A has distinct characteristic roots, there exists a nonsingular matrix P such that $A = P^{-1}\Lambda P$ and $PAP^{-1} = \Lambda$.
- If A is symmetric, then $A = H\Lambda H'$ and $H'AH = \Lambda$, and the characteristic roots are all real.
- The characteristic roots of A^{-1} are $\lambda_1^{-1}, \lambda_2^{-1}, \dots, \lambda_k^{-1}$.

The decomposition $A = H\Lambda H'$ is sometimes called the **spectral decomposition** of a matrix. We define the **rank** of a square matrix as the number of its non-zero characteristic roots.

1.8 Positive Definite Matrices

We say that a square matrix A is **positive semi-definite** if for all non-zero c , $c'Ac \geq 0$. This is written as $A \geq 0$. We say that A is **positive definite** if for all non-zero c , $c'Ac > 0$. This is written as $A > 0$.

If A is positive definite, then A is non-singular and A^{-1} exists. Furthermore, $A^{-1} > 0$.

We say that X is $n \times k$, $k < n$, has full rank k if there is no non-zero c such that $Xc = 0$. In this case, $X'X$ is symmetric and positive definite.

If A is symmetric, then $A > 0$ if and only if all its characteristic roots are positive.

1.9 Idempotent Matrices

A square matrix A is **idempotent** if $AA = A$.

If A is also symmetric (most idempotent matrices are) then all its characteristic roots equal either zero or one. To see this, note that we can write $A = H\Lambda H'$ where H is orthogonal and Λ contains the (real) characteristic roots. Then

$$A = AA = H\Lambda H' H\Lambda H' = H\Lambda^2 H'.$$

By the uniqueness of the characteristic roots, we deduce that $\Lambda^2 = \Lambda$ and $\lambda_i^2 = \lambda_i$ for $i = 1, \dots, k$. Hence they must equal either 0 or 1.

It follows that if A is symmetric and idempotent, then $\text{tr}(A) = \text{rank}(A)$.

1.10 Projection Matrices

Let X be an $n \times k$ matrix, $k < n$. Two **projection matrices** are

$$\begin{aligned} P &= X(X'X)^{-1}X' \\ M &= I_n - P \\ &= I_n - X(X'X)^{-1}X'. \end{aligned}$$

They are called projection matrices due to the property that for any matrix Z which can be written as $Z = X\Gamma$ for some matrix Γ , (we say that Z lies in the **range space** of X) then

$$PZ = PX\gamma = X(X'X)^{-1}X'X\gamma = X\gamma = Z$$

and

$$MZ = (I_n - P)Z = Z - PZ = Z - Z = 0.$$

As an important example of this property, partition the matrix X into two matrices X_1 and X_2 , so that

$$X = [X_1 \quad X_2].$$

Then $PX_1 = X_1$ and $MX_1 = 0$.

P and M are symmetric:

$$\begin{aligned}
P' &= \left(X (X'X)^{-1} X' \right)' \\
&= (X')' \left((X'X)^{-1} \right)' (X)' \\
&= X \left((X'X)' \right)^{-1} X' \\
&= X \left((X)' (X')' \right)^{-1} X' \\
&= P
\end{aligned}$$

and

$$M' = (I_n - P)' = I_n' - P' = I_n - P = M.$$

The projection matrices P and M are idempotent:

$$\begin{aligned}
PP &= \left(X (X'X)^{-1} X' \right) \left(X (X'X)^{-1} X' \right) \\
&= X (X'X)^{-1} X' X (X'X)^{-1} X' \\
&= X (X'X)^{-1} X' = P,
\end{aligned}$$

and

$$\begin{aligned}
MM &= (I_n - P)(I_n - P) \\
&= I_n I_n - P I_n - I_n P + PP \\
&= I_n - P - P + P \\
&= I_n - P = M.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
M + P &= I_n - P + P = I_n \\
MP &= (I_n - P)P = P - PP = P - P = 0.
\end{aligned}$$

Another useful property is that

$$\text{tr } P = k \tag{1.1}$$

$$\text{tr } M = n - k. \tag{1.2}$$

Indeed,

$$\begin{aligned}
\text{tr } P &= \text{tr} \left(X (X'X)^{-1} X' \right) \\
&= \text{tr} \left((X'X)^{-1} X'X \right) \\
&= \text{tr} (I_k) \\
&= k,
\end{aligned}$$

and

$$\text{tr } M = \text{tr} (I_n - P) = \text{tr} (I_n) - \text{tr} (P) = n - k.$$

From this, we deduce that the ranks of P and M are k and $n - k$, respectively. Since M is symmetric and idempotent, its spectral decomposition takes the form

$$M = H \begin{bmatrix} I_{n-k} & 0 \\ 0 & 0 \end{bmatrix} H' \tag{1.3}$$

with $H'H = I_n$.

Chapter 2

Probability Theory

This chapter is presented for review and reference.

2.1 Random Vectors

A random vector X takes values in R^k according to a probability distribution defined by

$$F(x) = P(X \leq x).$$

When $k > 1$, the inequality $X \leq x$ means that $X_j \leq x_j$ for $j = 1, \dots, k$. If the distribution is **discrete**, then there are a countable set of vectors τ_1, \dots, τ_r which X can take, with associated probabilities π_1, \dots, π_r . Specifically,

$$\begin{aligned} P(X = \tau_j) &= \pi_j, & j = 1, \dots, r \\ \sum_{j=1}^r \pi_j &= 1 \end{aligned}$$

If the distribution is **continuous**, then $P(X = \tau) = 0$ for all $\tau \in R^k$. Instead we represent the relative probabilities by the density function

$$f(x) = \frac{d^k}{dx_1 \cdots dx_k} F(x).$$

For any $h : R^k \rightarrow R^q$, we define the expectation $Eh(X)$ as follows. If X is discrete,

$$Eh(X) = \sum_{j=1}^r h(\tau_j) \pi_j,$$

and if X is continuous

$$Eh(X) = \int_{R^k} h(x) f(x) dx_1 \cdots dx_k.$$

2.2 Inequalities

Definition 1 *The Euclidean norm of an $m \times 1$ vector a is*

$$|a| = (a'a)^{1/2} = \left(\sum_{i=1}^m a_i^2 \right)^{1/2}.$$

If A is a $m \times n$ matrix, then its Euclidean norm is

$$|A| = \text{tr}(A'A)^{1/2} = (\text{vec}(A)' \text{vec}(A))^{1/2} = \left(\sum_{i=1}^m \sum_{j=1}^n a_{ij}^2 \right)^{1/2}.$$

Definition 2 The L^p norm, $p \geq 1$, of a random matrix X is

$$\|X\|_p = (E |X|^p)^{1/p}.$$

Lemma 3 *Triangle Inequality.*

$$|X + Y| \leq |X| + |Y|.$$

Lemma 4 *C^r Inequality.*

For $0 < r \leq 1$,

$$|X + Y|^r \leq |X|^r + |Y|^r.$$

For $r \geq 1$,

$$|X + Y|^r \leq 2^{r-1} (|X|^r + |Y|^r).$$

Lemma 5 *Jensen's Inequality.* If $\phi(\cdot) : R \rightarrow R$ is convex, then

$$\phi(E(X)) \leq E(\phi(X)).$$

Proof. By definition, ϕ is convex if and only if for all $s < t < u$,

$$\frac{\phi(t) - \phi(s)}{t - s} \leq \frac{\phi(u) - \phi(t)}{u - t}. \quad (2.1)$$

Set $t = E(X)$, and define

$$C = \sup_{s < t} \frac{\phi(t) - \phi(s)}{t - s}.$$

Hence for all $s < t$,

$$C \geq \frac{\phi(t) - \phi(s)}{t - s},$$

or re-arranging terms,

$$\phi(s) - \phi(t) - C(s - t) \geq 0. \quad (2.2)$$

By (2.1), for all $u > t$,

$$C \leq \frac{\phi(u) - \phi(t)}{u - t},$$

which implies that

$$\phi(u) - \phi(t) - C(u - t) \geq 0. \quad (2.3)$$

Equations (2.2) and (2.3) combine to yield that for all y ,

$$\phi(y) - \phi(t) - C(y - t) \geq 0. \quad (2.4)$$

Let $y = X$, and using $t = E(X)$, (2.4) implies

$$\phi(X) - \phi(E(X)) - C(X - E(X)) \geq 0.$$

Taking expectations yields

$$E\phi(X) - E\phi(E(X)) - CE(X - E(X)) = E\phi(X) - E\phi(E(X)) \geq 0,$$

or

$$E\phi(X) \geq E\phi(E(X)).$$

■

Lemma 6 . *Geometric Means.* If $\alpha_1 + \alpha_2 = 1$ and $A, B > 0$, then

$$A^{\alpha_1} B^{\alpha_2} \leq \alpha_1 A + \alpha_2 B.$$

Proof. An easy way to prove this result is to use Jensen's inequality. Consider a random variable X which takes the value $\ln A$ with probability α_1 , and the value $\ln B$ with probability α_2 . (Since $\alpha_1 + \alpha_2 = 1$ this is a valid probability measure). The exponential function is convex, so Jensen's inequality yields

$$\exp(E(X)) \leq E(\exp(X)). \quad (2.5)$$

Now we just calculate the values on the LHS and RHS of (2.5). First,

$$\begin{aligned} \exp(E(X)) &= \exp[\alpha_1 \ln A + \alpha_2 \ln B] = \exp[\ln A^{\alpha_1} + \ln B^{\alpha_2}] \\ &= \exp[\ln(A^{\alpha_1} B^{\alpha_2})] = A^{\alpha_1} B^{\alpha_2}. \end{aligned}$$

Second,

$$E(\exp(X)) = \alpha_1 \exp(\ln A) + \alpha_2 \exp(\ln B) = \alpha_1 A + \alpha_2 B.$$

■

Lemma 7 *Holder's Inequality.* If $p > 1$ and $q > 1$ and $\frac{1}{p} + \frac{1}{q} = 1$, then

$$E|XY| \leq \|X\|_p \|Y\|_q.$$

Proof. By the theorem of geometric means

$$|XY| = (|X|^p)^{1/p} (|Y|^q)^{1/q} \leq \frac{1}{p} |X|^p + \frac{1}{q} |Y|^q.$$

Taking expectations yields

$$E|XY| \leq \frac{1}{p} E|X|^p + \frac{1}{q} E|Y|^q. \quad (2.6)$$

Now set

$$W = \frac{X}{\|X\|_p}, \quad Z = \frac{Y}{\|Y\|_q}.$$

By (2.6),

$$E|WZ| \leq \frac{1}{p}E|W|^p + \frac{1}{q}E|Z|^q$$

which equals

$$\frac{1}{p}E \left| \frac{X}{\|X\|_p} \right|^p + \frac{1}{q}E \left| \frac{Y}{\|Y\|_q} \right|^q = \frac{1}{p} \frac{EX^p}{\|X\|_p^p} + \frac{1}{q} \frac{EY^q}{\|Y\|_q^q} = \frac{1}{p} + \frac{1}{q} = 1. \quad (2.7)$$

But

$$E|WZ| = E \left| \frac{X}{\|X\|_p} \frac{Y}{\|Y\|_q} \right| = \frac{E|XY|}{\|X\|_p \|Y\|_q},$$

so (2.7) implies that

$$\frac{E|XY|}{\|X\|_p \|Y\|_q} \leq 1,$$

which is the desired result. \blacksquare

Lemma 8 *Cauchy-Schwarz Inequality.*

$$E|XY| \leq \|X\|_2 \|Y\|_2.$$

Proof. This is just Holder's inequality with $p = q = 2$.

Lemma 9 *Minkowski's Inequality.* For $p \geq 1$,

$$\|X + Y\|_p \leq \|X\|_p + \|Y\|_p.$$

Proof. If $p = 1$, the triangle inequality yields $|X + Y| \leq |X| + |Y|$, and then take expectations.

For $p > 1$, define its conjugate $q = p/(p-1)$ (so $1/p + 1/q = 1$). By the triangle inequality,

$$\begin{aligned} |X + Y|^p &= |X + Y|^{p-1} |X + Y| \leq |X + Y|^{p-1} (|X| + |Y|) \\ &= |X + Y|^{p-1} |X| + |X + Y|^{p-1} |Y|. \end{aligned}$$

Thus taking expectations,

$$\|X + Y\|_p^p = E|X + Y|^p \leq E(|X + Y|^{p-1} |X|) + E(|X + Y|^{p-1} |Y|),$$

then applying Holder's inequality,

$$\begin{aligned} &\leq \left\| |X + Y|^{p-1} \right\|_q \|X\|_p + \left\| |X + Y|^{p-1} \right\|_q \|Y\|_p = \|X + Y\|_{(p-1)q}^{(p-1)} (\|X\|_p + \|Y\|_p) \\ &= \|X + Y\|_p^{p-1} (\|X\|_p + \|Y\|_p), \end{aligned}$$

where the final inequality holds since $(p - 1)q = p$. We have shown that

$$\|X + Y\|_p^p \leq \|X + Y\|_p^{p-1} (\|X\|_p + \|Y\|_p).$$

Multiplying both sides by $\|X + Y\|_p^{1-p}$ yields the result. ■

The above proofs are taken from Rudin (1987).

Lemma 10 *Markov's Inequality.* For any strictly increasing function $g(X) \geq 0$,

$$P(g(X) > \alpha) \leq \alpha^{-1} E g(X).$$

Proof. Set $Y = g(X)$, and let $\{\cdot\}$ denote the indicator function. Then

$$\alpha P(Y > \alpha) = \alpha E\{Y > \alpha\} = \alpha \int \{Y > \alpha\} dP = \int_{\{Y > \alpha\}} \alpha dP,$$

and since for the entire region of integration $\alpha < Y$, this is bounded by

$$\leq \int_{\{Y > \alpha\}} Y dP \leq \int Y dP = E(Y).$$

Hence $P(Y > \alpha) \leq \alpha^{-1} E(Y)$ and we are done. ■

2.3 Random Sampling

A set of random vectors $\{X_1, \dots, X_n\}$ are **independent and identically distributed** if they are mutually independent and are drawn from a common distribution F . This is typically abbreviated as IID.

We say that the set $\{X_1, \dots, X_n\}$ is a **random sample** if they are IID random vectors. We call the observed realization $\{X_1, \dots, X_n\}$ the **sample**, and the distribution F the **population**. The goal in statistical inference is to learn about characteristics of the population from a random sample. That is, we want to learn about F , the true distribution of the random variables. Sometimes this is a source of confusion, but the theory is based on the concept of a infinitely large population.

2.4 Weak Law of Large Numbers

Let $\{X_1, \dots, X_n\}$ be IID and define the sample mean

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

We say that Z_n converges in probability to Z as $n \rightarrow \infty$ denoted $Z_n \rightarrow_p Z$ as $n \rightarrow \infty$, if for all $\delta > 0$,

$$\lim_{n \rightarrow \infty} P(|Z_n - Z| > \delta) = 0.$$

This is a probabilistic way of generalizing the mathematical definition of a limit.

Theorem 11 If $E|X| < \infty$ then $\bar{X}_n \rightarrow_p E(X)$.

Proof. Without loss of generality, we can set $E(X) = 0$ (by recentering X_i on its expectation). Under the stronger assumption $EX^2 < \infty$ we have the following trivial proof by Markov's inequality: For any $\delta > 0$

$$P(|\bar{X}_n| > \delta) = P(\bar{X}_n^2 > \delta^2) \leq \delta^{-2} E\bar{X}_n^2 = \frac{EX^2}{\delta^2 n} \rightarrow 0.$$

To escape the unnecessary variance restriction we need a more delicate proof.

To prove the result, we need to show that for all $\delta > 0$ and $\eta > 0$ there is some $\bar{n} < \infty$ so that for all $n \geq \bar{n}$, $P(|\bar{X}_n| > \delta) \leq \eta$.

Fix δ and η . Set $\varepsilon = \delta\eta/3$. Pick $C < \infty$ large enough so that

$$E(|X|1(|X| > C)) \leq \varepsilon \tag{2.8}$$

(where $1(\cdot)$ is the indicator function) which is possible since $E|X| < \infty$. Then set

$$\bar{n} \geq 4C^2/\varepsilon^2. \tag{2.9}$$

Define the random vectors

$$\begin{aligned} W_i &= X_i 1(|X_i| \leq C) - E(X_i 1(|X_i| \leq C)) \\ Z_i &= X_i 1(|X_i| > C) - E(X_i 1(|X_i| > C)). \end{aligned}$$

Since X_i is iid, W_i and Z_i are also.

By Jensen's inequality and (2.8),

$$|E(X_i 1(|X_i| > C))| \leq E(|X_i| 1(|X_i| > C)) \leq \varepsilon.$$

By the triangle inequality and (2.8),

$$E|\bar{Z}_n| \leq E|Z_i| \leq E|X_i| 1(|X_i| > C) + |E(X_i 1(|X_i| > C))| \leq 2\varepsilon.$$

Note that $|W_i| \leq 2C$. Thus (crudely) $EW_i^2 \leq 4C^2$. Since the W_i are iid and mean zero,

$$E\bar{W}_n^2 = \frac{EW_i^2}{n} \leq \frac{4C^2}{n} \leq \varepsilon^2$$

the final inequality holding for $n \geq \bar{n}$ by (2.9). Thus by Jensen's inequality

$$(E|\bar{W}_n|)^2 \leq E\bar{W}_n^2 \leq \varepsilon^2.$$

Finally, by Markov's inequality, the fact that $\bar{X}_n = \bar{W}_n + \bar{Z}_n$, the triangle inequality, and these two bounds,

$$P(|\bar{X}_n| > \delta) \leq \frac{E|\bar{X}_n|}{\delta} \leq \frac{E|\bar{W}_n| + E|\bar{Z}_n|}{\delta} \leq \frac{3\varepsilon}{\delta} = \eta,$$

the equality by the definition of ε . We have shown that for any $\delta > 0$ and $\eta > 0$ there is some $\bar{n} < \infty$ so that for all $n \geq \bar{n}$, $P(|\bar{X}_n| > \delta) \leq \eta$, as needed. \blacksquare

The above proof is based one given in Andrews (1988).

2.5 Central Limit Theorem

Definition 2.5.1 Let Z_n be a random variable with distribution $F_n(x) = P(Z_n \leq x)$. We say that Z_n converges in distribution to Z as $n \rightarrow \infty$, denoted $Z_n \rightarrow_d Z$, where Z has distribution $F(x) = P(Z \leq x)$, if for all x at which $F(x)$ is continuous, $F_n(x) \rightarrow F(x)$ as $n \rightarrow \infty$.

Theorem 2.5.1 Central Limit Theorem (CLT). If $X_i \in R^k$ is iid and $E|X_i|^2 < \infty$, then as $n \rightarrow \infty$

$$\sqrt{n}(\bar{X}_n - \mu) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (X_i - \mu) \rightarrow_d N(0, V).$$

where $\mu = EX$ and $V = E(X - \mu)(X - \mu)'$.

Proof: Without loss of generality, it is sufficient to consider the case $\mu = 0$ and $V = I_k$. For ease of presentation, we assume that the moment generating function (MGF) of X exists, although the general case may be similarly examined using the characteristic function (which always exists). For $\lambda \in R^k$, set

$$\begin{aligned} m(\lambda) &= E \exp(\lambda'X) \\ g(\lambda) &= \log m(\lambda). \end{aligned}$$

Then observe

$$\begin{aligned} \frac{\partial}{\partial \lambda} m(\lambda) &= E(X \exp(\lambda'X)) \\ \frac{\partial^2}{\partial \lambda \partial \lambda'} m(\lambda) &= E(XX' \exp(\lambda'X)) \end{aligned}$$

so when evaluated at $\lambda = 0$

$$\begin{aligned} m(0) &= 1 \\ \frac{\partial}{\partial \lambda} m(0) &= E(X) = 0 \\ \frac{\partial^2}{\partial \lambda \partial \lambda'} m(0) &= E(XX') = I_k. \end{aligned}$$

Furthermore,

$$\begin{aligned} g_\lambda(\lambda) &= \frac{\partial}{\partial \lambda} g(\lambda) = m(\lambda)^{-1} \frac{\partial}{\partial \lambda} m(\lambda) \\ g_{\lambda\lambda}(\lambda) &= \frac{\partial^2}{\partial \lambda \partial \lambda'} g(\lambda) = m(\lambda)^{-1} \frac{\partial^2}{\partial \lambda \partial \lambda'} m(\lambda) - m(\lambda)^{-2} \frac{\partial}{\partial \lambda} m(\lambda) \frac{\partial}{\partial \lambda'} m(\lambda) \end{aligned}$$

so when evaluated at $\lambda = 0$

$$\begin{aligned} g(0) &= 0 \\ g_\lambda(0) &= 0 \\ g_{\lambda\lambda}(0) &= I_k. \end{aligned}$$

By a second-order Taylor series expansion of $g(\lambda)$ about $\lambda = 0$,

$$g(\lambda) = g(0) + g_\lambda(0)' \lambda + \frac{1}{2} \lambda' g_{\lambda\lambda}(\lambda^*) \lambda = \frac{1}{2} \lambda' g_{\lambda\lambda}(\lambda^*) \lambda$$

where λ^* lies on the line segment joining 0 and λ .

We now compute the log MGF of $\sqrt{n}\bar{X}_n$. For $\alpha \in R^k$, by the properties of the exponential function, the independence of the X_i , the definition of $g(\lambda)$

$$\begin{aligned}
G_n(\alpha) &= \log E \exp(\alpha' \sqrt{n}\bar{X}_n) \\
&= \log E \exp\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \alpha' X_i\right) \\
&= \log E \prod_{i=1}^n \exp\left(\frac{1}{\sqrt{n}} \alpha' X_i\right) \\
&= \log \prod_{i=1}^n E \exp\left(\frac{1}{\sqrt{n}} \alpha' X_i\right) \\
&= ng\left(\frac{\alpha}{\sqrt{n}}\right) \\
&= n\frac{1}{2}\left(\frac{\alpha}{\sqrt{n}}\right)' g_{\lambda\lambda}(\lambda_n^*) \left(\frac{\alpha}{\sqrt{n}}\right) \\
&= \frac{1}{2}\alpha' g_{\lambda\lambda}(\alpha_n)\alpha
\end{aligned}$$

where α_n lies on the line segment joining 0 and α/\sqrt{n} . Thus $\alpha_n \rightarrow 0$ as $n \rightarrow \infty$, and $g_{\lambda\lambda}(\alpha_n) \rightarrow g_{\lambda\lambda}(0) = I_k$. We see that as $n \rightarrow \infty$,

$$G_n(\alpha) \rightarrow \frac{1}{2}\alpha'\alpha.$$

It follows that the MGF of $\sqrt{n}\bar{X}_n$ tends to $\exp\left(\frac{1}{2}\alpha'\alpha\right)$ as $n \rightarrow \infty$, which is the MGF of the $N(0, I_k)$ distribution. This is sufficient to establish the theorem. ■

2.6 Continuous Mapping Theorem

There are two useful results, both referred to as the Continuous Mapping Theorem (CMT)

Theorem 12 *If $Z_n \rightarrow_p c$ as $n \rightarrow \infty$ and $g(\cdot)$ is continuous at c , then $g(Z_n) \rightarrow_p g(c)$ as $n \rightarrow \infty$.*

Theorem 13 *If $Z_n \rightarrow_d Z$ as $n \rightarrow \infty$ and $g(\cdot)$ is continuous, then $g(Z_n) \rightarrow_d g(Z)$ as $n \rightarrow \infty$.*

Chapter 3

The Random Sampling Model, Regression, and Projection

3.1 Random Sampling

An econometrician has the observational data

$$\{w_1, w_2, \dots, w_i, \dots, w_n\},$$

where each w_i is a vector of data on an individual (e.g., household or firm).

If this data is *cross-sectional* (each observation is a different individual) it is considered reasonable to assume they are mutually independent. If the data are symmetrically gathered (e.g., randomly), it is reasonable to model each observation as a random draw from the same probability distribution. Thus, the data are *independent and identically distributed*, or IID. We call this a *random sample*.

3.2 Regression

Partition

$$w_i = (y_i, x_i)$$

where $y_i \in R$, $x_i \in R^k$.

In regression, we want to find the central tendency of the conditional distribution of y_i given x_i . A standard measure of central tendency is the mean. The conditional analog is the *conditional mean*. To define the conditional mean, recall that if (y_i, x_i) have a joint density $f(y, x)$, then the conditional density

$$f(y | x) = \frac{f(y, x)}{f_x(x)}$$

exists, where $f_x(x) = \int_{-\infty}^{\infty} f(y, x) dy$ is the marginal density of x_i . The conditional mean is defined as the function

$$m(x) = E(y_i | x_i = x) = \int_{-\infty}^{\infty} y f(y | x) dy.$$

While this definition requires the existence of densities, it is possible to extend this definition so that $g(x)$ is well defined whenever $E|y_i| < \infty$.

In general, $m(x)$ can take any form. For the remainder of this section, we explore the general regression framework without imposing any particular form on $g(x)$.

The conditional mean $m(x) = E(y_i | x_i = x)$ is a function, meaning that when x_i equals x , then the expected value of y_i is $m(x)$. If you ask what is the conditional mean of y_i given the random vector x_i , the answer is $m(x_i)$, which is also denoted as $E(y_i | x_i)$. This conditional expectation is a random variable, as it is a function of the random vector x_i .

The regression error e_i is defined to be the difference between y_i and its conditional mean:

$$e_i = y_i - m(x_i).$$

By construction, this yields the formula

$$y_i = m(x_i) + e_i. \tag{3.1}$$

It is worth emphasizing that no assumptions have been used to develop (3.1), other than that (y_i, x_i) have a joint distribution and $E|y_i| < \infty$.

Proposition 3.2.1 *Properties of the regression error e_i*

1. $E(e_i | x_i) = 0$.
2. $E(e_i) = 0$.
3. $E(h(x_i)e_i) = 0$ for any function $h(\cdot)$.
4. $E(x_i e_i) = 0$.

Proof:

1. By the definition of e_i and the linearity of conditional expectations,

$$\begin{aligned} E(e_i | x_i) &= E((y_i - m(x_i)) | x_i) \\ &= E(y_i | x_i) - E(m(x_i) | x_i) \\ &= m(x_i) - m(x_i) \\ &= 0. \end{aligned}$$

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

$$\begin{aligned} E(e_i) &= E(E(e_i | x_i)) \\ &= E(0) \\ &= 0. \end{aligned}$$

3. By essentially the same argument,

$$\begin{aligned} E(h(x_i)e_i) &= E(E(h(x_i)e_i | x_i)) \\ &= E(h(x_i)E(e_i | x_i)) \\ &= E(h(x_i) \bullet 0) \\ &= 0. \end{aligned}$$

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

The final result is very useful. This implies that e_i and x_i are *uncorrelated*. It is important to understand that despite being uncorrelated, in general e_i need not be independent of x_i . We discuss this implication further in the next section.

Equation (3.1) plus the first result of Proposition 3.2.1 are often stated jointly as the regression framework:

$$\begin{aligned} y_i &= m(x_i) + e_i \\ E(e_i | x_i) &= 0. \end{aligned} \tag{3.2}$$

It is important to understand that this is a framework, not a model, because no restrictions have been placed on the joint distribution of the data. These equations hold true by definition. A regression model, as we explore in the following sections, imposes further restrictions on the joint distribution; most typically, restrictions on the permissible class of regression functions $m(x)$.

3.3 Linear Models

While $m(x)$ in general can take any shape, we typically pick a parametric family $\{m(x, \beta) : \beta \in R^l\}$ to simplify estimation and interpretation. Sometimes, the form of $m(x, \theta)$ is given by an economic theory or model. Most often, however, we use a linear form for convenience and data coherence.

For convenience, we revise the notation so that the vector x_i is presumed to include a constant, that is, the first element of x_i equals 1. We denote the k -element regressor x_i as

$$x_i = \begin{pmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{ki} \end{pmatrix} = \begin{pmatrix} 1 \\ x_{2i} \\ \vdots \\ x_{ki} \end{pmatrix}.$$

A linear model for $g(x)$ then takes the form

$$m(x_i) = \beta_1 + x_{2i}\beta_2 + \cdots + x_{ki}\beta_k$$

where β_1 through β_k are parameters. The parameter vector β is written as

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

In vector notation,

$$m(x_i) = x_i' \beta.$$

The regression model is then

$$\begin{aligned} y_i &= x_i' \beta + e_i \\ E(e_i | x_i) &= 0. \end{aligned} \tag{3.3}$$

This is a model because m has been restricted to the linear form.

While linearity is substantively restrictive, it has still a great deal of flexibility. For example, if x_i is real-valued and

$$m(x_i) = \beta_1 + x_i\beta_2 + x_i^2\beta_3 + \cdots + x_i^{k-1}\beta_k$$

is a polynomial, then a linear regression model still holds, by the redefinition of x_i as $(1, x_i, x_i^2, \dots, x_i^{k-1})$.

3.4 Linear Projection

The linear regression model (3.3) implies $E(x_i e_i) = 0$. This latter condition is sufficient for many asymptotic results. It is interesting to observe that in linear models, there is always a vector β such that this equation holds. This vector β may be called the linear projection coefficient or linear predictor.

Proposition 3.4.1 *For any random variables (y_i, x_i) , let*

$$\beta = (E(x_i x_i'))^{-1} E(x_i y_i) \tag{3.4}$$

and

$$e_i = y_i - x_i' \beta.$$

Then

$$E(x_i e_i) = 0.$$

Proof:

$$E(x_i e_i) = E(x_i (y_i - x_i' (E(x_i x_i'))^{-1} E(x_i y_i))) = 0.$$

■

If β is defined as in (3.4), then $E(x_i e_i) = 0$ holds true by construction. It does not necessarily follow that $E(e_i | x_i) = 0$. This only holds if the true conditional mean of y_i is $x_i' \beta$, which is a substantive restriction. Thus the linear regression assumption that $E(e_i | x_i) = 0$ is more restrictive than the linear projection construction.

It turns out that for most issues in statistical inference, the projection assumption is sufficient. Therefore we will adopt the more general assumption $E(x_i e_i) = 0$ wherever possible for our analysis.

For econometric practice, however, it is typically desired to estimate a regression equation rather than a projection equation. That is, it is typically desirable for $x_i' \beta$ to represent the conditional mean of y_i , rather than a simple linear projection. So while it is not necessary for inference on β , it may be necessary for inference on an economic relationship of interest.

3.5 Assumptions on the Regression Error

While the regression motivation leads naturally to the model (3.3), at times it is more convenient to adopt assumptions which are either more restrictive or less restrictive. We now describe the standard types of models considered by econometricians, and discuss their strengths and weaknesses. All the models are based on the decomposition

$$y_i = x_i' \beta + e_i. \tag{3.5}$$

In addition, all models normalize the error so that $E(e_i) = 0$ and presume a finite variance $E(e_i^2) = \sigma^2 < \infty$.

Definition 3.5.1 *The Linear Projection Model is (3.5) plus*

$$E(x_i e_i) = 0.$$

The advantage of the linear projection model is that it is true by construction, and many inferential results hold under this broad condition. The disadvantage is that the coefficients β may not have useful economic interpretations without additional structure.

Definition 3.5.2 *The Linear Regression Model is (3.5) plus*

$$E(e_i | x_i) = 0.$$

This model leads naturally from the derivation of the conditional mean function. The primary advantage is that the parameter β is easily interpretable.

Definition 3.5.3 *The Homoskedastic Regression Model is the Linear Regression model plus*

$$E(e_i^2 | x_i) = \sigma^2. \tag{3.6}$$

This model adds the auxiliary assumption (3.6) that the regression is *conditionally homoskedastic*. This assumption greatly simplifies many theoretical arguments and calculations, and is therefore very useful in pedagogical (illustrative) arguments. Many formulae simplify under this assumption, and as a result, alternative estimators and techniques are utilized. The danger in this assumption is that these simplifications result in incorrect answers and inferences if indeed the homoskedasticity assumption is false. As a result, we try to avoid making this assumption wherever possible, since the empirical validity of conditional homoskedasticity is always doubtful.

Another meaningful justification for assumption (3.6) is that while it may not be precisely true in the data, it may be approximately true, and in some applications the cost of imposing homoskedasticity on the estimates may be less than the cost of using the more general techniques appropriate for the linear regression model.

Definition 3.5.4 *The Classical Regression Model is (3.5) plus that e_i is independent of x_i .*

This model is more restrictive than the homoskedastic regression model, and is a common starting point in classical econometrics textbooks.

Definition 3.5.5 *The Normal Regression Model is (3.5) plus that e_i is independent of x_i and distributed as $N(0, \sigma^2)$.*

This model is a further restriction, and is also described in classical econometrics textbooks. The primary advantage of this model is that exact distributional results are available for the OLS estimators and test statistics. These exact results are less popular in current econometric practice, since the independence and normality assumptions are rarely considered viable.

The five models listed above are strictly nested, with the first (the linear projection model) the least restrictive, and the last (the normal regression model) the most restrictive. In this class, we will typically restrict attention to the first three models, with most attention being devoted to the second (the linear regression model). At many points in the analysis, we will discuss the implications of the homoskedasticity assumption (3.6), as many results are substantively different under this condition. Since these differences are important, it is worthwhile to briefly discuss this assumption.

The conditional variance function is

$$Var(y_i | x_i = x) = E(e_i^2 | x_i = x) = \sigma^2(x)$$

which is (potentially) a function of x . Just as the conditional mean function may take any form, so may the conditional variance function (other than the restriction that it is non-negative). Given the random variable x_i , the conditional variance is $\sigma_i^2 = \sigma^2(x_i)$.

In the general case where $\sigma^2(x)$ is not a constant function, so σ_i^2 is different across i , we say that the error e_i is *heteroskedastic*. In the special case of (3.6), where the function $\sigma^2(x)$ is a constant, and the conditional variances σ_i^2 all equal the same constant value σ^2 , we say that the error e_i is *homoskedastic*.

Chapter 4

Multivariate Linear Regression

4.1 Model in Matrix Notation

The linear regression model was introduced in equation (3.3). This is a system of n equations, one for each observation. We can stack these n equations together as

$$\begin{aligned}y_1 &= x'_1\beta + e_1 \\y_2 &= x'_2\beta + e_2 \\&\vdots \\y_n &= x'_n\beta + e_n.\end{aligned}$$

Let

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} x'_1 \\ x'_2 \\ \vdots \\ x'_n \end{pmatrix}, \quad e = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}.$$

Observe that Y and e are $n \times 1$ vectors, and X is an $n \times k$ matrix. Using this matrix notation, the n equations may be written in the single equation

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

4.2 Method of Moments Estimation

In all cases, we see that the model satisfies

$$E(x_i e_i) = 0. \tag{4.1}$$

Let β_0 denote the true value of β , where the latter denotes a generic value of the parameter. Then

$$E(x_i (y_i - x'_i \beta_0)) = 0.$$

Another way to write this is to define the “moment function”

$$g_i(\beta) = x_i (y_i - x'_i \beta)$$

and observe that

$$E(g_i(\beta_0)) = 0.$$

For values of $\beta \neq \beta_0$, $E(g_i(\beta)) \neq 0$.

The empirical, or sample, analog of a moment $E(X)$ is the sample moment $\frac{1}{n} \sum_{i=1}^n X_i$. Similarly, the empirical analog of $E(g_i(\beta))$ is

$$\begin{aligned}\bar{g}_n(\beta) &= \frac{1}{n} \sum_{i=1}^n g_i(\beta) \\ &= \frac{1}{n} \sum_{i=1}^n x_i (y_i - x_i' \beta) \\ &= \frac{1}{n} (X'Y - X'X\beta)\end{aligned}$$

The method of moments estimator (MME) of β is the value $\hat{\beta}$ such that $\bar{g}_n(\hat{\beta}) = 0$. Thus $\hat{\beta}$ is defined to mimic, as closely as possible, the orthogonality properties (4.1). Thus

$$0 = \frac{1}{n} (X'Y - X'X\hat{\beta})$$

which implies

$$X'Y = X'X\hat{\beta}.$$

This equation is called the “normal equations”. The solution is the MME of β .

Proposition 4.2.1 *The MME solution to $\bar{g}_n(\hat{\beta}) = 0$ is*

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

We define the *predicted value* $\hat{y}_i = x_i' \hat{\beta}$ and the *residual*

$$\begin{aligned}\hat{e}_i &= y_i - \hat{y}_i \\ &= y_i - x_i' \hat{\beta}.\end{aligned}$$

Another way of writing the relationship between the predicted value, residual, and observation is

$$y_i = \hat{y}_i + \hat{e}_i.$$

In vector notation, $\hat{Y} = X\hat{\beta}$, $\hat{e} = Y - X\hat{\beta}$, and $Y = \hat{Y} + \hat{e}$.

Note that by definition,

$$0 = \bar{g}_n(\hat{\beta}) = \frac{1}{n} X' \hat{e}$$

Thus

$$X' \hat{e} = 0$$

and the residual vector \hat{e} is orthogonal to the columns of the regressor matrix X .

Since the first column of X is a vector of ones, $X' \hat{e} = 0$ implies that $\sum_{i=1}^n \hat{e}_i = 0$.

Using the projection matrices defined earlier,

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

and

$$\hat{e} = Y - X\hat{\beta} = Y - PY = (I_n - P)Y = MY.$$

Another way of writing this is

$$\begin{aligned} Y &= (P + M)Y \\ &= PY + MY \\ &= \hat{Y} + \hat{e}. \end{aligned}$$

This decomposition is *orthogonal*, that is

$$\begin{aligned} \hat{Y}'\hat{e} &= (PY)'(MY) \\ &= Y'PMY \\ &= 0. \end{aligned}$$

4.3 MME is OLS

The MME estimator $\hat{\beta}$ is also known as the Ordinary Least Squares (OLS) estimator. The sum-of-squared errors (SSE) function is

$$\begin{aligned} S_n(\beta) &= \sum_{i=1}^n (y_i - x_i'\beta)^2 \\ &= Y'Y - 2Y'X\beta + \beta'X'X\beta \end{aligned}$$

The definition of OLS is the vector $\hat{\beta}$ which minimizes $S_n(\beta)$. Vector calculus gives the first-order conditions for minimization:

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

which yields the same normal equations $X'Y = X'X\hat{\beta}$ as for the MME estimator.

Proposition 4.3.1 *The solution to $\min_{\beta} S_n(\beta)$ is $\hat{\beta} = (X'X)^{-1}(X'Y)$.*

Following convention, we will call $\hat{\beta}$ the OLS estimator of β .

4.4 Gaussian Quasi-MLE

In this section, we relate a traditional motivation for the OLS estimator. The normal regression model is $y_i = x_i'\beta + e_i$ with $e_i \sim N(0, \sigma^2)$. The likelihood for a single observation is

$$L_i(\beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{1}{2\sigma^2} (y_i - x_i'\beta)^2\right).$$

Thus the log-likelihood for this full sample is

$$\begin{aligned} l_n(\beta, \sigma^2) &= \sum_{i=1}^n \log L_i(\beta, \sigma^2) \\ &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - x_i'\beta)^2 \\ &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} S_n(\beta). \end{aligned}$$

The MLE $(\hat{\beta}, \hat{\sigma}^2)$ maximize $l_n(\beta, \sigma^2)$. But since $l_n(\beta, \sigma^2)$ is a function of β only through $S_n(\beta)$, the MLE minimizes $S_n(\beta)$, which is the OLS estimator.

Proposition 4.4.1 *In the normal regression model, $\hat{\beta}_{MLE} = \hat{\beta}_{OLS}$.*

Due to this equality, the OLS estimator $\hat{\beta}$ is frequently referred to as the “Gaussian MLE” or the “Gaussian Quasi-MLE”. The term “quasi” is used to refer the context where the normality assumption has been used to construct the likelihood and the estimator, but the normality assumption is not believed to be true.

4.5 Semiparametric Efficiency

Take the model

$$\begin{aligned} y_i &= x_i' \beta + e_i \\ E(x_i e_i) &= 0. \end{aligned}$$

Under these assumptions, is OLS efficient? This question was answered in the affirmative by Chamberlain (*Journal of Econometrics*, 1987). We present this result for the broad class of method-of-moments estimation (MME).

MME applies when we have observations w_i , $k \times 1$ parameter vector β , and a $k \times 1$ vector function $g(w, \beta)$ such that $Eg(w_i, \beta_0) = 0$. The MME $\hat{\beta}_{MME}$ is (by definition) the solution to the equation $\bar{g}_n(\hat{\beta}_{MME}) = 0$, where $\bar{g}_n(\beta) = n^{-1} \sum_{i=1}^n g(w_i, \beta)$.

