

James B. McDonald
Brigham Young University
12/2011

I. Matrix algebra

1. Basic definitions
2. Matrix operations
 - a. Definitions
 - b. Properties
3. Partitioned matrices
4. Quadratic forms
5. Kronecker products
6. Characteristic roots and vectors
7. Vector and matrix differentiation with applications to optimization problems
8. Problem sets

I. Matrix algebra

1. Basic definitions:

matrix (order of matrix)

square matrix

transpose of a matrix

symmetric matrix

diagonal matrix

identity matrix

null or zero matrix

row or column vector

determinant of a square matrix

inverse of a square matrix

idempotent matrix

orthogonal matrix

trace of a square matrix

rank of a matrix

2. Matrix operations

a. Definitions

(1) scalar multiplication

(2) addition of matrices

(3) multiplication of matrices

(4) inverses of nonsingular matrices

- b. Some properties of matrix operations. Let c denote a real number and A, B, C denote matrices. The following properties are conditional on the operations being defined for the case in point.

(1) Scalar multiplication:

$$c(A + B) = cA + cB = (A + B)c$$

(2) Addition:

$$\begin{aligned} A + B &= B + A \\ A + (B + C) &= (A + B) + C \\ A + 0 &= 0 + A = A \end{aligned}$$

(3) Multiplication:

$$\begin{aligned} *AB &\neq BA \\ *AB = AC &\Rightarrow (for\ all\ cases)\ B = C \\ A(BC) &= (AB)C \\ A(B + C) &= AB + AC \\ (B + C)A &= BA + CA \\ A0 &= 0A = 0 \\ AI &= IA = A \end{aligned}$$

(4) Transposes and inverses:

$$\begin{aligned} (A')' &= A \\ (ABC)' &= C'B'A' \\ (A + B)' &= A' + B' \\ (A^{-1})^{-1} &= A \\ (ABC)^{-1} &= C^{-1}B^{-1}A^{-1}, \text{ if } A, B, \text{ and } C \text{ have inverses} \\ *(A + B)^{-1} &\neq A^{-1} + B^{-1} \\ (A^{-1})' &= (A')^{-1} \end{aligned}$$

(5) Trace:

$$\begin{aligned} \text{Trace } (ABC) &= \text{Trace } (CAB) = \text{Trace } (BCA) \\ \text{Trace } (A + B) &= \text{Trace } (A) + \text{Trace } (B) \end{aligned}$$

3. Partitioned Matrices (Greene, 4th ed., pp. 33-34)

Let $A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$ be an $m \times n$ matrix

where A_{11} is $m_1 \times n_1$

A_{22} is $m_2 \times n_2$

A_{12} is $m_1 \times n_2$

A_{21} is $m_2 \times n_1$

$n_1 + n_2 = n$

$m_1 + m_2 = m$

Also let B and C denote $m \times n$ and $n \times q$ matrices which are conformably partitioned as

$$B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$$

and

$$C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}$$

so that the operations to be discussed are defined.

Addition of partitioned matrices:

$$\begin{aligned} A + B &= \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} + \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} \\ &= \begin{pmatrix} A_{11} + B_{11} & A_{12} + B_{12} \\ A_{21} + B_{21} & A_{22} + B_{22} \end{pmatrix} \end{aligned}$$

* A_{ij} and B_{ij} must be of the same dimension.

Multiplication of partitioned matrices:

$$\begin{aligned} AC &= \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix} \\ &= \begin{pmatrix} A_{11}C_{11} + A_{12}C_{21} & A_{11}C_{12} + A_{12}C_{22} \\ A_{21}C_{11} + A_{22}C_{21} & A_{21}C_{12} + A_{22}C_{22} \end{pmatrix} \end{aligned}$$

*The number of columns in A_{ij} must be the same as the number of rows in B_{jk}

Inverses of partitioned matrices:

Let A be a partitioned square matrix of order $m \times m$

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

where A_{11} and A_{22} are nonsingular square matrices of order m_1 and m_2 , respectively ($m_1 + m_2 = m$).

The determinant of A can be expressed by either of the following relations:

$$\begin{aligned} A &= |A_{11}| |A_{22} - A_{21}A_{11}^{-1}A_{12}| \\ &= |A_{22}| |A_{11} - A_{12}A_{22}^{-1}A_{21}| \end{aligned}$$

If A is also nonsingular and

$$A^{-1} = B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}, \text{ then}$$

the partitioned inverse of A, B can be expressed as

$$B_{11} = (A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1}$$

$$B_{12} = -B_{11}A_{12}A_{22}^{-1}$$

$$B_{21} = -A_{22}^{-1}A_{21}B_{11}$$

$$B_{22} = A^{-1}_{22} + A^{-1}_{22} A_{21} B^{-1}_{11} A_{12} A_{22}$$

and an alternative form for the B_{ij} (inverse of A) is given by

$$B_{11} = A^{-1}_{11} + A^{-1}_{11} A_{12} B^{-1}_{22} A_{21} A_{11}$$

$$B_{12} = -A^{-1}_{11} A_{12} B_{22}$$

$$B_{21} = -B_{22} A_{21} A^{-1}_{11}$$

$$B_{22} = (A_{22} - A_{21} A^{-1}_{11} A_{12})^{-1}$$

Note that if the matrix A is

$$\text{Block diagonal } A = \begin{pmatrix} A_{11} & 0 \\ 0 & A_{22} \end{pmatrix}$$

$$\text{or Block triangular } A = \begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{pmatrix} \text{ or } A = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}$$

then the expressions for the determinants and inverses just considered are simplified considerably.

4. Quadratic forms and their classification.

Let A denote an $n \times n$ symmetric matrix with real entries and let X denote an $n \times 1$ column vector.

$Q = X'AX$ is said to be a quadratic form. Note that

$$Q = X'AX = (x_1 \dots x_n) \begin{pmatrix} a_{11} \dots a_{1n} \\ \vdots \\ a_{n1} \dots a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

$$\begin{aligned}
 &= (x_1 \dots x_n) \begin{pmatrix} \sum a_{11} x_i \\ \vdots \\ \sum a_{nn} x_i \end{pmatrix} \\
 &= a_{11} x_1^2 + a_{12} x_1 x_2 + \dots + a_{1n} x_1 x_n \\
 &\quad + a_{21} x_2 x_1 + a_{22} x_2^2 + \dots + a_{2n} x_2 x_n \\
 &\quad \cdot \quad \cdot \quad \cdot \\
 &\quad \cdot \quad \cdot \quad \cdot \\
 &\quad \cdot \quad \cdot \quad \cdot \\
 &\quad + a_{n1} x_n x_1 + a_{n2} x_n x_2 + \dots + a_{nn} x_n^2
 \end{aligned}$$

Classification of the quadratic form $Q = X'AX$:

negative definite: $Q < 0$ if $x \neq 0$

negative semidefinite: $Q \leq 0$ for all x and $Q = 0$ for some $x \neq 0$

positive definite: $Q > 0$ if $x \neq 0$

positive semidefinite: $Q \geq 0$ for all x and $Q = 0$ for some $x \neq 0$

indefinite: $Q > 0$ for some x and $Q < 0$ for some other x

A necessary and sufficient condition for positive or negative definiteness (A is symmetric):

Positive Definite

$$a_{11} > 0$$

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} > 0$$

.

.

.

$$\begin{vmatrix} a_{11} a_{12} \dots a_{1n} \\ a_{21} a_{22} \dots a_{2n} \\ \vdots & \vdots \\ a_{n1} a_{n2} \dots a_{nn} \end{vmatrix} > 0$$

(all positive determinants)

Negative Definite

$$a_{11} < 0$$

$$> 0$$

.

.

.

(has sign of $(-1)^n$)

(first negative, then alternate in sign)

Note:

In the discussion on regression theory it will be useful to note that if X is $N \times K$ ($N > K$) and if $(X'X)^{-1}$ exists, then $(X'X)$ is positive definite.

5. Kronecker Products

Let $A = (a_{ij})$ $m \times n$

$B = (b_{ij})$ $p \times q$,

then the Kronecker product of A and B is denoted by $A \otimes B$ and is defined by

$$A \otimes B = \begin{pmatrix} a_{11}B & a_{12}B & \dots & a_{1n}B \\ a_{21}B & a_{22}B & \dots & a_{2n}B \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ a_{m1}B & a_{m2}B & \dots & a_{mn}B \end{pmatrix}_{mp \times nq}$$

Properties: (matrices are assumed to be conformable)

1. $(A \otimes C)(B \otimes D) = (AB) \otimes (CD)$
2. $(P \otimes Q)^{-1} = P^{-1} \otimes Q^{-1}$, P and Q square and nonsingular
3. $(M \otimes N)' = M' \otimes N'$
4. $|A \otimes B| = |A|^p |B|^m$ A and B are square of order $m \times m$ and $p \times p$.
5. $\text{trace}(A \otimes B) = \text{trace}(A) \text{trace}(B)$

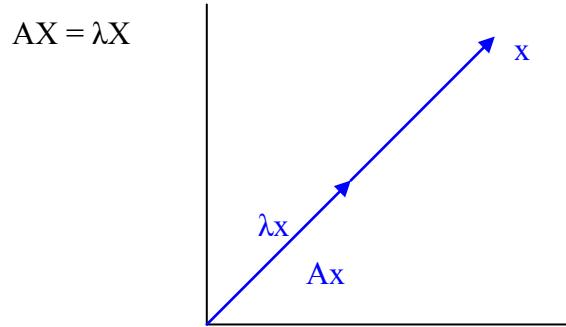
Example:

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \otimes \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix} = \begin{pmatrix} \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix} & 2 \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix} \\ 3 \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix} & 4 \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix} \end{pmatrix}$$

$$= \begin{pmatrix} 5 & 6 & 10 & 12 \\ 7 & 8 & 14 & 16 \\ 15 & 18 & 20 & 24 \\ 21 & 24 & 28 & 32 \end{pmatrix}$$

6. Characteristic Roots and Vectors

Let A denote an $n \times n$ matrix. Consider the matrix equation:



Given the $n \times n$ matrix A, it is frequently useful to determine ($n \times 1$) vectors X and corresponding scalars such that $AX = \lambda X$, i.e.

$$AX - \lambda X = (A - \lambda I)X = 0$$

The only way in which we can obtain a nontrivial solution ($X \neq 0$) to this system of equations is for $|A - \lambda I| = 0$.

The characteristic equation associated with the matrix A is defined by

$$|A - \lambda I| = 0.$$

The characteristic equation is an n^{th} degree polynomial in λ and will, by the fundamental theorem of algebra, have n solutions ($\lambda_1, \dots, \lambda_n$) which are referred to as the characteristic (latent or eigen) roots of the matrix A. The λ_i are also referred to as eigen values and may be real or possibly imaginary. The characteristic roots will be real if A is a real symmetric matrix.

The characteristic (latent or eigen) vectors associated with A are determined by solving

$$AV_i = \lambda_i V_i \quad i = 1, 2, \dots, n$$

where λ_i denotes the associated characteristic root and V_i is the i^{th} characteristic vector. Note: $A(cV_i) = \lambda_i(cV_i)$ and the characteristic vectors are only unique up to a scalar multiple.

Some Important Properties. Let $A = (a_{ij})$ be an $n \times n$ matrix with characteristic roots λ_i ($i = 1, 2, \dots, n$)

$$(1) \quad \lambda_1 + \lambda_2 + \dots + \lambda_n = \text{trace}(A) = a_{11} + \dots + a_{nn}$$

$$(2) \quad \lambda_1 \lambda_2 \dots \lambda_n = |A|$$

Some additional properties are useful for symmetric matrices.