Now suppose that the distribution of w_i is discrete. That is, for finite r ,

$$P(w_i = \tau_j) = \pi_j, \quad j = 1, \dots, r$$

for some constant vectors τ_j and constants π_j . Assume that the τ_j are known, but the π_j are unknown. (We know the values w_i can take, but we don't know the probabilities.)

In this discrete setting, the moment condition $Eg(w_i, \beta_0) = 0$ can be rewritten as

$$\sum_{j=1}^r \pi_j g(\tau_j, \beta_0) = 0. \quad (4.2)$$

By the implicit function theorem, we can write β as a function of the probabilities π_j :

$$\beta = h(\pi_1, \dots, \pi_r).$$

That is, for any π_1, \dots, π_r we can solve the k equations (4.2) to find the $k \times 1$ vector β . As the π_j change (the distribution of the data changes), the vector β changes. The function h is implicitly defined by this mapping.

As the data are multinomial, we can construct the maximum likelihood estimator (MLE). For $j = 1, \dots, r$

$$\hat{\pi}_j = \frac{1}{n} \sum_{i=1}^n 1(w_i = \tau_j)$$

where $1(\cdot)$ is the indicator function. That is, $\hat{\pi}_j$ is the percentage of the observed w_i which equal τ_j . The MLE for β is then

$$\hat{\beta}_{MLE} = h(\hat{\pi}_1, \dots, \hat{\pi}_r).$$

By the definition of the function h , this means that $\hat{\beta}_{MLE}$ solves the set of equations

$$\sum_{j=1}^r \hat{\pi}_j g(\tau_j, \hat{\beta}_{MLE}) = 0.$$

Substituting in the expressions for $\hat{\pi}_j$,

$$\begin{aligned}
0 &= \sum_{j=1}^r \left(\frac{1}{n} \sum_{i=1}^n 1(w_i = \tau_j) \right) g(\tau_j, \hat{\beta}_{MLE}) \\
&= \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^r 1(w_i = \tau_j) g(\tau_j, \hat{\beta}_{MLE}) \\
&= \frac{1}{n} \sum_{i=1}^n g(w_i, \hat{\beta}_{MLE}) \\
&= \bar{g}_n(\hat{\beta}_{MLE}).
\end{aligned}$$

Hence the MLE solves $0 = \bar{g}_n(\hat{\beta}_{MLE})$. But as this is the definition of the MME, it follows that in this setting the MLE and MME are identical.

In regular parametric models, the Cramer-Rao Theorem asserts that the MLE is asymptotically efficient – no other consistent estimator has a lower asymptotic variance. In the setting we just examined, it follows that the MME is asymptotically efficient.

Chamberlain's Theorem extends this argument to the case of continuously-distributed w_i . He observes that the above argument holds for all multinomial distributions, and any continuous distribution can be arbitrarily well approximated by a multinomial distribution. He then shows that this allows us to prove that generically the MME is asymptotically efficient.

This includes the linear model

$$\begin{aligned}
y_i &= x_i' \beta + e_i \\
E(x_i e_i) &= 0.
\end{aligned}$$

Hence in this setting, OLS is asymptotically efficient.

4.6 Frisch-Waugh-Lovell (FWL) Theorem

Partition

$$X = [X_1 \quad X_2]$$

and

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

Then the regression model can be rewritten as

$$Y = X_1 \beta_1 + X_2 \beta_2 + e. \tag{4.3}$$

Observe that the OLS estimator of $\beta = (\beta_1, \beta_2)$ can be obtained by regression of Y on $X = [X_1 \quad X_2]$. OLS estimation can be written as

$$Y = X_1 \hat{\beta}_1 + X_2 \hat{\beta}_2 + \hat{e} \tag{4.4}$$

where $\hat{e} = MY$ with $M = I - P$ and $P = X(X'X)^{-1}X'$.

Now define $P_1 = X_1(X_1'X_1)^{-1}X_1'$ and $M_1 = I - P_1$. Observe that since P_1 lies in the range space of X , $PP_1 = P_1$, so $P_1P = P_1$ and

$$\begin{aligned}
M_1M &= (I - P_1)(I - P) \\
&= I - P_1 - P + P_1P \\
&= I - P = M.
\end{aligned}$$

Now pre-multiply both sides of (4.4) by M_1 (and note that $M_1X_1 = 0$) to obtain

$$\begin{aligned} M_1Y &= M_1X_1\hat{\beta}_1 + M_1X_2\hat{\beta}_2 + M_1\hat{e} \\ &= M_1X_2\hat{\beta}_2 + M_1MY \\ &= M_1X_2\hat{\beta}_2 + MY. \end{aligned} \tag{4.5}$$

Next pre-multiply by X_2' , and note that since $MX_2 = 0$, we obtain

$$\begin{aligned} X_2'M_1Y &= X_2'M_1X_2\hat{\beta}_2 + X_2'MY \\ &= X_2'M_1X_2\hat{\beta}_2. \end{aligned}$$

Inverting $(X_2'M_1X_2)$, we find a new expression for the OLS estimate $\hat{\beta}_2$:

$$\hat{\beta}_2 = (X_2'M_1X_2)^{-1} (X_2'M_1Y). \tag{4.6}$$

Since M_1 is idempotent, we can rewrite this as

$$\begin{aligned} \hat{\beta}_2 &= (X_2'M_1M_1X_2)^{-1} (X_2'M_1M_1Y) \\ &= (\tilde{X}_2'\tilde{X}_2)^{-1} (\tilde{X}_2'\tilde{Y}), \end{aligned}$$

where $\tilde{X}_2 = M_1X_2$ and $\tilde{Y} = M_1Y$. We have shown that the OLS estimate $\hat{\beta}_2$ can alternatively be computed from a simple OLS regression of \tilde{Y} on \tilde{X}_2 . Interestingly, \tilde{Y} is the OLS residual from a regression of Y on X_1 , and \tilde{X}_2 is the OLS residual from a regression of X_2 on X_1 .

Furthermore, (4.5) says that

$$\tilde{Y} = \tilde{X}_2\hat{\beta}_2 + \hat{e}. \tag{4.7}$$

Since we have shown that $\hat{\beta}_2$ is the OLS estimate from a regression of \tilde{Y} on \tilde{X}_2 , the OLS residual from this regression is $\tilde{Y} - \tilde{X}_2\hat{\beta}_2$, and (4.7) shows that this equals \hat{e} , the OLS residual from regression on Y on X . We have proven the following theorem.

Theorem 4.6.1 (*Frisch-Waugh-Lovell*). *In the model (4.3), the OLS estimator of β_2 and the OLS residuals \hat{e} may be equivalently computed by either the OLS regression (4.4) or via the following algorithm:*

1. Regress Y on X_1 , obtain residuals \tilde{Y} ;
2. Regress X_2 on X_1 , obtain residuals \tilde{X}_2 ;
3. Regress \tilde{Y} on \tilde{X}_2 , obtain LS estimates $\hat{\beta}_2$ and residuals \hat{e} .

In some contexts, the FWL theorem can be used to speed computation, but in most cases there is little computational advantage to using the two-step algorithm. Rather, the theorem's primary use is theoretical.

Application: Demeaning

A common application of the FWL theorem, which you may have seen in an introductory econometrics course, is the demeaning formula for regression.

Partition $X = [X_1 \ X_2]$ where $X_1 = \iota$ is a vector of ones, and X_2 is the vector of observed regressors. In this case,

$$M_1 = I - \iota(\iota'\iota)^{-1}\iota'.$$

Observe that

$$\begin{aligned}\tilde{X}_2 &= M_1 X_2 = X_2 - \iota (\iota' \iota)^{-1} \iota' X_2 \\ &= X_2 - \bar{X}_2\end{aligned}$$

and

$$\begin{aligned}\tilde{Y} &= M_1 Y = Y - \iota (\iota' \iota)^{-1} \iota' Y \\ &= Y - \bar{Y},\end{aligned}$$

which are “demeaned”.

The FWL theorem says that $\hat{\beta}_2$ is the OLS estimate from a regression of \tilde{Y} on \tilde{X}_2 , or $y_i - \bar{y}$ on $x_{2i} - \bar{x}_2$:

$$\hat{\beta}_2 = \left(\sum_{i=1}^n (x_{2i} - \bar{x}_2)(x_{2i} - \bar{x}_2)' \right)^{-1} \left(\sum_{i=1}^n (x_{2i} - \bar{x}_2)(y_i - \bar{y}) \right).$$

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

4.7 Analysis of Variance

By definition,

$$Y = \hat{Y} + \hat{e}.$$

Subtracting \bar{y} from the two sides,

$$Y - \bar{y} = (\hat{Y} - \bar{y}) + \hat{e}.$$

Thus

$$(Y - \bar{y})'(Y - \bar{y}) = (\hat{Y} - \bar{y})'(\hat{Y} - \bar{y}) + 2(\hat{Y} - \bar{y})'\hat{e} + \hat{e}'\hat{e}.$$

But $\hat{Y}'\hat{e} = Y'PMY = 0$ and $\bar{y}'\hat{e} = \bar{y}\iota'\hat{e} = 0$ when the model contains an intercept (more generally, if ι lies in the range space of the matrix X). Thus

$$(Y - \bar{y})'(Y - \bar{y}) = (\hat{Y} - \bar{y})'(\hat{Y} - \bar{y}) + \hat{e}'\hat{e}.$$

or alternatively

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n \hat{e}_i^2.$$

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

$$TSS = ESS + RSS,$$

where TSS, ESS, and RSS stand for “Total sum of squares”, “Equation sum of squares” and “Residual sum of squares”, respectively. The equation R^2 is defined as

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}.$$

The equation R^2 measures the percentage of the variance of y_i which is “explained” or accounted for in the variation of the predicted value \hat{y}_i . R^2 is typically reported in applied work. It is frequently

referenced as a “measure of fit”. Unfortunately, this is an inappropriate label, as R^2 does not measure the adequacy or “fit” of the model.

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

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

Another interesting fact about R^2 is that it necessarily increases as regressors are added to the model. This is because by the definition of OLS, the parameter estimates minimize the sum of squared errors, or RSS. By adding additional regressors, the RSS cannot increase; it either can stay the same, or decrease. But TSS is unaffected by adding regressors, so the ratio RSS/TSS can only decrease. Thus R^2 either increases or stays constant. To counter-act this effect, Theil proposed an adjustment, typically called \bar{R}^2 (or “r-bar-squared”) which penalizes model dimensionality. While this statistic is often reported by statistical packages, and used to be reported frequently in applied work, it is not used much today, as better model evaluation criterion (such as the AIC and BIC, which will be discussed later) have been developed.

4.8 Bias

Proposition 4.8.1 *If $Y = X\beta + e$, then $\hat{\beta} - \beta = (X'X)^{-1} X'e$.*

Proof: Since $Y = X\beta + e$,

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1} X'y \\ &= (X'X)^{-1} X'(X\beta + e) \\ &= (X'X)^{-1} X'X\beta + (X'X)^{-1} X'e \\ &= \beta + (X'X)^{-1} X'e.\end{aligned}$$

■

We now explore some implications of the regression restriction $E(e_i | x_i) = 0$.

Proposition 4.8.2 *In the linear regression model, $E(e | X) = 0$.*

Proof: Since the observations are independent, $E(e_i | X) = E(e_i | x_i) = 0$. Thus

$$E(e | X) = \begin{pmatrix} E(e_1 | X) \\ E(e_2 | X) \\ \vdots \\ E(e_n | X) \end{pmatrix} = \begin{pmatrix} E(e_1 | x_1) \\ E(e_2 | x_2) \\ \vdots \\ E(e_n | x_n) \end{pmatrix} = 0.$$

■

Proposition 4.8.3 *In the linear regression model, $E(\hat{\beta} - \beta | X) = 0$ and $E\hat{\beta} = \beta$.*

Proof: By the previous result,

$$\begin{aligned} E(\hat{\beta} - \beta | X) &= E\left((X'X)^{-1} X'e | X\right) \\ &= (X'X)^{-1} X'E(e | X) \\ &= 0. \end{aligned}$$

Then applying the law of iterated expectations, $E\hat{\beta} = E\left(E(\hat{\beta} | X)\right) = \beta$. \blacksquare

Thus $\hat{\beta}$ is *unbiased* for β . Indeed, it is *conditionally unbiased*, conditional upon X , which is a stronger result.

4.9 Variance-Covariance Matrix of Regression Error

The conditional variance-covariance matrix of the regression error vector e is

$$D = E(ee' | X).$$

This $n \times n$ matrix plays an important role in our theory, so we start by exploring its structure.

Writing out the matrix ee' , we find

$$ee' = \begin{bmatrix} e_1^2 & e_1e_2 & e_1e_3 & \cdots & e_1e_n \\ e_2e_1 & e_2^2 & e_2e_3 & \cdots & e_2e_n \\ e_3e_1 & e_3e_2 & e_3^2 & \cdots & e_3e_n \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ e_ne_1 & e_ne_2 & e_ne_3 & \cdots & e_n^2 \end{bmatrix}.$$

If the data are a random sample then (x_i, e_i) is independent of (x_j, e_j) for $i \neq j$, so

$$E(e_i^2 | X) = E(e_i^2 | x_i) = \sigma_i^2,$$

where σ_i^2 was introduced at the end of Section 1. Again by the independence of (x_i, e_i) from (x_j, e_j) for $i \neq j$,

$$E(e_ie_j | X) = E(e_i | x_i)E(e_j | x_j) = 0.$$

Thus in general

$$\begin{aligned} D &= E(ee' | X) \\ &= \begin{bmatrix} E(e_1^2 | X) & E(e_1e_2 | X) & E(e_1e_3 | X) & \cdots & E(e_1e_n | X) \\ E(e_2e_1 | X) & E(e_2^2 | X) & E(e_2e_3 | X) & \cdots & E(e_2e_n | X) \\ E(e_3e_1 | X) & E(e_3e_2 | X) & E(e_3^2 | X) & \cdots & E(e_3e_n | X) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ E(e_ne_1 | X) & E(e_ne_2 | X) & E(e_ne_3 | X) & \cdots & E(e_n^2 | X) \end{bmatrix} \\ &= \begin{bmatrix} \sigma_1^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma_3^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}, \end{aligned} \tag{4.8}$$

which is a diagonal matrix with the conditional variances on the diagonal.

This is a case where the homoskedasticity restriction (3.6) implies a simplification. Under (3.6), $E(e_i^2 | x_i) = \sigma^2$ for all i , so D simplifies to

$$D = \begin{bmatrix} \sigma^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = I_n \sigma^2. \quad (4.9)$$

4.10 Covariance Matrix of OLS Estimator

The conditional covariance matrix for $\hat{\beta}$ is

$$V_n = E \left[(\hat{\beta} - \beta) (\hat{\beta} - \beta)' | X \right].$$

Since $\hat{\beta} - \beta = (X'X)^{-1} X'e$ (Proposition 4.8.1),

$$\begin{aligned} V_n &= E \left[(X'X)^{-1} X'ee'X (X'X)^{-1} | X \right] \\ &= (X'X)^{-1} X'E[ee' | X] X (X'X)^{-1} \\ &= (X'X)^{-1} X'DX (X'X)^{-1}, \end{aligned}$$

where D is defined in (4.8). It may be helpful to observe that the central portion of this expression may be written alternatively as

$$X'DX = \sum_{i=1}^n x_i x_i' \sigma_i^2.$$

In the special case of (3.6), then $\sigma_i^2 = \sigma^2$, so $X'DX = X'X\sigma^2$ and V_n simplifies to

$$\begin{aligned} V_n &= (X'X)^{-1} X'X\sigma^2 (X'X)^{-1} \\ &= \sigma^2 (X'X)^{-1} \\ &\equiv V_n^0. \end{aligned}$$

In general, however, V_n and V_n^0 differ.

We have shown the following.

Theorem 4.10.1 *In the linear regression model,*

$$V_n = (X'X)^{-1} X'DX (X'X)^{-1}. \quad (4.10)$$

If (3.6) holds,

$$V_n = \sigma^2 (X'X)^{-1} \equiv V_n^0. \quad (4.11)$$

The expression $V_n = (X'X)^{-1} X'DX (X'X)^{-1}$ is often called a “sandwich formula”, because the central variance matrix $X'DX$ is “sandwiched” between the moment matrices $(X'X)^{-1}$.

4.11 GLS and the Gauss-Markov Theorem

The linear regression model

$$\begin{aligned} y_i &= x_i' \beta + e_i \\ E(e_i | x_i) &= 0 \end{aligned}$$

imposes the condition of zero conditional mean. This is stronger than the orthogonality condition $E(x_i e_i) = 0$. This stronger condition can be exploited to improve estimation efficiency.

The *Generalized Least Squares* (GLS) estimator of β is

$$\tilde{\beta} = (X'D^{-1}X)^{-1} (X'D^{-1}Y). \quad (4.12)$$

The GLS estimator is sometimes called the Aitken estimator.

Since $Y = X\beta + e$, thus

$$\tilde{\beta} = \beta + (X'D^{-1}X)^{-1} (X'D^{-1}e).$$

Since D is a function of X , $E(\tilde{\beta} | X) = \beta$ and

$$\begin{aligned} \text{Var}(\tilde{\beta} | X) &= (X'D^{-1}X)^{-1} X'D^{-1}DD^{-1}X (X'D^{-1}X)^{-1} \\ &= (X'D^{-1}X)^{-1}. \end{aligned}$$

The class of unbiased linear estimators take the form

$$\tilde{\beta}_L = A(X)'Y, \quad A(X)'X = I_k$$

where $A(X)$, $n \times k$, is a function only of X . This is called *linear* because it is a linear function of Y , even though it is nonlinear in X . OLS is the case $A(X) = X(X'X)^{-1}$ and GLS is the case $A(X) = D^{-1}X (X'D^{-1}X)^{-1}$. Observe that

$$E(\tilde{\beta}_L | X) = A(X)'X\beta = \beta$$

so $\tilde{\beta}_L$ is unbiased. Thus $\tilde{\beta}_L = \beta + A(X)'e$, and its variance is

$$\text{Var}(\tilde{\beta}_L | X) = A(X)'DA(X).$$

The “best” estimator within this class is the one with the smallest variance.

Theorem 4.11.1 (*Gauss-Markov*). *The best (minimum-variance) unbiased linear estimator is GLS.*

Proof. Let $A^* = D^{-1}X (X'D^{-1}X)^{-1}$ and A be any other $n \times k$ function of X such that $A'X = I_k$. We need to show that $A'DA \geq A^*DA^*$.

Let $C = A - A^*$. Note that

$$\begin{aligned} C'X &= A'X - A^*X \\ &= I_k - I_k = 0 \end{aligned}$$

and

$$\begin{aligned} C'DA^* &= C'DD^{-1}X (X'D^{-1}X)^{-1} \\ &= C'X (X'D^{-1}X)^{-1} = 0. \end{aligned}$$

Then

$$\begin{aligned}
A'DA &= (C + A^*)' D (C + A^*) \\
&= C'DC + C'DA^* + A^{*'}DC + A^{*'}DA^* \\
&= C'DC + A^{*'}DA^* \geq A^*DA^{*'}.
\end{aligned}$$

■

Theoretically speaking, the Gauss-Markov theorem is not a very powerful theorem, because the restriction to linear estimators is quite unnatural. That is, perhaps a “nonlinear” estimator can do even better. However, at least the theorem points to the inefficiency of OLS in regression models.

Chamberlain (1987) established the general result. He showed that in the regression model, no regular consistent estimator can have a lower asymptotic variance than the GLS estimator. This establishes that the GLS estimator is asymptotically efficient. The proof of his theorem is quite deep and we cannot cover it here.

Except in the special case of homoskedastic errors (where $D = I\sigma^2$ and GLS=OLS), these results show that in the regression model, OLS is inefficient. Unfortunately, it is not trivial to achieve the efficiency improvement, as GLS is not feasible as the matrix D is unknown. This raises the difficult issue of implementation.

Earlier, we claimed that OLS is asymptotically efficient in the class of models with $E(x_i e_i) = 0$. In this section we claimed that OLS is inefficient (but GLS is efficient) if $E(e_i | x_i) = 0$. The gain of efficiency (through use of GLS) comes through the exploitation of the stronger conditional mean assumption, which has the cost of reduced robustness. If $E(e_i | x_i) \neq 0$ then the GLS estimator will be inconsistent for the projection coefficient β , but OLS will be consistent.

4.12 Estimation of Error Variance

The regression error e_i is mean-zero and has unconditional variance

$$Ee_i^2 = \sigma^2.$$

This parameter is well defined in both the homoskedastic and heteroskedastic regression contexts. In the homoskedastic case, $\sigma^2 = E(e_i^2 | x_i) = Ee_i^2$, and in the heteroskedastic case, $\sigma^2 = E[E(e_i^2 | x_i)]$ (by the law of iterated expectations). In either case, σ^2 measures the variation in the “unexplained” part of the regression. In the homoskedastic regression model, V_n^0 is proportional to σ^2 , so an estimate of σ^2 is required to obtain an estimate of V_n^0 . In either case, estimation of σ^2 is valuable.

A standard estimator for σ^2 is the sample average

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \hat{e}_i^2 = \frac{1}{n} \hat{e}'\hat{e}.$$

Since $Y = X\beta + e$ and $MX = 0$,

$$\hat{e} = MY = M(X\beta + e) = Me.$$

We see therefore that

$$n\hat{\sigma}^2 = \hat{e}'\hat{e} = e'MMe = e'Me, \tag{4.13}$$

and

$$\begin{aligned}
E(n\hat{\sigma}^2 | X) &= E(e'Me | X) \\
&= E(\text{tr}(e'Me) | X) \\
&= E(\text{tr}(Mee') | X) \\
&= \text{tr}[E(Mee' | X)] \\
&= \text{tr}[ME(ee' | X)] \\
&= \text{tr}[MD],
\end{aligned}$$

where D is defined in (4.8).

In the general case of a heteroskedastic regression model, this expression does not appear to simplify further, so we do not have a clear formula for the bias of $\hat{\sigma}^2$. However, in the special case of the homoskedastic regression (3.6), then D simplifies to $D = I_n\sigma^2$ (see (4.9)) and then

$$\begin{aligned}
E(n\hat{\sigma}^2 | X) &= \text{tr}[MD] \\
&= \text{tr}[MI_n\sigma^2] \\
&= \sigma^2 \text{tr}[M] \\
&= \sigma^2(n - k).
\end{aligned}$$

The final equality holds by (1.2). We have proved that:

Theorem 4.12.1 *In the homoskedastic regression model,*

$$E\hat{\sigma}^2 = \frac{(n - k)}{n}\sigma^2.$$

Since $\frac{(n-k)}{n} < 1$, $\hat{\sigma}^2$ is a biased estimator for σ^2 . The extent of the bias will depend on the ratio k/n . Based on this formula, it is simple to see how to construct an unbiased estimator:

$$\begin{aligned}
s^2 &= \frac{n}{n - k}\hat{\sigma}^2 \\
&= \frac{1}{n - k}\hat{e}'\hat{e}.
\end{aligned}$$

The estimator s^2 is the most typical estimate of the regression error variance.

This estimator is frequently referred to as “the unbiased estimator” because of Theorem 4.12.1. It is important to understand, however, that s^2 is unbiased only in the special case that (3.6) holds, and not generally.

4.13 Covariance Matrix Estimation

In a previous section, we derived two expressions for the covariance matrix of $\hat{\beta}$, V_n^0 and V_n , where the former is valid only under homoskedasticity.

Given our estimator s^2 for σ^2 , it is straightforward to construct an estimator of V_n^0 . It is

$$\hat{V}_n^0 = s^2 (X'X)^{-1}. \tag{4.14}$$

Estimating $V_n = (X'X)^{-1} X'DX (X'X)^{-1}$ is less straightforward, as $X'DX$ is unknown. Note, however, that we can compute its mean

$$\begin{aligned} \frac{1}{n}EX'DX &= E(x_i x_i' \sigma_i^2) \\ &= E(x_i x_i' E(e_i^2 | x_i)) \\ &= E(x_i x_i' e_i^2) \\ &\equiv \Omega. \end{aligned}$$

While Ω is unknown, we can easily estimate it by a sample average

$$\hat{\Omega} = \frac{1}{n} \sum_{i=1}^n x_i x_i' \hat{e}_i^2$$

where \hat{e}_i are the OLS residuals. Another way of writing this is

$$\hat{\Omega} = \frac{1}{n} X' \hat{D} X$$

where

$$\hat{D} = \begin{bmatrix} \hat{e}_1^2 & 0 & 0 & \cdots & 0 \\ 0 & \hat{e}_2^2 & 0 & \cdots & 0 \\ 0 & 0 & \hat{e}_3^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \hat{e}_n^2 \end{bmatrix},$$

which bears some algebraic similarity to the desired matrix $X'DX$.

The corresponding estimator of V_n is

$$\begin{aligned} \hat{V}_n &= (X'X)^{-1} n \hat{\Omega} (X'X)^{-1} \\ &= (X'X)^{-1} (X' \hat{D} X) (X'X)^{-1}. \end{aligned} \tag{4.15}$$

While equation (4.15) may be useful for understanding the motivation, it is inappropriate for computation as the $n \times n$ matrix \hat{D} is too large for storage in practice when n is large. A more useful computational formula is to define $\hat{u}_i = x_i \hat{e}_i$ and the $n \times k$ matrix

$$\hat{u} = \begin{pmatrix} \hat{u}'_1 \\ \hat{u}'_2 \\ \vdots \\ \hat{u}'_n \end{pmatrix}.$$

Then we can see that $X' \hat{D} X = \hat{u}' \hat{u}$, and

$$\hat{V}_n = (X'X)^{-1} (\hat{u}' \hat{u}) (X'X)^{-1}. \tag{4.16}$$

This estimator was introduced to the econometrics literature by White (1980).

The estimator \hat{V}_n^0 was the dominate covariance estimator used before 1980, and was still the standard choice in the 1980s. From my reading of the literature, the White estimate \hat{V}_n started to come in common use in the early 1990s, and by the late 1990s is quite commonly used, especially by younger researchers. When reading and reporting applied work, it is important to pay attention to

the distinction between \hat{V}_n^0 and \hat{V}_n , as it is not always clear which has been used. When \hat{V}_n is used rather than the traditional choice \hat{V}_n^0 , many authors will state that “their standard errors have been corrected for heteroskedasticity”, or that they use a “heteroskedasticity-robust covariance matrix estimator”, or that they use the “White formula”, the “Eicker-White formula”, the “Huber formula”, the “Huber-White formula” or the “GMM covariance matrix”. In most cases, these all mean the same thing.

4.14 Standard Errors

The variance estimators \hat{V}_n^0 and \hat{V}_n are estimates of the variance of the distribution of $\hat{\beta}$ (under homoskedasticity). A more easily interpretable measure of spread is its square root – the standard deviation. This motivates the definition of a standard error.

Definition 4.14.1 *A standard error $s(\hat{\beta})$ for an estimator $\hat{\beta}$ is an estimate of the standard deviation of the distribution of $\hat{\beta}$.*

When β is scalar, and $\hat{V}(\hat{\beta})$ is an estimator of the variance of $\hat{\beta}$, we set $s(\hat{\beta}) = \sqrt{\hat{V}(\hat{\beta})}$. When β is a vector, we focus on individual elements of β one-at-a-time, vis., β_j , $j = 1, \dots, k$. Given our estimator (4.16), the standard error for $\hat{\beta}_j$ is

$$s(\hat{\beta}_j) = \sqrt{[\hat{V}_n]_{jj}}$$

and similarly if (4.14) is used.

Generically, standard errors are not unique, as there may be more than one estimator of the variance of the estimator. It is therefore important to understand what formula and method is used by an author when studying their work. It is also important to understand that a particular standard error may be relevant under one set of model assumptions, but not under another set of assumptions, just as any other estimator.

From a computational standpoint, the standard method to calculate the standard errors is to first calculate \hat{V}_n , then take the diagonal elements, and then the square roots.

Chapter 5

Asymptotic Distribution Theory

In this section, we explore properties of the OLS estimator using a “large sample” approximation. For these results, we do not need the conditional mean restriction $E(e_i | x_i) = 0$. Rather, the orthogonal condition $E(x_i e_i) = 0$ is sufficient.

5.1 Consistency of the OLS Estimator

If $\hat{\beta}$ is an estimator for β , we say that $\hat{\beta}$ is *consistent* for β if $\hat{\beta} \rightarrow_p \beta$ as $n \rightarrow \infty$. This says that for a sufficiently large sample, the estimator will be arbitrarily close to the true value. It is generally believed that consistency is a good property for an estimator to possess. To prove the consistency of the OLS estimator, it turns out that the following assumption is appropriate.

Assumption 5.1.1 $E(x_i e_i) = 0$, $E e_i^2 < \infty$, $E x_i' x_i < \infty$, and $Q = E x_i x_i' > 0$.

The assumptions that $E x_i' x_i < \infty$ and $E e_i^2 < \infty$ mean that all elements of x_i and e_i have finite second moments, and all cross-moments are finite. To see this, first observe that since

$$E x_i' x_i = E x_{1i}^2 + \dots + E x_{ki}^2 < \infty,$$

then it is the case that for all $j = 1, \dots, k$, $E x_{ji}^2 < \infty$. By the Cauchy-Schwarz inequality, for each j and l ,

$$E |x_{ji} x_{li}| \leq E |x_{ji}^2|^{1/2} E |x_{li}^2|^{1/2} < \infty. \quad (5.1)$$

And for each j ,

$$E |x_{ji} e_i| \leq E |x_{ji}^2|^{1/2} E |e_i^2|^{1/2} < \infty. \quad (5.2)$$

We can now deduce the consistency of $\hat{\beta}$

Theorem 5.1.1 *Under Assumption 5.1.1, as $n \rightarrow \infty$*

1. $\frac{1}{n} X' X \rightarrow_p Q$;
2. $\frac{1}{n} X' e \rightarrow_p 0$;
3. $\hat{\beta} \rightarrow_p \beta$.

Proof: The bound (5.1) and the WLLN directly imply that for all j and l

$$\frac{1}{n} \sum_{i=1}^n x_{ji}^2 \rightarrow_p E x_{ji}^2$$

and

$$\frac{1}{n} \sum_{i=1}^n x_{ji} x_{li} \rightarrow_p E x_{ji} x_{li}.$$

Hence

$$\begin{aligned} \frac{1}{n} X'X &= \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n x_{1i}^2 & \frac{1}{n} \sum_{i=1}^n x_{1i} x_{2i} & \cdots & \frac{1}{n} \sum_{i=1}^n x_{1i} x_{ki} \\ \frac{1}{n} \sum_{i=1}^n x_{2i} x_{1i} & \frac{1}{n} \sum_{i=1}^n x_{2i}^2 & \cdots & \frac{1}{n} \sum_{i=1}^n x_{2i} x_{ki} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{1}{n} \sum_{i=1}^n x_{ki} x_{1i} & \frac{1}{n} \sum_{i=1}^n x_{ki} x_{2i} & \cdots & \frac{1}{n} \sum_{i=1}^n x_{ki}^2 \end{pmatrix} \\ &\rightarrow_p \begin{pmatrix} E x_{1i}^2 & E x_{1i} x_{2i} & \cdots & E x_{1i} x_{ki} \\ E x_{2i} x_{1i} & E x_{2i}^2 & \cdots & E x_{2i} x_{ki} \\ \vdots & \vdots & \cdots & \vdots \\ E x_{ki} x_{1i} & E x_{ki} x_{2i} & \cdots & E x_{ki}^2 \end{pmatrix} \\ &= E \begin{pmatrix} x_{1i}^2 & x_{1i} x_{2i} & \cdots & x_{1i} x_{ki} \\ x_{2i} x_{1i} & x_{2i}^2 & \cdots & x_{2i} x_{ki} \\ \vdots & \vdots & \cdots & \vdots \\ x_{ki} x_{1i} & x_{ki} x_{2i} & \cdots & x_{ki}^2 \end{pmatrix} \\ &= E (x_i x_i') = Q, \end{aligned}$$

which is the first result.

The bound (5.2) and the WLLN imply that for all j

$$\frac{1}{n} \sum_{i=1}^n x_{ji} e_i \rightarrow_p E (x_{ji} e_i) = 0.$$

Hence

$$\frac{1}{n} X'e = \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n x_{1i} e_i \\ \frac{1}{n} \sum_{i=1}^n x_{2i} e_i \\ \vdots \\ \frac{1}{n} \sum_{i=1}^n x_{ki} e_i \end{pmatrix} \rightarrow_p 0,$$

which is the second result.

Finally, by Proposition 4.8.1

$$\begin{aligned} \hat{\beta} &= \beta + (X'X)^{-1} (X'e) \\ &= \beta + \left(\frac{1}{n} X'X \right)^{-1} \left(\frac{1}{n} X'e \right) \\ &= \beta + g \left(\frac{1}{n} X'X, \frac{1}{n} X'e \right) \end{aligned}$$

where $g(A, b) = A^{-1}b$ is a continuous function of A and b , at all values of the arguments such that A^{-1} exist. Now by the first two parts of this theorem,

$$\left(\frac{1}{n} X'X, \frac{1}{n} X'e \right) \rightarrow_p (Q, 0).$$

Under Assumption 5.1.1, $Q > 0$, so Q^{-1} exists and $g(\cdot, \cdot)$ is continuous at $(Q, 0)$. Hence by the CMT,

$$g\left(\frac{1}{n}X'X, \frac{1}{n}X'e\right) \rightarrow_p g(Q, 0) = Q^{-1}0 = 0$$

so $\hat{\beta} = \beta + g\left(\frac{1}{n}X'X, \frac{1}{n}X'e\right) \rightarrow_p \beta + 0 = 0$, which is the third result. \blacksquare

5.2 Asymptotic Distribution of OLS Estimator

We have shown that in the context of the general regression model, $\hat{\beta}$ is a random vector with a conditional mean of β and a conditional variance of V_n . This alone gives a good sense of the accuracy of the estimator $\hat{\beta}$, but is not sufficient to construct accurate confidence intervals. For this, we need a good approximation to the distribution of $\hat{\beta}$.

Except in special cases, the exact distribution of $\hat{\beta}$ is unknown. Therefore we rely on approximations, and use a variety of techniques to assess the accuracy of these approximations. The dominate approximation technique is called asymptotic theory, and is based on calculating the limiting distribution of a normalized version of $\hat{\beta}$ as the sample size n tends to positive infinity. The heuristic idea is that this approximation is likely to be accurate in contexts where the sample size n is “large”, which is common in many econometric studies.

We need a stronger set of assumptions than for consistency. Define

$$\Omega = E(x_i x_i' \sigma_i^2) \tag{5.3}$$

$$\begin{aligned} &= E(x_i x_i' E(e_i^2 | x_i)) \\ &= E(E(x_i x_i' e_i^2 | x_i)) \\ &= E(x_i x_i' e_i^2), \end{aligned} \tag{5.4}$$

the last equality by the law of iterated expectations. Ω has finite elements (in general) if x_i and e_i have finite fourth moments, which is a strengthening of Assumption 5.1.1.

Assumption 5.2.1 $E(x_i e_i) = 0$, $E e_i^4 < \infty$, $E|x_i|^4 < \infty$, $Q = E x_i x_i' > 0$, and $\Omega = E(x_i x_i' e_i^2) > 0$.

The final statement, that $\Omega > 0$, is not needed for many results, but is important for the development of test statistics. If $Q > 0$ and $\sigma^2 > 0$, it is typically believed that requiring $\Omega > 0$ is not meaningfully restrictive.

The following result is helpful.

Theorem 5.2.1 Under Assumption 5.2.1, for $u_i = x_i e_i$,

1. u_i is iid, $E u_i = 0$, $E u_i' u_i < \infty$, and $E u_i u_i' = \Omega$;
2. $\frac{1}{\sqrt{n}} \sum_{i=1}^n u_i \rightarrow_d N(0, \Omega)$.

Proof: First, since (y_i, x_i) is iid, so is any function of (y_i, x_i) , including $u_i = x_i(y_i - x_i \beta)$. Second, $E u_i = E x_i e_i = 0$ under Assumption 5.2.1. Third, by the Cauchy-Schwarz inequality and Assumption 5.2.1,

$$\begin{aligned} E|u_i' u_i| &= E|u_i u_i'| = E|x_i x_i' e_i^2| \\ &\leq \left(E|x_i x_i'|^2\right)^{1/2} \left(E|e_i^4|\right)^{1/4} \\ &= \left(E|x_i|^4\right)^{1/2} \left(E|e_i^4|\right)^{1/4} < \infty. \end{aligned}$$

Fourth, $E u_i u_i' = E x_i x_i' e_i^2 = \Omega$ by (5.4). This establishes the first statement. The second statement follows by the CLT. ■

We now can give the asymptotic distribution of the OLS estimator.

Theorem 5.2.2 *Under Assumption 5.2.1,*

$$\sqrt{n}(\hat{\beta} - \beta) \rightarrow_d N(0, V)$$

where $V = Q^{-1}\Omega Q^{-1}$.

Proof: Using Proposition 4.8.1, Theorem 5.1.1, the CLT, and the CMT,

$$\begin{aligned} \sqrt{n}(\hat{\beta} - \beta) &= \left(\frac{1}{n}X'X\right)^{-1} \left(\frac{1}{\sqrt{n}}X'e\right) \\ &\rightarrow_d Q^{-1}N(0, \Omega) \\ &= N(0, Q^{-1}\Omega Q^{-1}). \end{aligned}$$

■

As V is the variance of the asymptotic distribution of $\sqrt{n}(\hat{\beta} - \beta)$, V is often referred to as the *asymptotic covariance matrix* of $\hat{\beta}$. The form $V = Q^{-1}\Omega Q^{-1}$ is called a *sandwich* form.

It is worth mentioning that under the homoskedasticity restriction (3.6), V simplifies. Indeed, (3.6) implies that

$$\Omega = E(x_i x_i' \sigma_i^2) = E(x_i x_i' \sigma^2) = \sigma^2 E(x_i x_i') = \sigma^2 Q,$$

so

$$V = Q^{-1}\Omega Q^{-1} = Q^{-1}(\sigma^2 Q) Q^{-1} = \sigma^2 Q^{-1} \equiv V^0.$$

In general, however, the two asymptotic variance expressions differ.

5.3 Consistent Covariance Matrix Estimation

To study the consistency of $\hat{\sigma}^2$ and s^2 , we return to equation (4.13), which is an algebraic equality, and see that

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n}e'Me \\ &= \frac{1}{n}e'e - \left(\frac{1}{n}e'X\right) \left(\frac{1}{n}X'X\right)^{-1} \left(\frac{1}{n}X'e\right). \end{aligned} \tag{5.5}$$

A simple application of the WLLN and the CMT yields the consistency of $\hat{\sigma}^2$ and s^2 .

Theorem 5.3.1 *Under Assumption 5.1.1, as $n \rightarrow \infty$*

1. $\frac{1}{n}e'e \rightarrow_p \sigma^2$;
2. $\hat{\sigma}^2 \rightarrow_p \sigma^2$;
3. $s^2 \rightarrow_p \sigma^2$.

Proof: Assumption 5.1.1 specifies that $Ee_i^2 = \sigma^2 < \infty$, so an application of the WLLN yields

$$\frac{1}{n} \sum_{i=1}^n e_i^2 \rightarrow_p Ee_i^2 = \sigma^2$$

as $n \rightarrow \infty$, which is the first result.

Observe that (5.5) can be written as

$$\hat{\sigma}^2 = g \left(\frac{1}{n} e'e, \frac{1}{n} X'e, \frac{1}{n} X'X \right),$$

where $g(a, b, c) = a - b'c^{-1}b$ is a continuous function for all values of its arguments such that c^{-1} exists. Theorem 5.1.1 and the first result of this theorem show that as $n \rightarrow \infty$

$$\left(\frac{1}{n} e'e, \frac{1}{n} X'e, \frac{1}{n} X'X \right) \rightarrow_p (\sigma^2, 0, Q).$$

Since Q^{-1} exists under Assumption 5.1.1, the function g is continuous at the limit $(\sigma^2, 0, Q)$, so we can apply the CMT to discover that

$$\hat{\sigma}^2 = g \left(\frac{1}{n} e'e, \frac{1}{n} X'e, \frac{1}{n} X'X \right) \rightarrow_p g(\sigma^2, 0, Q) = \sigma^2 - 0'Q^{-1}0 = \sigma^2$$

as $n \rightarrow \infty$, which is the second stated result.

Finally, since $\frac{n}{n-k} = \frac{1}{1-k/n} \rightarrow 1$ as $n \rightarrow \infty$,

$$s^2 = \frac{n}{n-k} \hat{\sigma}^2 \rightarrow 1 \cdot \sigma^2 = \sigma^2,$$

which is the final result. ■

Theorem 5.3.1 helps illustrate the distinctions between unbiased estimators and consistent estimators. In general, neither $\hat{\sigma}^2$ nor s^2 are necessarily unbiased, but under the homoskedasticity restriction (3.6), s^2 is unbiased and $\hat{\sigma}^2$ is biased. In general, however, both are consistent for σ^2 . This consistency property is driven by the fact that for sufficiently large n , the scale factor $\frac{n}{n-k}$ cannot have a big effect.

The practical distinction between s^2 and $\hat{\sigma}^2$ arises for cases where k/n is relatively large, which occurs when either the sample size n is quite small or when the parameterization is tied to the sample size, as in fixed effects panel regressions (discussed later in this course).

In earlier sections, we found the exact conditional covariance matrix of $\hat{\beta}$, and the variance of the asymptotic distribution of $\hat{\beta}$. In order to use these results to conduct inference on β , we need to estimate this covariance matrix.

In Chapter 4, we introduced the White covariance matrix estimator \hat{V}_n . We now demonstrate that it is a good estimator of the variance of $\hat{\beta}$.

Theorem 5.3.2 *Under Assumption 5.2.1, as $n \rightarrow \infty$,*

1. $\hat{\Omega} \rightarrow_p \Omega$;
2. $n\hat{V} \rightarrow_p V$.

Proof: The main job is to show that $\hat{\Omega} \rightarrow_p \Omega$. Since

$$\begin{aligned}\hat{e}_i &= y_i - x_i' \hat{\beta} \\ &= e_i - x_i' (\hat{\beta} - \beta),\end{aligned}$$

then

$$\begin{aligned}\hat{e}_i^2 &= \left(e_i - x_i' (\hat{\beta} - \beta) \right)^2 \\ &= e_i^2 - 2 (\hat{\beta} - \beta)' x_i e_i + (\hat{\beta} - \beta)' x_i x_i (\hat{\beta} - \beta).\end{aligned}$$

Hence

$$\begin{aligned}\hat{\Omega} &= \frac{1}{n} \sum_{i=1}^n x_i x_i' \hat{e}_i^2 \\ &= \frac{1}{n} \sum_{i=1}^n x_i x_i' e_i^2 - \frac{2}{n} \sum_{i=1}^n x_i x_i' (\hat{\beta} - \beta)' x_i e_i + \frac{1}{n} \sum_{i=1}^n x_i x_i' (\hat{\beta} - \beta)' x_i x_i (\hat{\beta} - \beta).\end{aligned}\quad (5.6)$$

First, Theorem 5.2.1 showed that $E |x_i x_i' e_i^2| < \infty$, so we can apply the WLLN to find that

$$\frac{1}{n} \sum_{i=1}^n x_i x_i' e_i^2 \rightarrow_p E (x_i x_i' e_i^2) = \Omega.$$

Second, by Holder's inequality

$$E \left(|x_i|^3 |e_i| \right) \leq \left(E |x_i|^4 \right)^{3/4} \left(E |e_i|^4 \right)^{1/4} < \infty,$$

so

$$\frac{1}{n} \sum_{i=1}^n |x_i|^3 |e_i| \rightarrow_p E \left(|x_i|^3 |e_i| \right),$$

and thus since $|\hat{\beta} - \beta| \rightarrow_p 0$,

$$\left| \frac{1}{n} \sum_{i=1}^n x_i x_i' (\hat{\beta} - \beta)' x_i e_i \right| \leq |\hat{\beta} - \beta| \left(\frac{1}{n} \sum_{i=1}^n |x_i|^3 |e_i| \right) \rightarrow_p 0.$$

Third, by the WLLN

$$\frac{1}{n} \sum_{i=1}^n |x_i|^4 \rightarrow_p E |x_i|^4,$$

so

$$\left| \frac{1}{n} \sum_{i=1}^n x_i x_i' (\hat{\beta} - \beta)' x_i x_i (\hat{\beta} - \beta) \right| \leq |\hat{\beta} - \beta|^2 \frac{1}{n} \sum_{i=1}^n |x_i|^4 \rightarrow_p 0.$$

Together, we find that the RHS of (5.6) converges in probability to Ω , which is the first result.

Then applying Theorem 5.1.1,

$$n\hat{V} = \left(\frac{1}{n} X'X \right)^{-1} \hat{\Omega} \left(\frac{1}{n} X'X \right)^{-1} \rightarrow_p Q^{-1} \Omega Q^{-1} = V,$$

as stated. ■

5.4 Studentized Statistic

For simplicity, suppose that β is real-valued. Equivalently, in the vector case, we are focusing on one element of β .

The asymptotic distribution of $\hat{\beta}$ depends on the unknown variance V . Thus the distribution of $\hat{\beta}$ is not suitable for testing or construction of confidence intervals. Gosset had the brilliant suggestion to study the studentized statistic

$$t_n(\beta) = \frac{\hat{\beta} - \beta}{s(\hat{\beta})}$$

where $s(\hat{\beta}) = \sqrt{\hat{V}}$ is a standard error for $\hat{\beta}$, since this statistic has a distribution which does not (asymptotically) depend on unknowns. In special cases (such as the normal regression model), the statistic t_n has an exact t distribution, and is therefore exactly free of unknowns. In this case, we say that t_n is a *pivotal* statistic. More generally, the exact distribution of t_n is unknown, but its asymptotic distribution is known. In this case, we say that t is *asymptotically pivotal*. Formally:

Theorem 5.4.1 *Under Assumption 5.2.1, $t_n(\beta) \rightarrow_d Z \sim N(0, 1)$ as $n \rightarrow \infty$.*

Proof. Theorems 5.2.2 and 5.3.2 showed that

$$\begin{aligned} \sqrt{n}(\hat{\beta} - \beta) &\rightarrow_d N(0, V) \\ n\hat{V} &\rightarrow_p V \end{aligned}$$

and by definition, $s(\hat{\beta}) = \sqrt{\hat{V}}$. Thus

$$\begin{aligned} t_n(\beta) &= \frac{\hat{\beta} - \beta}{s(\hat{\beta})} \\ &= \frac{\sqrt{n}(\hat{\beta} - \beta)}{\sqrt{n\hat{V}}} \\ &\rightarrow_d \frac{N(0, V)}{\sqrt{V}} = N(0, 1). \end{aligned}$$

■

5.5 Asymptotic Confidence Interval

A confidence interval C_n is an interval estimate of β , and is a function of the data and hence is random. It is designed to cover β with high probability. Either $\beta \in C_n$ or $\beta \notin C_n$. The coverage probability is $P(\beta \in C_n)$.

We typically cannot calculate the exact coverage probability $P(\beta \in C_n)$. However we often can calculate $\lim_{n \rightarrow \infty} P(\beta \in C_n)$. We call this the asymptotic coverage probability. We say that C_n has asymptotic $(1 - \alpha)\%$ coverage for β if $P(\beta \in C_n) \rightarrow 1 - \alpha$ as $n \rightarrow \infty$.

When $\hat{\beta}$ is asymptotically normally distributed and has standard error $s(\hat{\beta})$, the standard $(1 - \alpha)\%$ asymptotic confidence interval for β is

$$\begin{aligned} C_n &= \left[\hat{\beta} - z_{\alpha/2}s(\hat{\beta}), \hat{\beta} + z_{\alpha/2}s(\hat{\beta}) \right] \\ &= \left[\hat{\beta} \pm z_{\alpha/2}s(\hat{\beta}) \right] \end{aligned} \tag{5.7}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution. That is, if $Z \sim N(0, 1)$, $P(Z > z_{\alpha/2}) = \alpha/2$. For example, $z_{.025} = 1.96$ and $z_{.05} = 1.645$.

Theorem 5.5.1 *If C_n is defined by (5.7), then $P(\beta \in C_n) \rightarrow 1 - \alpha$ as $n \rightarrow \infty$.*

Proof. Since $t_n(\beta) \rightarrow_d Z \sim N(0, 1)$ by Theorem 5.4.1,

$$\begin{aligned}
 P(\beta \in C_n) &= P\left(\beta \in \left[\hat{\beta} - z_{\alpha/2}s(\hat{\beta}), \hat{\beta} + z_{\alpha/2}s(\hat{\beta})\right]\right) \\
 &= P\left(\hat{\beta} - z_{\alpha/2}s(\hat{\beta}) \leq \beta \leq \hat{\beta} + z_{\alpha/2}s(\hat{\beta})\right) \\
 &= P\left(-z_{\alpha/2} \leq \frac{\beta - \hat{\beta}}{s(\hat{\beta})} \leq z_{\alpha/2}\right) \\
 &= P(|t_n(\beta)| \leq z_{\alpha/2}) \\
 &\rightarrow P(|Z| \leq z_{\alpha/2}) \\
 &= 1 - 2P(Z > z_{\alpha/2}) \\
 &= 1 - \alpha.
 \end{aligned}$$

■

While there is no hard-and-fast guideline for the coverage probability $1 - \alpha$, the most common professional choice is 95%, or $\alpha = .05$. This corresponds to selecting the confidence interval $[\hat{\beta} \pm 1.96s(\hat{\beta})] \approx [\hat{\beta} \pm 2s(\hat{\beta})]$. Thus values of β within two standard errors of the estimated $\hat{\beta}$ are considered “reasonable” candidates for the true value β , and values of β outside two standard errors of the estimated $\hat{\beta}$ are considered unlikely or unreasonable candidates for the true value.

5.6 Functions of Parameters

Sometimes we are interested in some function of the parameter vector. Let $h : R^k \rightarrow R^q$, and

$$\theta = h(\beta).$$

We will assume from now on that $h(\beta)$ is continuously differentiable at the true value of β .

The estimate of θ is

$$\hat{\theta} = h(\hat{\beta}).$$

What is an appropriate standard error for $\hat{\theta}$?

The following result is often referred to as “the delta method”.

Theorem 5.6.1 *If $\sqrt{n}(\hat{\beta} - \beta) \rightarrow_d N(0, V)$ then*

$$\sqrt{n}(\hat{\theta} - \theta) \rightarrow_d N(0, V_\theta)$$

where

$$V_\theta = H'_\beta V H_\beta$$

and

$$H_\beta = \frac{\partial}{\partial \beta} h(\beta_0). \quad k \times q.$$

Proof. By a first-order Taylor series approximation:

$$h(\hat{\beta}) \simeq h(\beta) + H'_\beta (\hat{\beta} - \beta).$$

Thus

$$\begin{aligned} \sqrt{n}(\hat{\theta} - \theta) &= \sqrt{n}(h(\hat{\beta}) - h(\beta)) \\ &\simeq H'_\beta \sqrt{n}(\hat{\beta} - \beta) \\ &\rightarrow_d H'_\beta N(0, V) \\ &= N(0, H'_\beta V H_\beta). \end{aligned}$$

■

If \hat{V} is the estimated covariance matrix for $\hat{\beta}$, then the natural estimate for the variance of $\hat{\theta}$ is

$$\hat{V}_\theta = \hat{H}'_\beta \hat{V} \hat{H}_\beta$$

where

$$\hat{H}_\beta = \frac{\partial}{\partial \beta} h(\hat{\beta}).$$

In many cases, the function $h(\beta)$ is linear:

$$h(\beta) = R'\beta$$

for some $k \times q$ matrix R . In this case, $H_\beta = R$ and $\hat{H}_\beta = R$, so $\hat{V}_\theta = R'\hat{V}R$.

For example, if R is a “selector matrix”

$$R = \begin{pmatrix} I \\ 0 \end{pmatrix}$$

so that if $\beta = (\beta_1, \beta_2)$, then $\theta = R'\beta = \beta_1$ and

$$\hat{V}_\theta = \begin{pmatrix} I & 0 \end{pmatrix} \hat{V} \begin{pmatrix} I \\ 0 \end{pmatrix} = \hat{V}_{11},$$

the upper-left block of \hat{V} .

When $q = 1$ (so $h(\beta)$ is real-valued), the standard error for $\hat{\theta}$ is the square root of \hat{V}_θ , that is, $s(\hat{\theta}) = \sqrt{\hat{H}'_\beta \hat{V} \hat{H}_\beta}$. An asymptotic 95% confidence interval for θ is

$$\left[\hat{\theta} \pm 2s(\hat{\theta}) \right] = \left[h(\hat{\beta}) \pm 2\sqrt{\hat{H}'_\beta \hat{V} \hat{H}_\beta} \right].$$

Example: Conditional Mean

In the linear regression model,

$$g(x) = E(y_i | x_i = x) = x'\beta.$$

In some cases, we want to estimate $g(x)$ at a particular point x . Notice that this is a (linear) function of β . Letting $h(\beta) = x'\beta$ and $\theta = h(\beta)$, we see that $\hat{g}(x) = \hat{\theta} = x'\hat{\beta}$ and $H_\beta = x$, so $s(\hat{\theta}) = \sqrt{x'\hat{V}x}$. Thus an asymptotic 95% confidence interval for $g(x)$ is

$$\left[x'\hat{\beta} \pm 2\sqrt{x'\hat{V}x} \right].$$

It is interesting to observe that if this is viewed as a function of x , the width of the confidence set is dependent on x .

Example: Forecast Intervals

For a given value of $x_i = x$, we may want to forecast (guess) y_i out-of-sample. A reasonable guess is the conditional mean $g(x)$, and indeed this is the mean-square-minimizing decision rule. Thus a point forecast is $\hat{g}(x) = x'\hat{\beta}$, the estimated conditional mean, as discussed in the previous sub-section. We would also like a measure of uncertainty for the forecast.

The forecast error is $\hat{e}_i = y_i - \hat{g}(x) = e_i - x'(\hat{\beta} - \beta)$. As the out-of-sample error e_i is independent of the in-sample estimate $\hat{\beta}$, this has variance

$$\begin{aligned} E\hat{e}_i^2 &= E(e_i^2 | x_i = x) + x'E(\hat{\beta} - \beta)(\hat{\beta} - \beta)'x \\ &= \sigma^2(x) + x'V_n x. \end{aligned}$$

Assuming $E(e_i^2 | x_i) = \sigma^2$, the natural estimate of this variance is $\hat{\sigma}^2 + x'\hat{V}x$, so a standard error for the forecast is $\sqrt{\hat{\sigma}^2 + x'\hat{V}x}$. Notice that this is different from the standard error for the conditional mean.

It would appear natural to conclude that an asymptotic 95% forecast interval for y_i is

$$\left[x'\hat{\beta} \pm 2\sqrt{\hat{\sigma}^2 + x'\hat{V}x} \right],$$

but this turns out to be incorrect. In general, the validity of an asymptotic confidence interval is based on the asymptotic normality of the studentized ratio. In the present case, this would require the asymptotic normality of the ratio

$$\frac{e_i - x'(\hat{\beta} - \beta)}{\sqrt{\hat{\sigma}^2 + x'\hat{V}x}}.$$

But no such asymptotic approximation can be made. The only special exception is the case where e_i has the exact distribution $N(0, \sigma^2)$, which is generally invalid.

To get an accurate forecast interval, we need to estimate the conditional distribution of e_i given $x_i = x$, which is a much more difficult task. Given the difficulty, most applied forecasters focus on the simple and unjustified interval $\left[x'\hat{\beta} \pm 2\sqrt{\hat{\sigma}^2 + x'\hat{V}x} \right]$.

5.7 t tests

A simple null and composite hypothesis takes the form

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta \neq \theta_0 \end{aligned}$$

where θ_0 is some pre-specified value, and $\theta = h(\beta)$ is some function of the parameter vector. (For example, θ could be a single element of β).

The standard test for H_0 against H_1 is the t-statistic (or studentized statistic)

$$t_n = t_n(\theta_0) = \frac{\hat{\theta} - \theta_0}{s(\hat{\theta})}.$$

Under H_0 , $t_n \rightarrow_d N(0, 1)$. So a test of asymptotic size α rejects H_0 if $|t_n| \geq z_{\alpha/2}$, else does not reject, or “accepts” H_0 . This is because

$$\begin{aligned} P(\text{reject } H_0 \mid H_0) &= P(|t_n| \geq z_{\alpha/2} \mid \theta = \theta_0) \\ &\rightarrow P(|Z| \geq z_{\alpha/2}) = \alpha. \end{aligned}$$

The rejection/acceptance dichotomy is associated with the Neyman-Pearson approach to hypothesis testing.

An alternative approach, associated with Fisher, is to report an asymptotic p-value. The asymptotic p-value for the above statistic is constructed as follows. Define the tail probability, or asymptotic p-value function

$$p(t) = P(|Z| \geq |t|) = 2(1 - \Phi(|t|)).$$

Then the asymptotic p-value of the statistic t_n is

$$p_n = p(t_n).$$

If the p-value p_n is small (close to zero) then the evidence against H_0 is strong. In a sense, p-values and hypothesis tests are equivalent since $p_n \leq \alpha$ if and only if $|t_n| \geq z_{\alpha/2}$, thus an equivalent statement of a Neyman-Pearson test is to reject at the $\alpha\%$ level iff $p_n \leq \alpha$. The p-value is more general, however, in that the reader is allowed to pick the level of significance (α), in contrast to Neyman-Pearson rejection/acceptance reporting, where the researcher picks the level.

Another helpful observation is that the p-value function has simply made a unit-free transformation of the test statistic. That is, under H_0 , $p_n \rightarrow_d U[0, 1]$, so the “unusualness” of the test statistic can be compared to the easy-to-understand uniform distribution, regardless of the complication of the distribution of the original test statistic. To see this fact, note that the asymptotic distribution of $|t_n|$ is $F(x) = 1 - p(x)$. Thus

$$\begin{aligned} P(1 - p_n \leq u) &= P(1 - p(t_n) \leq u) \\ &= P(F(t_n) \leq u) \\ &= P(|t_n| \leq F^{-1}(u)) \\ &\rightarrow F(F^{-1}(u)) = u, \end{aligned}$$

establishing that $1 - p_n \rightarrow_d U[0, 1]$, from which it follows that $p_n \rightarrow_d U[0, 1]$.

It may be helpful to note that in the GAUSS language, the function $p(t)$ may be computed by the expression $p = 2 * cdfnc(t)$.

5.8 Wald Tests

Sometimes $\theta = h(\beta)$ is a $q \times 1$ vector, and it is desired to test the joint restrictions simultaneously. In this case the t-statistic approach does not work. We still have the null and alternative

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta \neq \theta_0. \end{aligned}$$

The natural estimate of θ is $\hat{\theta} = h(\hat{\beta})$ and has asymptotic covariance matrix estimate

$$\hat{V}_{\hat{\theta}} = \hat{H}'_{\hat{\beta}} \hat{V}_{\hat{\beta}} \hat{H}_{\hat{\beta}}$$

where

$$\hat{H}_\beta = \frac{\partial}{\partial \beta} h(\hat{\beta}).$$

The Wald statistic for H_0 against H_1 is

$$\begin{aligned} W_n &= (\hat{\theta} - \theta_0)' \hat{V}_\theta^{-1} (\hat{\theta} - \theta_0) \\ &= (h(\hat{\beta}) - \theta_0)' (\hat{H}'_\beta \hat{V} \hat{H}_\beta)^{-1} (h(\hat{\beta}) - \theta_0). \end{aligned}$$

When h is a linear function of β , $h(\beta) = R'\beta$, then the Wald statistic takes the form

$$W_n = (R'\hat{\beta} - \theta_0)' (R'\hat{V}R)^{-1} (R'\hat{\beta} - \theta_0).$$

The following fact is useful.

Theorem 5.8.1 *If $Z \sim N(0, A)$ and $A > 0$, then $Z'A^{-1}Z \sim \chi_q^2$, where $q = \dim(Z)$.*

Proof. The fact that $A > 0$ means that we can write $A = CC'$ where C is non-singular. Then $A^{-1} = C^{-1'}C^{-1}$ and

$$\begin{aligned} Z^* &= C^{-1}Z \\ &\sim N(0, C^{-1}AC^{-1'}) \\ &= N(0, C^{-1}CC'C^{-1'}) \\ &= N(0, I_q). \end{aligned}$$

Thus

$$Z'A^{-1}Z = Z'C^{-1'}C^{-1}Z = Z^*Z^* \sim \chi_q^2.$$

■

Theorem 5.8.2 *Under H_0 and Assumption 5.2.1, if $\text{rank}(H_\beta) = q$, then $W_n \rightarrow_d \chi_q^2$, a chi-square random variable with q degrees of freedom.*

Proof. Theorem 5.6.1 showed that $\sqrt{n}(\hat{\theta} - \theta) \rightarrow_d Z \sim N(0, V_\theta)$, and Theorem 5.3.2 showed that $n\hat{V} \rightarrow_p V$. Furthermore, $H_\beta(\beta)$ is a continuous function of β , so by the continuous mapping theorem, $H_\beta(\hat{\beta}) \rightarrow_p H_\beta$. Thus $n\hat{V}_\theta = \hat{H}'_\beta (n\hat{V}) \hat{H}_\beta \rightarrow_p H'_\beta V H_\beta = V_\theta$ and the latter has full rank q . Hence

$$\begin{aligned} W_n &= (\hat{\theta} - \theta_0)' \hat{V}_\theta^{-1} (\hat{\theta} - \theta_0) \\ &= \sqrt{n}(\hat{\theta} - \theta_0)' (n\hat{V}_\theta)^{-1} \sqrt{n}(\hat{\theta} - \theta_0) \\ &\rightarrow_d Z'V_\theta^{-1}Z = \chi_q^2, \end{aligned}$$

by Theorem 5.8.1.

■

An asymptotic Wald test rejects H_0 in favor of H_1 if W_n exceeds $\chi_q^2(\alpha)$, the upper- α quantile of the χ_q^2 distribution. For example, $\chi_1^2(.05) = 3.84 = z_{.025}^2$. The Wald test fails to reject if W_n is less than $\chi_q^2(\alpha)$. The asymptotic p-value for W_n is $p_n = p(W_n)$, where $p(x) = P(\chi_q^2 \geq x)$ is the tail probability function of the χ_q^2 distribution. As before, the test rejects at the $\alpha\%$ level iff $p_n \leq \alpha$, and p_n is asymptotically $U[0, 1]$ under H_0 . In addition, it may be helpful to note that in the GAUSS language, the function $p(t)$ may be computed by the expression $p = \text{cdfchic}(t)$.

5.9 F Tests

Take the linear model

$$Y = X_1\beta_1 + X_2\beta_2 + e$$

where X_1 is $n \times k_1$ and X_2 is $n \times k_2$ and $k = k_1 + k_2$. The null hypothesis is

$$H_0 : \beta_2 = 0.$$

In this case, $\theta = \beta_2$, and there are $q = k_2$ restrictions. Also $h(\beta) = R'\beta$ is linear with $R = \begin{pmatrix} 0 \\ I \end{pmatrix}$ a selector matrix. We know that the Wald statistic takes the form

$$\begin{aligned} W_n &= \hat{\theta}' \hat{V}_\theta^{-1} \hat{\theta} \\ &= \hat{\beta}_2' \left(R' \hat{V} R \right)^{-1} \hat{\beta}_2. \end{aligned}$$

What we will show in this section is that if \hat{V} is replaced with $\hat{V}^0 = s^2 (X'X)^{-1}$, the covariance matrix estimator valid under homoskedasticity, then the Wald statistic can be written in the form

$$W_n = (n - k) \frac{(\tilde{\sigma}^2 - \hat{\sigma}^2)}{\hat{\sigma}^2} \quad (5.8)$$

where

$$\tilde{\sigma}^2 = \frac{1}{n} \tilde{e}' \tilde{e}, \quad \tilde{e} = Y - X_1 \tilde{\beta}_1, \quad \tilde{\beta}_1 = (X_1' X_1)^{-1} X_1' Y$$

are from OLS of Y on X_1 , and

$$\hat{\sigma}^2 = \frac{1}{n} \hat{e}' \hat{e}, \quad \hat{e} = Y - X \hat{\beta}, \quad \hat{\beta} = (X'X)^{-1} X'Y$$

are from OLS of Y on $X = (X_1, X_2)$.

The elegant feature about (5.8) is that it is directly computable from the standard output from two simple OLS regressions, as the sum of square errors RSS is a typical output from statistical packages. This statistic is typically reported as an ‘‘F-statistic’’ which is defined as

$$F = \frac{W_n}{k_2} = \frac{(\tilde{\sigma}^2 - \hat{\sigma}^2) / k_2}{\hat{\sigma}^2 / (n - k)}.$$

While it should be emphasized that equality (5.8) only holds if $\hat{V}^0 = s^2 (X'X)^{-1}$, still this formula often finds good use in reading applied papers.

We now derive expression (5.8). First, note that

$$R' \hat{V}^0 R = s^2 [(X'X)^{-1}]_{22},$$

where $[A]_{22}$ refers to the lower-right block of the matrix A . I claim that

$$[(X'X)^{-1}]_{22} = (X_2' M_1 X_2)^{-1}, \quad (5.9)$$

where $M_1 = I - X_1(X_1'X_1)^{-1}X_1'$. We prove (5.9) below. But taking (5.9) as given, it implies that

$$\left(R' \hat{V}^0 R \right)^{-1} = \left(s^2 (X_2' M_1 X_2)^{-1} \right)^{-1} = s^{-2} (X_2' M_1 X_2),$$

so

$$\begin{aligned} W_n &= \hat{\beta}'_2 \left(R' \hat{V}^0 R \right)^{-1} \hat{\beta}_2 \\ &= \frac{\hat{\beta}'_2 (X'_2 M_1 X_2) \hat{\beta}_2}{s^2}. \end{aligned}$$

To simplify this expression further, note that if we regress Y on X_1 alone, the residual is $\tilde{e} = M_1 Y$. Now consider the residual regression of \tilde{e} on $\tilde{X}_2 = M_1 X_2$. By the FWL theorem, $\tilde{e} = \tilde{X}_2 \hat{\beta}_2 + \hat{e}$ and $\tilde{X}'_2 \hat{e} = 0$. Thus

$$\begin{aligned} \tilde{e}' \tilde{e} &= \left(\tilde{X}_2 \hat{\beta}_2 + \hat{e} \right)' \left(\tilde{X}_2 \hat{\beta}_2 + \hat{e} \right) \\ &= \hat{\beta}'_2 \tilde{X}'_2 \tilde{X}_2 \hat{\beta}_2 + \tilde{e}' \hat{e} \\ &= \hat{\beta}'_2 X'_2 M_1 X_2 \hat{\beta}_2 + \tilde{e}' \hat{e}, \end{aligned}$$

or alternatively,

$$\hat{\beta}'_2 X'_2 M_1 X_2 \hat{\beta}_2 = \tilde{e}' \tilde{e} - \tilde{e}' \hat{e}.$$

Also, since

$$s^2 = \frac{1}{n-k} \tilde{e}' \tilde{e}$$

we conclude that

$$W_n = \frac{\tilde{e}' \tilde{e} - \tilde{e}' \hat{e}}{\frac{1}{n-k} \tilde{e}' \tilde{e}} = (n-k) \left(\frac{\tilde{e}' \tilde{e} - \tilde{e}' \hat{e}}{\tilde{e}' \tilde{e}} \right),$$

as claimed.

It remains to show (5.9). This can be done by partitioned matrix inversion. We use an alternative trick based on the FWL theorem. Suppose that $E(e_i^2 | x_i) = 1$. Then we know that $Var(\hat{\beta} | X) = (X'X)^{-1}$. Thus

$$\begin{aligned} Var(\hat{\beta}_2 \mid X) &= Var(R' \hat{\beta} \mid X) \\ &= R' (X'X)^{-1} R \\ &= [(X'X)^{-1}]_{22}. \end{aligned}$$

By the FWL theorem, we also know that

$$\begin{aligned} \hat{\beta}_2 &= (X'_2 M_1 X_2)^{-1} X'_2 M_1 Y \\ &= (X'_2 M_1 X_2)^{-1} X'_2 M_1 (X_1 \beta_1 + X_2 \beta_2 + e) \\ &= \beta_2 + (X'_2 M_1 X_2)^{-1} X'_2 M_1 e. \end{aligned}$$

Thus

$$\begin{aligned} Var(\hat{\beta}_2 \mid X) &= Var\left((X'_2 M_1 X_2)^{-1} X'_2 M_1 e \mid X \right) \\ &= (X'_2 M_1 X_2)^{-1} X'_2 M_1 M_1 X_2 (X'_2 M_1 X_2)^{-1} \\ &= (X'_2 M_1 X_2)^{-1}. \end{aligned}$$

Since these two expressions must be the same, we can conclude that (5.9) is true. This completes our proof of (5.8).

In many statistical packages, when an OLS regression is reported, an “F statistic” is reported. This is

$$F = \frac{(\tilde{\sigma}_y^2 - \hat{\sigma}^2)/(k-1)}{\hat{\sigma}^2/(n-k)}.$$

where

$$\tilde{\sigma}_y^2 = \frac{1}{n} (y - \bar{y})' (y - \bar{y})$$

is the sample variance of y_i , equivalently the residual variance from an intercept-only model. This special F statistic is testing the hypothesis that *all* slope coefficients (other than the intercept) are zero. This was a popular statistic in the early days of econometric reporting, when sample sizes were very small and researchers wanted to know if there was “any explanatory power” to their regression. This is rarely an issue today, as sample sizes are typically sufficiently large that this F statistic is highly “significant”. Certainly, there are special cases where this F statistic is useful, but these cases are no longer typical.

5.10 Quasi-LR Tests

We showed before that the Gaussian quasi-MLE $\hat{\beta}$ is OLS. Recall, the Gaussian log-likelihood evaluated at the MLE $\hat{\beta}$ is

$$l_n(\hat{\beta}, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \hat{e}'\hat{e}.$$

Thus the MLE for σ^2 is found by solving the first-order-conditions:

$$\frac{\partial}{\partial \sigma^2} l_n(\hat{\beta}, \hat{\sigma}^2) = -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2(\hat{\sigma}^2)^2} \hat{e}'\hat{e} = 0.$$

Hence

$$\hat{\sigma}^2 = \frac{1}{n} \hat{e}'\hat{e}.$$

The log-likelihood evaluated at the parameter estimates is

$$\begin{aligned} \hat{l}_n &= -\frac{n}{2} \log(2\pi\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2} \hat{e}'\hat{e} \\ &= -\frac{n}{2} \log(\hat{\sigma}^2) - \frac{n}{2} \log(2\pi) - \frac{n}{2}. \end{aligned}$$

Similarly, the MLE of the model under the restriction

$$H_0 : \beta_2 = 0.$$

is OLS of y_i on x_{1i} with log-likelihood

$$\tilde{l}_n = -\frac{n}{2} \log(\tilde{\sigma}^2) - \frac{n}{2} \log(2\pi) - \frac{n}{2}.$$

The LR statistic for H_0 is

$$\begin{aligned} LR &= 2(\hat{l}_n - \tilde{l}_n) \\ &= n(\log(\hat{\sigma}^2) - \log(\tilde{\sigma}^2)) \\ &= n \log\left(\frac{\hat{\sigma}^2}{\tilde{\sigma}^2}\right). \end{aligned}$$

Note that both LR and the Wald statistic computed under homoskedasticity:

$$W_n = (n - k) \left(\frac{\tilde{\sigma}^2}{\hat{\sigma}^2} - 1 \right)$$

are functions of the data only through $\tilde{\sigma}^2/\hat{\sigma}^2$. They are numerically close, since

$$LR = n \log \left(1 + \frac{\tilde{\sigma}^2}{\hat{\sigma}^2} - 1 \right) \simeq n \left(\frac{\tilde{\sigma}^2}{\hat{\sigma}^2} - 1 \right) \simeq W_n.$$

Chapter 6

Exact Distribution Theory

6.1 Normal Regression Model

Some of the distribution results simplify dramatically under the assumption that e_i is independent of x_i and normally distributed. Much of classical econometric theory was derived under this set of assumptions.

Fairly directly, this implies that the vector e is independent of the matrix X , and has the distribution $N(0, I_n \sigma^2)$. Since linear functions of normals are also normal, this implies that conditional on X

$$\begin{pmatrix} \hat{\beta} - \beta \\ \hat{e} \end{pmatrix} = \begin{pmatrix} (X'X)^{-1} X' \\ M \end{pmatrix} e \sim N \left(0, \begin{pmatrix} \sigma^2 (X'X)^{-1} & 0 \\ 0 & \sigma^2 M \end{pmatrix} \right)$$

where $M = I - X(X'X)^{-1}X'$. Since uncorrelated normal variables are independent, it follows that $\hat{\beta}$ is independent of any function of the OLS residuals, including the estimated error variance s^2 .

The spectral decomposition of M yields

$$M = H \begin{bmatrix} I_{n-k} & 0 \\ 0 & 0 \end{bmatrix} H'$$

(see equation (1.3)) where $H'H = I_n$. Let $u = \sigma^{-1}H'e \sim N(0, H'H) \sim N(0, I_n)$. Then

$$\begin{aligned} \frac{(n-k)s^2}{\sigma^2} &= \frac{1}{\sigma^2} \hat{e}'\hat{e} \\ &= \frac{1}{\sigma^2} e'Me \\ &= \frac{1}{\sigma^2} e'H \begin{bmatrix} I_{n-k} & 0 \\ 0 & 0 \end{bmatrix} H'e \\ &= u' \begin{bmatrix} I_{n-k} & 0 \\ 0 & 0 \end{bmatrix} u \\ &\sim \chi_{n-k}^2, \end{aligned}$$

a chi-square distribution with $n - k$ degrees of freedom. Furthermore, if standard errors are calculated using the homoskedastic formula (4.14)

$$\frac{\hat{\beta}_j - \beta_j}{s(\hat{\beta}_j)} = \frac{\hat{\beta}_j - \beta_j}{s \sqrt{[(X'X)^{-1}]_{jj}}} \sim \frac{N\left(0, \sigma^2 [(X'X)^{-1}]_{jj}\right)}{\sqrt{\frac{\sigma^2}{n-k} \chi_{n-k}^2} \sqrt{[(X'X)^{-1}]_{jj}}} = \frac{N(0, 1)}{\sqrt{\frac{\chi_{n-k}^2}{n-k}}} \sim t_{n-k}$$

a t distribution with $n - k$ degrees of freedom.

We summarize these findings

Theorem 14 *If e_i is independent of x_i and distributed $N(0, \sigma^2)$, and standard errors are calculated using the homoskedastic formula (4.14) then*

- $\hat{\beta} \sim N\left(0, \sigma^2 (X'X)^{-1}\right)$
- $\frac{(n-k)s^2}{\sigma^2} \sim \chi_{n-k}^2$,
- $\frac{\hat{\beta}_j - \beta_j}{s(\hat{\beta}_j)} \sim t_{n-k}$

Theorem 5.4.1 showed that generally in large samples, $\hat{\beta}$ and t are approximately normally distributed. In contrast, Theorem 14 shows that under the strong assumption of normality, $\hat{\beta}$ has an exact normal distribution and t has an exact t distribution. As inference (confidence intervals) are based on the t -ratio, the notable distinction is between the $N(0, 1)$ and t_{n-k} distributions. The critical values are quite close if $n - k \geq 30$, so as a practical matter it does not matter which distribution is used. (Unless the sample size is unreasonably small.)

While Theorem 14 is quite remarkable, the result is difficult to generalize. Econometric models and questions typically fall outside the class of problems covered by this Theorem.

6.2 Monte Carlo Simulation

It will be helpful to define some general concepts so our discussion will be at a higher level of generality than the linear regression framework. Let w_i be the data and let $F(w) = P(w_i \leq w)$ be the cumulative distribution function (CDF) of w_i . Let F denote a general CDF, and let F_0 denote the true value. Let θ be some parameter of interest and let $T_n = T_n(w_1, \dots, w_n, \theta)$ be a statistic of interest. For example, in a linear regression model, $w_i = (y_i, x_i)$, $\theta = h(\beta)$ is a function of the parameter vector, and $T_n = \hat{\theta} - \theta$ where $\hat{\theta} = h(\hat{\beta})$ with $\hat{\beta}$ the OLS estimator. Or alternatively $T_n = (\hat{\theta} - \theta) / s(\hat{\theta})$.

For inference on θ , we want to know the sampling distribution of T_n . The exact CDF of T_n when the data are sampled from the distribution F is

$$G_n(x, F) = P(T_n \leq x \mid F)$$

Given the structure of the problem, G_n is a function only of F the distribution of w_i . In general, $G_n(x, F)$ will depend on F , meaning that if F as changes, G_n will change.

Ideally, inference on θ would be based on $G_n(x, F_0)$, the true value of the sampling distribution. This is (generally) impossible for two reasons. First, the function $G_n(x, F)$ is unknown. Second, F_0 is unknown.

The idea of Monte Carlo simulation is to solve the first problem through numerical simulation. The idea is that for any given F , the distribution function $G_n(x, F)$ can be calculated by numerical simulation. Since F is unknown, this does not solve the problem of inference. Instead, the method is typically used to assess the adequacy of statistical methods in practical settings.

The name Monte Carlo derives from the famous Meditterrean gambling resort, where games of chance are played.

The method of Monte Carlo is quite simple to describe. The researcher chooses F (the distribution of the data) and the sample size n . A “true” value of θ is implied by this choice, or equivalently the value of θ is selected directly by the researcher.

Then the following experiment is conducted

- n independent random vectors w_i^* , $i = 1, \dots, n$, are drawn from the distribution F using the computer’s random number generator.
- The statistic $T_n^* = T_n(w_1^*, \dots, w_n^*, \theta)$

For step 1, most computer packages have built-in procedures for generating $U[0, 1]$ and $N(0, 1)$ random numbers, and from these most random variables can be constructed. (For example, a chi-square can be generated by sums of squares of normals.)

For step 2, it is important that the statistic be evaluated at the “true” value of θ corresponding to the choice of F .

The above experiment creates one random draw from the distribution $G_n(x, F)$. This is one observation from an unknown distribution. Clearly, from one observation very little can be said. So the researcher repeats the experiment B times, where B is a large number. Typically, we set $B = 1000$ or $B = 5000$, and we will discuss this choice later.

Notationally, let the b ’th experiment result in the draw T_{nb}^* , $b = 1, \dots, B$. These results are stored. They constitute a random sample of size B from the distribution of $G_n(x, F)$.

From a random sample, we can estimate any feature of interest. For example:

Suppose we are interested in the bias, mean-squared error (mse), or variance of the distribution of $\hat{\theta} - \theta$. We then set $T_n = \hat{\theta} - \theta$, run the above experiment, and calculate

$$\begin{aligned}\widehat{Bias}(\hat{\theta}) &= \frac{1}{B} \sum_{b=1}^B T_{nb}^* \\ \widehat{MSE}(\hat{\theta}) &= \frac{1}{B} \sum_{b=1}^B (T_{nb}^*)^2 \\ \widehat{Var}(\hat{\theta}) &= \widehat{MSE}(\hat{\theta}) - \left(\widehat{Bias}(\hat{\theta}) \right)^2\end{aligned}$$

Suppose we are interested in the Type I error associated with an asymptotic 5% two-sided t-test. We would then set $T_n = \left| \hat{\theta} - \theta \right| / s(\hat{\theta})$ and calculate

$$\hat{P} = \frac{1}{B} \sum_{b=1}^B 1(T_{nb}^* \geq 1.96), \quad (6.1)$$

the percentage of the simulated t-ratios which exceed the asymptotic 5% critical value.

Suppose we are interested in the 5% and 95% quantile of $\hat{\theta}$ or $\left(\hat{\theta} - \theta \right) / s(\hat{\theta})$. We then set T_n to either choice, and compute the 10% and 90% sample quantiles of the sample $\{T_{nb}^*\}$. The $\alpha\%$ sample quantile is a number q_α such that $\alpha\%$ of the sample are less than q_α . A simple way to compute sample quantiles is to sort the sample $\{T_{nb}^*\}$ from low to high. Then q_α is the N ’th number in this ordered sequence, where $N = (B + 1)\alpha$. It is therefore convenient to pick B so that N is an integer. For example, if we set $B = 999$, then the 5% sample quantile is 50’th sorted value and the 95% sample quantile is the 950’th sorted value.

The typical purpose of a monte carlo simulation is to investigate the performance of a statistical procedure (estimator or test) in realistic settings. Generally, the performance will depend on n and F . In many cases, an estimator or test may perform wonderfully for some values, and poorly for others. It is therefore useful to conduct a variety of experiments, for a selection of choices of n and F .