- (1) $A(X'AX)$ is positive definite $\Leftrightarrow \lambda_i > 0$ for all i
- (2) $A(X'AX)$ is negative definite $\Leftrightarrow \lambda_i < 0$ for all i
- (3) If $\lambda_i \neq \lambda_j$ and A is symmetric, then $\nabla_i \cdot \nabla_j = 0$
- (4) If A is symmetric and if the characteristic vectors are chosen to have length one, then the matrix $V = (V_1, \dots, V_n)$ is orthogonal and

$$AV = V\Lambda \text{ where } \Lambda = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

It follows that

$$V'AV = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

Note:

- (1) This may be true even if not all of the characteristic roots are unique.
- (2) This allows for the diagonalization of quadratic forms

$$\begin{aligned} Q &= X'AX = (X'V\Lambda V'X) \\ &= Z'\Lambda Z = \sum \lambda_i z_i^2 \end{aligned}$$

where $Z = X'V$. This is a very important result.
- (3) Rank $(A) =$ the number of nonzero characteristic roots

- (4) Characteristic roots of A^k are λ^k .
- (5) If A is an idempotent matrix, then its characteristic roots are zero or one and
$$\text{Rank}(A) = \text{trace}(A)$$
- (6) The condition number of the matrix A is defined to be

$$\left(\frac{\text{max root}}{\text{min root}} \right)^{1/2}$$

Large values of the condition number can indicate near singularity of the matrix A and is sometimes used in econometrics where A is the correlation matrix associated with the explanatory variables.

7. Vector and matrix differentiation with applications to constrained and unconstrained optimization problems.

a. Basic definitions

Let $f(X) = f(x_1, x_2, \dots, x_n) = y$ be a real valued function of the vector X .

The derivative of $f(X)$ with respect to the vector X will be defined by

$$\frac{df}{dX} = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots, \frac{\partial f}{\partial x_n} \right)'$$

and the second derivative of $f(x)$ with respect to X is defined by

$$\frac{d^2 f}{dX^2} = \frac{d}{dX} \left(\frac{df}{dX} \right) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \dots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \dots & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{pmatrix}$$

which is known as the Hessian matrix. This suggests an obvious definition of the derivative of a real valued function with respect to a matrix, i.e., the derivative of a real valued function of a matrix is defined to be the matrix of derivatives of corresponding elements.

b. Unconstrained optimization

Consider the optimization problem:

$$\underset{X}{\text{maximize}} \quad f(X).$$

The necessary (first order) conditions for a maximum or minimum are given by

$$\frac{df}{dX} = 0$$

The sufficient or second order condition for a maximum (minimum) is that

$$\frac{d^2 f}{dX^2}$$

is negative (positive) definite.

These results are readily obtained from a Taylor Series expansion of $f(x)$,

$$f(x) \approx f(x_0) + \left(\frac{df}{dx} \right)(x - x_0) + \frac{1}{2}(x - x_0)' \frac{d^2 f}{dx^2}(x - x_0).$$

Hint : x_0 is selected to satisfy the necessary condition.

c. Constrained optimization

Consider the constrained optimization problem defined by

$$\begin{array}{ll} \text{maximize} & f(X) \\ X & \end{array}$$

$$\text{subject to } g(X) = 0.$$

where $g(X) = 0$ denotes a $m \times 1$ vector of constraints, $m < n$. The solution can be obtained from the Lagrangian function

$$L(X; \lambda) = f(X) + \lambda' g(X)$$

$$\text{where } \lambda' = (\lambda_1, \dots, \lambda_m).$$

The necessary (first order) conditions for a solution to this problem are that

$$\frac{\partial L}{\partial X} = f_X + (g_X)' \lambda = 0$$

$$\frac{\partial L}{\partial \lambda} = g(X) = 0.$$

Sufficient conditions for a maximum or minimum can be stated in terms of the Hessian of the Lagrangian function (or bordered Hessian)

$$\begin{aligned} H &= \frac{d^2 L}{d^2(\lambda, X)} = \begin{bmatrix} \frac{\partial^2 L}{\partial \lambda^2} & \frac{\partial^2 L}{\partial \lambda \partial X} \\ \frac{\partial^2 L}{\partial X \partial \lambda} & \frac{\partial^2 L}{\partial X^2} \end{bmatrix} \\ &= \begin{bmatrix} 0 & g_X' \\ g_X & L_{XX} \end{bmatrix} \end{aligned}$$

A sufficient condition for a minimum is that the determinants of border preserving principal minors of H have sign $(-1)^m$ or zero. If $|H|$ is of sign $(-1)^n$ and the determinants of border preserving principal minors are zero or alternate in sign, then a sufficient condition for a maximum is satisfied.

Some Useful Results (Matrix Cookbook, <http://matrixcookbook.com>)

Let $a = (a_1, \dots, a_n)'$,

$$A = (a_{ij})_{n \times n} \text{ and } X = (x_1, \dots, x_n)'.$$

$$1. \quad \frac{d(a'X)}{dX} = a$$

$$2. \quad \frac{d(X'AX)}{dX} = 2AX \text{ (A is symmetric); } = (A + A')X \text{ otherwise}$$

$$3. \quad \frac{d^2(X'AX)}{dX^2} = 2A \text{ (A is symmetric); } (A + A') \text{ otherwise.}$$

$$4. \quad \frac{\partial \text{trace}(A)}{\partial A} = I$$

$$5. \quad \frac{\partial |A|}{\partial A} = \begin{cases} |A|(A')^{-1} & \text{if } |A| \neq 0 \\ 0 & \text{if } |A| = 0 \end{cases}$$

$$6. \quad \frac{\partial \log|A|}{\partial A} = (A')^{-1}$$

Hint: $|A| = \sum a_{ij} |C_{ij}|$

$$7. \quad \frac{dA^{-1}(x)}{dx} = -A^{-1} \frac{dA(x)}{dx} A^{-1}$$

$$8. \quad \frac{\partial(X'AX)}{\partial A} = XX'$$

Example

Consider $f(X) = (X - \mu)'B(X - \mu)$ where B is positive definite and symmetric. Determine the value of X which minimizes $f(X)$. Check the sufficient conditions.

$$f(X) = (X - \mu)'B(X - \mu) = X'BX - 2X'B\mu + \mu'B\mu$$

$$\frac{df}{dx} = 2BX - 2B\mu = 2B(X - \mu) = 0$$

$$\frac{d^2f}{dx^2} = 2B$$

Since $B \neq 0$, the solution is $X = \mu$. Sufficient conditions for a minimum are satisfied if B is positive definite.

8. Problems sets

1. Expand $(A + B)(A - B)$ and $(A - B)(A + B)$. Are these expansions the same? If not, why not?

2. Given $A = \begin{pmatrix} 1 & 0 & 3 \\ 2 & -1 & 1 \end{pmatrix}$, $B = \begin{pmatrix} 3 & 4 & 1 \\ 0 & -1 & 5 \\ 0 & 2 & -2 \end{pmatrix}$, $C = \begin{pmatrix} 2 \\ -1 \\ 4 \end{pmatrix}$

Calculate $(AB)'$, $B'A'$, $C'A'$ and $(AC)'$.

3. Prove that diagonal matrices of the same order are commutative in multiplication with each other.

4. Prove that

$$\begin{vmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ 0 & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & a_{nn} \end{vmatrix} = a_{11}a_{22}\dots a_{nn}$$

5. Show that the matrix

$$Q = \begin{bmatrix} 1/\sqrt{6} & 2/\sqrt{5} & 1/\sqrt{30} \\ -2/\sqrt{6} & 1/\sqrt{5} & -2/\sqrt{30} \\ 1/\sqrt{6} & 0 & -5/\sqrt{30} \end{bmatrix}$$

is orthogonal, i.e., $Q' = Q^{-1}$.

6. Determine whether the following quadratic forms are positive definite:

(a) $6x_1^2 + 49x_2^2 + 51x_3^2 - 82x_2x_3 + 20x_1x_3 - 4x_1x_2$

(b) $4x_1^2 + 9x_2^2 + 2x_3^2 + 8x_2x_3 + 6x_3x_1 + 6x_1x_2$

7. Prove that

$$A = \begin{pmatrix} 1/2 & 1 \\ 1/4 & 1/2 \end{pmatrix}$$

is a nonsymmetric, idempotent matrix.

8. Let X denote an $N \times K$ ($N > K$) matrix. Demonstrate that $B = I_N - X(X'X)^{-1}X'$ is symmetric and idempotent. You will see this matrix again.

9. Obtain the characteristic roots of the matrix

$$C = \begin{pmatrix} 5/2 & 1/2 \\ 1/2 & 5/2 \end{pmatrix}$$

Can you determine the sign of $(x_1, x_2) C (x_1, x_2)'$ for arbitrary $X = (x_1, x_2) \neq 0$? Defend your answer.

10. Using the technique of inverses of partitioned matrices, determine the inverse of

$$D = \left[\begin{array}{c|c} 10 & 0 \\ \hline 0 & 5 \\ 0 & 2 \\ \hline 0 & 2 \end{array} \right] = \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix}$$

11. Let

$$A = \begin{pmatrix} 5 & 2 \\ 2 & 1 \end{pmatrix} \text{ and } B = \begin{pmatrix} 3 & 5 \\ 1 & 2 \end{pmatrix}$$

Evaluate:

$$A \otimes B$$

$$(A \otimes B)^{-1}$$

$$(A \otimes B)'$$

$$\text{trace}(A \otimes B)$$

$$|A \otimes B|$$

Some additional fun exercises:

12. Let A and B be square matrices. Prove that

$$\text{trace}(AB) = \text{trace}(BA)$$

Hint: The ij^{th} element in the matrix product AB is given by $\sum_k a_{ik} b_{kj}$.

13. Prove:

a. $(P \otimes Q)^{-1} = P^{-1} \otimes Q^{-1}$

b. $\|A \otimes B\| = \|A\|^p \|B\|^m$ where A and B are respectively $m \times m$ and $p \times p$ matrices

c. $\text{Trace}(A \otimes B) = \text{trace}(A) \text{trace}(B)$

14. Prove that the characteristic vectors corresponding to unique (unequal) characteristic roots of a symmetric matrix A are orthogonal.

15. Prove that the characteristic roots of a real symmetric matrix are real.

16. Prove $\frac{\partial \text{trace}(A)}{\partial A} = I$.

17. a. Prove $\sum \lambda_i = \text{trace}(A)$ and $\prod \lambda_i = |A|$.

- b. Prove that the characteristic roots of A^k are λ^k .
- c. Prove that the characteristic roots of an idempotent matrix are either zero or one.

18.. Determine an upper triangular matrix T, such that

$$TT = \begin{pmatrix} 1 & 2 \\ 2 & 13 \end{pmatrix}$$

Any symmetric positive definite matrix (A) can be written as a product of a lower triangular matrix and its transpose (upper triangular matrix). This is referred to as the Cholesky factorization of A.

References

Green, William H., Econometric Analysis, 4th edition, New Jersey: Prentice Hall, 2000.

Hadley, G., Linear Algebra, Palo Alto: Addison-Wesley, 1961.

Magnus, Jan R. and Heinz Neudecker, Matrix Differential Calculus with Application in Statistics and Econometrics, New York: J. Wiley & Sons, 1988.

Petersen, K. B. and M. S. Pedersen, The Matrix Cookbook, <http://matrixcookbook.com>.

Searle, Shayler R, Matrix Algebra useful for Statistics, New York: J. Wiley and Sons, 1982.