As discussed above, the researcher must select the number of experiments, B . Often this is called the number of **replications**. Quite simply, a larger B results in more precise estimates of the features of interest of G_n , but requires more computational time. In practice, therefore, the choice of B is often guided by the computational demands of the statistical procedure. However, it should be recognized that the results of a monte carlo experiment are all estimates computed from a random sample of size B , and therefore it is straightforward to calculate standard errors for any quantity of interest. If the standard error is too large to make a reliable inference, then B will have to be increased.

In particular, it is simple to make inferences about rejection probabilities from statistical tests, such as the percentage estimate reported in (6.1). The random variable $1(T_{nb}^* \geq 1.96)$ is iid Bernoulli, equalling 1 with probability $P = E1(T_{nb}^* \geq 1.96)$. The average (6.1) is therefore an unbiased estimator of P with standard error $s(\hat{P}) = \sqrt{P(1-P)/B}$. As P is unknown, this may be approximated using the estimated value $s(\hat{P}) = \sqrt{\hat{P}(1-\hat{P})/B}$ or using a hypothesized value. For example, if we are assessing an asymptotic 5% test, then we can set $s(\hat{P}) = \sqrt{(0.05)(.95)/B} \simeq .22/\sqrt{B}$. Hence the standard errors for $B = 100, 1000, \text{ and } 5000$, are, respectively, $s(\hat{P}) = .022, .007, \text{ and } .003$.

6.3 An Example

Here we illustrate Monte Carlo simulation to investigate estimation and tests on a nonlinear function of regression parameters.

Our model is an iid sample $\{y_i, x_{1i}, x_{2i}\}$ by the linear Gaussian regression

$$\begin{aligned} y_i &= \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + e_i \\ \begin{pmatrix} x_{1i} \\ x_{2i} \end{pmatrix} &\sim N(0, I_2) \\ e_i &\sim N(0, \sigma^2) \end{aligned}$$

We set $\sigma = 3, \beta_0 = 0, \beta_1 = 1, \beta_2 = .5$, and $n = 300$. In this example, the distribution of the data, F , is determined by the above choices. We don't have to be more explicit about F , as the above equations are sufficient to specify the joint distribution.

The parameter of interest is the ratio of the regression slopes

$$\theta = \frac{\beta_1}{\beta_2}.$$

The goal is to estimate θ , construct confidence intervals for θ , and test the hypothesis that $\theta = \theta_0$. Note that $\theta_0 = 2$.

We estimate the parameters of the model by OLS, and then estimate θ by

$$\hat{\theta} = \frac{\hat{\beta}_1}{\hat{\beta}_2}.$$

The standard error for $\hat{\theta}$ is calculated by the “delta method”:

$$se(\hat{\theta}) = \left(\hat{H}'_{\beta} \hat{V} \hat{H}_{\beta} \right)^{1/2}, \quad \hat{H}_{\beta} = \begin{pmatrix} 0 \\ 1 \\ \hat{\beta}_2 \\ -\hat{\beta}_1 \\ \hat{\beta}_2^2 \end{pmatrix},$$

where \hat{V} is the White covariance matrix estimate.

The asymptotic approximation to the distribution of $\hat{\theta}$ is $N(\theta, AVar(\hat{\theta}))$, where

$$AVar(\hat{\theta}) = n^{-1} \left(H'_{\beta} (Ex_i x'_i)^{-1} H_{\beta} \right) Ee_i^2 \simeq 0.6.$$

Thus the asymptotic approximation is $\hat{\theta} \sim N(2, 0.6)$.

We first investigate the distribution of $\hat{\theta}$. We set $B = 100,000$, which is extremely high, simply because the computation time was minimal. Consequently, the error due to the simulation is minimal.

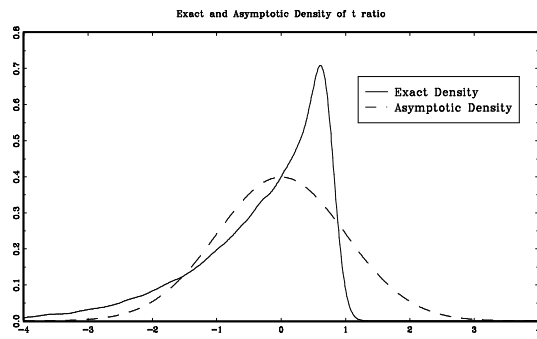
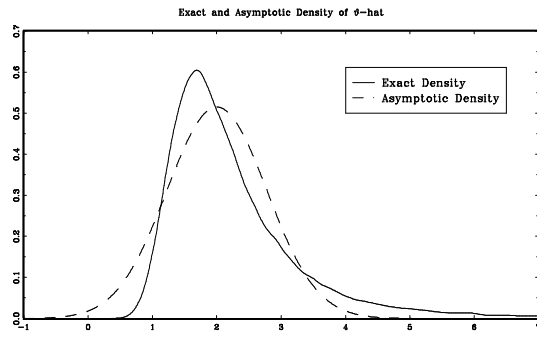
First, I nonparametrically estimated the density of $\hat{\theta}$ (using a kernel density estimator). This density, along with the asymptotic distribution, is displayed in the top section of Figure 1. The divergence between the exact and asymptotic densities is quite dramatic. The exact density is skewed and thick-tailed. Next I estimated¹ the mean and standard deviation of $\hat{\theta}$, finding $E(\hat{\theta}) = 2.32$ and $sd(\hat{\theta}) = 1.28$. Note that these are quite different from the asymptotic approximation (which are 2.0 and 0.77, respectively).

Inference is typically based on the t-ratio $t_n = (\hat{\theta} - \theta_0)/s(\hat{\theta})$. The asymptotic approximation to t_n is $N(0, 1)$. Using the same simulated samples, we estimated the exact density of t_n , which is displayed along with the asymptotic distribution in the bottom section of Figure 1.

Once again we find that the divergence between the exact and asymptotic distributions to be dramatic. The exact distribution is highly skewed and non-normal. The accuracy of hypothesis testing and confidence interval coverage depends on tail probabilities. We calculate the exact probabilities $P(t_n > 1.645) = 0.00$, $P(t_n < 1.645) = .115$ and $P(|t_n| > 1.96) = .084$, meaning that both one-tailed and two-tailed tests and confidence intervals based on asymptotic approximations have significant Type I error. (The nominal, or asymptotic, Type I error for each of these tests is .050).

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

¹Actually, these are trimmed means and standard deviations, calculated on a sample where the upper and lower 0.5% of the sample has been trimmed off. This was done to obtain a robust measure of mean and variance, due to the presence of extreme outliers.



Chapter 7

The Bootstrap

7.1 Definition of the Bootstrap

We use the same notation as introduced in the section on Monte Carlo simulation. The observations are w_i , i, \dots, n , with CDF $F(w) = P(w_i \leq w)$ and true value F_0 . Let $T_n = T_n(w_1, \dots, w_n, \theta)$ be a statistic of interest with CDF $G_n(x, F) = P(T_n \leq x \mid F)$. For exact inference on θ , we desire $G_n(x, F_0)$. This is impossible since F_0 is unknown.

In a seminal contribution, Efron (1979) observed that while F_0 is unknown, it can be estimated, and this can be plugged into $G_n(x, F)$ to get more accurate inference. Let F_n denote some estimate of F . Then the bootstrap estimator of $G_n(x, F_0)$ is

$$G_n^*(x) = G_n(x, F_n).$$

Bootstrap inference is based on $G_n^*(x)$.

The most common choice for F_n is the empirical distribution function (EDF) of the data, which we define in the next section. In this case $G_n^*(x)$ is called the nonparametric bootstrap. Other estimates of F_n are possible. Some are discussed later.

The bootstrap distribution substitutes F_n for F_0 in the formula $G_n(x, F)$. As such, it not only pretends that the distribution of w_i is F_n rather than F_0 , but it also pretends that the true value of the parameter¹ is the sample estimate $\hat{\theta}$, rather than θ_0 .

Let w_i^* be a random variable with distribution F_n . Let $T_n^* = T_n(w_1^*, \dots, w_n^*, \hat{\theta})$ be a random variable with distribution G_n^* . That is,

$$\begin{aligned} P(w_i^* \leq x) &= F_n(x) \\ P(T_n^* \leq x) &= G_n^*(x). \end{aligned}$$

We call w_i^* and T_n^* the bootstrap distribution of the data and statistic. T_n^* is the correct analog of T_n when the true CDF is F_n , as the data w_i^* are sampled from the CDF F_n and the parameter $\hat{\theta}_n$ is the true value under F_n .

7.2 The Empirical Distribution Function

Note that $F_0(w) = P(w_i \leq w) = E1(w_i \leq w)$, where $1(\cdot)$ is the indicator function, so $F_0(w)$ can be expressed as a population moment. A natural estimate is therefore the corresponding sample

¹More precisely, the bootstrap pretends that the true value of θ is the value consistent with F_n , the estimate of F used to construct the bootstrap. In most cases, and the ones we consider, this value of θ is the sample estimate $\hat{\theta}$.

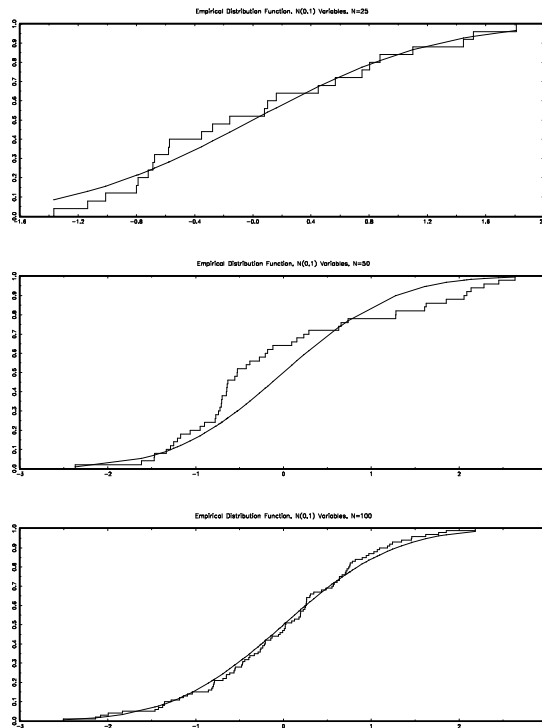
moment:

$$F_n(w) = \frac{1}{n} \sum_{i=1}^n 1(w_i \leq w).$$

$F_n(w)$ is called the empirical distribution function (EDF). F_n is a nonparametric estimate of F_0 . Note that while F_0 may be either discrete or continuous, F_n is by construction a (discontinuous) step function.

The EDF is a consistent estimator of the CDF. To see this, note that for any w , $1(w_i \leq w)$ is an iid random variable with expectation $F_0(w)$. Thus by the WLLN, $F_n(w) \rightarrow_p F_0(w)$. It also converges at rate \sqrt{n} . By the CLT,

$$\sqrt{n}(F_n(w) - F_0(w)) \rightarrow^d N(0, F_0(w)(1 - F_0(w))).$$



To see the effect of sample size on the EDF, in Figure 2, I have plotted the EDF and true CDF for three random samples of size $n = 25, 50,$ and 100 . The random draws are from the $N(0,1)$ distribution. For $n = 25$, the EDF is only a crude approximation to the CDF, but the approximation appears to improve for the large n . Yet even for $n = 100$, there is a significant divergence between the EDF and CDF around $w = -0.6$ in this sample. In general, as the sample size gets larger, the EDF step function gets uniformly close to the true CDF.

The EDF is a valid probability distribution. It is the discrete probability distribution which puts probability mass $1/n$ at each point $w_i = (y_i, x_i)$, $i = 1, \dots, n$. Notationally, it will be helpful

to think of a random variable w^* with the distribution F_n . We can easily calculate the moments of w^* :

$$\begin{aligned} Eh(w^*) &= \int h(w)dF_n(w) \\ &= \sum_{i=1}^n h(w_i)P(w^* = w_i) \\ &= \frac{1}{n} \sum_{i=1}^n h(w_i), \end{aligned}$$

the empirical sample average.

7.3 Computation

Since the EDF F_n is a multinomial (with n support points), in principle the distribution of T_n^* could be calculated by direct methods. Since there are 2^n possible samples $\{w_1^*, \dots, w_n^*\}$, however, such a calculation is computationally infeasible unless n is very small. The popular alternative is to use Monte Carlo simulation to approximate the distribution. The algorithm is identical to our discussion of Monte Carlo simulation, with the following points of clarification:

- The sample size n used for the simulation should be the same as the sample size
- The random vectors w_i^* are drawn randomly from the distribution function F_n
- $T_n^* = T_n(w_1^*, \dots, w_n^*, \hat{\theta})$ is evaluated at the sample estimate $\hat{\theta}$.

Implement of step 2, the generation of random vectors from F_n , depends on the form of F_n . When F_n equals the EDF, this is particularly simple. Recall that F_n is a discrete probability distribution putting probability mass $1/n$ at each sample point w_i . Thus a random draw from F_n is just a random draw from the sample $\{w_1, \dots, w_n\}$. For a bootstrap sample we need n independent random draws from F_n . This requires that we make n independent random draws from the sample $\{w_1, \dots, w_n\}$, **with replacement**. In consequence, a bootstrap sample $\{w_1^*, \dots, w_n^*\}$ will necessarily have some ties and multiple values, which is generally not a problem.

A theory for the determination of the number of bootstrap replications B has been developed by Andrews and Buchinsky (2000).

7.4 Bootstrap Estimation of Bias

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

$$\tau_n = E(\hat{\theta} - \theta_0).$$

Let $T_n(\theta) = \hat{\theta} - \theta$. Then

$$\tau_n = E(T_n(\theta_0)).$$

The bootstrap counterparts are $\hat{\theta}^* = \hat{\theta}(w_1^*, \dots, w_n^*)$ and $T_n^* = \hat{\theta}^* - \theta_n = \hat{\theta}^* - \hat{\theta}$. The bootstrap estimate of τ_n is

$$\tau_n^* = E(T_n^*).$$

If this is calculated by the simulation described in the previous subsection, the estimate of τ_n^* is

$$\begin{aligned}\hat{\tau}_n^* &= \frac{1}{B} \sum_{b=1}^B T_{nb}^* \\ &= \frac{1}{B} \sum_{b=1}^B \hat{\theta}_b^* - \hat{\theta} \\ &= \overline{\hat{\theta}^*} - \hat{\theta}.\end{aligned}$$

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

$$\tilde{\theta} = \hat{\theta} - \tau_n,$$

but τ_n is unknown. The (estimated) bootstrap biased-corrected estimator is

$$\begin{aligned}\tilde{\theta}^* &= \hat{\theta} - \hat{\tau}_n^* \\ &= \hat{\theta} - (\overline{\hat{\theta}^*} - \hat{\theta}) \\ &= 2\hat{\theta} - \overline{\hat{\theta}^*}.\end{aligned}$$

Note, in particular, that the biased-corrected estimator is *not* $\overline{\hat{\theta}^*}$. Intuitively, the bootstrap makes the following experiment. Suppose that $\hat{\theta}$ is the truth. Then what is the average value of $\hat{\theta}$ calculated from such samples? The answer is $\overline{\hat{\theta}^*}$. If this is lower than $\hat{\theta}$, this suggests that the estimator is *downward-biased*, so a biased-corrected estimator of θ should be *larger* than $\hat{\theta}$, and the best guess is the difference between $\hat{\theta}$ and $\overline{\hat{\theta}^*}$. Similarly if $\overline{\hat{\theta}^*}$ is higher than $\hat{\theta}$, then the estimator is upward-biased and the biased-corrected estimator should be lower than $\hat{\theta}$.

7.5 Bootstrap Estimation of Variance

Let $T_n = \hat{\theta}$. The variance of $\hat{\theta}$ is

$$V_n = E(T_n - ET_n)^2.$$

Let $T_n^* = \hat{\theta}^*$. It has variance

$$V_n^* = E(T_n^* - ET_n^*)^2.$$

The simulation estimate is

$$\hat{V}_n^* = \frac{1}{B} \sum_{b=1}^B (\hat{\theta}_b^* - \overline{\hat{\theta}^*})^2.$$

A bootstrap standard error for $\hat{\theta}$ is the square root of the bootstrap estimate of variance, $s(\hat{\beta}) = \sqrt{\hat{V}_n^*}$.

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

7.6 Efron's Percentile Interval

For a distribution function $G_n(x, F)$, let $q_n(\alpha, F)$ denote its quantile function. This is the function which solves

$$G_n(q_n(\alpha, F), F) = \alpha.$$

[When $G_n(x, F)$ is discrete, $q_n(\alpha, F)$ may be non-unique, but we will ignore such complications.] Let $q_n(\alpha) = q_n(\alpha, F_0)$ denote the quantile function of the true sampling distribution, and $q_n^*(\alpha) = q_n(\alpha, F_n)$ denote the quantile function of the bootstrap distribution. Note that this function will change depending on the underlying statistic T_n whose distribution is G_n .

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

$$C_1 = [q_n^*(\alpha/2), q_n^*(1 - \alpha/2)].$$

This is often called the *percentile confidence interval*.

Computationally, the quantile $q_n^*(x)$ is estimated by $\hat{q}_n^*(x)$, the x 'th sample quantile of the simulated statistics $\{T_{n1}^*, \dots, T_{nB}^*\}$, as discussed in the section on Monte Carlo simulation. The $(1 - \alpha)\%$ Efron percentile interval is then $[\hat{q}_n^*(\alpha/2), \hat{q}_n^*(1 - \alpha/2)]$.

The interval C_1 is a popular bootstrap confidence interval often used in empirical practice. This is because it is easy to compute, simple to motivate, was popularized by Efron early in the history of the bootstrap, and also has the feature that it is translation invariant. That is, if we define $\phi = f(\theta)$ as the parameter of interest for a monotonic function f , then percentile method applied to this problem will produce the confidence interval $[f(q_n^*(\alpha/2)), f(q_n^*(1 - \alpha/2))]$, which is a naturally good property.

However, as we show now, C_1 is in a deep sense very poorly motivated.

It will be useful if we introduce an alternative definition C_1 . Let $T_n(\theta) = \hat{\theta} - \theta$ and let $q_n(\alpha)$ be the quantile function of its distribution. (These are the original quantiles, with θ subtracted.) Then C_1 can alternatively be written as

$$C_1 = [\hat{\theta} + q_n^*(\alpha/2), \hat{\theta} + q_n^*(1 - \alpha/2)].$$

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

$$C_1^0 = [\hat{\theta} + q_n(\alpha/2), \hat{\theta} + q_n(1 - \alpha/2)].$$

The latter has coverage probability

$$\begin{aligned} P(\theta_0 \in C_1^0) &= P(\hat{\theta} + q_n(\alpha/2) \leq \theta_0 \leq \hat{\theta} + q_n(1 - \alpha/2)) \\ &= P(-q_n(1 - \alpha/2) \leq \hat{\theta} - \theta_0 \leq -q_n(\alpha/2)) \\ &= G_n(-q_n(\alpha/2), F_0) - G_n(-q_n(1 - \alpha/2), F_0) \end{aligned}$$

which generally is not $1 - \alpha$! There is one important exception. If $\hat{\theta} - \theta_0$ has a symmetric distribution, then $G_n(-x, F_0) = 1 - G_n(x, F_0)$, so

$$\begin{aligned} P(\theta_0 \in C_1^0) &= G_n(-q_n(\alpha/2), F_0) - G_n(-q_n(1 - \alpha/2), F_0) \\ &= (1 - G_n(q_n(\alpha/2), F_0)) - (1 - G_n(q_n(1 - \alpha/2), F_0)) \\ &= \left(1 - \frac{\alpha}{2}\right) - \left(1 - \left(1 - \frac{\alpha}{2}\right)\right) \\ &= 1 - \alpha \end{aligned}$$

and this idealized confidence interval is accurate. Therefore, C_1^0 and C_1 are designed for the case that $\hat{\theta}$ has a symmetric distribution about θ_0 .

When $\hat{\theta}$ does not have a symmetric distribution, C_1 may perform quite poorly.

However, by the translation invariance argument presented above, it also follows that if there exists some monotonic transformation $f(\cdot)$ such that $f(\hat{\theta})$ is symmetrically distributed about $f(\theta_0)$, then the idealized percentile bootstrap method will be accurate.

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

7.7 Alternative Percentile Interval

The problems with the percentile method outlined above can be circumvented by an alternative method.

Let $T_n(\theta) = \hat{\theta} - \theta$. Then

$$\begin{aligned} 1 - \alpha &= P(q_n(\alpha/2) \leq T_n(\theta_0) \leq q_n(1 - \alpha/2)) \\ &= P(\hat{\theta} - q_n(1 - \alpha/2) \leq \theta_0 \leq \hat{\theta} - q_n(\alpha/2)), \end{aligned}$$

so an exact $(1 - \alpha)\%$ confidence interval for θ_0 would be

$$C_2^0 = [\hat{\theta} - q_n(1 - \alpha/2), \hat{\theta} - q_n(\alpha/2)].$$

This motivates a bootstrap analog

$$C_2 = [\hat{\theta} - q_n^*(1 - \alpha/2), \hat{\theta} - q_n^*(\alpha/2)].$$

Notice that generally this is very different from the Efron interval C_1 ! They coincide in the special case that $G_n^*(x)$ is symmetric about $\hat{\theta}$, but otherwise they differ.

Computationally, this interval can be estimated from a bootstrap simulation by sorting the bootstrap statistics $T_n^* = (\hat{\theta}^* - \hat{\theta})$, which are centered at the sample estimate $\hat{\theta}$. These are sorted to yield the quantile estimates $\hat{q}_n^*(.025)$ and $\hat{q}_n^*(.975)$. The 95% confidence interval is then $[\hat{\theta} - \hat{q}_n^*(.975), \hat{\theta} - \hat{q}_n^*(.025)]$.

This confidence interval is discussed in most theoretical treatments of the bootstrap, but is not widely used in practice.

7.8 One-Sided Hypothesis Tests

Suppose we want to test $H_0 : \theta = \theta_0$ against $H_1 : \theta < \theta_0$ at size α . We would set $T_n(\theta) = (\hat{\theta} - \theta) / s(\hat{\theta})$ and reject H_0 in favor of H_1 if $T_n(\theta_0) < c$, where c would be selected so that

$$P(T_n(\theta_0) < c) = \alpha.$$

Thus $c = q_n(\alpha)$. Since this is unknown, a bootstrap test replaces $q_n(\alpha)$ with the bootstrap estimate $q_n^*(\alpha)$, and the test rejects if $T_n(\theta_0) < q_n^*(\alpha)$.

Similarly, if the alternative is $H_1 : \theta > \theta_0$, the bootstrap test rejects if $T_n(\theta_0) > q_n^*(1 - \alpha)$.

Computationally, these critical values can be estimated from a bootstrap simulation by sorting the bootstrap t-statistics $T_n^* = (\hat{\theta}^* - \hat{\theta}) / s(\hat{\theta}^*)$. Note, and this is important, that the bootstrap test statistic is centered at the estimate $\hat{\theta}$, and the standard error $s(\hat{\theta}^*)$ is calculated on the bootstrap sample. These t-statistics are sorted to find the estimated quantiles $\hat{q}_n^*(\alpha)$ and/or $\hat{q}_n^*(1 - \alpha)$.

7.9 Percentile-t Equal-Tailed Interval

Let $T_n(\theta) = (\hat{\theta} - \theta) / s(\hat{\theta})$. Then

$$\begin{aligned} 1 - \alpha &= P(q_n(\alpha/2) \leq T_n(\theta_0) \leq q_n(1 - \alpha/2)) \\ &= P\left(q_n(\alpha/2) \leq (\hat{\theta} - \theta_0) / s(\hat{\theta}) \leq q_n(1 - \alpha/2)\right) \\ &= P\left(\hat{\theta} - s(\hat{\theta})q_n(1 - \alpha/2) \leq \theta_0 \leq \hat{\theta} - s(\hat{\theta})q_n(\alpha/2)\right), \end{aligned}$$

so an exact $(1 - \alpha)\%$ confidence interval for θ_0 would be

$$C_3^0 = [\hat{\theta} - s(\hat{\theta})q_n(1 - \alpha/2), \quad \hat{\theta} - s(\hat{\theta})q_n(\alpha/2)].$$

This motivates a bootstrap analog

$$C_3 = [\hat{\theta} - s(\hat{\theta})q_n^*(1 - \alpha/2), \quad \hat{\theta} - s(\hat{\theta})q_n^*(\alpha/2)].$$

This is often called a *percentile-t confidence interval*. It is *equal-tailed* or *central* since the probability that θ_0 is below the left endpoint approximately equals the probability that θ_0 is above the right endpoint, each $\alpha/2$.

Computationally, this is based on the critical values from the one-sided hypothesis tests, discussed above.

7.10 Two-Sided Hypothesis Tests

Suppose we want to test $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ at size α . We would set $T_n(\theta) = (\hat{\theta} - \theta) / s(\hat{\theta})$ and reject H_0 in favor of H_1 if $|T_n(\theta_0)| > c$, where c would be selected so that

$$P(|T_n(\theta_0)| > c) = \alpha.$$

Note that

$$\begin{aligned} P(|T_n(\theta_0)| < c) &= P(-c < T_n(\theta_0) < c) \\ &= G_n(c) - G_n(-c) \\ &\equiv \bar{G}_n(c), \end{aligned}$$

which is a symmetric distribution function. The ideal critical value $c = q_n(\alpha)$ solves the equation

$$\bar{G}_n(q_n(\alpha)) = 1 - \alpha.$$

Equivalently, $q_n(\alpha)$ is the $1 - \alpha$ quantile of the distribution of $|T_n(\theta_0)|$.

The bootstrap estimate is $q_n^*(\alpha)$, the $1 - \alpha$ quantile of the distribution of $|T_n^*|$, or the number which solves the equation

$$\bar{G}_n^*(q_n^*(\alpha)) = G_n^*(q_n^*(\alpha)) - G_n^*(-q_n^*(\alpha)) = 1 - \alpha.$$

Computationally, $q_n^*(\alpha)$ is estimated from a bootstrap simulation by sorting the bootstrap t-statistics $|T_n^*| = |\hat{\theta}^* - \hat{\theta}| / s(\hat{\theta}^*)$, and taking the upper $\alpha\%$ quantile.

The bootstrap test rejects if $|T_n(\theta_0)| > q_n^*(\alpha)$.

7.11 Symmetric Percentile-t Intervals

Let

$$C_4 = [\hat{\theta} - s(\hat{\theta})q_n^*(\alpha), \hat{\theta} + s(\hat{\theta})q_n^*(\alpha)],$$

where $q_n^*(\alpha)$ is the bootstrap critical value for a two-sided hypothesis test, as defined above. C_4 is called the symmetric percentile-t interval. It is designed to work well since

$$\begin{aligned} P(\theta_0 \in C_4) &= P(\hat{\theta} - s(\hat{\theta})q_n^*(\alpha) \leq \theta_0 \leq \hat{\theta} + s(\hat{\theta})q_n^*(\alpha)) \\ &= P(|T_n(\theta_0)| < q_n^*(\alpha)) \\ &\simeq P(|T_n(\theta_0)| < q_n(\alpha)) \\ &= 1 - \alpha. \end{aligned}$$

7.12 Vector Tests

If θ is a vector, then to test $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ at size α , we would use a Wald statistic

$$W_n(\theta) = (\hat{\theta} - \theta)' \hat{V}_\theta^{-1} (\hat{\theta} - \theta)$$

or some other asymptotically chi-square statistic. Thus here $T_n(\theta) = W_n(\theta)$. The ideal test rejects if $W_n \geq q_n(\alpha)$, where $q_n(\alpha)$ is the $(1 - \alpha)\%$ quantile of the distribution of W_n . The bootstrap test rejects if $W_n \geq q_n^*(\alpha)$, where $q_n^*(\alpha)$ is the $(1 - \alpha)\%$ quantile of the distribution of

$$W_n^* = (\hat{\theta}^* - \hat{\theta})' \hat{V}_\theta^{*-1} (\hat{\theta}^* - \hat{\theta}).$$

Computationally, the critical value $q_n^*(\alpha)$ is found as the quantile from simulated values of W_n^* . Note in the simulation that the Wald statistic is a quadratic form in $(\hat{\theta}^* - \hat{\theta})$, not $(\hat{\theta}^* - \theta_0)$. [This is a typical mistake made by practitioners.]

7.13 Asymptotic Expansions

In this section, we assume that θ is real-value, and $T_n = (\hat{\theta} - \theta_0) / s(\hat{\theta})$. Then

$$T_n \rightarrow_d N(0, 1).$$

Another way of writing this is

$$\lim_{n \rightarrow \infty} G_n(x, F_0) = \Phi(x). \tag{7.1}$$

While (7.1) says that G_n converges to $\Phi(x)$ as $n \rightarrow \infty$, it says nothing, however, about the rate of convergence, or the size of the divergence for any particular sample size n . A better asymptotic approximation may be obtained through an *asymptotic expansion*.

The following notation will be helpful.

Definition 7.13.1 *Let a_n be a sequence. We say that $a_n = O(n^{-r})$ if $n^r |a_n|$ is uniformly bounded.*

Basically, $a_n = O(n^{-r})$ if it declines to zero like n^{-r} .

We say that a function $g(x)$ is *even* if $g(-x) = g(x)$, and a function $h(x)$ is *odd* if $h(-x) = -h(x)$. The derivative of an even function is odd, and vice-versa.

Theorem 7.13.1 *Under regularity conditions*

$$G_n(x, F) = \Phi(x) + \frac{1}{n^{1/2}}g_1(x, F) + \frac{1}{n}g_2(x, F) + O(n^{-3/2})$$

uniformly over x , where g_1 is an even function of x , and g_2 is an odd function of x . Moreover, g_1 and g_2 are differentiable functions of x and continuous in F relative to the supremum norm on the space of distribution functions.

We can interpret Theorem 7.13.1 as follows. First, $G_n(x, F_0)$ converges to the normal limit at rate $n^{1/2}$. To a second order of approximation,

$$G_n(x, F) \approx \Phi(x) + n^{-1/2}g_1(x, F).$$

Since the derivative of g_1 is odd, the density function is skewed. To a third order of approximation,

$$G_n(x, F) \approx \Phi(x) + n^{-1/2}g_1(x, F) + n^{-1}g_2(x, F)$$

which adds a symmetric non-normal component to the approximate density (for example, adding leptokurtosis).

7.14 One-Sided Tests

Using the expansion of Theorem 7.13.1, we can assess the accuracy of one-sided hypothesis tests and confidence regions based on a t-ratio.

To the second order, the exact distribution of the statistic is

$$P(T_n < x) = G_n(x, F_0) = \Phi(x) + \frac{1}{n^{1/2}}g_1(x, F_0) + O(n^{-1}).$$

An asymptotic test is based on $\Phi(x)$. The difference is

$$\Phi(x) - G_n(x, F_0) = \frac{1}{n^{1/2}}g_1(x, F_0) + O(n^{-1/2}) = O(n^{-1/2}),$$

so the order of the error is $O(n^{-1/2})$.

A bootstrap test is based on $G_n^*(x)$, which from Theorem 7.13.1 has the expansion

$$G_n^*(x) = G_n(x, F_n) = \Phi(x) + \frac{1}{n^{1/2}}g_1(x, F_n) + O(n^{-1}).$$

Because $\Phi(x)$ appears in both expansions, the difference between the bootstrap distribution and the true distribution is

$$G_n^*(x) - G_n(x, F_0) = \frac{1}{n^{1/2}}(g_1(x, F_n) - g_1(x, F_0)) + O(n^{-1}).$$

Since F_n converges to F at rate \sqrt{n} , and g_1 is continuous with respect to F , the difference $(g_1(x, F_n) - g_1(x, F_0))$ converges to 0 at rate \sqrt{n} . Heuristically,

$$\begin{aligned} g_1(x, F_n) - g_1(x, F_0) &\approx \frac{\partial}{\partial F}g_1(x, F_0)(F_n - F_0) \\ &= O(n^{-1/2}), \end{aligned}$$

The “derivative” $\frac{\partial}{\partial F}g_1(x, F)$ is only heuristic, as F is a function. We conclude that

$$G_n^*(x) - G_n(x, F_0) = O(n^{-1}),$$

or

$$P(T_n^* \leq x) = P(T_n \leq x) + O(n^{-1}),$$

which is an improved rate of convergence over the asymptotic test (which converged at rate $O(n^{-1/2})$). This rate can be used to show that one-tailed bootstrap inference based on the t -ratio achieves a so-called *asymptotic refinement* – the Type I error of the test converges at a faster rate than an analogous asymptotic test.

7.15 Symmetric Two-Sided Tests

If a random variable X has distribution function $H(x) = P(X \leq x)$, then the random variable $|X|$ has distribution function

$$\bar{H}(x) = H(x) - H(-x)$$

since

$$\begin{aligned} P(|X| \leq x) &= P(-x \leq X \leq x) \\ &= P(X \leq x) - P(X \leq -x) \\ &= H(x) - H(-x). \end{aligned}$$

For example, if $Z \sim N(0, 1)$, then $|Z|$ has distribution function

$$\bar{\Phi}(x) = \Phi(x) - \Phi(-x) = 2\Phi(x) - 1.$$

Similarly, if T_n has exact distribution $G_n(x, F)$, then $|T_n|$ has the distribution function

$$\bar{G}_n(x, F) = G_n(x, F) - G_n(-x, F).$$

A two-sided hypothesis test rejects H_0 for large values of $|T_n|$. Since $T_n \rightarrow_d Z$, then $|T_n| \rightarrow_d |Z| \sim \bar{\Phi}$. Thus asymptotic critical values are taken from the $\bar{\Phi}$ distribution, and exact critical values are taken from the $\bar{G}_n(x, F_0)$ distribution. From Theorem 7.13.1, we can calculate that

$$\begin{aligned} \bar{G}_n(x, F) &= G_n(x, F) - G_n(-x, F) \\ &= \left(\Phi(x) + \frac{1}{n^{1/2}}g_1(x, F) + \frac{1}{n}g_2(x, F) \right) \\ &\quad - \left(\Phi(-x) + \frac{1}{n^{1/2}}g_1(-x, F) + \frac{1}{n}g_2(-x, F) \right) + O(n^{-3/2}) \\ &= \bar{\Phi}(x) + \frac{2}{n}g_2(x, F) + O(n^{-3/2}), \end{aligned} \tag{7.2}$$

where the simplifications are because g_1 is even and g_2 is odd. Hence the difference between the asymptotic distribution and the exact distribution is

$$\bar{\Phi}(x) - \bar{G}_n(x, F_0) = \frac{2}{n}g_2(x, F_0) + O(n^{-3/2}) = O(n^{-1}).$$

The order of the error is $O(n^{-1})$.

Interestingly, the asymptotic two-sided test has a better coverage rate than the asymptotic one-sided test. This is because the first term in the asymptotic expansion, g_1 , is an even function, meaning that the errors in the two directions exactly cancel out.

Applying (7.2) to the bootstrap distribution, we find

$$\bar{G}_n^*(x) = \bar{G}_n(x, F_n) = \bar{\Phi}(x) + \frac{2}{n}g_2(x, F_n) + O(n^{-3/2}).$$

Thus the difference between the bootstrap and exact distributions is

$$\begin{aligned} \bar{G}_n^*(x) - \bar{G}_n(x, F_0) &= \frac{2}{n}(g_2(x, F_n) - g_2(x, F_0)) + O(n^{-3/2}) \\ &= O(n^{-3/2}), \end{aligned}$$

the last equality because F_n converges to F_0 at rate \sqrt{n} , and g_2 is continuous in F . Another way of writing this is

$$P(|T_n^*| < x) = P(|T_n| < x) + O(n^{-3/2})$$

so the error from using the bootstrap distribution (relative to the true unknown distribution) is $O(n^{-3/2})$. This is in contrast to the use of the asymptotic distribution, whose error is $O(n^{-1})$. Thus a two-sided bootstrap test also achieves an asymptotic refinement, similar to a one-sided test.

A reader might get confused between the two simultaneous effects. Two-sided tests have better rates of convergence than the one-sided tests, and bootstrap tests have better rates of convergence than asymptotic tests.

The analysis shows that there may be a trade-off between one-sided and two-sided tests. Two-sided tests will have more accurate size (Reported Type I error), but one-sided tests might have more power against alternatives of interest. Confidence intervals based on the bootstrap can be asymmetric if based on one-sided tests (equal-tailed intervals) and can therefore be more informative and have smaller length than symmetric intervals. Therefore, the choice between symmetric and equal-tailed confidence intervals is unclear, and needs to be determined on a case-by-case basis.

7.16 Percentile Confidence Intervals

To evaluate the coverage rate of the percentile interval, set $T_n = \sqrt{n}(\hat{\theta} - \theta_0)$. We know that $T_n \rightarrow_d N(0, V)$, which is not pivotal, as it depends on the unknown V . An analog of Theorem 7.13.1 holds, to a first-order approximation

$$G_n(x, F) = \Phi(x, V) + O(n^{-1/2}),$$

where $\Phi(x, V)$ is the distribution function of $N(0, V)$. This holds for the bootstrap as well:

$$G_n^*(x) = G_n(x, F_n) = \Phi(x, \hat{V}) + O(n^{-1/2}),$$

where $\hat{V} = V(F_n)$ is the bootstrap estimate of V . Since \hat{V} converges to V at rate \sqrt{n} , it follows that $\Phi(x, \hat{V})$ converges to $\Phi(x, V)$ at rate \sqrt{n} , so we have the order of approximation

$$G_n^*(x) = \Phi(x, V) + O(n^{-1/2}).$$

This says that bootstrap inference based on a non-pivotal statistic (such as the percentile intervals), is asymptotically correct, in that the bootstrap distribution converges to the asymptotic distribution, and that the rate of convergence is \sqrt{n} .

From a positive viewpoint, this is good news, because it means that the percentile-type methods (if appropriately used) can yield \sqrt{n} -convergent asymptotic inference, without even calculating standard errors! This means that in contexts where standard errors are not available or are difficult to calculate, the percentile bootstrap methods provide an attractive alternative.

From a negative viewpoint, the rate of convergence is disappointing. It is no better than the rate obtained from an asymptotic one-sided confidence region. Therefore if standard errors are available, it is unclear if there are any benefits from using the percentile bootstrap over simple asymptotic methods.

Based on these arguments, the theoretical literature (e.g. Hall, 1992, Horowitz, 2002) tends to advocate the use of the percentile-t bootstrap methods, rather than percentile methods.

7.17 Bootstrap Methods for Independent Errors

The bootstrap methods we have discussed have set $G_n^*(x) = G_n(x, F_n)$, where F_n is the EDF. Any other consistent estimate of F_0 may be used to define a feasible bootstrap estimator. The advantage of the EDF is that it is fully nonparametric, it imposes no conditions, and works in nearly any context. But since it is fully nonparametric, it may be inefficient in contexts where more is known about F . We discuss some bootstrap methods appropriate for the case of errors e_i which are independent of the regressors x_i . Thus the model is

$$y_i = x_i' \beta + e_i$$

with e_i independent of x_i .

To impose independence, it is sufficient to sample the x_i^* and e_i^* independently, and then create $y_i^* = x_i^{*'} \hat{\beta} + e_i^*$.

One method to generate the bootstrap errors e_i^* is to sample randomly from the OLS residuals $\{\hat{e}_1, \dots, \hat{e}_n\}$. This is sampling from the EDF of the residuals, and is a non-parametric estimate of the distribution of e_i .

Another method is to generate the bootstrap errors e_i^* from a parametric distribution, such as the normal $e_i^* \sim N(0, \hat{\sigma}^2)$. This is appropriate when the true errors are distributed with this parametric distribution.

For the regressors x_i^* , one method is to sample the x_i^* randomly from the sample values $\{x_1, \dots, x_n\}$. This is sampling from the EDF of the regressors, and is fully non-parametric.

Another approach is to sample x_i^* from an estimated parametric distribution.

Another approach is to set $x_i^* = x_i$. This is equivalent to treating the regressors as *fixed in repeated samples*. If this is done, then all inferential statements are made conditionally on the observed values of the regressors, which is a valid statistical approach. It does not really matter, however, whether or not the x_i are really “fixed” or random.

The methods described in this section require that e_i and x_i are independent. If the independence assumption is incorrect, then the estimate F_n of F_0 will be inconsistent, and the bootstrap estimate $G_n(x, F_n)$ will be inconsistent for the asymptotic distribution $\lim_{n \rightarrow \infty} G_n(x, F_0)$.

Chapter 8

Functional Form

8.1 Dummy Variables

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

For example, a dummy variable may be used to denote the gender (male/female) of an individual. The interest is the regression $E(Wage | Gender)$. There are several equivalent ways to write this down. One is to define a dummy variable

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

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

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

and the model $W_i = \beta_0 + \beta_1 d_{2i} + e_i$. In this model, $\beta_0 = E(Wage | Female)$ and $\beta_0 + \beta_1 = E(Wage | Male)$. Third, we could define both d_{1i} and d_{2i} as above, and write the model as $W_i = \beta_1 d_{1i} + \beta_2 d_{2i} + e_i$. Here, $\beta_1 = E(Wage | Female)$ and $\beta_2 = E(Wage | Male)$. These three models are equivalent.

A standard easy mistake is to include an intercept, d_{1i} and d_{2i} all in the regression. These are clearly collinear so this cannot be done.

If the equation of interest is $E(Wage | Education, Gender)$, a typical regression model is

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

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

A regression model allowing for slope differences is

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

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

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

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

$$Y_i = \beta_1 d_{1i} + \beta_2 d_{2i} + e_i.$$

We can write this in matrix notation as

$$Y = X\beta + e$$

where $\beta = (\beta_1 \ \beta_2)'$ and

$$X = [D_1 \ D_2].$$

Thus,

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1} X'Y \\ &= \begin{bmatrix} D_1' D_1 & D_1' D_2 \\ D_2' D_1 & D_2' D_2 \end{bmatrix}^{-1} \begin{bmatrix} D_1' Y \\ D_2' Y \end{bmatrix} \\ &= \begin{bmatrix} \sum_{i=1}^n d_{1i}^2 & \sum_{i=1}^n d_{1i} d_{2i} \\ \sum_{i=1}^n d_{1i} d_{2i} & \sum_{i=1}^n d_{2i}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^n d_{1i} y_i \\ \sum_{i=1}^n d_{1i} y_i \end{bmatrix} \\ &= \begin{bmatrix} n_1 & 0 \\ 0 & n_2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^n d_{1i} y_i \\ \sum_{i=1}^n d_{1i} y_i \end{bmatrix} \\ &= \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \end{bmatrix} \end{aligned}$$

where n_1 and n_2 are the number of observations with $d_{1i} = 1$ and $d_{2i} = 1$, respectively, and \bar{y}_1 and \bar{y}_2 are the sample means among those respective observations.

Now consider the variance estimates.

$$\hat{V}^0 = (X'X)^{-1} \hat{\sigma}^2 = \begin{bmatrix} \frac{\hat{\sigma}^2}{n_1} & 0 \\ 0 & \frac{\hat{\sigma}^2}{n_2} \end{bmatrix}$$

where

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \hat{e}_i^2,$$

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

$$\begin{aligned} \hat{V} &= \begin{bmatrix} (D_1' D_1)^{-1} \hat{\sigma}_1^2 & 0 \\ 0 & (D_2' D_2)^{-1} \hat{\sigma}_2^2 \end{bmatrix} \\ &= \begin{bmatrix} \frac{\hat{\sigma}_1^2}{n_1} & 0 \\ 0 & \frac{\hat{\sigma}_2^2}{n_2} \end{bmatrix} \end{aligned}$$

where

$$\hat{\sigma}_1^2 = \frac{1}{n_1} \sum_{i=1}^n d_{1i} \hat{e}_i^2$$

and

$$\hat{\sigma}_2^2 = \frac{1}{n_2} \sum_{i=1}^n d_{2i} \hat{e}_i^2$$

are the estimates of σ^2 based on observations with $d_{1i} = 1$ and $d_{2i} = 1$, respectively. Thus the conventional estimate imposes the restriction that the conditional variance of e_i is the same for the two groups, and the White estimate estimates the conditional variance separately for each group.

8.2 NonLinearity in Regressors

Suppose we are interested in $E(y_i | x_i = x) = g(x)$, $x \in R$, and the form of g is unknown. A common approach is to consider a polynomial approximation:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k + e_i.$$

Letting $\beta = (\beta_0, \beta_1, \dots, \beta_k)$ and $z_i = (1, x_i, x_i^2, \dots, x_i^k)$, this is $y_i = z_i' \beta + e_i$, which is a linear regression model. Typically, the polynomial order k is kept quite small.

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

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i}^2 + \beta_4 x_{2i}^2 + \beta_5 x_{1i} x_{2i} + e_i.$$

As the dimensionality of x increases, such approximations can become quite non-parsimonious! In practice, therefore, most applications do appear to use more than quadratic terms. Some applications add cubics without interactions:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i}^2 + \beta_4 x_{2i}^2 + \beta_5 x_{1i}^3 + \beta_6 x_{2i}^3 + \beta_7 x_{1i} x_{2i} + e_i.$$

Non-linear approximations can also be made using alternative *basis functions*, such as Fourier series (sins and cosines), splines, “neural nets”, or “wavelets”.

Since these non-linear models are linear in the parameters, they can be estimated by OLS, and inference is convention. However, the model is non-linear so interpretation must take this into account. For example, in the cubic model given above, the slope with respect to x_{1i} is

$$\frac{\partial}{\partial x_{1i}} E(y_i | x_i) = \beta_1 + 2\beta_3 x_{1i} + 3\beta_5 x_{1i}^2 + \beta_7 x_{2i},$$

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

$$\frac{\partial}{\partial x_{1i}} E(y_i | x_i) |_{x_i = \bar{x}} = \beta_1 + 2\beta_3 \bar{x}_1 + 3\beta_5 \bar{x}_1^2 + \beta_7 \bar{x}_2$$

and the average derivative

$$\frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial x_{1i}} E(y_i | x_i) = \beta_1 + 2\beta_3 \bar{x}_1 + 3\beta_5 \frac{1}{n} \sum_{i=1}^n x_{1i}^2 + \beta_7 \bar{x}_2.$$

8.3 Testing for Omitted NonLinearity

If the goal is to estimate the conditional expectation $E(y_i | x_i)$, it is useful to know how to test a given specification. Many such tests have been proposed. Here we discuss two simple tests.

One simple test for neglected nonlinearity is to add nonlinear functions of the regressors to the regression, and test their significance using a Wald test. Thus, if the model $y_i = x_i' \hat{\beta} + \hat{e}_i$ has been fit by OLS, let $z_i = h(x_i)$ denote functions of x_i which are not linear functions of x_i (perhaps squares of non-binary regressors) and then fit $y_i = x_i' \tilde{\beta} + z_i' \tilde{\gamma} + \tilde{e}_i$ by OLS, and form a Wald statistic for $\gamma = 0$.

Ramsey (1969) introduced the RESET test. The null model is

$$y_i = x_i' \beta + e_i$$

which is estimated by OLS, yielding predicted values $\hat{y}_i = x_i' \hat{\beta}$. Now let

$$z_i = \begin{pmatrix} \hat{y}_i^2 \\ \vdots \\ \hat{y}_i^m \end{pmatrix}$$

be an $(m - 1)$ -vector of powers of \hat{y}_i . Then run the auxiliary regression

$$y_i = x_i' \tilde{\beta} + z_i' \tilde{\gamma} + \tilde{e}_i \tag{8.1}$$

by OLS, and form the Wald statistic W_n for $\gamma = 0$. It is easy (although involves a somewhat lengthy derivation) to show that under the null hypothesis, $W_n \rightarrow_d \chi_{m-1}^2$. Thus the null is rejected at the $\alpha\%$ level if W_n exceeds the upper $\alpha\%$ tail critical value of the χ_{m-1}^2 distribution.

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

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

$$y_i = G(x_i' \beta) + e_i$$

where $G(\cdot)$ is a smooth “link” function. To see why this is the case, note that (8.1) may be written as

$$y_i = x_i' \tilde{\beta} + \left(x_i' \hat{\beta}\right)^2 \tilde{\gamma}_1 + \left(x_i' \hat{\beta}\right)^3 \tilde{\gamma}_2 + \cdots + \left(x_i' \hat{\beta}\right)^m \tilde{\gamma}_{m-1} + \tilde{e}_i$$

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

8.4 $\log(Y)$ versus Y as Dependent Variable

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

The plain truth is that either regression is “okay”, in the sense that both $E(y_i | x_i)$ and $E(\log(y_i) | x_i)$ are well-defined (so long as $y_i > 0$). It is perfectly valid to estimate either or both regressions. They are *different* regression functions, neither is more nor less valid than the other. To *test* one specification versus the other, or select one specification over the other, requires the imposition of additional structure, such as the assumptions that the conditional expectation is linear in x_i , and $e_i \sim N(0, \sigma^2)$.

There still may be good reasons for preferring the $\log(Y)$ regression over the Y regression. First, it may be the case that $E(\log(y_i) | x_i)$ is roughly linear in x_i over the support of x_i , while the regression $E(y_i | x_i)$ is non-linear, and linear models are easier to report and interpret. Second, it may be the case that the errors in $e_i = \log(y_i) - E(\log(y_i) | x_i)$ may be less heteroskedastic than the errors from the linear specification (although the reverse may be true!). Finally, and this may be most important reason, if the distribution of y_i is highly skewed, the conditional mean $E(y_i | x_i)$ may not be a useful measure of central tendency, and estimates will be undesirably influenced by extreme observations (“outliers”). In this case, the conditional mean-log $E(\log(y_i) | x_i)$ may be a better measure of central tendency, and hence more interesting to estimate and report.

8.5 Multicollinearity

If $\text{rank}(X'X) < k$, then $\hat{\beta}$ is not defined. This is defined as multicollinearity. It happens if $\text{rank}(X) < k$, which holds if and only if the columns of X are linearly dependent, i.e., there is some α such that $X\alpha = 0$. Most commonly, this arises when sets of regressors are included which are identically related. For example, if X includes the logs of two prices and the log of the relative prices $\log(p_1)$, $\log(p_2)$ and $\log(p_1/p_2)$. When this happens, the applied researcher quickly discovers the error, as the statistical software will be unable to construct $(X'X)^{-1}$. Since the error is discovered quickly, this is rarely a *problem* for applied econometric practice.

The more relevant issue is *near multicollinearity*, which is often called “multicollinearity” for brevity. This is the situation when the $X'X$ matrix is *near* singular, when the columns of X are *close* to linearly dependent. Notice that this definition is not precise, because we have not said what it means for a matrix to be “near singular”. This is one difficulty with the definition and interpretation of multicollinearity.

One implication of near singularity of matrices is that the numerical reliability of the calculations is reduced. It is more likely that the reported calculations will be in error due to floating-point calculation difficulties.

The other implication of near multicollinearity is that estimation precision of individual coefficients will be reduced. To see this, consider a simple example. Suppose that

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

where $E(e_i^2 | x_i) = \sigma^2$ and

$$X'X = \begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

When $\rho = 1$, then $X'X$ is singular, and when $\rho \approx 1$, then $X'X$ is close to singular. Then

$$\begin{aligned} \text{Var}(\hat{\beta}_1) &= (X_1'M_2X_1)^{-1} \\ &= \left(X_1'X_1 - X_1'X_2(X_2'X_2)^{-1}X_2'X_1 \right)^{-1} \\ &= \frac{1}{1 - \rho^2}. \end{aligned}$$

As $\rho \rightarrow 1$, $\text{Var}(\hat{\beta}_1) \rightarrow \infty$. Thus the more “collinear” are the regressors, the worse the precision of $\hat{\beta}_1$. (Similarly $\hat{\beta}_2$).

Basically, what is happening is that when X_1 and X_2 are highly dependent, it is statistically difficult to disentangle the impact of β_1 from that of β_2 . The precision of individual estimates are reduced.

Is there a simple solution? Basically, *No*. Fortunately, multicollinearity does not lead to errors in inference. The asymptotic distribution is still valid. Regression estimates are asymptotically normal, and estimated standard errors are consistent for the asymptotic variance. So reported confidence intervals are not inherently misleading. They will be *large*, correctly indicating the inherent uncertainty about the true parameter value.

8.6 Omitted Variables

Let x_{1i} and x_{2i} be two sets of regressors. We can define

$$g_1(x_1) = E(y_i | x_{1i} = x_1)$$

and

$$g_2(x_1, x_2) = E(y_i | x_{1i} = x_1, x_{2i} = x_2).$$

Both of these functions exist and are well defined. Given data, either can be estimated. Thus, if the function $g_1(x_1)$ is estimated by regression of y_i on x_{1i} only, there is no “bias” due to the omission of x_{2i} . In this sense, there is no such thing as “omitted variable bias”.

However, the function g_1 may not be of interest. Rather, the function g_2 may be of interest. Thus if g_1 is estimated, when the true relationship of interest is g_2 , then there will be estimation bias. That is, what may be of interest is the effect of x_{1i} on the conditional mean of y_i , holding x_{2i} constant, namely

$$\frac{\partial}{\partial x_1} g_2(x_1, x_2) \neq \frac{\partial}{\partial x_1} g_1(x_1).$$

In this sense, omission of x_{2i} from the regression can induce bias.

Another way to see this is by focusing on the linear regression model. Suppose that

$$\begin{aligned} y_i &= x'_{1i}\beta_1 + x'_{2i}\beta_2 + e_i \\ E(e_i | x_{1i}, x_{2i}) &= 0. \end{aligned}$$

Then

$$\begin{aligned} E(y_i | x_{1i}) &= E(x'_{1i}\beta_1 + x'_{2i}\beta_2 + e_i | x_{1i}) \\ &= x'_{1i}\beta_1 + E(x_{2i} | x_{1i})' \beta_2 \\ &\neq x'_{1i}\beta_1. \end{aligned}$$

Thus a regression of y_i on x_{1i} does not yield the coefficient β_1 , unless $E(x_{2i} | x_{1i}) = 0$ or $\beta_2 = 0$.

Furthermore, suppose $E(x_{2i} | x_{1i}) = \Gamma x_{1i}$. Then

$$\begin{aligned} E(y_i | x_{1i}) &= x'_{1i}\beta_1 + (\Gamma x_{1i})' \beta_2 \\ &= x'_{1i}(\beta_1 + \Gamma' \beta_2). \end{aligned}$$

So a regression of y_i on x_{1i} will consistently estimate $\beta_1 + \Gamma' \beta_2$. β_1 cannot be uncovered from this regression, unless $\Gamma = 0$ or $\beta_2 = 0$, and thus the regression is “biased”, if the parameter β_1 is of interest.

Notice that the omitted variable bias problem disappears if $\Gamma = 0$ (so x_{1i} and x_{2i} are uncorrelated) or if $\beta_2 = 0$ (so x_{2i} does not enter the joint regression). The first can be assessed by examining the correlation between x_{1i} and x_{2i} , but the second can only be assessed by computing the joint regression. Therefore the standard advice is when in doubt, to always estimate the more general model, since it is by construction free of the omitted variables problem.

8.7 Irrelevant Variables

In the model

$$\begin{aligned} y_i &= x'_{1i}\beta_1 + x'_{2i}\beta_2 + e_i \\ E(e_i | x_{1i}, x_{2i}) &= 0, \end{aligned}$$

x_{2i} is “irrelevant” if β_1 is the parameter of interest and $\beta_2 = 0$. That is, the truth can be written as

$$\begin{aligned} y_i &= x'_{1i}\beta_1 + e_i \\ E(e_i | x_{1i}, x_{2i}) &= 0. \end{aligned}$$

One estimator of β_1 is to regress y_i on x_{1i} alone, denoted $\tilde{\beta}_1$. Another is to regress y_i on x_{1i} and x_{2i} , yielding $(\hat{\beta}_1, \hat{\beta}_2)$. Under which conditions is $\hat{\beta}_1$ or $\tilde{\beta}_1$ superior?

First, it is easy to see that both are unbiased and consistent for β_1 . So in comparison with the problem of omitted variables, we see that the presence (or absence) of irrelevant variables is relatively less important.

Second, we can consider the relative efficiency of $\tilde{\beta}_1$ versus $\hat{\beta}_1$. It is harder to make comparisons in the general case, so we focus on the homoskedastic case $E(e_i^2 | x_{1i}, x_{2i}) = \sigma^2$. Then

$$\lim_{n \rightarrow \infty} nVar(\tilde{\beta}_1 | X) = (Ex_{1i}x'_{1i})^{-1} \sigma^2 = Q_{11}^{-1} \sigma^2,$$

say, and

$$\lim_{n \rightarrow \infty} nVar(\hat{\beta}_1 | X) = (Ex_{1i}x'_{1i} - Ex_{1i}x'_{2i}(Ex_{2i}x'_{2i})^{-1}Ex_{2i}x'_{1i})^{-1} \sigma^2 = (Q_{11} - Q_{121})^{-1} \sigma^2,$$

say. If $Ex_{1i}x'_{2i} = 0$ (so the variables are uncorrelated) then these two variance matrices equal, and the two estimators have equal asymptotic efficiency.

Proposition 8.7.1 *If $Ex_{1i}x'_{2i} = 0$, $\hat{\beta}_1$ and $\tilde{\beta}_1$ are both consistent and have equal asymptotic variances.*

When $Ex_{1i}x'_{2i} \neq 0$, however, then $Q_{121} > 0$ and

$$Q_{11} = Q_{11} - Q_{121} + Q_{121} > Q_{11} - Q_{121},$$

so

$$Q_{11}^{-1} < (Q_{11} - Q_{121})^{-1},$$

meaning that $\tilde{\beta}_1$ has a lower asymptotic variance matrix than $\hat{\beta}_1$. The inclusion of irrelevant variables reduces efficiency if these variables are correlated with the relevant variables.

Proposition 8.7.2 *If $Ex_{1i}x'_{2i} \neq 0$, $\hat{\beta}_1$ and $\tilde{\beta}_1$ are both consistent, and*

$$\lim_{n \rightarrow \infty} nVar(\tilde{\beta}_1 | X) < \lim_{n \rightarrow \infty} nVar(\hat{\beta}_1 | X).$$

8.8 Model Selection

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

In many cases the problem of model selection can be reduced to the comparison of two nested models, as the larger problem can be written as a sequence of such comparisons. We thus consider the question of the inclusion of X_2 in the linear regression

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

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

$$\begin{aligned} \mathcal{M}_1 &: Y = X_1\beta_1 + e, & E(e | X_1, X_2) &= 0 \\ \mathcal{M}_2 &: Y = X_1\beta_1 + X_2\beta_2 + e, & E(e | X_1, X_2) &= 0. \end{aligned}$$

Note that $\mathcal{M}_1 \subset \mathcal{M}_2$. To be concrete, we say that \mathcal{M}_2 is true if $\beta_2 \neq 0$.

To fix notation, models 1 and 2 are estimated by OLS, with residual vectors \hat{e}_1 and \hat{e}_2 , estimated variances $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$, etc., respectively. To simplify some of the statistical discussion, we will on occasion use the homoskedasticity assumption $E(e_i^2 | x_{1i}, x_{2i}) = \sigma^2$.

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

$$\begin{aligned} P(\widehat{\mathcal{M}} = \mathcal{M}_1 | \mathcal{M}_1) &\rightarrow 1 \\ P(\widehat{\mathcal{M}} = \mathcal{M}_2 | \mathcal{M}_2) &\rightarrow 1 \end{aligned}$$

We now discuss a number of possible model selection methods.

Selection Based on Fit

Natural measures of fit of a regression are the residual sum of squares $\hat{e}'\hat{e}$, $R^2 = 1 - (\hat{e}'\hat{e})/\hat{\sigma}_y^2$ or Gaussian log-likelihood $l = -(n/2) \log \hat{\sigma}^2$. It might therefore be thought attractive to base a model selection procedure on one of these measures of fit. The problem is that each of these measures are necessarily monotonic between nested models, namely $\hat{e}'_1\hat{e}_1 \geq \hat{e}'_2\hat{e}_2$, $R_1^2 \leq R_2^2$, and $l_1 \leq l_2$, so model \mathcal{M}_2 would always be selected, regardless of the actual data and probability structure. This is clearly an inappropriate decision rule!

Selection based on Testing

A common approach to model selection is to base the decision on a statistical test such as the Wald W_n or Gaussian Likelihood Ratio

$$\begin{aligned} LR_n &= 2(l_2 - l_1) = n(\log \hat{\sigma}_1^2 - \log \hat{\sigma}_2^2) = n \left(\log \left(\frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} \right) \right) \\ &\simeq n \left(\frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} - 1 \right) = n \left(\frac{\hat{\sigma}_1^2 - \hat{\sigma}_2^2}{\hat{\sigma}_2^2} \right) = W_n. \end{aligned}$$

The model selection rule is as follows, expositied for the Wald test. For some critical level α , let c_α satisfy $P(\chi_{k_2}^2 > c_\alpha)$. Then select \mathcal{M}_1 if $W_n \leq c_\alpha$, else select \mathcal{M}_1 .

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

Adjusted R-squared

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

$$\begin{aligned}\bar{R}^2 &= 1 - \frac{(\hat{e}'\hat{e}) / (n - k)}{\hat{\sigma}_y^2} \\ &= 1 - \frac{s^2}{\hat{\sigma}_y^2}.\end{aligned}$$

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

$$\begin{aligned}\log(s^2) &= \log\left(\hat{\sigma}^2 \frac{n}{n - k}\right) \\ &= \log(\hat{\sigma}^2) + \log\left(1 + \frac{k}{n - k}\right) \\ &\simeq \log(\hat{\sigma}^2) + \frac{k}{n - k} \\ &\simeq \log(\hat{\sigma}^2) + \frac{k}{n},\end{aligned}$$

(the first approximation is $\log(1 + x) \simeq x$ for small x). Thus selecting based on \bar{R}^2 is the same as selecting based on $\log(\hat{\sigma}^2) + \frac{k}{n}$, which is a particular choice of penalized likelihood criteria. It turns out that model selection based on any criterion of the form

$$\log(\hat{\sigma}^2) + c\frac{k}{n}, \quad c > 0, \tag{8.2}$$

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

$$LR_n = n(\log \hat{\sigma}_1^2 - \log \hat{\sigma}_2^2) \simeq W_n \rightarrow_d \chi_{k_2}^2, \tag{8.3}$$

$$\begin{aligned}P(\widehat{\mathcal{M}} = \mathcal{M}_1 | \mathcal{M}_1) &= P(\bar{R}_1^2 > \bar{R}_2^2 | \mathcal{M}_1) \\ &\simeq P(n \log(s_1^2) < n \log(s_2^2) | \mathcal{M}_1) \\ &\simeq P(n \log(\hat{\sigma}_1^2) + ck_1 < n \log(\hat{\sigma}_2^2) + c(k_1 + k_2) | \mathcal{M}_1) \\ &= P(LR_n < ck_2 | \mathcal{M}_1) \\ &\rightarrow P(\chi_{k_2}^2 < ck_2) < 1.\end{aligned}$$

Akaike Information Criterion

Akaike proposed an information criterion which takes the form (8.2) with $c = 2$:

$$AIC = \log(\hat{\sigma}^2) + 2\frac{k}{n}. \quad (8.4)$$

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

Since the AIC criterion (8.4) takes the form (8.2), it is an inconsistent model selection criterion, and tends to overfit.

Schwarz Criterion

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

$$BIC = \log(\hat{\sigma}^2) + \log(n)\frac{k}{n}. \quad (8.5)$$

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

In contrast to the other methods studied above, BIC model selection is consistent. Indeed, since (8.3) holds under \mathcal{M}_1 ,

$$\frac{LR_n}{\log(n)} \rightarrow_p 0,$$

so

$$\begin{aligned} P(\widehat{\mathcal{M}} = \mathcal{M}_1 | \mathcal{M}_1) &= P(BIC_1 < BIC_2 | \mathcal{M}_1) \\ &= P(LR_n < \log(n)k_2 | \mathcal{M}_1) \\ &= P\left(\frac{LR_n}{\log(n)} < k_2 | \mathcal{M}_1\right) \\ &\rightarrow P(0 < k_2) = 1. \end{aligned}$$

Also under \mathcal{M}_2 , one can show that

$$\frac{LR_n}{\log(n)} \rightarrow_p \infty,$$

thus

$$\begin{aligned} P(\widehat{\mathcal{M}} = \mathcal{M}_2 | \mathcal{M}_2) &= P\left(\frac{LR_n}{\log(n)} > k_2 | \mathcal{M}_2\right) \\ &\rightarrow 1. \end{aligned}$$

Selection Among Multiple Regressors

We have discussed model selection between two models. The methods extend readily to the issue of selection among multiple regressors. The general problem is the model

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + e_i, \quad E(e_i | x_i) = 0$$

and the question is which subset of the coefficients are non-zero (equivalently, which regressors enter the regression).

There are two leading cases: ordered regressors and unordered.

In the ordered case, the models are

$$\begin{aligned}
 \mathcal{M}_1 & : \beta_1 \neq 0, \beta_2 = \beta_3 = \dots = \beta_K = 0 \\
 \mathcal{M}_2 & : \beta_1 \neq 0, \beta_2 \neq 0, \beta_3 = \dots = \beta_K = 0 \\
 & \quad \vdots \\
 \mathcal{M}_K & : \beta_1 \neq 0, \beta_2 \neq 0, \dots, \beta_K \neq 0.
 \end{aligned}$$

which are nested. The AIC selection criteria estimates the K models by OLS, stores the residual variance $\hat{\sigma}^2$ for each model, and then selects the model with the lowest AIC (8.4). Similarly for the BIC, selecting based on (8.5).

In the unordered case, a model consists of any possible subset of the regressors $\{x_{1i}, \dots, x_{Ki}\}$, and the AIC or BIC in principle can be implemented by estimating all possible subset models. However, there are 2^K such models, which can be a very large number. For example, $2^{10} = 1024$, and $2^{20} = 1,048,576$. In the latter case, a full-blown implementation of the BIC selection criterion would seem computationally prohibitive.

Chapter 9

NonLinear Regression

9.1 NonLinear Regression Models

We say that the regression function $g(x, \theta) = E(y_i | x_i = x)$ is nonlinear in the parameters if it cannot be written as $g(x, \theta) = z(x)' \theta$ for some function $z(x)$. Examples of nonlinear regression functions include

$$\begin{aligned}g(x, \theta) &= \theta_1 + \theta_2 \frac{x}{1 + \theta_3 x} \\g(x, \theta) &= \theta_1 + \theta_2 x^{\theta_3} \\g(x, \theta) &= \theta_1 + \theta_2 \exp(\theta_3 x) \\g(x, \theta) &= G(x' \theta), \quad G \text{ known} \\g(x, \theta) &= \theta_1 + \theta_2 x_1 + (\theta_3 + \theta_4 x_1) \Phi\left(\frac{x_2 - \theta_5}{\theta_6}\right) \\g(x, \theta) &= \theta_1 + \theta_2 x + \theta_4 (x - \theta_3) \mathbf{1}(x > \theta_3) \\g(x, \theta) &= (\theta_1 + \theta_2 x_1) \mathbf{1}(x_2 < \theta_3) + (\theta_4 + \theta_5 x_1) \mathbf{1}(x_2 > \theta_3)\end{aligned}$$

In the first five examples, $g(x, \theta)$ is (generically) differentiable in the parameters θ . In the final two examples, g is not differentiable with respect to θ_3 , which alters some of the analysis. When it exists, let

$$g_\theta(x, \theta) = \frac{\partial}{\partial \theta} g(x, \theta).$$

Nonlinear regression is frequently adopted because the functional form $g(x, \theta)$ is suggested by an economic model. In other cases, it is adopted as a flexible approximation to an unknown regression function.

9.2 NLLS Estimation

The least squares estimator $\hat{\theta}$ minimizes the sum-of-squared-errors

$$S_n(\theta) = \sum_{i=1}^n (y_i - g(x_i, \theta))^2.$$

When the regression function is nonlinear, we call this the *nonlinear least squares* (NLLS) estimator.

The NLLS residuals are $\hat{\varepsilon}_i = y_i - g(x_i, \hat{\theta})$.

Since the problem is non-linear, $\hat{\theta}$ must be found by numerical methods. A common method to minimize the function $S_n(\theta)$ is the Gauss-Newton method or one of its variants. When $g(x, \theta)$ is differentiable, then the FOC for minimization are

$$0 = \sum_{i=1}^n g_{\theta}(x_i, \hat{\theta}) \hat{e}_i. \quad (9.1)$$

9.3 Concentration

A major simplification can be achieved through concentration. This can be done when we partition $\theta = (\beta, \gamma)$ so that

$$g(x_i, \theta) = \beta' x_i(\gamma)$$

where $x_i(\gamma)$ is a $k \times 1$ function of x_i and γ . In all the examples, this can be done with γ of much smaller dimension than β . In many cases, γ is scalar.

The SSE function is $S_n(\theta) = S_n(\beta, \gamma)$ and thus

$$\min_{\theta} S_n(\theta) = \min_{\gamma} \min_{\beta} S_n(\beta, \gamma).$$

Since β enters the model linearly, we see that

$$\begin{aligned} \hat{\beta}(\gamma) &= \arg \min_{\beta} S_n(\beta, \gamma) \\ &= (X(\gamma)' X(\gamma))^{-1} X(\gamma)' Y \end{aligned} \quad (9.2)$$

where $X(\gamma)$ is the $n \times k$ matrix of the stacked $x_i(\gamma)'$.

Now set

$$S_n(\gamma) = S_n(\hat{\beta}(\gamma), \gamma)$$

which is the concentrated sum of squared errors. We have $\hat{\gamma} = \arg \min S_n(\gamma)$ and $\hat{\beta} = \hat{\beta}(\hat{\gamma})$. The pair $(\hat{\beta}, \hat{\gamma})$ are the joint NLLS estimates of (β, γ) .

The main benefit of concentration is that the dimension of the numerical optimization is typically reduced dramatically. When γ is scalar, the final minimization over γ can be done by a grid search, for example.

9.4 Computation Using Linearization

A linearized regression can be used to find the NLLS estimator $\hat{\theta}$. It is an iterative technique, meaning that we start with an initial guess $\hat{\theta}_1$, and then define an iteration rule $\hat{\theta}_j \rightarrow \hat{\theta}_{j+1}$, stopping when the iteration “converges”, meaning in practice that the difference $|\hat{\theta}_{j+1} - \hat{\theta}_j|$ is smaller than some pre-specified level.

We now define the iteration rule

$$\hat{\theta}_{j+1} = \hat{\theta}_j + d_j. \quad (9.3)$$

where the “direction” d_j is a function of $\hat{\theta}_j$. Let

$$\begin{aligned} g_{\theta i}(j) &= g_{\theta}(x_i, \hat{\theta}_j), \\ \hat{e}_i(j) &= y_i - g(x_i, \hat{\theta}_j), \end{aligned}$$

and

$$d_j = \left(\sum_{i=1}^n g_{\theta_i}(j) g_{\theta_i}(j)' \right)^{-1} \left(\sum_{i=1}^n g_{\theta_i}(j) \hat{e}_i(j) \right).$$

Convergence requires $d_j = 0$, which requires $\sum_{i=1}^n g_{\theta_i} \hat{e}_i = 0$, which is the same as the first-order condition for NLLS minimization (9.1). Thus if (9.3) converges, it yields the NLLS estimator.

One problem is that the updating rule (9.3) may tend to overshoot and thus fail to converge. The algorithm can be easily modified to correct for this, by substituting for (9.3) the rule

$$\hat{\theta}_{j+1} = \hat{\theta}_j + \lambda d_j,$$

where $\lambda > 0$ is a scalar “step length”. Rules for determining the step length are discussed in the numerical optimization literature. The goal is to find λ so that $S_n^*(\lambda) = S_n(\hat{\theta}_j + \lambda d_j)$ is minimized. One simple rule is the “half” rule. Essentially, try the sequence $\lambda = 1, \frac{1}{2}, \frac{1}{4}, \dots$, until a value of λ is found which reduces the criterion $S_n^*(\lambda)$. Specifically, first compute $S_n^*(1)$. If $S_n^*(1) < S_n^*(0) = S_n(\hat{\theta}_j)$, then set $\hat{\theta}_{j+1} = \hat{\theta}_j + d_j$. If not, compute $S_n^*(1/2)$. If $S_n^*(1/2) < S_n^*(0)$, then set $\hat{\theta}_{j+1} = \hat{\theta}_j + \frac{1}{2}d_j$. This is continued until a value of λ yields an “improvement” in the criterion.

9.5 Asymptotic Distribution

Let $g_{\theta_i} = g_{\theta}(x_i, \theta_0)$.

Theorem 9.5.1 *If the model is identified and $g(x, \theta)$ is differentiable with respect to θ ,*

$$\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow_d N(0, V)$$

$$V = (E(g_{\theta_i} g_{\theta_i}'))^{-1} (E(g_{\theta_i} g_{\theta_i}' e_i^2)) (E(g_{\theta_i} g_{\theta_i}'))^{-1}.$$

“Proof”. First, it must be shown that $\hat{\theta} \rightarrow_p \theta_0$. This can be done using arguments for optimization estimators, but we won’t cover that argument here.

Since $\hat{\theta} \rightarrow_p \theta_0$, $\hat{\theta}$ is close to θ_0 for n large, so the minimization of $S_n(\theta)$ only needs to be examined for θ close to θ_0 .

Let

$$y_i^0 = e_i + g_{\theta_i}' \theta_0.$$

For θ close to the true value θ_0 , by a first-order Taylor series approximation,

$$g(x_i, \theta) \simeq g(x_i, \theta_0) + g_{\theta_i}'(\theta - \theta_0).$$

Thus

$$\begin{aligned} y_i - g(x_i, \theta) &\simeq (e_i + g(x_i, \theta_0)) - (g(x_i, \theta_0) + g_{\theta_i}'(\theta - \theta_0)) \\ &= e_i - g_{\theta_i}'(\theta - \theta_0) \\ &= y_i^0 - g_{\theta_i}'\theta. \end{aligned}$$

Hence the sum of squared errors function is

$$S_n(\theta) = \sum_{i=1}^n (y_i - g(x_i, \theta))^2 \simeq \sum_{i=1}^n (y_i^0 - g_{\theta_i}'\theta)^2$$

and the right-hand-side is the SSE function for a linear regression of y_i^0 on $g_{\theta i}$. Thus the NLLS estimator $\hat{\theta}$ has the same asymptotic distribution as the (infeasible) OLS regression of y_i^0 on $g_{\theta i}$, which is that stated in the theorem. ■

Based on Theorem 9.5.1, an estimate of the variance matrix for $\hat{\theta}$ is

$$\hat{V} = \left(\sum_{i=1}^n \hat{g}_{\theta i} \hat{g}'_{\theta i} \right)^{-1} \left(\sum_{i=1}^n \hat{g}_{\theta i} \hat{g}'_{\theta i} \hat{e}_i^2 \right) \left(\sum_{i=1}^n \hat{g}_{\theta i} \hat{g}'_{\theta i} \right)^{-1}$$

where $\hat{g}_{\theta i} = g_{\theta}(x_i, \hat{\theta})$ and $\hat{e}_i = y_i - g(x_i, \hat{\theta})$.

9.6 Identification

Identification is often tricky in nonlinear regression models. Suppose that

$$g(x_i, \theta) = \beta_1' z_i + \beta_2' x_i(\gamma).$$

The model is linear when $\beta_2 = 0$, and this is often a useful hypothesis (sub-model) to consider. Thus we want to test

$$H_0 : \beta_2 = 0.$$

However, under H_0 , the model is

$$y_i = \beta_1' z_i + e_i$$

and both β_2 and γ have dropped out. This means that under H_0 , γ is not identified. This renders the distribution theory presented in the previous section invalid. Thus when the truth is that $\beta_2 = 0$, the parameter estimates are not asymptotically normally distributed. Furthermore, tests of H_0 do not have asymptotic normal or chi-square distributions.

The asymptotic theory of such tests have been worked out by Andrews and Ploberger (1994) and B. Hansen (1996). In particular, Hansen shows how to use simulation (similar to the bootstrap) to construct the asymptotic critical values (or p-values) in a given application.

Chapter 10

Feasible GLS

Earlier, we showed that in regression models, OLS is inefficient relative to GLS, but the latter is infeasible. Now we discuss feasible approximate GLS estimation.

10.1 Skedastic Regression

Suppose that the conditional variance takes the parametric form

$$\begin{aligned}\sigma_i^2 &= \alpha_0 + z'_{1i}\alpha_1 \\ &= \alpha'z_i,\end{aligned}$$

where z_{1i} is some $q \times 1$ function of x_i . Typically, z_{1i} are squares (and perhaps levels) of some (or all) elements of x_i . Often the functional form is kept simple for parsimony.

Let $\eta_i = e_i^2$. Then

$$E(\eta_i | x_i) = \alpha_0 + z'_{1i}\alpha_1$$

and we have the regression equation

$$\begin{aligned}\eta_i &= \alpha_0 + z'_{1i}\alpha_1 + \xi_i \\ E(\xi_i | x_i) &= 0.\end{aligned}\tag{10.1}$$

It is helpful to think about the regression error ξ_i . It has conditional variance

$$\begin{aligned}Var(\xi_i | x_i) &= Var(e_i^2 | x_i) \\ &= E\left((e_i^2 - E(e_i^2 | x_i))^2 | x_i\right) \\ &= E(e_i^4 | x_i) - (E(e_i^2 | x_i))^2.\end{aligned}$$

If e_i is heteroskedastic, then $Var(\xi_i | x_i)$ will depend on x_i . In contrast, when e_i is independent of x_i then it is a constant

$$Var(\xi_i | x_i) = E(e_i^4) - \sigma^4$$

and under normality it simplifies to

$$Var(\xi_i | x_i) = 2\sigma^4.\tag{10.2}$$

10.2 Estimation of Skedastic Regression

Suppose e_i (and thus η_i) were observed. Then we could estimate α by OLS:

$$\hat{\alpha} = (Z'Z)^{-1} Z'\eta \rightarrow_p \alpha$$

and

$$\sqrt{n}(\hat{\alpha} - \alpha) \rightarrow_d N(0, V_\alpha)$$

where

$$V_\alpha = (E(z_i z_i'))^{-1} E(z_i z_i' \xi_i^2) (E(z_i z_i'))^{-1}. \quad (10.3)$$

While e_i is not observed, we have the OLS residual $\hat{e}_i = y_i - x_i' \hat{\beta} = e_i - x_i'(\hat{\beta} - \beta)$. Thus

$$\begin{aligned} \hat{\eta} - \eta_i &= \hat{e}_i^2 - e_i^2 \\ &= -2e_i x_i' (\hat{\beta} - \beta) + (\hat{\beta} - \beta)' x_i x_i' (\hat{\beta} - \beta) \\ &= \phi_i, \end{aligned}$$

say. Note that

$$\begin{aligned} \frac{1}{\sqrt{n}} \sum_{i=1}^n z_i \phi_i &= \frac{-2}{n} \sum_{i=1}^n z_i e_i x_i' \sqrt{n} (\hat{\beta} - \beta) + \frac{1}{n} \sum_{i=1}^n z_i (\hat{\beta} - \beta)' x_i x_i' (\hat{\beta} - \beta) \sqrt{n} \\ &\rightarrow_p 0 \end{aligned}$$

Let

$$\tilde{\alpha} = (Z'Z)^{-1} Z'\hat{\eta} \quad (10.4)$$

be from OLS regression of $\hat{\eta}_i$ on z_i . Then

$$\begin{aligned} \sqrt{n}(\tilde{\alpha} - \alpha) &= \sqrt{n}(\hat{\alpha} - \alpha) + (n^{-1}Z'Z)^{-1} n^{-1/2} Z'\phi \\ &\rightarrow_d N(0, V_\alpha) \end{aligned} \quad (10.5)$$

Thus the fact that η_i is replaced with $\hat{\eta}_i$ is asymptotically irrelevant. We may call (10.4) the *skedastic* regression, as it is estimating the conditional variance of the regression of y_i on x_i .

We have shown that α is consistently estimated by a simple procedure, and hence we can estimate $\sigma_i^2 = z_i' \alpha$ by $\tilde{\sigma}_i^2 = z_i' \tilde{\alpha}$. We now discuss how to use these results to test hypotheses on α , and construct a FGLS estimator for β .

10.3 Testing for Heteroskedasticity

The hypothesis of homoskedasticity is that $E(e_i^2 | x_i) = \sigma^2$, or equivalently that

$$H_0 : \alpha_1 = 0$$

in the regression (10.1). We may therefore test this hypothesis by the estimation (10.4) and constructing a Wald statistic.

This hypothesis does not imply that ξ_i is independent of x_i . Typically, however, we impose the stronger hypothesis and test the hypothesis that e_i is independent of x_i , in which case ξ_i is independent of x_i and the asymptotic variance (10.3) for $\tilde{\alpha}$ simplifies to

$$V_\alpha = (E(z_i z_i'))^{-1} E(\xi_i^2). \quad (10.6)$$

Hence the standard test of H_0 is a classic F (or Wald) test for exclusion of all regressors from the skedastic regression (10.4). The asymptotic distribution (10.5) and the asymptotic variance (10.6) under independence show that this test has an asymptotic chi-square distribution.