II. Statistics

1. Univariate distributions

- a. Models for positive random variables
- b. Models for real-valued random variables
- c. Hazard functions
- d. Moments and characteristics functions
- e. Transformations of random variables

2. Multivariate distributions

- a. Expectations
- b. Multivariate normal
- c. Distribution results
 - (1) Chi-square
 - (2) Some useful theorems
 - (3) Example
 - (4) t-distribution
 - (5) F-distribution
- d. Some generalized multivariate distributions

3. Estimation

- a. Maximum likelihood
- b. Method of moments
- c. Kernel estimation

4. Asymptotic distribution theory

- a. Modes of convergence
- b. Law of large numbers

5. Problem sets

James B. McDonald
Brigham Young University
12/2012

II. STATISTICS

Statistical distributions play an important role in many areas of economics. These include descriptive models for the distribution of such economic variables as income, firm size, prices, and stock returns to mention but a few. Assumptions about underlying distributions also provide the theoretical basis for such common estimation techniques as least squares. It can be shown that many of the common distributions are closely related to each other and are special cases of a few generalized density functions. These general univariate distributions will be considered in the first section. Some multivariate distributions will be considered in the second section. Section three will discuss principles of maximum likelihood estimation and the distribution of these estimators under correct model specification. If the density function is misspecified, these (maximum likelihood) estimators will be referred to as quasi maximum likelihood (QMLE) and in some cases can still be consistent and asymptotically normal; however, in some important cases misspecification of the density can lead to inconsistent estimators. Section four reviews different statistical concepts of convergence and important theorems which arise in asymptotic distribution theory.

1. Some Univariate Distributions

A few statistical distributions have played a dominant role in empirical and theoretical work. Some of the most widely used models in economics include the normal, t, lognormal, Pareto, gamma and beta. Often these models have been adopted without testing their validity or considering the consequences of model misspecification. In selecting a statistical distribution it is important to consider distributional characteristics such as the mean, variance, skewness and kurtosis. For example, the selection of an exponential distribution implies that the associated variance equals the mean squared. We first consider some probability density functions (pdf) for positive random variables and then some pdf's for random variables which can assume positive and negative values. Hazard functions are then discussed. It is important to be able to determine the distribution of transformations of random variables with known distribution functions. The methods of change of variable and using moment generating or cumulant generating functions will be discussed and illustrated with examples.

a. **Models for Positive Random Variables:**

The generalized beta (GB), generalized beta of the first and second kind (GB1, GB2), and generalized gamma (GG) are density functions for positive random variables and are defined by

$$(1.1) \quad GB(y; a, b, c, p, q) = \frac{|a| y^{ap-1} (1 - (1-c)(y/b)^a)^{q-1}}{b^{ap} B(p, q) (1 + c(y/b)^a)^{p+q}} \quad \text{for } 0 < y^a < \frac{b^a}{1-c}$$

$$(1.2) \quad GB1(y; a, b, p, q) = \frac{|a| y^{ap-1} (1 - (y/b)^a)^{q-1}}{b^{ap} B(p, q)} \quad 0 < y < b,$$

$$(1.3) \quad GB2(y; a, b, p, q) = \frac{|a| y^{ap-1}}{b^{ap} B(p, q) (1 + (y/b)^a)^{p+q}} \quad 0 < y,$$

$$(1.4) \quad GG(y; a, \beta, p) = \begin{cases} \frac{|a| y^{ap-1} e^{-(y/\beta)^a}}{\beta^{ap} \Gamma(p)} & 0 < y \\ 0 & \text{otherwise.} \end{cases}$$

These distributions can be shown to include the beta of the first kind (B1), the beta of the second kind (B2), or Burr type 12 (BR12), Burr type 3 (BR3), Power (P), Lognormal (LN), Weibull (W), gamma (GA), F, Lomax (L), Fisk (Fisk), Uniform (U), Rayleigh (R), exponential (EXP), Chi Square (χ^2) and half Normal and half t distributions as special or limiting cases. These relationships are depicted in the form of a distribution tree in Figure 1 and are developed in detail in McDonald [1984], McDonald and Butler [1987], McDonald and Richards [1987], and McDonald and Xu [1991]. The GB2 distribution is referred to as a generalized F by Kalbfleisch and Prentice [1980] and a modified version (with a non-zero threshold) as a Feller-Pareto distribution, Arnold [1983].

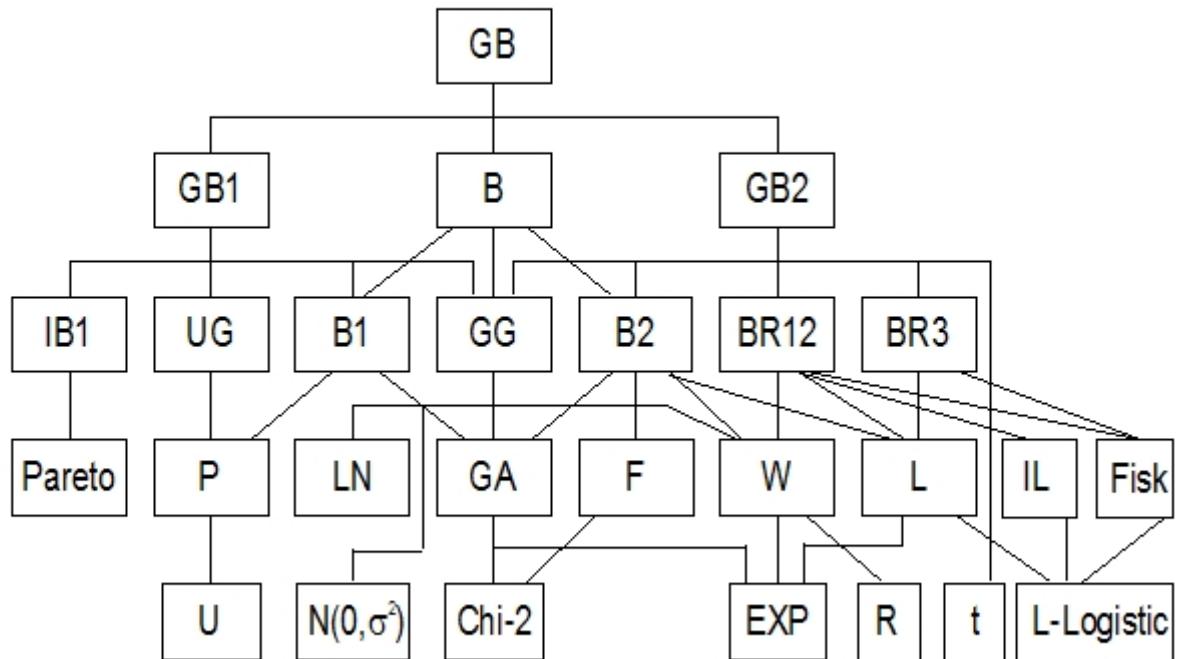
The parameters in these distributions generally determine the shape and location of the density in a complex manner. A few comments may be helpful. The parameters b and β are merely scale parameters and depend upon the units of measurement. The generalized beta type I and generalized gamma have defined moments ($E(y^h)$) of all integer order for $a > 0$ or more generally as long as $p+h/a > 0$; whereas, the generalized beta type II has integer moments of order up to " h " where $-p < h/a < q$. Consequently, the GB2 density permits the analysis of situations characterized by infinite variance. Generally speaking, the larger the value of "a" or "q," the "thinner" the tails of the density function. In fact, for "large" values of the parameter "a" the corresponding GB2 density function is characterized by the probability mass being concentrated near the value of the parameter "b." This can be verified by noting that for "large" values of "a" the mean is approximately "b" and the variance is near zero. The relative

values of the parameters "p" and "q" play an important role in determining the value of skewness and permit positive or negative skewness. This is in contrast to such distributions as the lognormal which is always positively skewed.

The interrelationships between many of these distributions are summarized in Figure 1, where the notation is formalized in Table 1. From Figure 1, we observe that the GB (eqn. 1.1) includes the GB1 (eqn. 1.2) and GB2 (eqn. 1.3) corresponding to $c = 0, 1$, respectively.

Figure 1

The Generalized Beta Family



see McDonald and Xu, "A Generalization of the Beta Distribution with Applications," *Journal of Econometrics* 66 (1995) pp.133-152

Table 1. Key to notation used in the distribution tree in figure 1.

Symbol	Distribution
$GB = Beta(y; a, b, c, p, q)$	Generalized Beta
$GB1 = GB1(y; a, b, p, q)$	Generalized beta type I
$GB2 = GB2(y; a, b, p, q)$	Generalized beta type II
$IB1 = IB1(y; b, p, q)$	Inverse beta type I
$UG = UG(y; b, \delta, q)$	Unit gamma
$B1 = B1(y; b, p, q)$	Beta of first kind
$LT = LT(y; \mu, \sigma^2, q)$	Log t
$GG = GG(y; a, \beta, p)$	Generalized gamma
$B2 = B2(y; b, p, q)$	Beta of second kind
$BR12 = BR12(y; a, b, q)$	Burr(type 12), Beta-P or Singh Maddala
$BR3 = BR3(y; a, b, p)$	Kappa (K-3), Beta-K or Burr (type 3)
$Pareto = Pareto(y; b, p)$	Pareto
$P = P(y; b, p)$	Power
$LN = LN(y; \mu, \sigma^2)$	Lognormal
$W = W(y; a, \beta)$	Weibull, Rosin-Rammler
$GA = GA(y; \beta, p)$	Gamma
$F = F(y; u, v)$	Snedecor F
$L = L(y; b, q)$	Lomax
$IL = IL(y; b, q)$	Inverse Lomax
$Fisk = FISK(y; a, b)$	Fisk
$U = U(y; b)$	Uniform
$\frac{1}{2}(N) = \frac{1}{2}(N(O, \sigma^2))$	Half normal
$\chi^2 = \chi^2(y; u)$	Chi square
$Exp = E(y; \beta)$	Exponential
$R = R(y; \beta)$	Rayleigh
$\frac{1}{2}(t) = \frac{1}{2}(t(y; v))$	Half t
Logistic	Logistics

The generalized beta distributions (GB1 and GB2) includes the generalized gamma as a limiting case,

$$\begin{aligned} \text{GG}(y; a, \beta, p) &= \lim_{q \rightarrow \infty} \text{GB1}(y; a, \beta q^{1/a}, p, q) \\ &= \lim_{q \rightarrow \infty} \text{GB2}(y; a, \beta q^{1/a}, p, q), \end{aligned}$$

and hence include all the special cases of the generalized gamma as limiting cases. See McDonald [1984] and McDonald and Richards [1987] and McDonald and Xu [1992] for details. Additional relationships can be more formally written as

$$\text{LN}(y; \mu, \sigma^2) = \underset{a \rightarrow 0}{\text{Limit}} \text{GG}\left(y; a, \beta = (\sigma^2 a^2)^{1/a}, p = \frac{a\mu+1}{\sigma^2 a^2}\right)$$

$$\begin{aligned} \text{GA}(y; \beta, p) &= \text{GG}(y; a = 1, \beta, p) \\ \text{W}(y; a, \beta) &= \text{GG}(y; a, \beta, p = 1) \\ \text{BR12}(y; a, b, q) &= \text{GB2}(y; a, b, p = 1, q) \\ \text{W}(y; a, \beta) &= \lim_{q \rightarrow \infty} \text{BR12}(y; a, b = \beta q^{1/a}, q). \end{aligned}$$

We see from figure 1 and the previous discussion that both the generalized gamma and Burr type 12 are more general models than the widely used Weibull (W).

The generalized beta of the second kind involves four parameters and is a particularly useful family of distributions. It includes the generalized gamma, beta of the second kind, the Burr type 12 or Beta-P and the Burr type 3, the three parameter Kappa or Beta-K distribution, and all the previously mentioned associated special cases as members. The distribution of the Snedecor F statistic (variance ratio) is also a special case of the generalized beta of the second kind, as are the half normal and half t. The hierarchical relationships among the distributions can be used in testing the nested hypotheses.

Table 2 is included for completeness and includes expressions for the density, cumulative distribution and moments. The notation is defined in McDonald [1984] and McDonald and Richards [1987].