Theorem 10.3.1 *Under H_0 , and e_i independent of x_i , the Wald test of H_0 is asymptotically χ_q^2 .*

Most tests for heteroskedasticity take this basic form. The main differences between popular “tests” is which transformations of x_i enter z_i . Motivated by the form of the asymptotic variance of the OLS estimator $\hat{\beta}$, White (1980) proposed that the test for heteroskedasticity be based on setting z_i to equal all non-redundant elements of x_i , its squares, and all cross-products. Breusch-Pagan (1979) proposed what might appear to be a distinct test, but the only difference is that they allowed for general choice of z_i , and used an assumption of normality to use the simplification (10.2) for their test. If this simplification is replaced by the standard formula (under independence of the error), the two tests coincide.

10.4 Feasible GLS Estimation

Let

$$\tilde{\sigma}_i^2 = \tilde{\alpha}' z_i.$$

Suppose that $\tilde{\sigma}_i^2 > 0$ for all i . Then set

$$\tilde{D} = \text{diag}\{\tilde{\sigma}_1^2, \dots, \tilde{\sigma}_n^2\}$$

and

$$\tilde{\beta} = \left(X' \tilde{D}^{-1} X \right)^{-1} X' \tilde{D}^{-1} Y.$$

This is the feasible GLS, or FGLS, estimator of β .

Since there is not a unique specification for the conditional variance the FGLS estimator is not unique, and will depend on the model (and estimation method) for the skedastic regression.

One typical problem with implementation of FGLS estimation is that in a linear regression specification, there is no guarantee that $\tilde{\sigma}_i^2 > 0$ for all i . If $\tilde{\sigma}_i^2 < 0$ for some i , then the FGLS estimator is not well defined. Furthermore, if $\tilde{\sigma}_i^2 \approx 0$ for some i , then the FGLS estimator will force the regression equation through the point (y_i, x_i) , which is typically undesirable. This suggests that there is a need to bound the estimated variances away from zero. A trimming rule might make sense:

$$\bar{\sigma}_i^2 = \max[\tilde{\sigma}_i^2, \underline{\sigma}^2]$$

for some $\underline{\sigma}^2 > 0$.

It is possible to show that if the skedastic regression is correctly specified, then FGLS is asymptotically equivalent to GLS, but the proof of this is tricky in our notational structure. We just state the result without proof.

Theorem 10.4.1 *If the skedastic regression is correctly specified,*

$$\sqrt{n} \left(\tilde{\beta}_{GLS} - \tilde{\beta}_{FGLS} \right) \rightarrow_p 0,$$

and thus

$$\sqrt{n} \left(\tilde{\beta}_{FGLS} - \beta \right) \rightarrow_d N(0, V),$$

where

$$V = \left(E \left(\sigma_i^{-2} x_i x_i' \right) \right)^{-1}.$$

10.5 Covariance Matrix Estimation

Examining the asymptotic distribution of Theorem 10.4.1, the natural estimator of $Var(\tilde{\beta})$ is

$$\tilde{V}^0 = \left(\sum_{i=1}^n \tilde{\sigma}_i^{-2} x_i x_i' \right)^{-1} = \left(X' \tilde{D}^{-1} X \right)^{-1}.$$

It is consistent for V in the sense that $n\tilde{V}^0 \rightarrow_p V$ as $n \rightarrow \infty$.

\tilde{V}^0 is appropriate when the skedastic regression (10.1) is correctly specified.

It may be the case that $\alpha'z_i$ is only an approximation to the true conditional variance $\sigma_i^2 = E(e_i^2 | x_i)$. In this case we interpret $\alpha'z_i$ as a linear projection of e_i^2 on z_i . $\tilde{\beta}$ should perhaps be called a quasi-FGLS estimator of β . Its asymptotic variance is not that given in Theorem 10.4.1. Instead,

$$V = \left(E \left((\alpha'z_i)^{-1} x_i x_i' \right) \right)^{-1} \left(E \left((\alpha'z_i)^{-2} \sigma_i^2 x_i x_i' \right) \right) \left(E \left((\alpha'z_i)^{-1} x_i x_i' \right) \right)^{-1}.$$

V takes a sandwich form, similar to the covariance matrix of the OLS estimator. Unless $\sigma_i^2 = \alpha'z_i$, \tilde{V}^0 is inconsistent for V .

An appropriate solution is to use a White-type estimator in place of \tilde{V}^0 . This may be written as

$$\begin{aligned} \tilde{V} &= \left(\sum_{i=1}^n \tilde{\sigma}_i^{-2} x_i x_i' \right)^{-1} \left(\sum_{i=1}^n \tilde{\sigma}_i^{-4} \hat{e}_i^2 x_i x_i' \right) \left(\sum_{i=1}^n \tilde{\sigma}_i^{-2} x_i x_i' \right)^{-1} \\ &= \left(X' \tilde{D}^{-1} X \right)^{-1} \left(X' \tilde{D}^{-1} \hat{D} \tilde{D}^{-1} X \right) \left(X' \tilde{D}^{-1} X \right)^{-1} \end{aligned}$$

where $\hat{D} = \text{diag}\{\hat{e}_1^2, \dots, \hat{e}_n^2\}$. This is an estimator which is robust to misspecification of the conditional variance, and was proposed by Cragg (*Journal of Econometrics*, 1992).

10.6 Commentary: FGLS versus OLS

In a regression model, FGLS is asymptotically superior to OLS. Why then do we not exclusively estimate regression models by FGLS? This is a good question. There are three reasons.

First, FGLS estimation depends on specification and estimation of the skedastic regression. Since the form of the skedastic regression is unknown, and it may be estimated with considerable error, the estimated conditional variances may contain more noise than information about the true conditional variances. In this case, FGLS will do worse than OLS in practice.

Second, individual estimated conditional variances may be negative, and this requires trimming to solve. This introduces an element of arbitrariness which is unsettling to empirical researchers.

Third, OLS is a more robust estimator of the parameter vector. It is consistent not only in the regression model, but also under the assumptions of linear projection. The GLS and FGLS estimators, on the other hand, require the assumption of a correct conditional mean. If the equation of interest is a linear projection, and not a conditional mean, then the OLS and FGLS estimators will converge in probability to different limits, as they will be estimating two different projections. And the FGLS probability limit will depend on the particular function selected for the skedastic regression. The point is that the efficiency gains from FGLS are built on the stronger assumption of a correct conditional mean, and the cost is a reduction of robustness to misspecification.

Chapter 11

Generalized Method of Moments

11.1 Endogeneity

We say that there is a problem of endogeneity in the linear model $y = z_i'\beta + e_i$ if β is the parameter of interest and $E(z_i e_i) \neq 0$. This cannot happen if β is defined by linear projection, so requires a structural interpretation. The coefficient β must have meaning separately from the definition of a conditional mean or linear projection.

Example: Measurement error in the regressor. Suppose that (y_i, x_i^*) are joint random variables, $E(y_i | x_i^*) = x_i^{*'}\beta$ is linear, β is the parameter of interest, and x_i^* is not observed. Instead we observe $x_i = x_i^* + u_i$ where u_i is an $k \times 1$ measurement error, independent of y_i and x_i^* . Then

$$\begin{aligned}y_i &= x_i^{*'}\beta + e_i \\ &= (x_i - u_i)'\beta + e_i \\ &= x_i'\beta + v_i\end{aligned}$$

where

$$v_i = e_i - u_i'\beta.$$

The problem is that

$$E(x_i v_i) = E[(x_i^* + u_i)(e_i - u_i'\beta)] = -E(u_i u_i')\beta \neq 0$$

if $\beta \neq 0$ and $E(u_i u_i') \neq 0$. It follows that if $\hat{\beta}$ is the OLS estimator, then

$$\hat{\beta} \rightarrow_p \beta - (E(x_i x_i'))^{-1} E(u_i u_i')\beta \neq \beta.$$

This is called **measurement error bias**.

Example: Supply and Demand. The variables q_i and p_i (quantity and price) are determined jointly by the demand equation

$$q_i = -\beta_1 p_i + e_{1i}$$

and the supply equation

$$q_i = \beta_2 p_i + e_{2i}.$$

Assume that $e_i = \begin{pmatrix} e_{1i} \\ e_{2i} \end{pmatrix}$ is iid, $E e_i = 0$ and $E e_i e_i' = I_2$ (the latter for simplicity). The question is, if we regress q_i on p_i , what happens?

It is helpful to solve for q_i and p_i in terms of the errors. In matrix notation,

$$\begin{bmatrix} 1 & \beta_1 \\ 1 & -\beta_2 \end{bmatrix} \begin{pmatrix} q_i \\ p_i \end{pmatrix} = \begin{pmatrix} e_{1i} \\ e_{2i} \end{pmatrix}$$

so

$$\begin{aligned} \begin{pmatrix} q_i \\ p_i \end{pmatrix} &= \begin{bmatrix} 1 & \beta_1 \\ 1 & -\beta_2 \end{bmatrix}^{-1} \begin{pmatrix} e_{1i} \\ e_{2i} \end{pmatrix} \\ &= \frac{1}{\beta_1 + \beta_2} \begin{bmatrix} \beta_2 & \beta_1 \\ 1 & -1 \end{bmatrix} \begin{pmatrix} e_{1i} \\ e_{2i} \end{pmatrix} \\ &= \begin{pmatrix} \frac{\beta_2}{\beta_1 + \beta_2} e_{1i} + \frac{\beta_1}{\beta_1 + \beta_2} e_{2i} \\ \frac{1}{\beta_1 + \beta_2} (e_{1i} - e_{2i}) \end{pmatrix}. \end{aligned}$$

The regression of q_i on p_i yields

$$\begin{aligned} \hat{\beta} &= \frac{\sum_{i=1}^n p_i q_i}{\sum_{i=1}^n p_i^2} \\ &\xrightarrow{p} \frac{E(p_i q_i)}{E(p_i^2)} \\ &= \frac{E\left(\left(\frac{1}{\beta_1 + \beta_2} (e_{1i} - e_{2i})\right) \left(\frac{\beta_2}{\beta_1 + \beta_2} e_{1i} + \frac{\beta_1}{\beta_1 + \beta_2} e_{2i}\right)\right)}{E\left(\frac{1}{\beta_1 + \beta_2} (e_{1i} - e_{2i})\right)^2} \\ &= \frac{\beta_2 - \beta_1}{2}, \end{aligned}$$

which does not equal either β_1 or β_2 . This is called **simultaneous equations bias**.

11.2 Instrumental Variables

Let the equation of interest be

$$y_i = z_i' \beta + e_i \tag{11.1}$$

where z_i is $k \times 1$, and assume that $E(z_i e_i) \neq 0$ so that there is a problem of *endogeneity*. We call (11.1) the structural equation. In matrix notation, this can be written as

$$Y = Z\beta + e. \tag{11.2}$$

Any solution to the problem of endogeneity requires additional information which we call *instruments*.

Definition 11.2.1 *The $l \times 1$ random vector x_i is an instrumental variable for (11.1) if $E(x_i e_i) = 0$.*

In a typical set-up, some regressors in z_i will be uncorrelated with e_i (for example, at least the intercept). Thus we make the partition

$$z_i = \begin{pmatrix} z_{1i} \\ z_{2i} \end{pmatrix} \begin{matrix} k_1 \\ k_2 \end{matrix} \tag{11.3}$$

where $E(z_{1i}e_i) = 0$ yet $E(z_{2i}e_i) \neq 0$. We call z_{1i} exogenous and z_{2i} endogenous. By the above definition, z_{1i} is an instrumental variable for (11.1), so should be included in x_i . So we have the partition

$$x_i = \begin{pmatrix} z_{1i} \\ x_{2i} \end{pmatrix} \begin{matrix} k_1 \\ l_2 \end{matrix} \quad (11.4)$$

where $z_{1i} = x_{1i}$ are the *included exogenous variables*, and x_{2i} are the *excluded exogenous variables*. That is x_{2i} are variables which could be included in the equation for y_i (in the sense that they are uncorrelated with e_i) yet can be *excluded*, as they would have true zero coefficients in the equation.

For reasons which will become clear later, we say that the model is *just-identified* if $l = k$ (i.e., if $l_2 = k_2$) and *over-identified* if $l > k$ (i.e., if $l_2 > k_2$). If $l < k$ then the model is not identified.

We have noted that any solution to the problem of endogeneity requires instruments. This does not mean that valid instruments actually exist.

11.3 Reduced Form

The reduced form relationship between the variables or “regressors” z_i and the instruments x_i is found by linear projection. Let

$$\Gamma = E(x_i x_i')^{-1} E(x_i z_i')$$

be the $l \times k$ matrix of coefficients from a projection of z_i on x_i , and define

$$u_i = z_i - x_i' \Gamma$$

as the projection error. Then the reduced form linear relationship between z_i and x_i is

$$z_i = x_i' \Gamma + u_i. \quad (11.5)$$

In matrix notation, we can write (11.5) as

$$Z = X\Gamma + u \quad (11.6)$$

where u is $n \times k$.

By construction,

$$E(x_i u_i') = 0,$$

so (11.5) is a projection and can be estimated by OLS:

$$\begin{aligned} Z &= X\hat{\Gamma} + \hat{u} \\ \hat{\Gamma} &= (X'X)^{-1} (X'Z). \end{aligned}$$

Substituting (11.6) into (11.2), we find

$$\begin{aligned} Y &= (X\Gamma + u)\beta + e \\ &= X\lambda + v, \end{aligned} \quad (11.7)$$

where

$$\lambda = \Gamma\beta \quad (11.8)$$

and

$$v = u\beta + e.$$

Observe that

$$E(x_i v_i) = E(x_i u_i') \beta + E(x_i e_i) = 0.$$

Thus (11.7) is a projection equation and may be estimated by OLS. This is

$$\begin{aligned} Y &= X\hat{\lambda} + \hat{v}, \\ \hat{\lambda} &= (X'X)^{-1}(X'Y) \end{aligned}$$

The equation (11.7) is the reduced form for Y . (11.6) and (11.7) together are the *reduced form equations* for the system

$$\begin{aligned} Y &= X\lambda + v \\ Z &= X\Gamma + u. \end{aligned}$$

As we showed above, OLS yields the reduced-form estimates $(\hat{\lambda}, \hat{\Gamma})$

11.4 Identification

The structural parameter β relates to (λ, Γ) through (11.8). The parameter β is *identified*, meaning that it can be recovered from the reduced form, if

$$\text{rank}(\Gamma) = k. \tag{11.9}$$

Assume that (11.9) holds. If $l = k$, then $\beta = \Gamma^{-1}\lambda$. If $l > k$, then for any $W > 0$, $\beta = (\Gamma'W\Gamma)^{-1}\Gamma'W\lambda$.

If (11.9) is not satisfied, then β cannot be recovered from (λ, Γ) . Note that a necessary (although not sufficient) condition for (11.9) is $l \geq k$.

Since X and Z have the common variables X_1 , we can rewrite some of the expressions. Using (11.3) and (11.4) to make the matrix partitions $X = [X_1, X_2]$ and $Z = [X_1, Z_2]$, we can partition Γ as

$$\begin{aligned} \Gamma &= \begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{bmatrix} \\ &= \begin{bmatrix} I & \Gamma_{12} \\ 0 & \Gamma_{22} \end{bmatrix} \end{aligned}$$

(11.6) can be rewritten as

$$\begin{aligned} Z_1 &= X_1 \\ Z_2 &= X_1\Gamma_{12} + X_2\Gamma_{22} + u_2. \end{aligned} \tag{11.10}$$

β is identified if $\text{rank}(\Gamma) = k$, which is true if and only if $\text{rank}(\Gamma_{22}) = k_2$ (by the upper-diagonal structure of Γ). Thus the key to identification of the model rests on the $l_2 \times k_2$ matrix Γ_{22} in (11.10).

11.5 Instrumental Variables Estimation

Suppose the model is just-identified ($k = l$). Then $\beta = \Gamma^{-1}\lambda$. This suggests the Indirect Least Squares (ILS) estimator:

$$\begin{aligned}\hat{\beta} &= \hat{\Gamma}^{-1}\hat{\lambda} \\ &= \left((X'X)^{-1} (X'Z) \right)^{-1} \left((X'X)^{-1} (X'Y) \right) \\ &= (X'Z)^{-1} (X'X) (X'X)^{-1} (X'Y) \\ &= (X'Z)^{-1} (X'Y).\end{aligned}$$

$\hat{\beta}$ is also called the *instrumental variables estimator* of β , where X is used as an instrument for Z .

Since $(\hat{\lambda}, \hat{\Gamma}) \rightarrow_p (\lambda, \Gamma)$ and Γ is invertible, $\hat{\beta}$ is consistent by the CMT. A more direct way to see consistency is to substitute the equation $Y = Z\beta + e$ into the equation for $\hat{\beta}$ to obtain

$$\hat{\beta} = (X'Z)^{-1} (X' (Z\beta + e)) = \beta + (X'Z)^{-1} (X'e) \rightarrow_p 0$$

if $E(x_i z_i')$ is invertible and $E(x_i e_i) = 0$. These are the identifying assumptions.

We can also derive the IV estimator $\hat{\beta}$ as a MME. The defining characteristic of x_i as an instrumental variable is that $E(x_i e_i) = 0$. This suggests the moment equation

$$g_i(\beta) = x_i (y_i - z_i' \beta) \tag{11.11}$$

so that at the true value β_0 we have the equality

$$E(g_i(\beta_0)) = 0. \tag{11.12}$$

The sample analog is

$$\bar{g}_n(\beta) = \frac{1}{n} \sum_{i=1}^n g_i(\beta) = \frac{1}{n} \sum_{i=1}^n x_i (y_i - z_i' \beta) = \frac{1}{n} (X'Y - X'Z\beta). \tag{11.13}$$

The MME $\hat{\beta}$ sets $\bar{g}_n(\hat{\beta}) = 0$, vis.,

$$0 = \bar{g}_n(\hat{\beta}) = \frac{1}{n} (X'Y - X'Z\hat{\beta})$$

which yields $\hat{\beta} = (X'Z)^{-1} (X'Y)$, the IV estimator.

11.6 GMM Estimator

In the overidentified case ($l > k$), then the simple IV estimator described above does not exist. We generalize the MME. As above, we have (11.11), (11.12), and (11.13). Since $l > k$, $\bar{g}(\beta)$ is $l \times 1$ while β is $k \times 1$, so there is (in general) no $\hat{\beta}$ such that $\bar{g}_n(\hat{\beta}) = 0$. The idea of the generalized method of moments (GMM) is to set this vector “close” to zero.

For some $l \times l$ weight matrix $W_n > 0$, let

$$J(\beta) = n \cdot \bar{g}_n(\beta)' W_n \bar{g}_n(\beta).$$

This is a non-negative measure of the “length” of the vector $\bar{g}_n(\beta)$. For example, if $W_n = I$, then, $J(\beta) = n \cdot \bar{g}_n(\beta)' \bar{g}_n(\beta) = n \cdot |\bar{g}_n(\beta)|^2$, the square of the Euclidean length. The GMM estimator minimizes $J(\beta)$.

Definition 11.6.1 $\hat{\beta}_{GMM} = \underset{\beta}{\operatorname{argmin}} J(\beta)$.

Note that if $k = l$, then $\bar{g}_n(\hat{\beta}_{IV}) = 0$, so $J(\hat{\beta}_{IV}) = 0$ and thus $\hat{\beta}_{GMM} = \hat{\beta}_{IV}$. The first order conditions for the definition of the GMM estimator are

$$\begin{aligned} 0 &= \frac{\partial}{\partial \beta} J(\hat{\beta}) \\ &= 2 \frac{\partial}{\partial \beta} \bar{g}_n(\hat{\beta})' W_n \bar{g}_n(\hat{\beta}) \\ &= -2 \frac{1}{n} Z' X W_n \frac{1}{n} X' (Y - Z \hat{\beta}) \end{aligned}$$

so

$$2Z' X W_n X' Z \hat{\beta} = 2Z' X W_n X' Y$$

or

Proposition 11.6.1

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

While the estimator depends on W_n , the dependence is only up to scale, for if W_n is replaced by cW_n for some $c > 0$, $\hat{\beta}_{GMM}$ does not change.

11.7 2SLS Estimator

Suppose $W_n = (X'X)^{-1}$. Then the GMM estimator equals

$$\hat{\beta}_{2SLS} = \left(Z' X (X'X)^{-1} X' Z \right)^{-1} Z' X (X'X)^{-1} X' Y$$

This is called the 2SLS estimator. It was originally proposed by Theil (1953) and Basman (1957), and is the classic estimator for linear equations with instruments.

It is useful to observe that writing

$$\begin{aligned} P_X &= X (X'X)^{-1} X', \\ \hat{Z} &= P_X Z = X (X'X)^{-1} X' Z, \end{aligned}$$

then

$$\begin{aligned} \hat{\beta} &= (Z' P_X Z)^{-1} Z' P_X Y \\ &= (\hat{Z}' \hat{Z})^{-1} \hat{Z}' Y. \end{aligned}$$

The source of the “two-stage” name is since it can be computed as follows

- First regress Z on X , vis., $\hat{\Gamma} = (X'X)^{-1} X' Z$ and $\hat{Z} = X \hat{\Gamma} = P_X Z$.
- Second, regress Y on \hat{Z} , vis., $\hat{\beta} = (\hat{Z}' \hat{Z})^{-1} \hat{Z}' Y$.

It is useful to scrutinize the projection \hat{Z} . Recall, $Z = [Z_1, Z_2]$ and $X = [Z_1, X_2]$. Then

$$\begin{aligned}\hat{Z} &= \begin{bmatrix} \hat{Z}_1 \\ \hat{Z}_2 \end{bmatrix} \\ &= [P_X Z_1, P_X Z_2] \\ &= [Z_1, P_X Z_2] \\ &= \begin{bmatrix} Z_1 \\ \hat{Z}_2 \end{bmatrix},\end{aligned}$$

since Z_1 lies in the span of X . Thus in the second stage, we regress Y on Z_1 and \hat{Z}_2 . So only the endogenous variables Z_2 are replaced by their fitted values:

$$\hat{Z}_2 = X_1 \hat{\Gamma}_{12} + X_2 \hat{\Gamma}_{22}.$$

11.8 Distribution of GMM Estimator

Assume that $W_n \rightarrow_p W > 0$. Let

$$Q = E(x_i z_i')$$

and

$$\Omega = E(x_i x_i e_i^2) = E(g_i g_i'),$$

where $g_i = x_i e_i$. Then

$$\left(\frac{1}{n} Z' X\right) W_n \left(\frac{1}{n} X' Z\right) \rightarrow_p Q' W Q$$

and

$$\left(\frac{1}{n} Z' X\right) W_n \left(\frac{1}{n} X' e\right) \rightarrow_p Q' W N(0, \Omega).$$

We conclude:

Theorem 11.8.1 $\sqrt{n}(\hat{\beta} - \beta) \rightarrow_d N(0, V),$

$$V = (Q' W Q)^{-1} (Q' W \Omega W Q) (Q' W Q)^{-1}.$$

In general, GMM estimators are asymptotically normal with “sandwich form” asymptotic variances.

11.9 Optimal Weight Matrix

The optimal weight matrix W_0 is one which minimizes V . This turns out to be $W_0 = \Omega^{-1}$. The proof is left as an exercise. This yields the *efficient GMM* estimator:

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

Thus we have

Theorem 15 *For the efficient GMM estimator, $\sqrt{n}(\hat{\beta} - \beta) \rightarrow_d N(0, (Q' \Omega^{-1} Q)^{-1})$.*

This estimator is efficient only in the sense that it is the best (asymptotically) in the class of GMM estimators with this set of moment conditions.

$W_0 = \Omega^{-1}$ is not known in practice, but it can be estimated consistently. For any $W_n \rightarrow_p W_0$, we still call $\hat{\beta}$ the efficient GMM estimator, as it has the same asymptotic distribution.

In the special case that $E(e_i^2 | x_i) = \sigma^2$ (homoskedasticity), then

$$\begin{aligned} W_0 &= \sigma^{-2} (E(x_i x_i'))^{-1} \\ &\propto (E(x_i x_i'))^{-1}. \end{aligned}$$

Recall, 2SLS sets $W_n = (X'X)^{-1}$. So under the homoskedasticity assumption 2SLS is *asymptotically efficient*.

In general, however, if $l > k$, then 2SLS is asymptotically inefficient.

11.10 Estimation of the Efficient Weight Matrix

Let

$$W_n = \left(\frac{1}{n} \sum_{i=1}^n \hat{g}_i \hat{g}_i' \right)^{-1}$$

where $\hat{g}_i = x_i \hat{e}_i$, and \hat{e}_i are consistent first-stage residuals, such as the 2SLS residuals. Then $W_n \rightarrow_p \Omega^{-1} = W_0$, and GMM using W_n is asymptotically efficient. This works in general, under homoskedasticity and heteroskedasticity.

Alastair Hall (2000) has shown that this estimator may have difficulties for testing purposes. In a testing context, the moment conditions may be violated, in which case $Eg_i \neq 0$, so W_n above will contain the sum of a bias and a variance component, which is undesirable. The solution is to re-center the moment conditions. Let

$$\bar{g}_n = \bar{g}_n(\hat{\beta}) = \frac{1}{n} \sum_{i=1}^n \hat{g}_i$$

and set

$$\begin{aligned} W_n &= \left(\frac{1}{n} \sum_{i=1}^n (\hat{g}_i - \bar{g}_n) (\hat{g}_i - \bar{g}_n)' \right)^{-1} \\ &= \left(\frac{1}{n} \sum_{i=1}^n \hat{g}_i \hat{g}_i' - \bar{g}_n \bar{g}_n' \right)^{-1} \end{aligned}$$

as the GMM weight matrix.

Here is a simple way to compute the efficient GMM estimator.

First, estimate $\hat{\beta}$ by 2SLS, and construct the residual $\hat{e}_i = y_i - z_i' \hat{\beta}$. Then set $\hat{g}_i = x_i \hat{e}_i$, and let \hat{g} be the associated $n \times l$ matrix. The efficient GMM estimator is either

$$\hat{\beta} = \left(Z'X (\hat{g}'\hat{g})^{-1} X'Z \right)^{-1} Z'X (\hat{g}'\hat{g})^{-1} X'Y.$$

or

$$\hat{\beta} = \left(Z'X (\hat{g}'\hat{g} - n\bar{g}_n\bar{g}_n')^{-1} X'Z \right)^{-1} Z'X (\hat{g}'\hat{g} - n\bar{g}_n\bar{g}_n')^{-1} X'Y.$$

In most cases, when we say “GMM”, we actually mean “efficient GMM”, as there is little point in using an inefficient GMM estimator, and it is so easy to compute.

An estimator of $Var(\hat{\beta})$ can be seen from the above formula. We set either

$$\hat{V} = \left(Z'X (\hat{g}'\hat{g})^{-1} X'Z \right)^{-1}.$$

or

$$\hat{V} = \left(Z'X (\hat{g}'\hat{g} - n\bar{g}_n\bar{g}'_n)^{-1} X'Z \right)^{-1}.$$

Asymptotic standard errors are given by the square roots of the diagonal elements of \hat{V} .

11.11 Over-Identification Test

If $l > k$, there are restrictions which are testable.

Theorem 11.11.1 (*Sargon-Hansen*). *Under the hypothesis of correct specification, and if the weight matrix is asymptotically efficient,*

$$J = J(\hat{\beta}) \rightarrow_d \chi^2_{l-k}.$$

The proof of the theorem is left as an exercise. This result was established by Sargon (1958) for the 2SLS estimator under conditional homoskedasticity, and by L. Hansen (1982) for the general case.

Note that the degrees of freedom of the asymptotic distribution are the number of overidentifying restrictions.

If the statistic J exceeds the chi-square critical value, we can reject the model. Based on this information alone, it is unclear what is wrong, but it is typically cause for concern. The GMM overidentification test is a very useful by-product of the GMM methodology, and it is advisable to report the statistic J whenever GMM is the estimation method.

It may be helpful to observe that

$$\begin{aligned} J &= n \bar{g}'_n W_n \bar{g}_n \\ &= n^2 \bar{g}'_n (\hat{g}'\hat{g} - n\bar{g}_n\bar{g}'_n)^{-1} \bar{g}_n. \end{aligned}$$

When over-identified models are estimated by GMM, it is customary to report the J statistic as a general test of model adequacy.

11.12 GMM: The General Case

In its most general form, GMM applies whenever an economic or statistical model implies the $l \times 1$ moment condition

$$E(g_i(\beta)) = 0.$$

Often, this is *all* that is known. Identification requires $l \geq k = \dim(\beta)$. The GMM estimator minimizes

$$J(\beta) = n \cdot \bar{g}_n(\beta)' W_n \bar{g}_n(\beta)$$

where

$$\bar{g}_n(\beta) = \frac{1}{n} \sum_{i=1}^n g_i(\beta)$$

and either

$$W_n = \left(\frac{1}{n} \sum_{i=1}^n \hat{g}_i \hat{g}_i' \right)^{-1},$$

or

$$W_n = \left(\frac{1}{n} \sum_{i=1}^n \hat{g}_i \hat{g}_i' - \bar{g}_n \bar{g}_n' \right)^{-1},$$

with $\hat{g}_i = g_i(\tilde{\beta})$ constructed using $\tilde{\beta}$, a preliminary consistent estimator, perhaps obtained by first setting $W_n = I$. Since the GMM estimator depends upon the first-stage estimator, often the weight matrix W_n is updated, and then $\hat{\beta}$ recomputed. This estimator can be iterated if needed.

Theorem 11.12.1 *Under general regularity conditions, $\sqrt{n}(\hat{\beta} - \beta) \rightarrow_d N(0, (G'\Omega^{-1}G)^{-1})$, where $\Omega = (E(g_i g_i'))^{-1}$ and $G = E\frac{\partial}{\partial \beta'} g_i(\beta)$. The variance of $\hat{\beta}$ may be estimated by $(\hat{G}'\hat{\Omega}^{-1}\hat{G})^{-1}$ where $\hat{\Omega} = n^{-1} \sum_i \hat{g}_i \hat{g}_i'$ or $\hat{\Omega} = n^{-1} \sum_i \hat{g}_i \hat{g}_i' - \bar{g}_n \bar{g}_n'$, and $\hat{G} = n^{-1} \sum_i \frac{\partial}{\partial \beta'} g_i(\hat{\beta})$.*

The general theory of GMM estimation and testing was exposted by L. Hansen (1982).

11.13 Hypothesis Testing: The Distance Statistic

We described before how to construct estimates of the asymptotic covariance matrix of the GMM estimates. These may be used to construct Wald tests of statistical hypotheses.

If the hypothesis is non-linear, a better 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).

For a given weight matrix W_n , the GMM criterion function is

$$J(\beta) = n \cdot \bar{g}_n(\beta)' W_n \bar{g}_n(\beta)$$

For $h : R^k \rightarrow R^r$, the hypothesis is

$$H_0 : h(\beta) = 0.$$

The estimates under H_1 are

$$\hat{\beta} = \arg \min_{\beta} J(\beta)$$

and those under H_0 are

$$\tilde{\beta} = \arg \min_{h(\beta)=0} J(\beta).$$

The two minimizing criterion functions are $J(\hat{\beta})$ and $J(\tilde{\beta})$. The GMM distance statistic is the difference

$$D = J(\tilde{\beta}) - J(\hat{\beta}).$$

Proposition 11.13.1 *If the same weight matrix W_n is used for both null and alternative,*

1. $D \geq 0$
2. $D \rightarrow_d \chi_r^2$

3. If h is linear in β , then D equals the Wald statistic.

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 W_n 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.

This test shares the useful feature of LR tests is that it is a natural by-product of the computation of alternative models.

11.14 GMM as Semiparametrically Efficient

We have described the estimator $\hat{\beta}$ as “efficient GMM” if the optimal (variance minimizing) weight matrix is selected. This is a weak concept of optimality, as we are only considering alternative weight matrices W_n . However, it turns out that the GMM estimator is semiparametrically efficient, as shown by Gary Chamberlain (1987).

If it is known that $E(g_i(\beta)) = 0$, and this is all that is known, this is a semi-parametric problem, as the distribution of the data is unknown. Chamberlain showed that in this context, no semiparametric estimator (one which is consistent globally for the class of models considered) can have a smaller asymptotic variance than $(G'\Omega^{-1}G)^{-1}$. Since the GMM estimator has this asymptotic variance, it is semiparametrically efficient.

This results shows that in the linear model, no estimator has greater asymptotic efficiency than the efficient linear GMM estimator. No estimator can do better (in this first-order asymptotic sense), without imposing additional assumptions.

11.15 Conditional Moment Restrictions

In many contexts, the model implies more than an unconditional moment restriction of the form $Eg_i(\beta) = 0$. It implies a conditional moment restriction of the form

$$E(e_i(\beta) | x_i) = 0$$

where $e_i(\beta)$ is some $s \times 1$ function of the observation and the parameters. In many cases, $s = 1$.

It turns out that this conditional moment restriction is much more powerful, and restrictive, than the unconditional moment restriction discussed above.

Our linear model $y_i = z_i'\beta + e_i$ with instruments x_i falls into this class under the stronger assumption $E(e_i | x_i) = 0$. Then $e_i(\beta) = y_i - z_i'\beta$.

It is also helpful to realize that conventional regression models also fall into this class, except that in this case $z_i = x_i$. For example, in linear regression, $e_i(\beta) = y_i - x_i'\beta$, while in a nonlinear regression model $e_i(\beta) = y_i - g(x_i, \beta)$. In a joint model of the conditional mean and variance

$$e_i(\beta, \gamma) = \begin{cases} y_i - x_i'\beta \\ (y_i - x_i'\beta)^2 - f(x_i)'\gamma \end{cases}.$$

Here $s = 2$.

Given a conditional moment restriction, an unconditional moment restriction can always be constructed. That is for any $l \times 1$ function $\phi(x_i, \beta)$, we can set $g_i(\beta) = \phi(x_i, \beta)e_i(\beta)$ which satisfies

$Eg_i(\beta) = 0$ and hence defines a GMM estimator. The obvious problem is that the class of functions ϕ is infinite. Which should be selected?

This is equivalent to the problem of selection of the best instruments. If x_i is a valid instrument satisfying $E(e_i | x_i) = 0$, then x_i, x_i^2, x_i^3, \dots , etc., are all valid instruments. Which both be used?

Once solution is to construct an infinite list of potent instruments, and then use the first k instruments. How is k to be determined? This is an area of theory still under development. A recent study of this problem is Donald and Newey (2001).

Another approach is to construct the *optimal instrument*. The form was uncovered by Chamberlain (1987). Take the case $s = 1$. Let

$$R_i = E \left(\frac{\partial}{\partial \beta} e_i(\beta) \mid x_i \right)$$

and

$$\sigma_i^2 = E(e_i(\beta)^2 \mid x_i).$$

Then the “optimal instrument” is

$$A_i = \sigma_i^{-2} R_i$$

so the optimal moment is

$$g_i(\beta) = A_i e_i(\beta).$$

Setting $g_i(\beta)$ to be this choice (which is $k \times 1$, so is just-identified) yields the best GMM estimator possible.

In practice, A_i is unknown, but its form does help us think about construction of optimal instruments.

In the linear model $e_i(\beta) = y_i - z_i' \beta$, note that

$$R_i = -E(z_i \mid x_i)$$

and

$$\sigma_i^2 = E(e_i^2 \mid x_i),$$

so

$$A_i = \sigma_i^{-2} E(z_i \mid x_i).$$

In the case of linear regression, $z_i = x_i$, so $A_i = \sigma_i^{-2} x_i$. Hence efficient GMM is GLS, as we discussed earlier in the course.

In the case of endogenous variables, note that the efficient instrument A_i involves the estimation of the conditional mean of z_i given x_i . In other words, to get the best instrument for z_i , we need the best conditional mean model for z_i given x_i , not just an arbitrary linear projection. The efficient instrument is also inversely proportional to the conditional variance of e_i . This is the same as the GLS estimator; namely that improved efficiency can be obtained if the observations are weighted inversely to the conditional variance of the errors.

11.16 Continuously-Updated GMM

Let the weight matrix be considered as a function of β . Set

$$W_n(\beta) = \left(\frac{1}{n} \sum_{i=1}^n g_i(\beta) g_i(\beta)' \right)^{-1}$$

or

$$W_n(\beta) = \left(\frac{1}{n} \sum_{i=1}^n g_i^*(\beta) g_i^*(\beta)' - \bar{g}_n(\beta) \bar{g}_n(\beta)' \right)^{-1}.$$

Then define the criterion function

$$J(\beta) = n \cdot \bar{g}_n(\beta)' W_n(\beta) \bar{g}_n(\beta).$$

The $\hat{\beta}$ which minimizes this function is called the “continuously-updated GMM estimator”, and was introduced by L. Hansen, Heaton and Yaron (1996).

The estimator appears to have some better properties than traditional GMM, but can be numerically tricky to obtain in some cases. This is a current area of research in econometrics.

11.17 Empirical Likelihood

A new alternative to GMM but with great similarities is empirical likelihood. The idea is due to Art Owen (1988, 2001) and has been extended to the GMM context by Qin and Lawless (1994). It provides a non-parametric analog of likelihood estimation suitable for the GMM context.

The idea is as follows. Let $w_i = (y_i, z_i, x_i)$, which has some distribution F . This distribution may be well-approximated non-parametrically by the empirical distribution function F_n , which places probability $1/n$ at each observation w_i . Now consider a general multinomial distribution which places probability p_i at each observation w_i , with the constraint that $\sum_{i=1}^n p_i = 1$. The log-likelihood is

$$EL_n = \sum_{i=1}^n \log(p_i) \quad s.t. \quad \sum_{i=1}^n p_i = 1.$$

This “empirical likelihood” is maximized by setting $p_i = 1/n$ for all i , which yields $EL_n = -n \log(n)$.

Our model specifies that

$$E(g_i(\beta)) = 0. \tag{11.14}$$

where g_i is $l \times 1$ and β is $k \times 1$. The empirical likelihood function for any β is

$$EL_n(\beta) = \max_{p_1, \dots, p_n} \sum_{i=1}^n \log(p_i) \quad s.t. \quad \sum_{i=1}^n p_i = 1 \quad \text{and} \quad \sum_{i=1}^n p_i g_i(\beta) = 0.$$

The condition $\sum_{i=1}^n p_i g_i(\beta) = 0$ imposes the moment condition (11.14).

This looks like a difficult maximization problem, but it turns out that it is quite simple to solve using a $l \times 1$ vector of Lagrange multipliers λ which are dependent on β . Let $\lambda(\beta)$ satisfy

$$\sum_{i=1}^n \frac{g_i(\beta)}{1 + \lambda(\beta)' g_i(\beta)} = 0$$

which is found by a numerical search. Then setting

$$p_i(\beta) = \frac{1}{n(1 + \lambda(\beta)' g_i(\beta))}$$

we see that

$$\sum_{i=1}^n p_i(\beta) g_i(\beta) = 0$$

as desired.

The resulting empirical log-likelihood is

$$\begin{aligned} EL_n(\beta) &= \sum_{i=1}^n \log \left(\frac{1}{n(1 + \lambda(\beta)'g_i(\beta))} \right) \\ &= -n \log n - \sum_{i=1}^n \log (1 + \lambda(\beta)'g_i(\beta)) \end{aligned}$$

The empirical likelihood estimate of β numerically maximizes this function

$$\hat{\beta} = \arg \max EL_n(\beta).$$

As a by-product, one obtains the probabilities

$$\hat{p}_i = \frac{1}{n(1 + \lambda(\hat{\beta})'g_i(\hat{\beta}))}$$

and the empirical likelihood

$$EL_n = - \sum_{i=1}^n \log \hat{p}_i.$$

We will discuss the EL probabilities \hat{p}_i later in our coverage of the GMM bootstrap.

This estimator has the same asymptotic distribution as the efficient GMM estimator, so standard errors can be calculated similarly.

There is mounting evidence that the empirical likelihood estimator has better sampling properties than the GMM estimator. This is a current area of research in econometrics, including Kitamura and Stutzer (1997), Imbens (1997), Imbens, Spady and Johnson (1998), and Kitamura (2001).

For testing, empirical likelihood lends itself nicely to LR-type test statistics, rather than Wald test statistics. This requires the maximization of $EL_n(\beta)$ subject to the hypothesis, and is similar to the GMM distance statistic discussed before.

11.18 Bootstrap GMM Inference

Let $w_i = (y_i, x_i, z_i)$. Using the EDF of w_i , we can apply the bootstrap methods discussed before to compute estimates of the bias and variance of functions of the 2SLS estimator $\hat{\beta}$, and construct confidence intervals for $\theta = h(\beta)$, identically as in the regression model. However, caution should be applied when interpreting such results.

A straightforward application of the nonparametric bootstrap works in the sense of consistently achieving the first-order asymptotic distribution. This has been shown by Hahn (1996). However, it fails to achieve an asymptotic refinement when the model is over-identified, jeopardizing the theoretical justification for percentile-t methods. Furthermore, the bootstrap applied J test will yield the wrong answer.

The problem is that in the sample, $\hat{\beta}$ is the “true” value and yet $\bar{g}_n(\hat{\beta}) \neq 0$. Thus according to random variables w_i^* drawn from the EDF F_n ,

$$E \left(g_i \left(\hat{\beta} \right) \right) = \bar{g}_n(\hat{\beta}) \neq 0.$$

This means that w_i^* do not satisfy the same moment conditions as the population distribution.

A correction suggested by Hall and Horowitz (1996) can solve the problem. Given the bootstrap sample (Y^*, X^*, Z^*) , define the bootstrap GMM criterion

$$J^*(\beta) = n \cdot \left(\bar{g}_n^*(\beta) - \bar{g}_n(\hat{\beta}) \right)' W_n^* \left(\bar{g}_n^*(\beta) - \bar{g}_n(\hat{\beta}) \right)$$

where $\bar{g}_n(\hat{\beta})$ is from the in-sample data, not from the bootstrap data.

Let $\hat{\beta}^*$ minimize $J^*(\beta)$, and define all statistics and tests accordingly. In the linear model, this implies that the bootstrap estimator is

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

where $\hat{e} = Y - Z\hat{\beta}$ are the in-sample residuals. The bootstrap J statistic is $J^*(\hat{\beta}^*)$.

Brown and Newey (2001) have an alternative solution. They note that we can sample from the observations $\{w_1, \dots, w_n\}$ with the empirical likelihood probabilities $\{\hat{p}_i\}$ mentioned before. Since $\sum_{i=1}^n \hat{p}_i g_i(\hat{\beta}) = 0$, this sampling scheme preserves the moment conditions of the model, so no recentering or adjustments are needed. Brown and Newey argue that this bootstrap procedure will be more efficient than the Hall-Horowitz GMM bootstrap.

To date, there are very few empirical applications of bootstrap GMM, as this is a very new area of research.

11.19 Bekker Asymptotics

Bekker (1994) used an alternative asymptotic framework to analyze the finite-sample bias in the 2SLS estimator. Here we present a simplified version of one of his results. In our notation, the model is

$$Y = Z\beta + e \tag{11.15}$$

$$Z = X\Gamma + u \tag{11.16}$$

$$\xi = (e, u)$$

$$E(\xi | X) = 0$$

$$E(\xi\xi' | X) = S$$

As before, X is $n \times l$ so there are l instruments.

First, let's analyze the approximate bias of OLS applied to (11.15). Using (11.16),

$$E\left(\frac{1}{n} Z' e\right) = E(z_i e_i) = \Gamma' E(x_i e_i) + E(u_i e_i) = S_{21}$$

and

$$\begin{aligned} E\left(\frac{1}{n} Z' Z\right) &= E(z_i z_i') \\ &= \Gamma' E(x_i x_i') \Gamma + E(u_i x_i') \Gamma + \Gamma' E(x_i u_i') + E(u_i u_i') \\ &= \Gamma' Q \Gamma + S_{22} \end{aligned}$$

where $Q = E(x_i x_i')$. Hence by a first-order approximation

$$\begin{aligned} E\left(\hat{\beta}_{OLS} - \beta\right) &\approx \left(E\left(\frac{1}{n}Z'Z\right)\right)^{-1} E\left(\frac{1}{n}Z'e\right) \\ &= (\Gamma'Q\Gamma + S_{22})^{-1} S_{21} \end{aligned} \quad (11.17)$$

which is zero only when $S_{21} = 0$ (when Z is exogenous).

We now derive a similar result for the 2SLS estimator.

$$\hat{\beta}_{2SLS} = (Z'P_X Z)^{-1} (Z'P_X Y).$$

Let $P_X = X(X'X)^{-1}X'$. By the spectral decomposition of an idempotent matrix, $P = H\Lambda H'$ where $\Lambda = \text{diag}(I_l, 0)$. Let $q = H'\xi S^{-1/2}$ which satisfies $Eqq' = I_n$ and partition $q = (q_1' \ q_2')$ where q_1 is $l \times 1$. Hence

$$\begin{aligned} E\left(\frac{1}{n}\xi'P_X\xi\right) &= \frac{1}{n}S^{1/2'}E(q'\Lambda q)S^{1/2} \\ &= \frac{1}{n}S^{1/2'}E\left(\frac{1}{n}q_1'q_1\right)S^{1/2} \\ &= \frac{l}{n}S^{1/2'}S^{1/2} \\ &= \alpha S \end{aligned}$$

where

$$\alpha = \frac{l}{n}.$$

Using (11.16) and this result,

$$\frac{1}{n}E(Z'P_X e) = \frac{1}{n}E(\Gamma'X'e) + \frac{1}{n}E(u'P_X e) = \alpha S_{21},$$

and

$$\begin{aligned} \frac{1}{n}E(Z'P_X Z) &= \Gamma'E(x_i x_i')\Gamma + \Gamma'E(x_i u_i) + E(u_i x_i')\Gamma + \frac{1}{n}E(u'P_X u) \\ &= \Gamma'Q\Gamma + \alpha S_{22}. \end{aligned}$$

Together

$$\begin{aligned} E\left(\hat{\beta}_{2SLS} - \beta\right) &\approx \left(E\left(\frac{1}{n}Z'P_X Z\right)\right)^{-1} E\left(\frac{1}{n}Z'P_X e\right) \\ &= \alpha (\Gamma'Q\Gamma + \alpha S_{22})^{-1} S_{21}. \end{aligned} \quad (11.18)$$

In general this is non-zero, except when $S_{21} = 0$ (when Z is exogenous). It is also close to zero when $\alpha = 0$. Becker (1994) pointed out that it also has the reverse implication – that when $\alpha = l/n$ is large, the bias in the 2SLS estimator will be large. Indeed as $\alpha \rightarrow 1$, the expression in (11.18) approaches that in (11.17), indicating that the bias in 2SLS approaches that of OLS as the number of instruments increases.

Becker (1994) showed further that under the alternative asymptotic approximation that α is fixed as $n \rightarrow \infty$ (so that the number of instruments goes to infinity proportionately with sample size) then the expression in (11.18) is the probability limit of $\hat{\beta}_{2SLS} - \beta$

11.20 Identification Failure

Recall the reduced form equation

$$Z_2 = X_1\Gamma_{12} + X_2\Gamma_{22} + u_2.$$

The parameter β fails to be identified if Γ_{22} has deficient rank. The consequences of identification failure for inference are quite severe.

Take the simplest case where $k = l = 1$ (so there is no X_1). Then the model may be written as

$$\begin{aligned} y_i &= z_i\beta + e_i \\ z_i &= x_i\gamma + u_i \end{aligned}$$

and $\Gamma_{22} = \gamma = E(x_i z_i) / E x_i^2$. We see that β is identified if and only if $\Gamma_{22} = \gamma \neq 0$, which occurs when $E(z_i x_i) \neq 0$. Thus identification hinges on the existence of correlation between the excluded exogenous variable and the included endogenous variable.

Suppose this condition fails, so $E(z_i x_i) = 0$. Then by the CLT

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n x_i e_i \rightarrow_d N_1 \sim N(0, E(x_i^2 e_i^2)) \quad (11.19)$$

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n x_i z_i = \frac{1}{\sqrt{n}} \sum_{i=1}^n x_i u_i \rightarrow_d N_2 \sim N(0, E(x_i^2 u_i^2)) \quad (11.20)$$

therefore

$$\hat{\beta} - \beta = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n x_i e_i}{\frac{1}{\sqrt{n}} \sum_{i=1}^n x_i z_i} \rightarrow_d \frac{N_1}{N_2} \sim \text{Cauchy},$$

since the ratio of two normals is Cauchy. This is particularly nasty, as the Cauchy distribution does not have a finite mean. This result carries over to more general settings, and was examined by Phillips (1989) and Choi and Phillips (1992).

Suppose that identification does not complete fail, but is *weak*. This occurs when Γ_{22} is full rank, but *small*. This can be handled in an asymptotic analysis by modeling it as local-to-zero, viz

$$\Gamma_{22} = n^{-1/2}C,$$

where C is a full rank matrix. The $n^{-1/2}$ is picked because it provides just the right balancing to allow a rich distribution theory.

To see the consequences, once again take the simple case $k = l = 1$. Here, the instrument x_i is weak for z_i if

$$\gamma = n^{-1/2}c.$$

Then (11.19) is unaffected, but (11.20) instead takes the form

$$\begin{aligned} \frac{1}{\sqrt{n}} \sum_{i=1}^n x_i z_i &= \frac{1}{\sqrt{n}} \sum_{i=1}^n x_i^2 \gamma + \frac{1}{\sqrt{n}} \sum_{i=1}^n x_i u_i \\ &= \frac{1}{n} \sum_{i=1}^n x_i^2 c + \frac{1}{\sqrt{n}} \sum_{i=1}^n x_i u_i \\ &\rightarrow_d Qc + N_2 \end{aligned}$$

therefore

$$\hat{\beta} - \beta \rightarrow_d \frac{N_1}{Qc + N_2}.$$

As in the case of complete identification failure, we find that $\hat{\beta}$ is inconsistent for β and the asymptotic distribution of $\hat{\beta}$ is non-normal. In addition, standard test statistics have non-standard distributions, meaning that inferences about parameters of interest can be misleading.

The distribution theory for this model was developed by Staiger and Stock (1997) and extended to nonlinear GMM estimation by Stock and Wright (2000). Further results on testing were obtained by Wang and Zivot (1998).

Furthermore, bootstrap inference cannot solve this problem. The bootstrap distribution is not consistent for the non-standard asymptotic distributions obtained under identification failure. Therefore the bootstrap cannot be viewed as a panacea for solving all distributional problems.

The bottom line is that it is highly desirable to avoid identification failure. Once again, the equation to focus on is the reduced form

$$Z_2 = X_1\Gamma_{12} + X_2\Gamma_{22} + u_2$$

and identification requires $\text{rank}(\Gamma_{22}) = k_2$. If $k_2 = 1$, this requires $\Gamma_{22} \neq 0$, which is straightforward to assess using a hypothesis test on the reduced form. Therefore in the case of $k_2 = 1$ (one RHS endogenous variable), one constructive recommendation is to explicitly estimate the reduced form equation for Z_2 , construct the test of $\Gamma_{22} = 0$, and at a minimum check that the test rejects $H_0 : \Gamma_{22} = 0$.

When $k_2 > 1$, $\Gamma_{22} \neq 0$ is not sufficient for identification. It is not even sufficient that each column of Γ_{22} is non-zero (each column corresponds to a distinct endogenous variable in X_2). So while a minimal check is to test that each columns of Γ_{22} is non-zero, this cannot be interpreted as definitive proof that Γ_{22} has full rank. Unfortunately, tests of deficient rank are difficult to implement. In any event, it appears reasonable to explicitly estimate and report the reduced form equations for X_2 , and attempt to assess the likelihood that Γ_{22} has deficient rank.

Chapter 12

Univariate Time Series

A time series y_t is a process observed in sequence over time, $t = 1, \dots, T$. To indicate the dependence on time, we adopt new notation, and use the subscript t to denote the individual observation, and T to denote the number of observations.

Because of the sequential nature of time series, we expect that Y_t and Y_{t-1} are *not* independent, so classical assumptions are not valid.

We can separate time series into two categories: univariate ($y_t \in R$ is scalar); and multivariate ($y_t \in R^m$ is vector-valued). The primary model for univariate time series is autoregressions (ARs). The primary model for multivariate time series is vector autoregressions (VARs).

12.1 Stationarity and Ergodicity

Definition 12.1.1 $\{Y_t\}$ is covariance (weakly) stationary if

$$E(Y_t) = \mu$$

is independent of t , and

$$\text{Cov}(Y_t, Y_{t-k}) = \gamma(k)$$

is independent of t for all k .

$\gamma(k)$ is called the autocovariance function.

Definition 12.1.2 $\{Y_t\}$ is strictly stationary if the joint distribution of (Y_t, \dots, Y_{t-k}) is independent of t for all k .

Definition 12.1.3 $\rho(k) = \gamma(k)/\gamma(0) = \text{Corr}(Y_t, Y_{t-k})$ is the autocorrelation function.

Definition 12.1.4 (loose). A stationary time series is ergodic if $\gamma(k) \rightarrow 0$ as $k \rightarrow \infty$.

The following two theorems are essential to the analysis of stationary time series. Their proofs are rather difficult, however.

Theorem 12.1.1 If Y_t is strictly stationary and ergodic and $X_t = f(Y_t, Y_{t-1}, \dots)$ is a random variable, then X_t is strictly stationary and ergodic.

Theorem 12.1.2 (*Ergodic Theorem*). If X_t is strictly stationary and ergodic and $E|X_t| < \infty$, then as $T \rightarrow \infty$,

$$\frac{1}{T} \sum_{t=1}^T X_t \rightarrow_p E(X_t).$$

This allows us to consistently estimate parameters using time-series moments:
The sample mean:

$$\hat{\mu} = \frac{1}{T} \sum_{t=1}^T Y_t$$

The sample autocovariance

$$\hat{\gamma}(k) = \frac{1}{T} \sum_{t=1}^T (Y_t - \hat{\mu})(Y_{t-k} - \hat{\mu}).$$

The sample autocorrelation

$$\hat{\rho}(k) = \frac{\hat{\gamma}(k)}{\hat{\gamma}(0)}.$$

Theorem 12.1.3 If Y_t is strictly stationary and ergodic and $EY_t^2 < \infty$, then as $T \rightarrow \infty$,

1. $\hat{\mu} \rightarrow_p E(Y_t)$;
2. $\hat{\gamma}(k) \rightarrow_p \gamma(k)$;
3. $\hat{\rho}(k) \rightarrow_p \rho(k)$.

Proof. Part (1) is a direct consequence of the Ergodic theorem. For Part (2), note that

$$\begin{aligned} \hat{\gamma}(k) &= \frac{1}{T} \sum_{t=1}^T (Y_t - \hat{\mu})(Y_{t-k} - \hat{\mu}) \\ &= \frac{1}{T} \sum_{t=1}^T Y_t Y_{t-k} - \frac{1}{T} \sum_{t=1}^T Y_t \hat{\mu} - \frac{1}{T} \sum_{t=1}^T Y_{t-k} \hat{\mu} + \hat{\mu}^2. \end{aligned}$$

By Theorem 12.1.1 above, the sequence $Y_t Y_{t-k}$ is strictly stationary and ergodic, and it has a finite mean by the assumption that $EY_t^2 < \infty$. Thus an application of the Ergodic Theorem yields

$$\frac{1}{T} \sum_{t=1}^T Y_t Y_{t-k} \rightarrow_p E(Y_t Y_{t-k}).$$

Thus

$$\hat{\gamma}(k) \rightarrow_p E(Y_t Y_{t-k}) - \mu^2 - \mu^2 + \mu^2 = E(Y_t Y_{t-k}) - \mu^2 = \gamma(k).$$

Part (3) follows by the continuous mapping theorem: $\hat{\rho}(k) = \hat{\gamma}(k)/\hat{\gamma}(0) \rightarrow_p \gamma(k)/\gamma(0) = \rho(k)$. ■

12.2 Autoregressions

In time-series, the series $\{\dots, Y_1, Y_2, \dots, Y_T, \dots\}$ are jointly random. We consider the conditional expectation

$$E(Y_t | I_{t-1})$$

where $I_{t-1} = \{Y_{t-1}, Y_{t-2}, \dots\}$ is the past history of the series.

An autoregressive (AR) model specifies that only a finite number of past lags matter:

$$E(Y_t | I_{t-1}) = E(Y_t | Y_{t-1}, \dots, Y_{t-k}).$$

A linear AR model (the most common type used in practice) specifies linearity:

$$E(Y_t | I_{t-1}) = \alpha + \rho_1 Y_{t-1} + \rho_2 Y_{t-2} + \dots + \rho_k Y_{t-k}.$$

Letting

$$e_t = Y_t - E(Y_t | I_{t-1}),$$

then we have the autoregressive model

$$\begin{aligned} Y_t &= \alpha + \rho_1 Y_{t-1} + \rho_2 Y_{t-2} + \dots + \rho_k Y_{t-k} + e_t \\ E(e_t | I_{t-1}) &= 0. \end{aligned}$$

The last property defines a special time-series process.

Definition 12.2.1 e_t is a martingale difference sequence (MDS) if $E(e_t | I_{t-1}) = 0$.

Regression errors are naturally a MDS. Some time-series processes may be a MDS as a consequence of optimizing behavior. For example, some versions of the life-cycle hypothesis imply that either changes in consumption, or consumption growth rates, should be a MDS. Most asset pricing models imply that asset returns should be the sum of a constant plus a MDS.

The MDS property for the regression error plays the same role in a time-series regression as does the conditional mean-zero property for the regression error in a cross-section regression. In fact, it is even more important in the time-series context, as it is difficult to derive distribution theories without this property.

A useful property of a MDS is that e_t is uncorrelated with any function of the lagged information I_{t-1} . Thus for $k > 0$, $E(Y_{t-k} e_t) = 0$.

12.3 Stationarity of AR(1) Process

A mean-zero AR(1) is

$$Y_t = \rho Y_{t-1} + e_t.$$

Assume that e_t is iid, $E(e_t) = 0$ and $E e_t^2 = \sigma^2 < \infty$.

By back-substitution, we find

$$\begin{aligned} Y_t &= e_t + \rho e_{t-1} + \rho^2 e_{t-2} + \dots \\ &= \sum_{k=0}^{\infty} \rho^k e_{t-k}. \end{aligned}$$

Loosely speaking, this series converges if the sequence $\rho^k e_{t-k}$ gets small as $k \rightarrow \infty$. This occurs when $|\rho| < 1$.

Theorem 12.3.1 *If $|\rho| < 1$ then Y_t is strictly stationary and ergodic.*

We can compute the moments of Y_t using the infinite sum:

$$EY_t = \sum_{k=0}^{\infty} \rho^k E(e_{t-k}) = 0$$

$$Var(Y_t) = \sum_{k=0}^{\infty} \rho^{2k} Var(e_{t-k}) = \frac{\sigma^2}{1 - \rho^2}.$$

If the equation for Y_t has an intercept, the above results are unchanged, except that the mean of Y_t can be computed from the relationship

$$EY_t = \alpha + \rho EY_{t-1},$$

and solving for $EY_t = EY_{t-1}$ we find $EY_t = \alpha/(1 - \rho)$.

12.4 Lag Operator

An algebraic construct which is useful for the analysis of autoregressive models is the lag operator.

Definition 12.4.1 *The lag operator L satisfies $LY_t = Y_{t-1}$.*

Defining $L^2 = LL$, we see that $L^2Y_t = LY_{t-1} = Y_{t-2}$. In general, $L^kY_t = Y_{t-k}$.

The AR(1) model can be written in the format

$$Y_t - \rho Y_{t-1} + e_t$$

or

$$(1 - \rho L)Y_{t-1} = e_t.$$

The operator $\rho(L) = (1 - \rho L)$ is a polynomial in the operator L . We say that the *root* of the polynomial is $1/\rho$, since $\rho(z) = 0$ when $z = 1/\rho$. We call $\rho(L)$ the autoregressive polynomial of Y_t .

From Theorem 12.3.1, an AR(1) is stationary iff $|\rho| < 1$. Note that an equivalent way to say this is that an AR(1) is stationary iff the root of the autoregressive polynomial is larger than one (in absolute value).

12.5 Stationarity of AR(k)

The AR(k) model is

$$Y_t = \rho_1 Y_{t-1} + \rho_2 Y_{t-2} + \cdots + \rho_k Y_{t-k} + e_t.$$

Using the lag operator,

$$Y_t - \rho_1 LY_t - \rho_2 L^2 Y_t - \cdots - \rho_k L^k Y_t = e_t,$$

or

$$\rho(L)Y_t = e_t$$

where

$$\rho(L) = 1 - \rho_1 L - \rho_2 L^2 - \cdots - \rho_k L^k.$$

We call $\rho(L)$ the autoregressive polynomial of Y_t .

The *Fundamental Theorem of Algebra* says that any polynomial can be factored as

$$\rho(z) = (1 - \lambda_1^{-1}z)(1 - \lambda_2^{-1}z) \cdots (1 - \lambda_k^{-1}z)$$

where the $\lambda_1, \dots, \lambda_k$ are the complex *roots* of $\rho(z)$, which satisfy $\rho(\lambda_j) = 0$.

We know that an AR(1) is stationary iff the absolute value of the root of its autoregressive polynomial is larger than one. For an AR(k), the requirement is that all roots are larger than one. Let $|\lambda|$ denote the modulus of a complex number λ .

Theorem 12.5.1 *The AR(k) is strictly stationary and ergodic if and only if $|\lambda_j| > 1$ for all j .*

One way of stating this is that “All roots lie outside the unit circle.”

If one of the roots equals 1, we say that $\rho(L)$, and hence Y_t , “has a unit root”. This is a special case of non-stationarity, and is of great interest in applied time series.

12.6 Estimation

Let

$$\begin{aligned} x_t &= (1 \quad Y_{t-1} \quad Y_{t-2} \quad \cdots \quad Y_{t-k})' \\ \beta &= (\alpha \quad \rho_1 \quad \rho_2 \quad \cdots \quad \rho_k)'. \end{aligned}$$

Then the model can be written as

$$y_t = x_t' \beta + e_t.$$

The OLS estimator is

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

To study $\hat{\beta}$, it is helpful to define the process $u_t = x_t e_t$. Note that u_t is a MDS, since

$$E(u_t | I_{t-1}) = E(x_t e_t | I_{t-1}) = x_t E(e_t | I_{t-1}) = 0.$$

By Theorem 12.1.1, it is also strictly stationary and ergodic. Thus

$$\frac{1}{T} \sum_{t=1}^T x_t e_t = \frac{1}{T} \sum_{t=1}^T u_t \rightarrow_p E(u_t) = 0. \quad (12.1)$$

Theorem 12.6.1 *If the AR(k) process Y_t is strictly stationary and ergodic and $EY_t^2 < \infty$, then $\hat{\beta} \rightarrow_p \beta$ as $T \rightarrow \infty$.*

Proof. The vector x_t is strictly stationary and ergodic, and by Theorem 12.1.1, so is $x_t x_t'$. Thus by the Ergodic Theorem,

$$\frac{1}{T} \sum_{t=1}^T x_t x_t' \rightarrow_p E(x_t x_t') = Q.$$

Combined with (12.1) and the continuous mapping theorem, we see that

$$\hat{\beta} = \beta + \left(\frac{1}{T} \sum_{t=1}^T x_t x_t' \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^T x_t e_t \right) \rightarrow_p Q^{-1} 0 = 0.$$

■

12.7 Asymptotic Distribution

Theorem 12.7.1 *MDS CLT. If u_t is a strictly stationary and ergodic MDS and $E(u_t u_t') = \Omega < \infty$, then as $T \rightarrow \infty$,*

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T u_t \rightarrow_d N(0, \Omega).$$

Since $x_t e_t$ is a MDS, we can apply Theorem 12.7.1 to see that

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T x_t e_t \rightarrow_d N(0, \Omega),$$

where

$$\Omega = E(x_t x_t' e_t^2).$$

Theorem 12.7.2 *If the $AR(k)$ process Y_t is strictly stationary and ergodic and $EY_t^4 < \infty$, then as $T \rightarrow \infty$,*

$$\sqrt{T}(\hat{\beta} - \beta) \rightarrow_d N(0, Q^{-1}\Omega Q^{-1}).$$

This is identical in form to the asymptotic distribution of OLS in cross-section regression. The implication is that asymptotic inference is the same. In particular, the asymptotic covariance matrix is estimated just as in the cross-section case.

12.8 Bootstrap for Autoregressions

In the non-parametric bootstrap, we constructed the bootstrap sample by randomly resampling from the data values $\{y_t, x_t\}$. This creates an iid bootstrap sample. Clearly, this cannot work in a time-series application, as this imposes inappropriate independence.

Briefly, there are two popular methods to implement bootstrap resampling for time-series data.

Method 1: Model-Based (Parametric) Bootstrap.

1. Estimate $\hat{\beta}$ and residuals \hat{e}_t .
2. Fix an initial condition $(Y_{-k+1}, Y_{-k+2}, \dots, Y_0)$.
3. Simulate iid draws e_i^* from the empirical distribution of the residuals $\{\hat{e}_1, \dots, \hat{e}_T\}$.
4. Create the bootstrap series Y_t^* by the recursive formula

$$Y_t^* = \hat{\alpha} + \hat{\rho}_1 Y_{t-1}^* + \hat{\rho}_2 Y_{t-2}^* + \dots + \hat{\rho}_k Y_{t-k}^* + e_t^*.$$

This construction imposes homoskedasticity on the errors e_i^* , which may be different than the properties of the actual e_i . It also presumes that the $AR(k)$ structure is the truth.

Method 1: Block Resampling

1. Divide the sample into T/m blocks of length m .

2. Resample complete blocks. For each simulated sample, draw T/m blocks.
3. Paste the blocks together to create the bootstrap time-series Y_t^* .
4. This allows for arbitrary stationary serial correlation, heteroskedasticity, and for model-misspecification.
5. The results may be sensitive to the block length, and the way that the data is partitioned into blocks.
6. May not work well in small samples.

12.9 Trend Stationarity

$$Y_t = \mu_0 + \mu_1 t + S_t \quad (12.2)$$

$$S_t = \rho_1 S_{t-1} + \rho_2 S_{t-2} + \cdots + \rho_k S_{t-k} + e_t, \quad (12.3)$$

or

$$Y_t = \alpha_0 + \alpha_1 t + \rho_1 Y_{t-1} + \rho_2 Y_{t-2} + \cdots + \rho_k Y_{t-k} + e_t. \quad (12.4)$$

There are two essentially equivalent ways to estimate the autoregressive parameters (ρ_1, \dots, ρ_k) .

- You can estimate (12.4) by OLS.
- You can estimate (12.2)-(12.3) sequentially by OLS. That is, first estimate (12.2), get the residual \hat{S}_t , and then perform regression (12.3) replacing S_t with \hat{S}_t . This procedure is sometimes called *Detrending*.

The reason why these two procedures are (essentially) the same is the Frisch-Waugh-Lovell theorem.

Seasonal Effects

There are three popular methods to deal with seasonal data.

- Include dummy variables for each season. This presumes that “seasonality” does not change over the sample.
- Use “seasonally adjusted” data. The seasonal factor is typically estimated by a two-sided weighted average of the data for that season in neighboring years. Thus the seasonally adjusted data is a “filtered” series. This is a flexible approach which can extract a wide range of seasonal factors. The seasonal adjustment, however, also alters the time-series correlations of the data.
- First apply a seasonal differencing operator. If s is the number of seasons (typically $s = 4$ or $s = 12$),

$$\Delta_s Y_t = Y_t - Y_{t-s},$$

or the season-to-season change. The series $\Delta_s Y_t$ is clearly free of seasonality. But the long-run trend is also eliminated, and perhaps this was of relevance.

12.10 Testing for Omitted Serial Correlation

For simplicity, let the null hypothesis be an AR(1):

$$Y_t = \alpha + \rho Y_{t-1} + u_t. \quad (12.5)$$

We are interested in the question if the error u_t is serially correlated. We model this as an AR(1):

$$u_t = \theta u_{t-1} + e_t \quad (12.6)$$

with e_t a MDS. The hypothesis of no omitted serial correlation is

$$\begin{aligned} H_0 &: \theta = 0 \\ H_1 &: \theta \neq 0. \end{aligned}$$

We want to test H_0 against H_1 .

To combine (12.5) and (12.6), we take (12.5) and lag the equation once:

$$Y_{t-1} = \alpha + \rho Y_{t-2} + u_{t-1}.$$

We then multiply this by θ and subtract from (12.5), to find

$$Y_t - \theta Y_{t-1} = \alpha - \theta\alpha + \rho Y_{t-1} - \theta\rho Y_{t-1} + u_t - \theta u_{t-1},$$

or

$$Y_t = \alpha(1 - \theta) + (\rho + \theta) Y_{t-1} - \theta\rho Y_{t-2} + e_t = AR(2).$$

Thus under H_0 , Y_t is an AR(1), and under H_1 it is an AR(2). H_0 may be expressed as the restriction that the coefficient on Y_{t-2} is zero.

An appropriate test of H_0 against H_1 is therefore a Wald test that the coefficient on Y_{t-2} is zero. (A simple exclusion test).

In general, if the null hypothesis is that Y_t is an AR(k), and the alternative is that the error is an AR(m), this is the same as saying that under the alternative Y_t is an AR($k+m$), and this is equivalent to the restriction that the coefficients on $Y_{t-k-1}, \dots, Y_{t-k-m}$ are jointly zero. An appropriate test is the Wald test of this restriction.

12.11 Model Selection

What is the appropriate choice of k in practice? This is a problem of model selection.

One approach to model selection is to choose k based on a Wald tests.

Another is to minimize the AIC or BIC information criterion, e.g.

$$AIC(k) = \log \hat{\sigma}^2(k) + \frac{2k}{T},$$

where $\hat{\sigma}^2(k)$ is the estimated residual variance from an AR(k)

One ambiguity in defining the AIC criterion is that the sample available for estimation changes as k changes. (If you increase k , you need more initial conditions.) This can induce strange behavior in the AIC. The best remedy is to fix a upper value \bar{k} , and then reserve the first \bar{k} as initial conditions, and then estimate the models AR(1), AR(2), ..., AR(\bar{k}) on this (unified) sample.

12.12 Autoregressive Unit Roots

The AR(k) model is

$$\begin{aligned}\rho(L)Y_t &= \mu + e_t \\ \rho(L) &= 1 - \rho_1L - \cdots - \rho_kL^k.\end{aligned}$$

As we discussed before, Y_t has a unit root when $\rho(1) = 0$, or

$$\rho_1 + \rho_2 + \cdots + \rho_k = 1.$$

In this case, Y_t is non-stationary. The ergodic theorem and MDS CLT do not apply, and test statistics are asymptotically non-normal.

A helpful way to write the equation is the so-called Dickey-Fuller reparameterization:

$$\Delta Y_t = \mu + \alpha_0 Y_{t-1} + \alpha_1 \Delta Y_{t-1} + \cdots + \alpha_{k-1} \Delta Y_{t-(k-1)} + e_t. \quad (12.7)$$

These models are equivalent linear transformations of one another. The DF parameterization is convenient because the parameter α_0 summarizes the information about the unit root, since $\rho(1) = -\alpha_0$. To see this, observe that the lag polynomial for the Y_t computed from (12.7) is

$$(1 - L) - \alpha_0 L - \alpha_1(L - L^2) - \cdots - \alpha_{k-1}(L^{k-1} - L^k)$$

But this must equal $\rho(L)$, as the models are equivalent. Thus

$$\rho(1) = (1 - 1) - \alpha_0 - (1 - 1) - \cdots - (1 - 1) = -\alpha_0.$$

Hence, the hypothesis of a unit root in Y_t can be stated as

$$H_0 : \alpha_0 = 0.$$

Note that the model is stationary if $\alpha_0 < 0$. So the natural alternative is

$$H_1 : \alpha_0 < 0.$$

Under H_0 , the model for Y_t is

$$\Delta Y_t = \mu + \alpha_1 \Delta Y_{t-1} + \cdots + \alpha_{k-1} \Delta Y_{t-(k-1)} + e_t,$$

which is an AR(k-1) in the first-difference ΔY_t . Thus if Y_t has a (single) unit root, then ΔY_t is a stationary AR process. Because of this property, we say that if Y_t is non-stationary but $\Delta^d Y_t$ is stationary, then Y_t is “integrated of order d ”, or $I(d)$. Thus a time series with unit root is $I(1)$.

Since α_0 is the parameter of a linear regression, the natural test statistic is the t-statistic for H_0 from OLS estimation of (12.7). Indeed, this is the most popular unit root test, and is called the Augmented Dickey-Fuller (ADF) test for a unit root.

It would seem natural to assess the significance of the ADF statistic using the normal table. However, under H_0 , Y_t is non-stationary, so conventional normal asymptotics are invalid. An alternative asymptotic framework has been developed to deal with non-stationary data. We do not have the time to develop this theory in detail, but simply assert the main results.

Theorem 12.12.1 (*Dickey-Fuller Theorem*). Assume $\alpha_0 = 0$. As $T \rightarrow \infty$,

$$T\hat{\alpha}_0 \rightarrow_d (1 - \alpha_1 - \alpha_2 - \cdots - \alpha_{k-1}) DF_\alpha$$

$$ADF = \frac{\hat{\alpha}_0}{s(\hat{\alpha}_0)} \rightarrow DF_t.$$

The limit distributions DF_α and DF_t are non-normal. They are skewed to the left, and have negative means.

The first result states that $\hat{\alpha}_0$ converges to its true value (of zero) at rate T , rather than the conventional rate of $T^{1/2}$. This is called a “super-consistent” rate of convergence.

The second result states that the t-statistic for $\hat{\alpha}_0$ converges to a limit distribution which is non-normal, but does not depend on the parameters α . This distribution has been extensively tabulated, and may be used for testing the hypothesis H_0 . Note: The standard error $s(\hat{\alpha}_0)$ is the conventional (“homoskedastic”) standard error. But the theorem does not require an assumption of homoskedasticity. Thus the Dickey-Fuller test is robust to heteroskedasticity.

Since the alternative hypothesis is one-sided, the ADF test rejects H_0 in favor of H_1 when $ADF < c$, where c is the critical value from the ADF table. If the test rejects H_0 , this means that the evidence points to Y_t being stationary. If the test does not reject H_0 , a common conclusion is that the data suggests that Y_t is non-stationary. This is not really a correct conclusion, however. All we can say is that there is insufficient evidence to conclude whether the data is stationary or not.

We have described the test for the setting of with an intercept. Another popular setting includes as well a linear time trend. This model is

$$\Delta Y_t = \mu_1 + \mu_2 t + \alpha_0 Y_{t-1} + \alpha_1 \Delta Y_{t-1} + \cdots + \alpha_{k-1} \Delta Y_{t-(k-1)} + e_t. \quad (12.8)$$

This is natural when the alternative hypothesis is that the series is stationary about a linear time trend. If the series has a linear trend (e.g. GDP, Stock Prices), then the series itself is non-stationary, but it may be stationary around the linear time trend. In this context, it is a silly waste of time to fit an AR model to the level of the series without a time trend, as the AR model cannot conceivably describe this data. The natural solution is to include a time trend in the fitted OLS equation. When conducting the ADF test, this means that it is computed as the t-ratio for α_0 from OLS estimation of (12.8).

If a time trend is included, the test procedure is the same, but different critical values are required. The ADF test has a different distribution when the time trend has been included, and a different table should be consulted.

Most texts include as well the critical values for the extreme polar case where the intercept has been omitted from the model. These are included for completeness (from a pedagogical perspective) but have no relevance for empirical practice where intercepts are always included.

Chapter 13

Multivariate Time Series

A multivariate time series Y_t is a vector process $m \times 1$. Let $I_{t-1} = (Y_{t-1}, Y_{t-2}, \dots)$ be all lagged information at time t . The typical goal is to find the conditional expectation $E(Y_t | I_{t-1})$. Note that since Y_t is a vector, this conditional expectation is also a vector.

13.1 Vector Autoregressions (VARs)

A VAR model specifies that the conditional mean is a function of only a finite number of lags:

$$E(Y_t | I_{t-1}) = E(Y_t | Y_{t-1}, \dots, Y_{t-k}).$$

A linear VAR specifies that this conditional mean is linear in the arguments:

$$E(Y_t | Y_{t-1}, \dots, Y_{t-k}) = A_0 + A_1 Y_{t-1} + A_2 Y_{t-2} + \dots + A_k Y_{t-k}.$$

Observe that A_0 is $m \times 1$, and each of A_1 through A_k are $m \times m$ matrices.

Defining the $m \times 1$ regression error

$$e_t = Y_t - E(Y_t | I_{t-1}),$$

we have the VAR model

$$\begin{aligned} Y_t &= A_0 + A_1 Y_{t-1} + A_2 Y_{t-2} + \dots + A_k Y_{t-k} + e_t \\ E(e_t | I_{t-1}) &= 0. \end{aligned}$$

Alternatively, defining the $mk + 1$ vector

$$x_t = \begin{pmatrix} 1 \\ Y_{t-1} \\ Y_{t-2} \\ \vdots \\ Y_{t-k} \end{pmatrix}$$

and the $m \times (mk + 1)$ matrix

$$A = (A_0 \quad A_1 \quad A_2 \quad \dots \quad A_k),$$

then

$$Y_t = Ax_t + e_t.$$

The VAR model is a system of m equations. One way to write this is to let a'_j be the j th row of A . Then the VAR system can be written as the k equations

$$Y_{jt} = a'_j x_t + e_{jt}.$$

Unrestricted VARs were introduced to econometrics by Sims (1980).

13.2 Estimation

Consider the moment conditions

$$E(x_t e_{jt}) = 0,$$

$j = 1, \dots, k$. These are implied by the VAR model, either as a regression, or as a linear projection.

The GMM estimator corresponding to these moment conditions is equation-by-equation OLS

$$\hat{a}_j = (X'X)^{-1}X'Y_j.$$

An alternative way to compute this is as follows. Note that

$$\hat{a}'_j = Y'_j X (X'X)^{-1}.$$

And if we stack these to create the estimate \hat{A} , we find

$$\begin{aligned} \hat{A} &= \begin{pmatrix} Y'_1 \\ Y'_2 \\ \vdots \\ Y'_{k+1} \end{pmatrix} X (X'X)^{-1} \\ &= Y' X (X'X)^{-1}, \end{aligned}$$

where

$$Y = (Y_1 \quad Y_2 \quad \cdots \quad Y_k)$$

the $T \times m$ matrix of the stacked y'_t .

This (system) estimator is known as the SUR (Seemingly Unrelated Regressions) estimator, and was originally derived by Zellner (1962)

13.3 Restricted VARs

The unrestricted VAR is a system of m equations, each with the same set of regressors. A restricted VAR imposes restrictions on the system. For example, some regressors may be excluded from some of the equations. Restrictions may be imposed on individual equations, or across equations. The GMM framework gives a convenient method to impose such restrictions on estimation.

13.4 Single Equation from a VAR

Often, we are only interested in a single equation out of a VAR system. This takes the form

$$Y_{jt} = a'_j x_t + e_t,$$

and x_t consists of lagged values of Y_{jt} and the other Y'_{it} s. In this case, it is convenient to re-define the variables. Let $y_t = Y_{jt}$, and Z_t be the other variables. Let $e_t = e_{jt}$ and $\beta = a_j$. Then the single equation takes the form

$$y_t = x'_t \beta + e_t, \quad (13.1)$$

and

$$x_t = \left[\left(1 \quad Y_{t-1} \quad \cdots \quad Y_{t-k} \quad Z'_{t-1} \quad \cdots \quad Z'_{t-k} \right)' \right].$$

This is just a conventional regression, with time series data.

13.5 Testing for Omitted Serial Correlation

Consider the problem of testing for omitted serial correlation in equation (13.1). Suppose that e_t is an AR(1). Then

$$\begin{aligned} y_t &= x'_t \beta + e_t \\ e_t &= \theta e_{t-1} + u_t \\ E(u_t | I_{t-1}) &= 0. \end{aligned} \quad (13.2)$$

Then the null and alternative are

$$H_0 : \theta = 0 \quad H_1 : \theta \neq 0.$$

Take the equation $y_t = x'_t \beta + e_t$, and subtract off the equation once lagged multiplied by θ , to get

$$\begin{aligned} y_t - \theta y_{t-1} &= (x'_t \beta + e_t) - \theta (x'_{t-1} \beta + e_{t-1}) \\ &= x'_t \beta - \theta x'_{t-1} \beta + e_t - \theta e_{t-1}, \end{aligned}$$

or

$$y_t = \theta y_{t-1} + x'_t \beta + x'_{t-1} \gamma + u_t, \quad (13.3)$$

which is a valid regression model.