Table 2. Some Probability Density Functions

Distribution	Domain of Y	pdf	Restrictions	Mean	Variance*	MGF
Poisson($y; \lambda$)	$\{0, 1, 2, \dots\}$	$e^{-\lambda} \lambda^y / y!$	$0 < \lambda$	λ	λ	$e^{\mu(\epsilon'-1)}$
Uniform($y; b$)	$\{0 < y < b\}$	$1/b$	$0 < b$	$b/2$	$b^2/12$	$(e^{bt} - 1)/(bt)$
$EXP(y; \beta)$	$\{0 < y\}$	$e^{-y/\beta} / \beta$	$0 < \beta$	β	β^2	$(1 - \beta t)^{-1}$
$\chi^2(d; d)$	$\{0 < y\}$	$y^{(d/2)-1} e^{-y/2} / (\Gamma(d/2) 2^{d/2})$	$d = 1, 2, \dots$	d	$2d$	$(1 - 2t)^{-d/2}$
$GA(y; \beta, p)$	$\{0 < y\}$	$y^{p-1} e^{-y/\beta} / (\Gamma(p) \beta^p)$	$0 < \beta, p$	βp	$\beta^2 p$	$(1 - \beta t)^{-p}$
$B1(y; b, p, q)$	$\{0 < y < b\}$	$y^{p-1} (1 - y/b)^{q-1} / (b^p B(p, q))$	$0 < b, p, q$	$\frac{bp}{p+q}$	$\frac{b^2 pq}{(p+q)^2 (p+q+1)}$	Involves an infinite series
$B2(y; b, p, q)$	$\{0 < y\}$	$y^{p-1} / (b^p B(p, q) (1 + y/b)^{p+q})$	$0 < b, p, q$	$\frac{bp}{q-1}$	$\frac{bB(p+h, q-h)}{B(p, q)}$	"
$F(y; p, q)$	$\{0 < y\}$	$y^{p-1} / (b^p B(p, q) (1 + y/b)^{p+q})$	$p = \frac{d_1}{2}, q = \frac{d_2}{2}$ $b = d_2/d_1$	$\frac{d_2}{d_2-2}$	B2 with appropriate substitutions	"
$GG(y; a, \beta, p)$	$\{0 < y\}$	$ a y^{p-1} e^{-(y/\beta)^a} / (\Gamma(p) \beta^p)$	$0 < \beta, p$	$\frac{\Gamma(p+1/a)}{\Gamma(p)}$	$\frac{\beta^a \Gamma(p+h/a)}{\Gamma(p)}$	"

$GB1(y; a, b, p, q)$	$\{0 < y < b\}$	$ a y^{a-1} \left(1 - (y/b)^a\right)^{q-1} / (B(p, q))$	$0 < b, p, q$	$\frac{b B(p+q, 1/a)}{B(p, 1/a)}$	$\frac{b^* B(p+q, h/a)}{B(p, h/a)}$	"
$GB2(y; a, b, p, q)$	$\{0 < y\}$	$ a y^{a-1} / \left(B(p, q) \left(1 + (y/b)^a\right)^{q-1} \right)$	$0 < b, p, q$	$\frac{b B\left(\frac{1}{a}, \frac{1}{a}\right)}{B(p, q)}$	$\frac{b^* B(p+h/a, q-h/a)}{B(p, q)} *$	"
$Laplace(y; s)$	$\{-\infty < y < \infty\}$	$\frac{e^{- y/s }}{2s}$	$0 < s$	0	$2s^2$	$\frac{1}{(1-s^2 t^2)}$
	$\{-\infty < y < \infty\}$	$\frac{e^{-(y-\mu)^2/2\sigma^2}}{\sqrt{2\pi}\sigma}$	$0 < \sigma^2$	μ	σ^2	$e^{\mu t + (\sigma^2/2)}$
$GED(y; \mu, s, p)$	$\{-\infty < y < \infty\}$	$\frac{pe^{- y-\mu /s^p}}{2s\Gamma(1/p)}$	$0 < p, s$	μ	$\frac{s^2\Gamma(3/p)}{\Gamma(1/p)}$	$N(y; \mu, \sigma^2)$
$SGT(y; \lambda, s, p, q)$ **	$\{-\infty < y < \infty\}$	$\frac{2}{2\lambda s^{1/p} B\left(\frac{1}{p}, q\right) \left(1 + \frac{ y ^p}{(1+\lambda s g_n(y))^p}\right)^{q+1/p}}$	$0 < s, p, q$ $0 \leq \lambda < 1$	$\frac{s\left(\frac{2}{p} + q - \frac{1}{p}\right)}{2\lambda s^{1/p} B\left(\frac{1}{p}, q\right)}$	ALC	ALC
$EGB2(y; m, s, p, q)$	$\{-\infty < y < \infty\}$	$\frac{e^{p(y-m)/s}}{s B(p, q) \left(1 + e^{(y-m)/s}\right)^{p+q}}$	$0 < s, p, q$	$s^{1/p} [\psi(p) - \psi(q)]$	$s^2 [\psi'(p) + \psi'(q)]$	$\frac{s^{(p+q-q-u)}}{s^u B(p, q)}$

* when an asterisk appears in the variance column, the entry denotes the h^{th} order moment about the origin

$\psi(s) = d \ln \Gamma(s) / ds$ is called the ψ' function. is the trigamma function.

ALC = a little complicated, ** Results for many special cases of the SGT can be obtained by referring to the SGT figure in the class notes. Also, y can be replaced by $y-m$.

b. Models for Real Valued (Positive and Negative) Random Variables:

Two other important families of distributions for random variables which can be positive or negative are the skewed generalized T (SGT) and exponential generalized beta (EGB) distributions.

(1) **The skewed generalized T (SGT) is defined by**

$$(1.5) \text{SGT}(\varepsilon; \mu, \sigma, n, k, \lambda) = \frac{C}{\sigma} \left(1 + \frac{|\varepsilon|^k}{((n-2)/k)(1 + \text{sign}(\varepsilon)\lambda)^k \theta^k \sigma^k} \right)^{-(n+1)/k}$$

for $-\infty < \varepsilon < \infty$ and for $n > 2$ where C and θ are rather complicated expressions involving beta functions and other parameters, see Theodossiou (1998). An alternative less restrictive formulation is given in (2.5)'

$$(1.5)' \quad \text{SGT}(\varepsilon; \lambda, s, p, q) = \frac{p}{2sq^{1/p}B\left(\frac{1}{p}, q\right) \left(1 + \frac{|\varepsilon|^p}{qs^p(1 + \lambda \text{sign}(\varepsilon))^p} \right)^{q+1/p}}$$

The moments for the SGT given in (1.5)' can be expressed as

$$E(\varepsilon_{\text{SGT}}^h) = \left(\frac{s^h}{2} \right) \left\{ \frac{q^{h/p} B\left(\frac{h+1}{p}, q - \frac{h}{p}\right)}{B\left(\frac{1}{p}, q\right)} \right\} \left((1+\lambda)^{h+1} + (-1)^h (1-\lambda)^{h+1} \right).$$

The product of the parameters p and q is the degrees of freedom of the SGT. An important limiting case of the SGT is the SGED (skewed generalized error distribution) which corresponds to the limit of (1.5)' as the parameter q grows indefinitely large. The pdf of the SGED, sometimes referred to as the skewed Box-Taio generalized error distribution, or skewed generalized power distribution is given by

$$\text{SGED}(\varepsilon; s, \lambda, p) = \frac{pe^{-\left(\frac{|s|^p}{(1+\lambda \text{sign}(\varepsilon))^p s^p}\right)}}{2s\Gamma\left(\frac{1}{p}\right)}$$

For large values of the parameter q , the second bracketed expression for the moments of the SGT approaches $(\Gamma((h+1)/p)/\Gamma(1/p))$ to yield expressions for the moments of the SGED..

The relationship between the SGT, GED, skewed t, the t, double exponential or Laplace, normal and Cauchy variates is given in Figure 2. The λ parameter is the skewness parameter where the probability of ε being

positive is given by $\left(\frac{1+\lambda}{2}\right)$ with $\lambda=0$ corresponding to a symmetric generalized t distribution (GT).

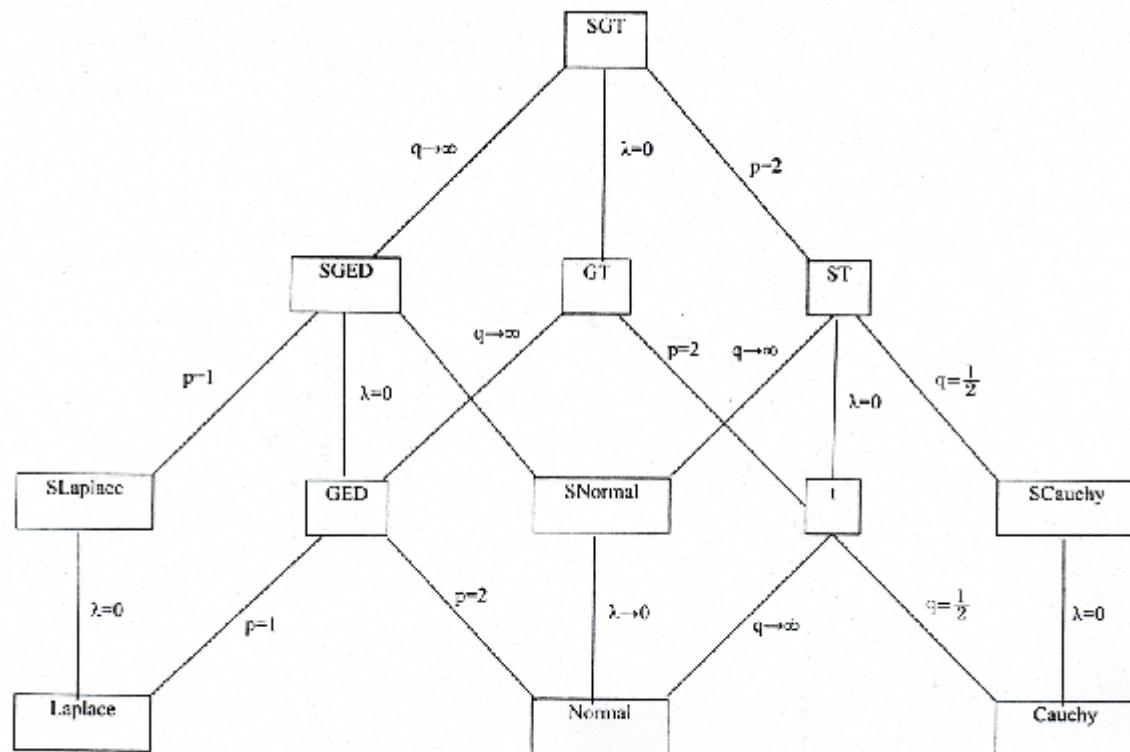


Figure 2

SGT Distribution Family

The GT distribution was introduced in McDonald and Newey [1988]. The GT and its special cases are all symmetric, but can accommodate tails which are thicker or thinner than the normal. They provide the basis for "robust" or partially adaptive estimation of regression and time series models.

(2) **The pdf for the exponential generalized beta** is given by

$$(1.6) \quad EGB(\varepsilon; \delta, \sigma, c, p, q) = \frac{e^{p(\varepsilon-\delta)/\sigma} (1 - (1-c)e^{(\varepsilon-\delta)/\sigma})^{q-1}}{|\sigma| B(p, q) (1 + ce^{(\varepsilon-\delta)/\sigma})^{p+q}}$$

The domain for ε is $-\infty < \frac{\varepsilon - \delta}{\sigma} < \ln\left(\frac{1}{1-c}\right)$ for the EGB.