So testing H_0 versus H_1 is equivalent to testing for the significance of adding (y_{t-1}, x_{t-1}) to the regression. This can be done by a Wald test. We see that an appropriate, general, and simple way to test for omitted serial correlation is to test the significance of extra lagged values of the dependent variable and regressors.

You may have heard of the Durbin-Watson test for omitted serial correlation, which once was very popular, and is still routinely reported by conventional regression packages. The DW test is appropriate only when regression $y_t = x'_t \beta + e_t$ is not dynamic (has no lagged values on the RHS), and e_t is iid $N(0, 1)$. Otherwise it is invalid.

Another interesting fact is that (13.2) is a special case of (13.3), under the restriction $\gamma = -\beta\theta$. This restriction, which is called a common factor restriction, may be tested if desired. If valid, the model (13.2) may be estimated by iterated GLS. (A simple version of this estimator is called Cochrane-Orcutt.) Since the common factor restriction appears arbitrary, and is typically rejected empirically, direct estimation of (13.2) is uncommon in recent applications.

13.6 Selection of Lag Length in an VAR

If you want a data-dependent rule to pick the lag length k in a VAR, you may either use a testing-based approach (using, for example, the Wald statistic), or an information criterion approach. The formula for the AIC and BIC are

$$\begin{aligned} AIC(k) &= \log \det \left(\hat{\Omega}(k) \right) + 2\frac{p}{T} \\ BIC(k) &= \log \det \left(\hat{\Omega}(k) \right) + \frac{p \log(T)}{T} \\ \hat{\Omega}(k) &= \frac{1}{T} \sum_{t=1}^T \hat{e}_t(k) \hat{e}_t(k)' \\ p &= m(km + 1) \end{aligned}$$

where p is the number of parameters in the model, and $\hat{e}_t(k)$ is the OLS residual vector from the model with k lags. The log determinant is the criterion from the multivariate normal likelihood.

13.7 Granger Causality

Partition the data vector into (Y_t, Z_t) . Define the two information sets

$$\begin{aligned} I_{1t} &= (Y_t, Y_{t-1}, Y_{t-2}, \dots) \\ I_{2t} &= (Y_t, Z_t, Y_{t-1}, Z_{t-1}, Y_{t-2}, Z_{t-2}, \dots) \end{aligned}$$

The information set I_{1t} is generated only by the history of Y_t , and the information set I_{2t} is generated by both Y_t and Z_t . The latter has more information.

We say that Z_t does not *Granger-cause* Y_t if

$$E(Y_t | I_{1,t-1}) = E(Y_t | I_{2,t-1}).$$

That is, conditional on information in lagged Y_t , lagged Z_t does not help to forecast Y_t . If this condition does not hold, then we say that Z_t Granger-causes Y_t .

The reason why we call this “Granger Causality” rather than “causality” is because this is not a physical or structure definition of causality. If Z_t is some sort of forecast of the future, such as a futures price, then Z_t may help to forecast Y_t even though it does not “cause” Y_t . This definition of causality was developed by Granger (1969) and Sims (1972).

In a linear VAR, the equation for Y_t is

$$Y_t = \alpha + \rho_1 Y_{t-1} + \dots + \rho_k Y_{t-k} + Z'_{t-1} \gamma_1 + \dots + Z'_{t-k} \gamma_k + e_t.$$

In this equation, Z_t does not Granger-cause Y_t if and only if

$$H_0 : \gamma_1 = \gamma_2 = \dots = \gamma_k = 0.$$

This may be tested using an exclusion (Wald) test.

This idea can be applied to blocks of variables. That is, Y_t and/or Z_t can be vectors. The hypothesis can be tested by using the appropriate multivariate Wald test.

If it is found that Z_t does not Granger-cause Y_t , then we deduce that our time-series model of $E(Y_t | I_{t-1})$ does not require the use of Z_t . Note, however, that Z_t may still be useful to explain other features of Y_t , such as the conditional variance.

13.8 Cointegration

The idea of cointegration is due to Granger (1981), and was articulated in detail by Engle and Granger (1987).

Definition 13.8.1 *The $m \times 1$ series Y_t is cointegrated if Y_t is $I(1)$ yet there exists β , $m \times r$, of rank r , such that $z_t = \beta'Y_t$ is $I(0)$. The r vectors in β are called the cointegrating vectors.*

If the series Y_t is not cointegrated, then $r = 0$. If $r = m$, then Y_t is $I(0)$. For $0 < r < m$, Y_t is $I(1)$ and cointegrated.

In some cases, it may be believed that β is known a priori. Often, $\beta = (1 \quad -1)'$. For example, if Y_t is a pair of interest rates, then $\beta = (1 \quad -1)'$ specifies that the spread (the difference in returns) is stationary. If $Y = (\log(\text{Consumption}) \quad \log(\text{Income}))'$, then $\beta = (1 \quad -1)'$ specifies that $\log(\text{Consumption}/\text{Income})$ is stationary.

In other cases, β may not be known.

If Y_t is cointegrated with a single cointegrating vector ($r = 1$), then it turns out that β can be consistently estimated by an OLS regression of one component of Y_t on the others. Thus $Y_t = (Y_{1t}, Y_{2t})$ and $\beta = (\beta_1 \quad \beta_2)$ and normalize $\beta_1 = 1$. Then $\hat{\beta}_2 = (Y_2'Y_2)^{-1}Y_2Y_1 \rightarrow_p \beta_2$. Furthermore this estimation is super-consistent: $T(\hat{\beta}_2 - \beta_2) \rightarrow_d \text{Limit}$, as first shown by Stock (1987). This is not, in general, a good method to estimate β , but it is useful in the construction of alternative estimators and tests.

We are often interested in testing the hypothesis of no cointegration:

$$\begin{aligned} H_0 &: r = 0 \\ H_1 &: r > 0. \end{aligned}$$

Suppose that β is known, so $z_t = \beta'Y_t$ is known. Then under H_0 z_t is $I(1)$, yet under H_1 z_t is $I(0)$. Thus H_0 can be tested using a univariate ADF test on z_t .

When β is unknown, Engle and Granger (1987) suggested using an ADF test on the estimated residual $\hat{z}_t = \hat{\beta}'Y_t$, from OLS of Y_{1t} on Y_{2t} . Their justification was Stock's result that $\hat{\beta}$ is super-consistent under H_1 . Under H_0 , however, $\hat{\beta}$ is not consistent, so the ADF critical values are not appropriate. The asymptotic distribution was worked out by Phillips and Ouliaris (1990).

When the data have time trends, it may be necessary to include a time trend in the estimated cointegrating regression. Whether or not the time trend is included, the asymptotic distribution of the test is affected by the presence of the time trend. The asymptotic distribution was worked out in B. Hansen (1992).

13.9 Cointegrated VARs

We can write a VAR as

$$\begin{aligned} A(L)Y_t &= e_t \\ A(L) &= I - A_1L - A_2L^2 - \dots - A_kL^k \end{aligned}$$

or alternatively as

$$\Delta Y_t = \Pi Y_{t-1} + D(L)\Delta Y_{t-1} + e_t$$

where

$$\begin{aligned} \Pi &= -A(1) \\ &= -I + A_1 + A_2 + \dots + A_k. \end{aligned}$$

Theorem 13.9.1 (*Granger Representation Theorem*). Y_t is cointegrated with $m \times r$ β if and only if $\text{rank}(\Pi) = r$ and $\Pi = \alpha\beta'$ where α is $m \times r$, $\text{rank}(\alpha) = r$.

Thus cointegration imposes a restriction upon the parameters of a VAR. The restricted model can be written as

$$\begin{aligned}\Delta Y_t &= \alpha\beta'Y_{t-1} + D(L)\Delta Y_{t-1} + e_t \\ \Delta Y_t &= \alpha z_{t-1} + D(L)\Delta Y_{t-1} + e_t.\end{aligned}$$

If β is known, this can be estimated by OLS of ΔY_t on z_{t-1} and the lags of ΔY_t .

If β is unknown, then estimation is done by “reduced rank regression”, which is least-squares subject to the stated restriction. Equivalently, this is the MLE of the restricted parameters under the assumption that e_t is iid $N(0, \Omega)$.

One difficulty is that β is not identified without normalization. When $r = 1$, we typically just normalize one element to equal unity. When $r > 1$, this does not work, and different authors have adopted different identification schemes.

In the context of a cointegrated VAR estimated by reduced rank regression, it is simple to test for cointegration by testing the rank of Π . These tests are constructed as likelihood ratio (LR) tests. As they were discovered by Johansen (1988, 1991, 1995), they are typically called the “Johansen Max and Trace” tests. Their asymptotic distributions are non-standard, and are similar to the Dickey-Fuller distributions.

Chapter 14

Limited Dependent Variables

A “limited dependent variable” Y is one which takes a “limited” set of values. The most common cases are

- Binary: $Y = \{0, 1\}$
- Multinomial: $Y = \{0, 1, 2, \dots, k\}$
- Integer: $Y = \{0, 1, 2, \dots\}$
- Censored: $Y = \{x : x \geq 0\}$

The traditional approach to the estimation of limited dependent variable (LDV) models is parametric maximum likelihood. A parametric model is constructed, allowing the construction of the likelihood function. A more modern approach is semi-parametric, eliminating the dependence on a parametric distributional assumption. We will discuss only the first (parametric) approach, due to time constraints. They still constitute the majority of LDV applications. If, however, you were to write a thesis involving LDV estimation, you would be advised to consider employing a semi-parametric estimation approach.

For the parametric approach, estimation is by MLE. A major practical issue is construction of the likelihood function.

14.1 Binary Choice

The dependent variable $Y_i = \{0, 1\}$. This represents a Yes/No outcome. Given some regressors x_i , the goal is to describe $P(Y_i = 1 | x_i)$, as this is the full conditional distribution.

The linear probability model specifies that

$$P(Y_i = 1 | x_i) = x_i' \beta.$$

As $P(Y_i = 1 | x_i) = E(Y_i | x_i)$, this yields the regression: $Y_i = x_i' \beta + e_i$ which can be estimated by OLS. However, the linear probability model does not impose the restriction that $0 \leq P(Y_i | x_i) \leq 1$. Even so estimation of a linear probability model is a useful starting point for subsequent analysis.

The standard alternative is to use a function of the form

$$P(Y_i = 1 | x_i) = F(x_i' \beta)$$

where $F(\cdot)$ is a known CDF, typically assumed to be symmetric about zero, so that $F(z) = 1 - F(-z)$. The two standard choices for F are

- Logistic: $F(u) = (1 + e^{-u})^{-1}$.
- Normal: $F(u) = \Phi(u)$.

If F is logistic, we call this the *logit* model, and if F is normal, we call this the *probit* model. This model is identical to the latent variable model

$$\begin{aligned} Y_i^* &= x_i' \beta + e_i \\ e_i &\sim F(\cdot) \\ Y_i &= \begin{cases} 1 & \text{if } Y_i^* < 0 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

For then

$$\begin{aligned} P(Y_i = 1 \mid x_i) &= P(Y_i^* > 0 \mid x_i) \\ &= P(x_i' \beta + e_i > 0 \mid x_i) \\ &= P(e_i > -x_i' \beta \mid x_i) \\ &= 1 - F(-x_i' \beta) \\ &= F(x_i' \beta). \end{aligned}$$

Estimation is by maximum likelihood. To construct the likelihood, we need the conditional distribution of an individual observation. Recall that if Y is Bernoulli, such that $P(Y = 1) = p$ and $P(Y = 0) = 1 - p$, then we can write the density of Y as

$$f(y) = p^y (1 - p)^{1-y}, \quad y = 0, 1.$$

In the Binary choice model, Y_i is conditionally Bernoulli with $P(Y_i = 1 \mid x_i) = p_i = F(x_i' \beta)$. Thus the conditional density is

$$\begin{aligned} f(y_i \mid x_i) &= p_i^{y_i} (1 - p_i)^{1-y_i} \\ &= F(x_i' \beta)^{y_i} (1 - F(x_i' \beta))^{1-y_i}. \end{aligned}$$

Hence the log-likelihood function is

$$\begin{aligned} l_n(\beta) &= \sum_{i=1}^n \log f(y_i \mid x_i) \\ &= \sum_{i=1}^n \log (F(x_i' \beta)^{y_i} (1 - F(x_i' \beta))^{1-y_i}) \\ &= \sum_{i=1}^n [y_i \log F(x_i' \beta) + (1 - y_i) \log(1 - F(x_i' \beta))] \\ &= \sum_{y_i=1} \log F(x_i' \beta) + \sum_{y_i=0} \log(1 - F(x_i' \beta)). \end{aligned}$$

The MLE $\hat{\beta}$ is the value of β which maximizes $l_n(\beta)$. Standard errors and test statistics are computed by asymptotic approximations. Details of such calculations are left to more advanced courses.

14.2 Count Data

If $Y = \{0, 1, 2, \dots\}$, a typical approach is to employ *Poisson regression*. This model specifies that

$$\begin{aligned}P(Y_i = k | x_i) &= \frac{\exp(-\lambda_i) \lambda_i^k}{k!}, & k = 0, 1, 2, \dots \\ \lambda_i &= \exp(x_i' \beta).\end{aligned}$$

The conditional density is the Poisson with parameter λ_i . The functional form for λ_i has been picked to ensure that $\lambda_i > 0$.

The log-likelihood function is

$$l_n(\beta) = \sum_{i=1}^n \log f(y_i | x_i) = \sum_{i=1}^n (-\exp(x_i' \beta) + y_i x_i' \beta - \log(y_i!)).$$

The MLE is the value $\hat{\beta}$ which maximizes $l_n(\beta)$.

Since

$$E(Y_i | x_i) = \lambda_i = \exp(x_i' \beta)$$

is the conditional mean, this motivates the label Poisson “regression.”

Also observe that the model implies that

$$\text{Var}(Y_i | x_i) = \lambda_i = \exp(x_i' \beta),$$

so the model imposes the restriction that the conditional mean and variance of Y_i are the same. This may be considered restrictive. A generalization is the negative binomial.

14.3 Censored Data

The idea of “censoring” is that some data above or below a threshold are mis-reported at the threshold. Thus the model is that there is some latent process y_i^* with unbounded support, but we observe only

$$y_i = \begin{cases} y_i^* & \text{if } y_i^* \geq 0 \\ 0 & \text{if } y_i^* < 0 \end{cases}. \quad (14.1)$$

(This is written for the case of the threshold being zero, any known value can substitute.) The observed data y_i therefore come from a mixed continuous/discrete distribution.

Censored models are typically applied when the data set has a meaningful proportion (say 5% or higher) of data at the boundary of the sample support. The censoring process may be explicit in data collection, or it may be a by-product of economic constraints.

An example of a data collection censoring is top-coding of income. In surveys, incomes above a threshold are typically reported at the threshold.

The first censored regression model was developed by Tobin (1958) to explain consumption of durable goods. Tobin observed that for many households, the consumption level (purchases) in a particular period was zero. He proposed the latent variable model

$$\begin{aligned}y_i^* &= x_i' \beta + e_i \\ e_i &\sim \text{iid } N(0, \sigma^2)\end{aligned}$$

with the observed variable y_i generated by the censoring equation (14.1). This model (now called the Tobit) specifies that the latent (or ideal) value of consumption may be negative (the household

would prefer to sell than buy). All that is reported is that the household purchased zero units of the good.

The naive approach to estimate β is to regress y_i on x_i . This does not work because regression estimates $E(Y_i | x_i)$, not $E(Y_i^* | x_i) = x_i'\beta$, and the latter is of interest. Thus OLS will be biased for the parameter of interest β .

[Note: it is still possible to estimate $E(Y_i | x_i)$ by LS techniques. The Tobit framework postulates that this is not inherently interesting, that the parameter of β is defined by an alternative statistical structure.]

Consistent estimation will be achieved by the MLE. To construct the likelihood, observe that the probability of being censored is

$$\begin{aligned} P(y_i = 0 | x_i) &= P(y_i^* < 0 | x_i) \\ &= P(x_i'\beta + e_i < 0 | x_i) \\ &= P\left(\frac{e_i}{\sigma} < -\frac{x_i'\beta}{\sigma} \mid x_i\right) \\ &= \Phi\left(-\frac{x_i'\beta}{\sigma}\right). \end{aligned}$$

The conditional distribution function above zero is Gaussian:

$$P(y_i = y | x_i) = \int_0^y \sigma^{-1} \phi\left(\frac{z - x_i'\beta}{\sigma}\right) dz, \quad y > 0.$$

Therefore, the density function can be written as

$$f(y | x_i) = \Phi\left(-\frac{x_i'\beta}{\sigma}\right)^{1(y=0)} \left[\sigma^{-1} \phi\left(\frac{z - x_i'\beta}{\sigma}\right)\right]^{1(y>0)},$$

where $1(\cdot)$ is the indicator function.

Hence the log-likelihood is a mixture of the probit and the normal:

$$\begin{aligned} l_n(\beta) &= \sum_{i=1}^n \log f(y_i | x_i) \\ &= \sum_{y_i=0} \log \Phi\left(-\frac{x_i'\beta}{\sigma}\right) + \sum_{y_i>0} \log \left[\sigma^{-1} \phi\left(\frac{y_i - x_i'\beta}{\sigma}\right)\right]. \end{aligned}$$

The MLE is the value $\hat{\beta}$ which maximizes $l_n(\beta)$.

14.4 Sample Selection

The problem of sample selection arises when the sample is a non-random selection of potential observations. This occurs when the observed data is systematically different from the population of interest. For example, if you ask for volunteers for an experiment, and they wish to extrapolate the effects of the experiment on a general population, you should worry that the people who volunteer may be systematically different from the general population. This has great relevance for the evaluation of anti-poverty and job-training programs, where the goal is to assess the effect of “training” on the general population, not just on the volunteers.

A simple sample selection model can be written as the latent model

$$\begin{aligned} y_i &= x_i' \beta + e_{1i} \\ T_i &= 1(z_i' \gamma + e_{0i} > 0) \end{aligned}$$

where $1(\cdot)$ is the indicator function. The dependent variable y_i is observed if (and only if) $T_i = 1$. Else it is unobserved.

For example, y_i could be a wage, which can be observed only if a person is employed. The equation for T_i is an equation specifying the probability that the person is employed.

The model is often completed by specifying that the errors are jointly normal

$$\begin{pmatrix} e_{0i} \\ e_{1i} \end{pmatrix} \sim N\left(0, \begin{pmatrix} 1 & \rho \\ \rho & \sigma^2 \end{pmatrix}\right).$$

It is presumed that we observe $\{x_i, z_i, T_i\}$ for all observations.

Under the normality assumption,

$$e_{1i} = \rho e_{0i} + v_i,$$

where v_i is independent of $e_{0i} \sim N(0, 1)$. A useful fact about the standard normal distribution is that

$$E(e_{0i} | e_{0i} > -x) = \lambda(x) = \frac{\phi(x)}{\Phi(x)},$$

and the function $\lambda(x)$ is called the inverse Mills ratio.

The naive estimator of β is OLS regression of y_i on x_i for those observations for which y_i is available. The problem is that this is equivalent to conditioning on the event $\{T_i = 1\}$. However,

$$\begin{aligned} E(e_{1i} | T_i = 1, Z_i) &= E(e_{1i} | \{e_{0i} > -z_i' \gamma\}, Z_i) \\ &= \rho E(e_{0i} | \{e_{0i} > -z_i' \gamma\}, Z_i) + E(v_i | \{e_{0i} > -z_i' \gamma\}, Z_i) \\ &= \rho \lambda(z_i' \gamma), \end{aligned}$$

which is non-zero. Thus

$$e_{1i} = \rho \lambda(z_i' \gamma) + u_i,$$

where

$$E(u_i | T_i = 1, Z_i) = 0.$$

Hence

$$y_i = x_i' \beta + \rho \lambda(z_i' \gamma) + u_i \tag{14.2}$$

is a valid regression equation for the observations for which $T_i = 1$.

Heckman (1979) observed that we could consistently estimate β and ρ from this equation, if γ were known. It is unknown, but also can be consistently estimated by a Probit model for selection. The ‘‘Heckit’’ estimator is thus calculated as follows

- Estimate $\hat{\gamma}$ from a Probit, using regressors z_i . The binary dependent variable is T_i .
- Estimate $(\hat{\beta}, \hat{\rho})$ from OLS of y_i on x_i and $\lambda(z_i' \hat{\gamma})$.
- The OLS standard errors will be incorrect, as this is a two-step estimator. They can be corrected using a more complicated formula. Or, alternatively, by viewing the Probit/OLS estimation equations as a large joint GMM problem.

The Heckit estimator is frequently used to deal with problems of sample selection. However, the estimator is built on the assumption of normality, and the estimator can be quite sensitive to this assumption. Some modern econometric research is exploring how to relax the normality assumption.

The estimator can also work quite poorly if $\lambda(z_i'\hat{\gamma})$ does not have much in-sample variation. This can happen if the Probit equation does not “explain” much about the selection choice. Another potential problem is that if $z_i = x_i$, then $\lambda(z_i'\hat{\gamma})$ can be highly collinear with x_i , so the second step OLS estimator will not be able to precisely estimate β . Based this observation, it is typically recommended to find a valid exclusion restriction: a variable should be in z_i which is not in x_i . If this is valid, it will ensure that $\lambda(z_i'\hat{\gamma})$ is not collinear with x_i , and hence improve the second stage estimator’s precision.

Chapter 15

Panel Data

A panel is a set of observations on individuals, collected over time. An observation is the pair $\{y_{it}, x_{it}\}$, where the i subscript denotes the individual, and the t subscript denotes time. A panel may be *balanced*:

$$\{y_{it}, x_{it}\} : t = 1, \dots, T; \quad i = 1, \dots, n,$$

or *unbalanced*:

$$\{y_{it}, x_{it}\} : \text{For } i = 1, \dots, n, \quad t = \underline{t}_i, \dots, \bar{t}_i.$$

15.1 Individual-Effects Model

The standard panel data specification is that there is an individual-specific effect which enters linearly in the regression

$$y_{it} = x'_{it}\beta + u_i + e_{it}.$$

The typical maintained assumptions are that the individuals i are mutually independent, that u_i and e_{it} are independent, that e_{it} is iid across individuals and time, and that e_{it} is uncorrelated with x_{it} .

OLS of y_{it} on x_{it} is called pooled estimation. It is consistent if

$$E(x_{it}u_i) = 0 \tag{15.1}$$

If this condition fails, then OLS is inconsistent. (15.1) fails if the individual-specific unobserved effect u_i is correlated with the observed explanatory variables x_{it} . This is often believed to be plausible if u_i is an omitted variable.

If (15.1) is true, however, OLS can be improved upon via a GLS technique. In either event, OLS appears a poor estimation choice.

Condition (15.1) is called the *random effects hypothesis*. It is a strong assumption, and most applied researchers try to avoid its use.

15.2 Fixed Effects

This is the most common technique for estimation of non-dynamic linear panel regressions.

The motivation is to allow u_i to be arbitrary, and have arbitrary correlated with x_i . The goal is to eliminate u_i from the estimator, and thus achieve invariance.

There are several derivations of the estimator.

First, let

$$d_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{else} \end{cases},$$

and

$$d_i = \begin{pmatrix} d_{i1} \\ \vdots \\ d_{in} \end{pmatrix},$$

an $n \times 1$ dummy vector with a “1” in the i 'th place. Let

$$u = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}.$$

Then note that

$$u_i = d_i' u,$$

and

$$y_{it} = x_{it}' \beta + d_i' u + e_{it}. \quad (15.2)$$

Observe that

$$E(e_{it} \mid x_{it}, d_i) = 0,$$

so (15.2) is a valid regression, with d_i as a regressor along with x_i .

OLS on (15.2) yields estimator $(\hat{\beta}, \hat{u})$. Conventional inference applies.

Observe that

- This is generally consistent.
- If x_{it} contains an intercept, it will be collinear with d_i , so the intercept is typically omitted from x_{it} .
- Any regressor in x_{it} which is constant over time for all individuals (e.g., their gender) will be collinear with d_i , so will have to be omitted.
- There are $n + k$ regression parameters, which is quite large as typically n is very large.

Computationally, you do not want to actually implement conventional OLS estimation, as the parameter space is too large. OLS estimation of β proceeds by the FWL theorem. Stacking the observations together:

$$Y = X\beta + Du + e,$$

then by the FWL theorem,

$$\begin{aligned} \hat{\beta} &= (X'(1 - P_D)X)^{-1} (X'(1 - P_D)Y) \\ &= (X^*X^*)^{-1} (X^*Y^*), \end{aligned}$$

where

$$\begin{aligned} Y^* &= Y - D(D'D)^{-1}D'Y \\ X^* &= X - D(D'D)^{-1}D'X. \end{aligned}$$

Since the regression of y_{it} on d_i is a regression onto individual-specific dummies, the predicted value from these regressions is the individual specific mean \bar{y}_i , and the residual is the demean value

$$y_{it}^* = y_{it} - \bar{y}_i.$$

The fixed effects estimator $\hat{\beta}$ is OLS of y_{it}^* on x_{it}^* , the dependent variable and regressors in deviation-from mean form.

Another derivation of the estimator is to take the equation

$$y_{it} = x_{it}'\beta + u_i + e_{it},$$

and then take individual-specific means by taking the average for the i 'th individual:

$$\frac{1}{T_i} \sum_{t=\underline{t}_i}^{\bar{t}_i} y_{it} = \frac{1}{T_i} \sum_{t=\underline{t}_i}^{\bar{t}_i} x_{it}'\beta + u_i + \frac{1}{T_i} \sum_{t=\underline{t}_i}^{\bar{t}_i} e_{it}$$

or

$$\bar{y}_i = \bar{x}_i'\beta + u_i + \bar{e}_i.$$

Subtracting, we find

$$y_{it}^* = x_{it}'\beta + e_{it}^*,$$

which is free of the individual-effect u_i .

15.3 Dynamic Panel Regression

A dynamic panel regression has a lagged dependent variable

$$y_{it} = \alpha y_{it-1} + x_{it}'\beta + u_i + e_{it}. \quad (15.3)$$

This is a model suitable for studying dynamic behavior of individual agents.

Unfortunately, the fixed effects estimator is inconsistent, at least if T is held finite as $n \rightarrow \infty$. This is because the sample mean of y_{it-1} is correlated with that of e_{it} .

The standard approach to estimate a dynamic panel is to combine first-differencing with IV or GMM. Taking first-differences of (15.3) eliminates the individual-specific effect:

$$\Delta y_{it} = \alpha \Delta y_{it-1} + \Delta x_{it}'\beta + \Delta e_{it}. \quad (15.4)$$

However, if e_{it} is iid, then it will be correlated with Δy_{it-1} :

$$E(\Delta y_{it-1} \Delta e_{it}) = E((y_{it-1} - y_{it-2})(e_{it} - e_{it-1})) = -E(y_{it-1} e_{it-1}) = -\sigma_e^2.$$

So OLS on (15.4) will be inconsistent.

But if there are valid instruments, then IV or GMM can be used to estimate the equation. Typically, we use lags of the dependent variable, two periods back, as y_{t-2} is uncorrelated with Δe_{it} . Thus values of y_{it-k} , $k \geq 2$, are valid instruments.

Hence a valid estimator of α and β is to estimate (15.4) by IV using y_{t-2} as an instrument for Δy_{t-1} (which is just identified). Alternatively, GMM using y_{t-2} and y_{t-3} as instruments (which is overidentified, but loses a time-series observation).

A more sophisticated GMM estimator recognizes that for time-periods later in the sample, there are more instruments available, so the instrument list should be different for each equation. This is conveniently organized by the GMM principle, as this enables the moments from the different time-periods to be stacked together to create a list of all the moment conditions. A simple application of GMM yields the parameter estimates and standard errors.

Bibliography

- [1] Aitken, A.C. (1935): “On least squares and linear combinations of observations,” *Proceedings of the Royal Statistical Society*, 55, 42-48.
- [2] Akaike, H. (1973): “Information theory and an extension of the maximum likelihood principle.” In B. Petroc and F. Csake, eds., *Second International Symposium on Information Theory*.
- [3] Anderson, T.W. and H. Rubin (1949): “Estimation of the parameters of a single equation in a complete system of stochastic equations,” *The Annals of Mathematical Statistics*, 20, 46-63.
- [4] Andrews, D.W.K. (1988): “Laws of large numbers for dependent non-identically distributed random variables,” *Econometric Theory*, 4, 458-467.
- [5] Andrews, D.W.K. (1993), “Tests for parameter instability and structural change with unknown change point,” *Econometrica*, 61, 821-8516.
- [6] Andrews, D.W.K. and M. Buchinsky: (2000): “A three-step method for choosing the number of bootstrap replications,” *Econometrica*, 68, 23-51.
- [7] Andrews, D.W.K. and W. Ploberger (1994): “Optimal tests when a nuisance parameter is present only under the alternative,” *Econometrica*, 62, 1383-1414.
- [8] Basmann, R. L. (1957): “A generalized classical method of linear estimation of coefficients in a structural equation,” *Econometrica*, 25, 77-83.
- [9] Bekker, P.A. (1994): “Alternative approximations to the distributions of instrumental variable estimators,” *Econometrica*, 62, 657-681.
- [10] Billingsley, P. (1968): *Convergence of Probability Measures*. New York: Wiley.
- [11] Billingsley, P. (1979): *Probability and Measure*. New York: Wiley.
- [12] Bose, A. (1988): “Edgeworth correction by bootstrap in autoregressions,” *Annals of Statistics*, 16, 1709-1722.
- [13] Breusch, T.S. and A.R. Pagan (1979): “The Lagrange multiplier test and its application to model specification in econometrics,” *Review of Economic Studies*, 47, 239-253.
- [14] Brown, B.W. and W.K. Newey (2001): “GMM, efficient bootstrapping, and improved inference,” <http://econ-www.mit.edu/faculty/wnewey/files/gmmjbes.pdf>.
- [15] Carlstein, E. (1986): “The use of subsamples methods for estimating the variance of a general statistic from a stationary time series,” *Annals of Statistics*, 14, 1171-1179.

- [16] Chamberlain, G. (1987): "Asymptotic efficiency in estimation with conditional moment restrictions," *Journal of Econometrics*, 34, 305-334.
- [17] Choi, I. and P.C.B. Phillips (1992): "Asymptotic and finite sample distribution theory for IV estimators and tests in partially identified structural equations," *Journal of Econometrics*, 51, 113-150.
- [18] Chow, G.C. (1960): "Tests of equality between sets of coefficients in two linear regressions," *Econometrica*, 28, 591-603.
- [19] Davidson, J. (1994): *Stochastic Limit Theory: An Introduction for Econometricians*. Oxford: Oxford University Press.
- [20] Davison, A.C. and D.V. Hinkley (1997): *Bootstrap Methods and their Application*. Cambridge University Press.
- [21] Dickey, D.A. and W.A. Fuller (1979): "Distribution of the estimators for autoregressive time series with a unit root," *Journal of the American Statistical Association*, 74, 427-431.
- [22] Donald S.G. and W.K. Newey (2001): "Choosing the number of instruments," *Econometrica*, 69, 1161-1191.
- [23] Dufour, J.M. (1997): "Some impossibility theorems in econometrics with applications to structural and dynamic models," *Econometrica*, 65, 1365-1387.
- [24] Efron, B. (1979): "Bootstrap methods: Another look at the jackknife," *Annals of Statistics*, 7, 1-26.
- [25] Efron, B. and R.J. Tibshirani (1993): *An Introduction to the Bootstrap*, New York: Chapman-Hall.
- [26] Eicker, F. (1963): "Asymptotic normality and consistency of the least squares estimators for families of linear regressions," *Annals of Mathematical Statistics*, 34, 447-456.
- [27] Engle, R.F. and C.W.J. Granger (1987): "Co-integration and error correction: Representation, estimation and testing," *Econometrica*, 55, 251-276.
- [28] Frisch, R. and F. Waugh (1933): "Partial time regressions as compared with individual trends," *Econometrica*, 1, 387-401.
- [29] Gallant, A.F. and D.W. Nychka (1987): "Seminonparametric maximum likelihood estimation," *Econometrica*, 55, 363-390.
- [30] Gallant, A.R. and H. White (1988): *A Unified Theory of Estimation and Inference for Non-linear Dynamic Models*. New York: Basil Blackwell.
- [31] Gauss, K.F. (1809): "Theoria motus corporum coelestium," in *Werke*, Vol. VII, 240-254.
- [32] Granger, C.W.J. (1969): "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, 37, 424-438.
- [33] Granger, C.W.J. (1981): "Some properties of time series data and their use in econometric specification," *Journal of Econometrics*, 16, 121-130.

- [34] Granger, C.W.J. and T. Teräsvirta (1993): *Modelling Nonlinear Economic Relationships*, Oxford University Press, Oxford.
- [35] Hall, A. R. (2000): "Covariance matrix estimation and the power of the overidentifying restrictions test," *Econometrica*, 68, 1517-1527,
- [36] Hall, P. (1992): *The Bootstrap and Edgeworth Expansion*, New York: Springer-Verlag.
- [37] Hall, P. (1994): "Methodology and theory for the bootstrap," *Handbook of Econometrics, Vol. IV*, eds. R.F. Engle and D.L. McFadden. New York: Elsevier Science.
- [38] Hall, P. and J.L. Horowitz (1996): "Bootstrap critical values for tests based on Generalized-Method-of-Moments estimation," *Econometrica*, 64, 891-916.
- [39] Hahn, J. (1996): "A note on bootstrapping generalized method of moments estimators," *Econometric Theory*, 12, 187-197.
- [40] Hansen, B.E. (1992): "Efficient estimation and testing of cointegrating vectors in the presence of deterministic trends," *Journal of Econometrics*, 53, 87-121.
- [41] Hansen, B.E. (1996): "Inference when a nuisance parameter is not identified under the null hypothesis," *Econometrica*, 64, 413-430.
- [42] Hansen, L.P. (1982): "Large sample properties of generalized method of moments estimators," *Econometrica*, 50, 1029-1054.
- [43] Hansen, L.P., J. Heaton, and A. Yaron (1996): "Finite sample properties of some alternative GMM estimators," *Journal of Business and Economic Statistics*, 14, 262-280.
- [44] Hausman, J.A. (1978): "Specification tests in econometrics," *Econometrica*, 46, 1251-1271.
- [45] Heckman, J. (1979): "Sample selection bias as a specification error," *Econometrica*, 47, 153-161.
- [46] Imbens, G.W. (1997): "One step estimators for over-identified generalized method of moments models," *Review of Economic Studies*, 64, 359-383.
- [47] Imbens, G.W., R.H. Spady and P. Johnson (1998): "Information theoretic approaches to inference in moment condition models," *Econometrica*, 66, 333-357.
- [48] Jarque, C.M. and A.K. Bera (1980): "Efficient tests for normality, homoskedasticity and serial independence of regression residuals," *Economic Letters*, 6, 255-259.
- [49] Johansen, S. (1988): "Statistical analysis of cointegrating vectors," *Journal of Economic Dynamics and Control*, 12, 231-254.
- [50] Johansen, S. (1991): "Estimation and hypothesis testing of cointegration vectors in the presence of linear trend," *Econometrica*, 59, 1551-1580.
- [51] Johansen, S. (1995): *Likelihood-Based Inference in Cointegrated Vector Auto-Regressive Models*, Oxford University Press.
- [52] Johansen, S. and K. Juselius (1992): "Testing structural hypotheses in a multivariate cointegration analysis of the PPP and the UIP for the UK," *Journal of Econometrics*, 53, 211-244.

- [53] Kitamura, Y. (2001): "Asymptotic optimality and empirical likelihood for testing moment restrictions," *Econometrica*, 69, 1661-1672.
- [54] Kitamura, Y. and M. Stutzer (1997): "An information-theoretic alternative to generalized method of moments," *Econometrica*, 65, 861-874..
- [55] Kunsch, H.R. (1989): "The jackknife and the bootstrap for general stationary observations," *Annals of Statistics*, 17, 1217-1241.
- [56] Kwiatkowski, D., P.C.B. Phillips, P. Schmidt, and Y. Shin (1992): "Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?" *Journal of Econometrics*, 54, 159-178.
- [57] Lafontaine, F. and K.J. White (1986): "Obtaining any Wald statistic you want," *Economics Letters*, 21, 35-40.
- [58] Lovell, M.C. (1963): "Seasonal adjustment of economic time series," *Journal of the American Statistical Association*, 58, 993-1010.
- [59] MacKinnon, J.G. (1990): "Critical values for cointegration," in Engle, R.F. and C.W. Granger (eds.) *Long-Run Economic Relationships: Readings in Cointegration*, Oxford, Oxford University Press.
- [60] MacKinnon, J.G. and H. White (1985): "Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties," *Journal of Econometrics*, 29, 305-325.
- [61] Magnus, J. R., and H. Neudecker (1988): *Matrix Differential Calculus with Applications in Statistics and Econometrics*, New York: John Wiley and Sons.
- [62] Muirhead, R.J. (1982): *Aspects of Multivariate Statistical Theory*. New York: Wiley.
- [63] Newey, W.K. and K.D. West (1987): "Hypothesis testing with efficient method of moments estimation," *International Economic Review*, 28, 777-787.
- [64] Owen, Art B. (1988): "Empirical likelihood ratio confidence intervals for a single functional," *Biometrika*, 75, 237-249.
- [65] Owen, Art B. (2001): *Empirical Likelihood*. New York: Chapman & Hall.
- [66] Phillips, P.C.B. (1989): "Partially identified econometric models," *Econometric Theory*, 5, 181-240.
- [67] Phillips, P.C.B. and S. Ouliaris (1990): "Asymptotic properties of residual based tests for cointegration," *Econometrica*, 58, 165-193.
- [68] Politis, D.N. and J.P. Romano (1996): "The stationary bootstrap," *Journal of the American Statistical Association*, 89, 1303-1313.
- [69] Potscher, B.M. (1991): "Effects of model selection on inference," *Econometric Theory*, 7, 163-185.
- [70] Qin, J. and J. Lawless (1994): "Empirical likelihood and general estimating equations," *The Annals of Statistics*, 22, 300-325.

- [71] Ramsey, J. B. (1969): "Tests for specification errors in classical linear least-squares regression analysis," *Journal of the Royal Statistical Society, Series B*, 31, 350-371.
- [72] Rudin, W. (1987): *Real and Complex Analysis*, 3rd edition. New York: McGraw-Hill.
- [73] Said, S.E. and D.A. Dickey (1984): "Testing for unit roots in autoregressive-moving average models of unknown order," *Biometrika*, 71, 599-608.
- [74] Shao, J. and D. Tu (1995): *The Jackknife and Bootstrap*. NY: Springer.
- [75] Sargan, J.D. (1958): "The estimation of economic relationships using instrumental variables," *Econometrica*, 26, 393-415.
- [76] Shin, Y. (1994): "A residual-based test of the null of cointegration against the alternative of no cointegration," *Econometric Theory*, 10, 91-115.
- [77] Silverman, B.W. (1986): *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.
- [78] Sims, C.A. (1972): "Money, income and causality," *American Economic Review*, 62, 540-552.
- [79] Sims, C.A. (1980): "Macroeconomics and reality," *Econometrica*, 48, 1-48.
- [80] Staiger, D. and J.H. Stock (1997): "Instrumental variables regression with weak instruments," *Econometrica*, 65, 557-586.
- [81] Stock, J.H. (1987): "Asymptotic properties of least squares estimators of cointegrating vectors," *Econometrica*, 55, 1035-1056.
- [82] Stock, J.H. (1991): "Confidence intervals for the largest autoregressive root in U.S. macroeconomic time series," *Journal of Monetary Economics*, 28, 435-460.
- [83] Stock, J.H. and J.H. Wright (2000): "GMM with weak identification," *Econometrica*, 68, 1055-1096.
- [84] Theil, H. (1953): "Repeated least squares applied to complete equation systems," The Hague, Central Planning Bureau, mimeo.
- [85] Tobin, J. (1958): "Estimation of relationships for limited dependent variables," *Econometrica*, 26, 24-36.
- [86] Wald, A. (1943): "Tests of statistical hypotheses concerning several parameters when the number of observations is large," *Transactions of the American Mathematical Society*, 54, 426-482.
- [87] Wang, J. and E. Zivot (1998): "Inference on structural parameters in instrumental variables regression with weak instruments," *Econometrica*, 66, 1389-1404.
- [88] White, H. (1980): "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity," *Econometrica*, 48, 817-838.
- [89] White, H. (1984): *Asymptotic Theory for Econometricians*, Academic Press.
- [90] Zellner, A. (1962): "An efficient method of estimating seemingly unrelated regressions, and tests for aggregation bias," *Journal of the American Statistical Association*, 57, 348-368.