If $Y \sim GB(y; a, b, c, p, q)$, then the random variable $\varepsilon = \ln(Y)$, $Y = e^\varepsilon$, will be said to be distributed as an exponential generalized beta (EGB). A case could be made for saying that ε is distributed as a log - GB2(LGB2). However, the convention introduced by the relationship between the lognormal and normal distribution will be adopted. Since the EGB and GB distributions are related by the logarithmic transformation, similar "exponential" distributions can be obtained by transforming each of the distributions shown in figure 1. Several of these distributions are of special interest. Consider, for example, the exponential generalized beta of the first and second kind (EGB1 and EGB2) and the exponential generalized gamma (EGG) defined by the probability density functions:

$$(1.7) \quad EGB1(Z; \delta, \sigma, p, q) = EGB(Z; \delta, \sigma, c=0, p, q) \\ = \frac{e^{p(z-\delta)/\sigma} (1 - e^{(z-\delta)/\sigma})^{q-1}}{|\sigma| B(p, q)};$$

$$(1.8) \quad EGB2(Z; \delta, \sigma, p, q) = EGB(Z; \delta, \sigma, c=1, p, q) \\ = \frac{e^{p(z-\delta)/\sigma}}{|\sigma| B(p, q) (1 + e^{(z-\delta)/\sigma})^{p+q}};$$

and

$$(1.9) \quad \text{EGG}(z; \delta, \sigma, p) = \underset{q \rightarrow \infty}{\text{Limit}} \text{ EGB}(Z; \delta * = \sigma \ln q + \delta, c, p, q)$$

$$= \frac{e^{p(z-\delta)/\sigma} e^{-e^{(z-\delta)/\sigma}}}{|\sigma| \Gamma(p)}.$$

The EGB1, EGB2 and EGG are merely alternative representations of the generalized exponential, logistic and Gompertz distributions reviewed in Johnson and Kotz [1970, vol. 2] and Patil et al. [1984]. If (δ, σ, p, q) in the EGB1, EGB2 and EGG is replaced with $(-\sigma \ln p, -\sigma, \delta, \theta)$, the notation found in Johnson and Kotz [p. 271] is obtained. The terminology for these and closely related distribution differs in the literature. For example, Patil et al. [1984] refer to generalized exponential and logistics distributions which are different from those referred to by Johnson and Kotz. The notation used here may help clarify some of the interrelationships. The generalized Gumbell corresponds to the EGB2 with $p = q$. The EBR3 is merely the Burr type 2 distribution. The exponential Weibull (EW) is the extreme value type I distribution. These distributions and interrelationships could be graphically summarized as in figure 1. In fact a figure similar to figure 1 could be constructed for the EGB distribution with each distributional type in figure 1 being preceded with an "E" to denote "exponential," and the parameter a replaced with $1/\sigma$ and b with e^δ . Thus the EGB would appear in place of GB. In the case of the lognormal, $\text{LN}(y; \mu, \sigma^2)$, ELN would correspond to the normal, $N(z; \mu, \sigma^2)$. The structure of this distributional tree would be the same as figure 1, but would include generalized forms of the exponential, logistics, Gompertz, Gumbell and extreme value distributions.

Rather than replicating figure 1 for the "exponential" distributions, we present an abbreviated form which includes some of the most well known members of the EGB family. A comparison of figures 1 and 3 suggests many other distributions which could have been included.

The EGB density with $c = 1$, the EGB2, can be skewed to the right or left or symmetric depending on the relative values of p and q and has tails which are thicker than the normal. This density also includes the normal as a special case as well as many models used in the reliability literature. See McDonald and Xu for additional details.

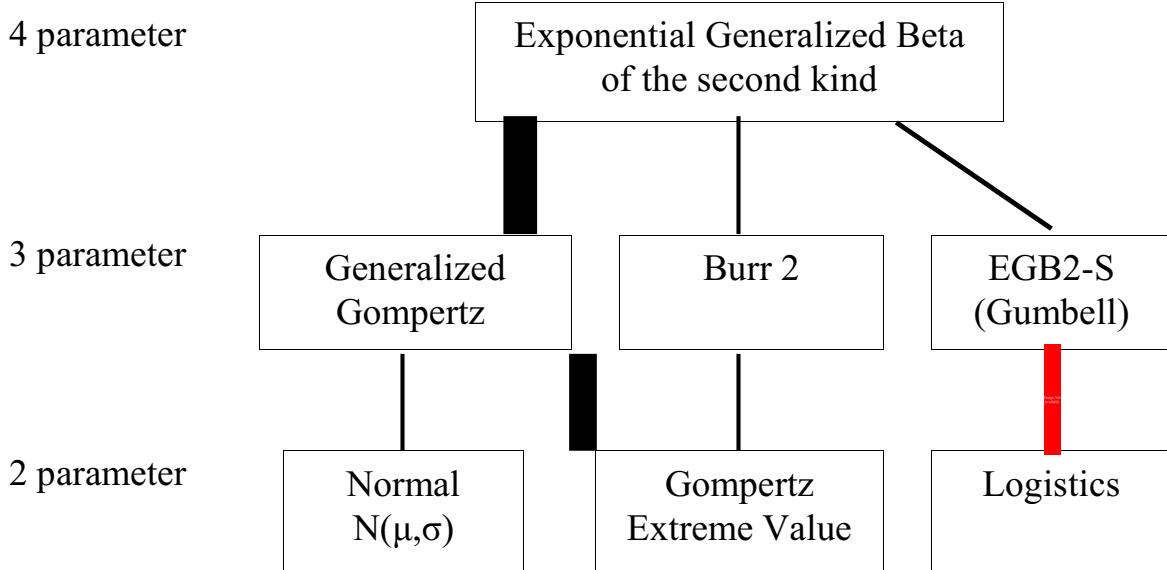


Figure 3
EGB Distribution Tree

$$*BR2(Z; \delta, \sigma, q) = EGB2(Z; \delta, \sigma, p=1, q) = EBR12(Z; \delta, \sigma, q)$$

$$EGB2(Z; \delta, -\sigma, p=1, q) = EBR3(Z; \delta, -\sigma, q)$$

** EGB2-S represents a S(symmetric) EGB2

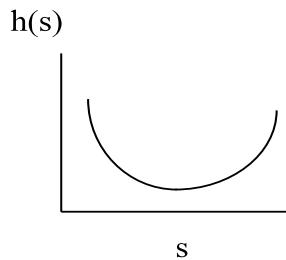
The density functions summarized in Figures 1, 2, and 3 include nearly all of the Pearson family and additional distributions as well. They have the advantage of starting from a few relatively tractable functional forms with the remaining distributions being special cases of the general forms. They are very flexible and can be "thick" or "thin" tailed. In fact, the GB2 and SGT define a finite number of integer moments. The EGB has finite integer moments of all order. $b(\beta)$ or σ are merely scale parameters. a, p, q are shape parameters. δ is a location parameter. The densities in Figure 2 (SGT and special cases) can be skewed or symmetric about the origin. The generalized beta density functions depicted in Figure 1 and the EGB (figure 3) can be skewed to the left or right depending on the relative values of p and q ; whereas, the lognormal is positively skewed for all parameter values.

c. **Hazard Functions**

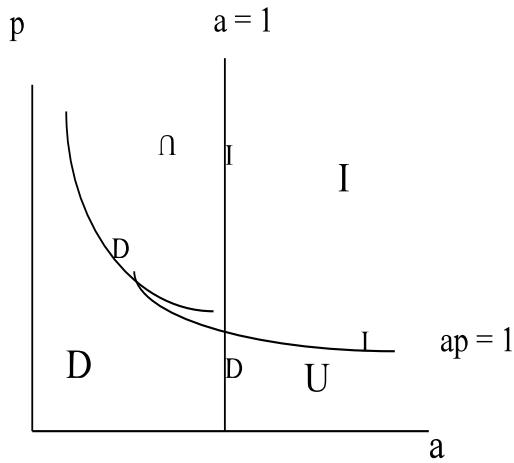
The hazard function is a very useful characterization of density functions and is defined by

$$(1.10) \quad h(s) = \frac{f(s)}{1 - F(s)}$$

and is important in many applications. For example, if s denotes the length of life, then $h(s)$ is the rate of death at age s , given that an individual has lived up to s . The hazard function for mortality data might be expected to appear as follows:



$h(s)$ exhibits decreasing mortality in the first few months of life, then a period of relatively constant mortality and finally an increasing probability of death at older ages. The notion of hazard functions is useful in modeling unemployment duration, length of time of first purchase of a new product or failure time of products. Many models in current use have very limited flexibility for hazard functions. However, many applications may require a "U" or "∩" shaped or strictly increasing (I) or decreasing (D) hazard functions. The generalized gamma is "U" shaped for $a > 1$ and $p < 1/a$; "∩" shaped for $a < 1$ and $p > 1/a$; I shaped for $a > 1$ and $p > 1/a$; and D shaped for $a < 1$ and $p < 1/a$, Glaser [1980]. This is compactly summarized in Figure 4, see McDonald and Richards [1987b]. Exercise: What are possible shapes of the hazard function for the gamma, Weibull, or exponential density? (Hint: refer to figure 1 and figure 4)



Hazard Functions for the Generalized Gamma
Figure 4

d. Moments, Moment Generating Functions, and Cumulant Generating Functions:

A knowledge of $M_X(t)$ can be used to obtain defined moments about the origin as follows (where the i th moment is defined)

$$E(x^i) = \mu'_i = \int x^i f(x) dx$$

$$= \frac{d^i M_x(t)}{dx^i} \Big|_{t=0}.$$

The proof of this results follows from the moment generating function

$$\begin{aligned} M_X(t) &= E(e^{tx}) = \int e^{tx} f(x) dx \\ &= \int \sum_{i=0}^{\infty} \frac{(tx)^i}{i!} f(x) dx \\ &= \sum_{i=0}^{\infty} \frac{t^i}{i!} E(x^i) \end{aligned}$$

which requires the assumption that the integral and summation can be interchanged.

Exercise: (1) Demonstrate that the moment generating function for the normal, $N(\mu, \sigma^2)$, is given by

$$M_X(t) = e^{\mu t + (\sigma^2 t^2 / 2)}$$

(2) Given this result derive the mean and variance for $N(\mu, \sigma^2)$

$$E(X) = (dM_X(t)/dt)|_{t=0}$$

$$\text{Var}(X) = E(X^2) - E^2(X)$$

$$= \frac{d^2 M_X(t)}{dt^2} |_{t=0} - E^2(X)$$

The cumulant generating function

$$\Psi_x(t) = \ln M_x(t)$$

is very useful for obtaining the first four defined moments about the mean. These results are obtained from

$$\frac{d\Psi_x(t)}{dt} |_{t=0} \quad \mu = \text{mean} = E(X) =$$

$$\frac{d^2\Psi_x(t)}{dt^2} |_{t=0} = \text{Var } \sigma^2 = E(X - E(X))^2 =$$

$$\frac{d^3\Psi_x(t)}{dt^3} |_{t=0} = \text{Skewness}(X) = E(X - E(X))^3$$

$$\frac{d^4\Psi_x(t)}{dt^4} |_{t=0} = \text{Excess Kurtosis} = E(X - E(X))^4 - 3(\text{Variance})^2$$

Dividing the expressions for skewness and excess kurtosis by σ^3 or σ^4 , respectively, yields standardized or normalized measures of skewness and excess kurtosis which are independent of the units of measurement.

Exercises:

- (1) Use this result to derive the variance, skewness
kurtosis $\left(\frac{E(X-E(X))^4}{\text{variance}^2} \right)$ and
 $\left(\frac{E(X-E(X))^3}{\text{variance}^{3/2}} \right)$ coefficients for the normal.

- (2) The moment generating function for the EGB2 can be shown to be

$$\frac{e^{\delta t} \Gamma(p+t\sigma) \Gamma(q-t\sigma)}{\Gamma(p) \Gamma(q)}$$

Show that the mean and variance of the EGB2 are given

$$\delta + \sigma[\psi(p) - \psi(q)]$$

$$\sigma^2[\psi'(p) + \psi'(q)]$$

where $\psi(a) = d \ln \Gamma(a)/da$ is the digamma function. Demonstrate that the EGB2 is symmetric if $p = q$.

e. Transformations of Random Variables

The relationships between the random variables considered in figures 1, 2 and 3 can also be considered as arising from transformations of random variables (moving from special cases to more general forms) rather than as special cases of the general forms. The techniques for determining the distribution of transformations are important to understand. The methods of change of variables and those based on moment generating functions are two of the most common.

(1) Change of Variable Approach:

The change of variable approach for determining the distribution can be summarized as follows.

Let $f(x)$ denote the known probability density of the random variable X . Let a new random variable be defined by $Y = g(X)$. Then the density of Y can be written as

$$(1.11) \quad h(y) = f(g^{-1}(y) = x) \left| \frac{dx}{dy} \right|$$

where $\left| \frac{dx}{dy} \right|$ denotes the Jacobian of the transformation $Y = g(X)$. The

result given by (1.11) can be easily extended to the case of multivariate distributions.

Consider the following applications of (1.11).

If $X \sim N(\mu, \sigma^2)$, then $Y = e^X \sim LN(\mu, \sigma^2)$ $Y^2 = (e^X)^2 \sim LN(2\mu, 4\sigma^2)$

If $X \sim N(\mu, \sigma^2)$, then

$$(1.13) \quad Z = \left[\frac{X - \mu}{\sigma} \right]^2 \sim \chi^2(1)$$

The proofs of (1.12) and (1.13) follow directly from (1.11) and have important implications in economics.

A variable Y is lognormally distributed $LN(\mu, \sigma^2)$, if the natural logarithm of that variable is $N(\mu, \sigma^2)$. The lognormal distribution has been used as a model for the size distribution of income. It is also important to know that the square of $N(0,1)$ is $\chi^2(1)$.

Two other probability density functions which have recently been applied in the finance literature are the IHS and the g-h distributions. The IHS (inverse hyperbolic sine distribution) was introduced in the literature by Johnson (1949) and is defined by the transformation:

$$Y_{IHS} = a + b \sinh(\lambda + z/k) = a + b \left(\frac{e^{\lambda+z/k} - e^{-\lambda-z/k}}{2} \right)$$

and the g- and h- random variable $(Y_{g,h})$,defined by John Tukey (1977) as follows:

$$Y_{g,h} = a + b \left(e^{gZ} - 1 \right) \left(\frac{e^{hZ^2/2}}{g} \right)$$

where, in each case, Z denotes a standard normal or z statistic. The IHS and g-and-h distributions are both four parameter distributions. The EGB2 and g- and h-distribution have been used as generalizations of the famous Black Scholes Option Pricing formula. Each of these four parameter distributions allows for skewness and kurtosis common in many data series, particularly in financial data. See the papers by Bookstaber and McDonald (1991), Dutta and Babble (2005), McDonald and Bookstaber (1987), and Mauer and McDonald (2012) for more details.

(2) Moment Generating Function Approach:

A distribution function is uniquely characterized by its moment generating function (where defined). This one-to-one relationship provides the basis for using moment generating functions to determine the distributions of some transformations of random variables. Recall that the moment generating function of the random variable X is defined by

$$(1.14) \quad M_X(t) = E(e^{tx}) = \int_{-\infty}^{\infty} e^{tx} f(x) dx$$

$$= \sum_{i=0}^{\infty} t^i E(x^i) / i! .$$

If X and Y denote two independent random variables, then

$$(1.15) \quad M_{X+Y}(t) = E(e^{t(x+y)}) = E(e^{tx+ty})$$

$$= E(e^{tx} \cdot e^{ty})$$

$$= E(e^{tx})E(e^{ty}) \text{ or}$$

$$\boxed{M_{X+Y}(t) = M_X(t)M_Y(t)}$$

(1.15) demonstrates that the moment generating function of the sum of two independent random variables is the product of the individual moment generating functions. The chi-square density provides an application of the moment generating function approach to determining the density of a transformation of random variables. If $X \sim \chi^2(v)$, then

$$(1.16) \quad M_X(t) = E(e^{tx}) = \int_0^{\infty} e^{tx} \left\{ \frac{e^{-\frac{x}{2}} x^{\frac{v}{2}-1}}{2^{v/2} \Gamma(v/2)} \right\} dx$$

$$= \int_0^{\infty} \frac{e^{-\frac{x}{2}(1-2t)} x^{\frac{v}{2}-1}}{2^{v/2} \Gamma(v/2)}$$

$$= \frac{1}{(1-2t)^{v/2}} \int_0^{\infty} \frac{e^{\frac{-x(1-2t)}{2}} (x(1-2t))^{\frac{v}{2}-1}}{2^{v/2} \Gamma(v/2)} (1-2t) dx$$

$$= \frac{1}{(1-2t)^{v/2}} \int_0^{\infty} \frac{e^{-s/2} s^{v/2-1}}{2^{v/2} \Gamma(v/2)} ds$$

$$= \frac{1}{(1-2t)^{v/2}}$$

which is the moment generating function for a $\chi^2(v)$.

(1.15) and (1.16) can be used to show that the sum of two independent chi-square variables is distributed as a chi-square with degrees of freedom equal to the sum of the degrees of freedom of the original independent chi-square variables. Let $Z = X + Y$ where X and Y are independently distributed as $\chi^2(v_x)$ and $\chi^2(v_y)$ respectively. The moment generating function of Z is given by

$$M_Z(t) = E(e^{t(x+y)}) = M_X(t)M_Y(t)$$

$$\begin{aligned} &= \left\{ \frac{1}{(1-2t)^{v_x/2}} \right\} \left\{ \frac{1}{(1-2t)^{v_y/2}} \right\} \\ &= \frac{1}{(1-2t)^{(v_x+v_y)/2}}; \end{aligned}$$

which is the moment generating function of a chi-square with degrees of freedom v_x+v_y . Hence, the sum of two independently distributed chi-square variables is distributed as a chi-square with degrees of freedom equal to the sum of the degrees of freedom of the individual random variables. These are important results which will be used in the next section and link univariate results to some important relationships involving multivariate distributions.

2. Multivariate Distributions

The distribution of a vector of random variables $\mathbf{Y} = (y_1, y_2, \dots, y_n)'$ is referred to as a multivariate distribution. Many multivariate distributions have been considered and include the multivariate gamma, beta, Wishart, t, generalized t, and normal distributions. The multivariate normal is probably the most widely used and has some very useful properties. We first review the notion of the expected value and variance of a vector of random variables and then summarize some important properties of the multivariate normal. We conclude by defining some other multivariate distributions.

a. **Expectations.** Let \mathbf{Y} denote an $n \times l$ vector of random variables and \mathbf{Z} be an $n \times m$ matrix of random variables

The expected value of \mathbf{Y} , mean vector, is defined by

$$\mathbb{E}(\mathbf{Y}) = (\mathbb{E}(y_1), \dots, \mathbb{E}(y_n))' = \boldsymbol{\mu}$$

The expected value of a matrix \mathbf{Z} is defined by

$$\begin{pmatrix} \mathbb{E}(z_{11}) & \dots & \mathbb{E}(z_{1m}) \\ \vdots & & \vdots \\ \mathbb{E}(z_{n1}) & \dots & \mathbb{E}(z_{nm}) \end{pmatrix}$$

The variance covariance matrix associated with \mathbf{Y} is defined by

$$\Sigma = \mathbb{E}(\mathbf{Y} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})'$$

$$= \begin{pmatrix} \mathbb{E}(y_1 - \mu_1)^2 & \dots & \mathbb{E}(y_1 - \mu_1)(y_n - \mu_n) \\ \mathbb{E}(y_2 - \mu_2)(y_1 - \mu_1) & \dots & \mathbb{E}(y_2 - \mu_2)(y_n - \mu_n) \\ \vdots & & \vdots \\ \mathbb{E}(y_n - \mu_n)(y_1 - \mu_1) & \dots & \mathbb{E}(y_n - \mu_n)^2 \end{pmatrix}$$

$$= \begin{pmatrix} \text{Var}(y_1) & \text{Cov}(y_1, y_2) & \dots & \text{Cov}(y_1, y_n) \\ \text{Cov}(y_2, y_1) & \text{Var}(y_2) & \dots & \text{Cov}(y_2, y_n) \\ \vdots & \vdots & & \vdots \\ \text{Cov}(y_n, y_1) & \text{Cov}(y_n, y_2) & \dots & \text{Var}(y_n) \end{pmatrix}$$

$$= \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & & \sigma_{2n} \\ \vdots & \vdots & & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_{nn} \end{pmatrix}$$

$$= \text{Var}(Y)$$

Note that the variance-covariance matrix includes the variances of the individual variables on the main diagonal and covariance between individual random variables in the corresponding off-diagonal elements.

b. Multivariate Normal: The vector Y_{nxl} is said to be distributed as the multivariate normal with mean vector μ and variance-covariance matrix Σ , denoted $Y \sim N(\mu, \Sigma)$, if the density of Y is given by

$$N(Y; \mu, \Sigma) = \frac{e^{-(Y-\mu)' \Sigma^{-1} (Y-\mu)/2}}{(2\pi)^{n/2} |\Sigma|^{1/2}}.$$

The multivariate normal includes the univariate ($n=1$) and bivariate ($n=2$) normal as special cases. For example, if $n = 1$,

$$Y = (y_1), \mu = (\mu), \Sigma = (\sigma^2) \text{ and}$$

$$\begin{aligned} N(y_1; \mu, \sigma^2) &= \frac{e^{-1/2(y_1-\mu)(1/\sigma^2)(y_1-\mu)}}{\sqrt{2\pi} \sqrt{\sigma^2}}. \\ &= \frac{e^{-(y_1-\mu)^2/2\sigma^2}}{\sqrt{2\pi} \sqrt{\sigma^2}} \end{aligned}$$

The reader may want to consider the case $n = 2$.

c. Distribution Theory for the normal (See Appendix A for proofs of the results contained in this section)

(1) Chi Square

(a) Let y_1, \dots, y_n be independently distributed as $N(0, \sigma^2)$, i.e.,

$$Y = (y_1, \dots, y_n)' \sim N(0, \sigma^2 I), \text{ then}$$

$$Y(\sigma^2 I)^{-1} Y = \frac{Y' Y}{\sigma^2} = \sum_{i=1}^n \frac{y_i^2}{\sigma^2} = \sum_{i=1}^n \left(\frac{y_i - 0}{\sigma} \right)^2 \sim \chi^2(n).$$

The sum of squares of n independent standard normal variables is distributed as a chi-squared variable with n degrees of freedom.

(b) If $Y \sim N(\mu, \sigma^2 I)$, then

$$(Y - \mu)' (\sigma^2 I)^{-1} (Y - \mu) = \frac{(Y - \mu)' (Y - \mu)}{\sigma^2} \sim \chi^2(n).$$

Note this is the quadratic form which appears in the density function for Y .

(c) **More generally, If $\mathbf{Y} \sim \mathbf{N}(\boldsymbol{\mu}, \Sigma)$, then**

$$(\mathbf{Y} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{Y} - \boldsymbol{\mu}) \sim \chi^2(n).$$

Note: This is the quadratic form which appears in the exponent of the multivariate normal. Hint: take the "square root" of the matrix Σ , Cholesky decomposition.

(2) Some Useful Theorems

(a) **If $\mathbf{Y} \sim \mathbf{N}(\boldsymbol{\mu}_y, \Sigma_y)$, then $\mathbf{Z} = \mathbf{AY} \sim \mathbf{N}(\boldsymbol{\mu}_z = \mathbf{A}\boldsymbol{\mu}_y; \Sigma_z = \mathbf{A}\Sigma_y\mathbf{A}')$**

where \mathbf{A} is a matrix of constants.

(b) **If $\mathbf{Y} \sim \mathbf{N}(\mathbf{O}, \mathbf{I})$ and \mathbf{A} is a symmetric idempotent matrix, then $\mathbf{Y}'\mathbf{AY} \sim \chi^2(m)$, where the degrees of freedom (m) = trace(\mathbf{A}).**

Proof: Diagonalize \mathbf{A} . Recall that the characteristic roots of an idempotent matrix are zero or one.

(c) **If $\mathbf{Y} \sim \mathbf{N}(\mathbf{O}, \mathbf{I})$, then the idempotent quadratic forms $\mathbf{Y}'\mathbf{AY}$ and $\mathbf{Y}'\mathbf{BY}$ are independently distributed χ^2 variables if $\mathbf{AB} = \mathbf{0}$.**

Proof: $\mathbf{Y}'\mathbf{AY} = \mathbf{Y}'\mathbf{A}'\mathbf{AY}$ and $\mathbf{Y}'\mathbf{BY} = \mathbf{Y}'\mathbf{B}'\mathbf{BY}$. The covariance matrix of \mathbf{AY} and \mathbf{BY} is \mathbf{AB} .

(d) **If $\mathbf{Y} \sim \mathbf{N}(\mathbf{O}, \mathbf{I})$ and \mathbf{L} is a $k \times n$ matrix of rank k , then \mathbf{LY} and the idempotent quadratic form $\mathbf{Y}'\mathbf{AY}$ are independently distributed if $\mathbf{LA} = \mathbf{0}$.**

Proof: Similar to (c). $\mathbf{Y}'\mathbf{AY} = \mathbf{Y}'\mathbf{A}'\mathbf{AY}$. The covariance matrix of \mathbf{LY} and \mathbf{AY} is \mathbf{LA} .

See Appendix A for more details.

(3) A simple example (see exercise 9, 10)

If $\mathbf{Y} \sim \mathbf{N}(\boldsymbol{\mu}, \sigma^2 \mathbf{I})$, then

$$(a) \frac{(\mathbf{Y} - \boldsymbol{\mu})' (\mathbf{Y} - \boldsymbol{\mu})}{\sigma^2} \sim \chi^2(n);$$

$$(b) \frac{(n-1)s^2}{\sigma^2} = \sum \frac{(y_t - \bar{Y})^2}{\sigma^2} = \mathbf{Y}' (\mathbf{I} - (1/n)\mathbf{i}\mathbf{i}') (\mathbf{I} - (1/n)\mathbf{i}\mathbf{i}') \mathbf{Y} / \sigma^2$$

$$= \mathbf{Y}' (\mathbf{I} - 1/n\mathbf{i}\mathbf{i}') \mathbf{Y} / \sigma^2 \sim \chi^2(n-1)$$

where \mathbf{i} denotes a $n \times 1$ column of 1's, thus $\mathbf{i}\mathbf{i}'$ is an $n \times n$ matrix of 1's. It is also important to show that $(\mathbf{I} - (1/n)\mathbf{i}\mathbf{i}')$ is symmetric and idempotent and has trace = $n-1$

Recall $\cdot E(\chi^2(d)) = d$ (degrees of freedom)

s^2 is an unbiased estimator of σ^2

$$\cdot E(n-1) \frac{s^2}{\sigma^2} = (n-1) \frac{E(s^2)}{\sigma^2} = n-1$$

(4) t distribution

The ratio of a standard normal variable divided by the square root of an independent chi-square variable divided by its degrees of freedom is distributed as a t statistic with the same degrees of freedom as the chi-square. This might be represented:

$$\frac{N(0, 1)}{\sqrt{\chi^2(d)/d}} \sim t(d)$$

The t-density is symmetric about zero, has variance $[d/(d-2)]$ and normalized kurtos $3 + 6/(d-4)$; hence, the tails of the t are thicker than the normal, but approach the normal $N(0,1)$ as $d \rightarrow \infty$.

(5) F distribution

The ratio of two independent chi-square variables divided by their respective degrees of freedom is distributed as an F statistic with degrees of freedom equal to those associated with the chi-square variables. This can be represented as

$$\frac{\chi^2(d_1)/d_1}{\chi^2(d_2)/d_2} \sim F(d_1, d_2)$$

Exercise: Show that the mean of the F distribution is given by $d_2/(d_2 - 2)$. Can you provide any

intuition why this is close to one for large d_2 ? Hint: $F = \left(\frac{d_2}{d_1} \right) (\chi^2(d_1)) (\chi^2(d_2))^{-1}$.

order moment of a chi-square variable is $E(\chi^2(d))^h = 2^h \Gamma(d/2+h)/\Gamma(d/2)$. Let $h=1$ for the first chi-square and $h=-1$ for the second. The same "trick" can be used to evaluate the moments of the $t(d)$.

For example,

$$\begin{aligned} \text{Var}(t(d)) &= E(t(d))^2 = E[(N(0,1))^2 (\chi^2(d))^{-1}]d \\ &= (1)(d)E(\chi^2(d))^{-1} = d(2)^{-1} \Gamma(d/2 - 1)/\Gamma(d/2) = d/(d-2). \end{aligned}$$

d. Multivariate generalized error and multivariate generalized t distributions

(1) **Multivariate generalized error distribution (MGED), multivariate power exponential (MPEXP), or multivariate Box-Tiao distribution (MBT)** of an $n \times 1$ vector Y is defined by the pdf

$$MGED(y, \mu, \Sigma, p) = \frac{n\Gamma(n/2)e^{-((y-\mu)' \Sigma^{-1} (y-\mu))^p/2}}{|\Sigma|^{1/2} \pi^{n/2} \Gamma\left(1 + \frac{n}{2p}\right) 2^{1+n/2p}}$$

If $n=1$, the univariate GED results. The multivariate normal corresponds to $p=2$. For additional details, see [Gomez, E., M.A. Gomez-Villegas, and J.M. Marin, 1998. "A Multivariate generalization of the power exponential family of distributions," Communications in Statistics: Theory and Methods, 27(3), 589-600.

(2) **Multivariate generalized t (MGT)**. Arslan (2001) extended the method used by McDonald and Newey in obtaining the generalized t to derive a multivariate generalized t. The pdf of the MGT is given by

$$MGT(y; \mu, \Sigma, \sigma, p, q) = \frac{p\Gamma\left(\frac{n}{2}\right)q^q}{\pi^{n/2} B(q, n/2p) \left(q + ((y - \mu)' \Sigma^{-1} (y - \mu) / \sigma)^p\right)^{q+n/2p}}$$

The MGED and MGT are members of elliptically contoured random variables, see (Fang, K. and Y. Zhang, 1990. Generalized Multivariate Analysis, Springer Verlag. Beijing) for additional details on elliptical distributions. When $\sigma = p = 1, \Sigma = 2\Sigma_1$, and $2q = v$

standard multivariate t-distribution with location and scatter parameters μ and Σ_1 and degrees of freedom v . The multivariate normal is obtained from the MGT by letting $p = \sigma = 1$ and $q \rightarrow \infty$. The MGED is obtained from the MGT by letting q grow indefinitely large. Finally, as

$\sigma \rightarrow \infty$ the MGT approaches a multivariate generalization of the uniform distribution.

Arskan(2001) reports expressions for the first four moments of MGT. We report the first two moments in these notes:

- $E(Y) = \mu$

- $Var(Y) = \left[\frac{\sigma q^{1/p} \Gamma\left(\frac{n+2}{2p}\right) \Gamma(q-1/p)}{n \Gamma\left(\frac{n}{2p}\right) \Gamma(q)} \right] \Sigma$

Reference: Arslan, Olcay, 2001. "Family of Multivariate Generalized t-Distributions," working paper, University of Cukurova, Turkey. Submitted to Journal of Multivariate Analysis.

(3) Multivariate generalized beta distributions (MGB1 and MGB2). There are often several ways multivariate generalizations of univariate distributions can be formed which yield the same marginal distributions. An extension of the GB1 and GB2 are the MGB1 and MGB2 given as follows

$$MGB1(y; a, b, p, q) = \frac{\left(\prod_{i=1}^n |a_i| y_i^{a_i p_i - 1} \right) \left(1 - \sum_{i=1}^n (y_i / b_i)^{a_i} \right)^{q-1}}{\left(\prod_{i=1}^n b_i^{a_i p_i} \right) B(p_1, p_2, \dots, p_n, q)}$$

for $0 < \sum_{i=1}^n (y_i / b_i)^{a_i} < 1$

$$MGB2(y; a, b, p, q) = \frac{\left(\prod_{i=1}^n |a_i| y_i^{a_i p_i - 1} \right)}{\left(\prod_{i=1}^n b_i^{a_i p_i} \right) B(p_1, p_2, \dots, p_n, q) \left(1 + \sum_{i=1}^n (y_i / b_i)^{a_i} \right)^{\bar{p}+q}}$$

for $0 < y_i$

where $B(p_1, p_2, \dots, p_n, q) = \frac{\Gamma(p_1)\Gamma(p_2)\dots\Gamma(p_n)\Gamma(q)}{\Gamma(\bar{p}+q)}$ with $\bar{p} = \sum_{i=1}^n p_i$.

Multivariate GG, EGB1, and EGB2 distributions can be defined in a similar manner. The pdf for the MEGB2 is given by

$$MEGB2(z; \delta, \sigma, p, q) = \frac{\left(\prod_{i=1}^n e^{p_i(z_i - \delta_i)/\sigma_i} \right)}{\left(\prod_{i=1}^n |\sigma_i| \right) B(p_1, p_2, \dots, p_n, q) \left(1 + \sum_{i=1}^n e^{(z_i - \delta_i)/\sigma_i} \right)^{p+q}}$$

for $-\infty < y_i < \infty$

Expressions for the moments are reported in McDonald (Some Multivariate Generalized Beta Distributions, Working Paper, February 1993). Other forms of multivariate beta distributions are summarized in several books by Johnson and Kotz dealing with continuous multivariate distributions.

3. Estimation

a. Maximum likelihood estimation: basic theory and results

Assume that the random variables $y_t, t = 1, 2, \dots, n$ are independently and identically distributed with the probability density function $f(y_t; \theta)$ where $\theta' = (\theta_1, \theta_2, \dots, \theta_k)$ denotes a vector of unknown parameters. We will review two important methods of estimating the unknown parameters. In comparing alternative estimators of the same parameter we will want to compare their statistical properties. **Three important criteria** to consider in

such a comparison are **bias**, **consistency**, and **relative efficiency**. Recall that $\hat{\theta}$ is an *unbiased* estimator if

$$E(\hat{\theta}) = \theta$$

Consistency requires that $p\lim_{n \rightarrow \infty} \hat{\theta} = \theta$ or $\lim_{n \rightarrow \infty} \Pr(|\hat{\theta} - \theta| \geq \varepsilon) = 0$ for any $\varepsilon > 0$

If the variance and bias of $\hat{\theta}$ approach zero, $\hat{\theta}$ will be consistent; however, there are cases in which the variance/bias may not be defined and the estimator is still consistent. An unbiased estimator is said to be efficient if its variance is less than the variance of any other unbiased estimator. In the case of a vector of estimators, one would investigate whether the difference of the variance covariance matrices are positive definite. In the case of biased estimators, comparisons can be made using the mean squared error which is equal to sum of the variance and square of the bias, i.e. $MSE(\hat{\theta}) = VAR(\hat{\theta}) + BIAS^2(\hat{\theta})$

We now turn to the development of maximum likelihood estimation and then we briefly summarize the notion of extremum estimators.

The likelihood function associated with the random sample $\{y_t, t = 1, 2, \dots, n\}$ is given by

$$(3.1) \quad L(Y; \theta) = \prod_{t=1}^n f(y_t; \theta)$$

with corresponding log likelihood function

$$(3.2) \quad \ell(Y; \theta) = \ln(L(Y; \theta))$$

$$= \sum_{t=1}^n \ln f(y_t; \theta)$$

$$= \sum_{t=1}^n \ell(y_t; \theta),$$

where $\ell(y_t; \theta) = \ln f(y_t; \theta)$.

The maximum likelihood estimators (MLE) of the parameters θ are implicitly defined by the equation

$$(3.3) \quad \frac{d\ell(Y; \theta)}{d\theta} = \begin{pmatrix} \frac{\partial \ell}{\partial \theta_1} \\ \vdots \\ \frac{\partial \ell}{\partial \theta_k} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} = 0.$$

These estimators will optimize the log likelihood function $\ell(y; \theta)$. It is possible that multiple solutions to (3.3) will exist; however, a unique maximum will exist if $\ell(Y; \theta)$ is concave in θ . The concavity can be investigated by determining whether the matrix

$$\frac{d^2\ell(Y; \theta)}{d\theta^2} = \begin{pmatrix} \ell_{11} & \cdots & \ell_{1k} \\ \vdots & & \vdots \\ \ell_{k1} & \cdots & \ell_{kk} \end{pmatrix}$$

is negative definite where $\ell_{ij} = \frac{\partial^2 \ell(Y; \theta)}{\partial \theta_i \partial \theta_j}$.

Under rather general regularity conditions the MLE of θ , $\hat{\theta}$, will be asymptotically normally distributed.

The regularity conditions will be given and the main results cited. The literature contains a number of alternative formulations of "regularity conditions."

Regularity Conditions:

Let Θ denote the set of permissible values of θ

(R.1) For almost all y_t

$$\frac{d\ell(y_t; \theta)}{d\theta}, \frac{d^2\ell(y_t; \theta)}{d\theta^2}, \frac{d^3\ell(y_t; \theta)}{d\theta^3}$$

exist for all $\theta \in \Theta$

(R.2) There exist integrable functions $F_1(y)$, $F_2(y)$ and $H(y)$ such that

$$\frac{d\ell(y_t; \theta)}{d\theta} < F_1(y_t)$$

$$\frac{d^2\ell(y_t; \theta)}{d\theta^2} < F_2(y_t)$$

$$\frac{d^3\ell(y_t; \theta)}{d\theta^3} < H(y_t)$$

for all $\theta \in \Theta$

and

$$E(H(y)) = \int_{-\infty}^{\infty} H(y) f(y; \theta) dy < M$$

where M is independent of θ .

$$(R.3) E\left[\frac{\partial\ell(y_t; \theta)}{\partial\theta}\right]^2 = \int_{-\infty}^{\infty} \left[\frac{\partial \ln f(y; \theta)}{\partial\theta}\right]^2 f(y; \theta) dy < +\infty.$$

These assumptions are important in demonstrating the asymptotic normality of MLE of θ . R.1 insures the existence of necessary Taylor Series expansions while R.2 permits differentiation under an integral and (R.3)

implies that the random variable(s) $\left\{ \frac{\partial \ln f(y_t; \theta)}{\partial \theta} \right\}$ will have a finite variance.

Theorem. Under the regularity conditions (R.1), (R.2) and (R.3) the MLE of θ will be consistent and asymptotically normal.

$$(3.4) \quad \theta_{MLE} \xrightarrow{a} N[\theta, \Sigma_{\theta_{MLE}}]$$

where

$$\begin{aligned} \Sigma_{\theta_{MLE}} &= -\left(\frac{E d^2 \ell}{d\theta d\theta'} \right)^{-1} \\ &= \left(E \left(\frac{d\ell}{d\theta} \right) \left(\frac{d\ell}{d\theta} \right) \right)^{-1}. \end{aligned}$$

This distributional result can be alternatively written

$$\sqrt{n}(\theta_{MLE} - \theta) \xrightarrow{a} N[0, n\Sigma_{\theta_{MLE}}].$$

Appendix B contains a proof of this important result:

The proof follows from the basic condition for probability density functions $\int L(Y; \theta) dY = 1$ and leads to useful intermediate results that

$$E\left(\frac{d\ell(Y; \theta)}{d\theta} \right) = 0$$

$$\text{Var}\left(\frac{d\ell(Y; \theta)}{d\theta} \right) = E\left(\frac{d\ell(Y; \theta)}{d\theta} \frac{d\ell(Y; \theta)}{d\theta'} \right) = \Sigma_{\theta_{MLE}}^{-1}$$

Two excellent references to this material are Rao [1973] and Theil [1971].

In another section , we will explore how computer programs such as STATA can be used to obtain MLE.

Examples of MLE estimators and their asymptotic distributions

We now consider two examples. The first example is based upon the **power density** function

$$f(x) = px^{p-1} \quad \text{for } 0 < x < 1$$

$$= 0 \quad \text{otherwise.}$$

The log likelihood function is

$$\begin{aligned} \ell &= \ln \left\{ \prod_{t=1}^n f(x_t) \right\} = \ln \left\{ p^n \left(\prod_{t=1}^n x_t \right)^{p-1} \right\} \\ &= n \ln p + (p-1) \sum_{t=1}^n \ln x_t. \end{aligned}$$

The necessary condition for a maximum of ℓ is

$$\frac{d\ell}{dp} = \frac{n}{p} + \sum_{t=1}^n \ln x_t = 0$$

which yields

$$\hat{p} = -[n / (\sum \ln x_t)].$$

The asymptotic distribution of \hat{p} depends on

$$-E \frac{d^2 \ell}{dp^2} = -E \left\{ \frac{-n}{p^2} \right\} = \frac{n}{p^2},$$

hence

$$\hat{p} \xrightarrow{a} N[p, p^2/n] \quad \text{or}$$

$$\sqrt{n}(\hat{p} - p) \xrightarrow{a} N[0, p^2].$$

(2) **The normal.** The second example involves the derivation of MLE of the two parameters

$$\text{in } N(X; \mu, \sigma^2) = \frac{e^{-(x-\mu)^2/2\sigma^2}}{\sqrt{2\pi}\sigma}.$$

The loglikelihood function is given by

$$\ell = \frac{-n}{2} \ln(2\pi) - \frac{n}{2} (\ln \sigma^2) - \frac{1}{2} \sum_t \left[\frac{(X_t - \mu)^2}{\sigma^2} \right].$$

Differentiating ℓ with respect to the parameters to be estimated yields:

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2} \sum_t (X_t - \mu) = 0$$

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_t (X_t - \mu)^2 = 0.$$

Solving these equations yields

$$\begin{aligned}\hat{\mu} &= \bar{X} && \text{and} \\ \hat{\sigma}^2 &= \sum_{t=1}^n (X_t - \bar{X})^2/n\end{aligned}$$

The asymptotic-variance covariance matrix of $\hat{\mu}$ and $\hat{\sigma}^2$ can be obtained from the Hessian matrix associated with ℓ .

$$H = \begin{pmatrix} \frac{d^2\ell}{d\mu^2} & \frac{d^2\ell}{d\sigma^2 d\mu} \\ \frac{d^2\ell}{d\mu d\sigma} & \frac{d^2\ell}{d\sigma^2} \end{pmatrix} = \begin{pmatrix} -n/\sigma^2 & \frac{-\sum_{t=1}^n (X_t - \mu)}{\sigma^4} \\ \frac{-\sum_{t=1}^n (X_t - \mu)}{\sigma^4} & \frac{n}{2\sigma^4} - \frac{2\sum_{t=1}^n (X_t - \mu)^2}{2\sigma^6} \end{pmatrix};$$

hence

$$E(H) = \begin{pmatrix} -n/\sigma^2 & 0 \\ 0 & -n/2\sigma^4 \end{pmatrix}.$$

The asymptotic variance matrix of $(\hat{\mu}, \hat{\sigma}^2)$ is

$$(-E(H))^{-1} = \begin{pmatrix} \sigma^2/n & 0 \\ 0 & \frac{2\sigma^4}{n} \end{pmatrix};$$

hence,

$$\sqrt{n} \begin{pmatrix} \hat{\mu} - \mu \\ \hat{\sigma}^2 - \sigma^2 \end{pmatrix} \stackrel{a}{\approx} N \begin{pmatrix} 0 \\ 0 \end{pmatrix}; \begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix}$$

or

$$\begin{pmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{pmatrix} \stackrel{a}{\approx} N \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}; \begin{pmatrix} \sigma^2/n & 0 \\ 0 & 2\sigma^4/n \end{pmatrix}.$$

b. Method of moments (MOM)

Let the data generating process (DGP) be defined by the pdf $f(y, \theta_i, i = 1, 2, \dots, p)$. Method of moments estimators of the unknown parameters are obtained by selecting parameter estimators which equate sample moments and theoretical moments. We generally use the same number of moments as parameters. If more moments are used than parameters, generalized methods of moments (GMM) can be used.

To formalize the method, let

$$\hat{m}_h = \frac{1}{n} \sum_{i=1}^n Y_i^h \text{ and } m_h(\theta) = E(Y^h)$$

The MOM estimators are then obtained by solving:

$$\sum Y_i^h / n = E(Y^h) = m_h(\hat{\theta}) \quad , h = 1, 2, \dots, p, \text{ for the vector of parameter estimates } \hat{\theta} .$$

Examples:

Exponential: (one parameter)

$$\text{Sample mean } (\bar{Y}) = E(Y) = \beta \text{ and solve}$$

Power distribution:

The method of moments estimator of p (\tilde{p}_{mom}) in the power distribution is found from

$$E(x) = \int_0^1 x \{px^{p-1}\} dx = \frac{\tilde{p}}{\tilde{p}+1} = \bar{x};$$

hence

$$\tilde{p}_{\text{mom}} = \frac{\bar{x}}{1-\bar{x}}.$$

The expected value and variance of X are given by $\left(\frac{p}{p+1} \right)$ and

$p/(p+1)^2(p+2)n$. Using the Slutsky Theorem: \tilde{p}_{mom} can be shown to be consistent

$$\text{plim } (\tilde{p}_{\text{mom}}) = \frac{\text{plim}(\bar{X})}{1-\text{plim}(\bar{X})}$$

$$= \frac{\frac{p}{p+1}}{1 - (p/p+1)} = p.$$

We will consider ways of determining its asymptotic distribution in another section.

Normal: two parameters

Set the sample mean = μ and the sample $E(Y^2) = \sigma^2 + \mu^2$ or Sample Var = σ^2

and solving for $\hat{\mu}$ and $\hat{\sigma}^2$

Gamma: two parameters

$$E(Y) = \beta \Gamma(p+1)/\Gamma(p) = \beta p$$

$$E(Y^2) = \beta^2 \Gamma(p+2)/\Gamma(p) = \beta^2(p+1)p$$

In summary, method of moments estimators are obtained by solving solve

$$\hat{m}_h = m_h(\hat{\theta}) \text{ for } h=1, 2, \dots, p$$

for $\hat{\theta}$ where there are p parameters. Method of Moment estimators are generally:

- Consistent
- Not always defined (equations may not have a solution)
- May involve nonlinear equations

Alternatively, let

$$g(\theta) = (m(\theta) - \hat{m})$$

where $\hat{m} = (\hat{m}_1, \hat{m}_2, \dots, \hat{m}_p)'$ & $m(\theta) = (m_1(\theta), m_2(\theta), \dots, m_p(\theta))'$

and, as before, $\hat{m}_h = \frac{1}{n} \sum_{i=1}^n Y_i^h$ and $m_h(\theta) = E(Y^h)$

can be obtained by solving any of the following problems:

$$(1) \quad g(\hat{\theta}) = 0$$

$$(2) \quad \min_{\theta} (g(\theta))^T g(\theta)$$

$$(3) \quad \min_{\theta} (g(\theta))^T W g(\theta)$$

where W is a positive definite weighting matrix. If there is a method of moments estimator (with the same number of moment restrictions as parameters) it is given as the solution to (1) and will be

the same as the solution to problems (2) or (3) which will yield a value of zero for the corresponding quadratic forms. If there are more moment conditions (m) than parameters (p), there will generally not be a solution to (1) and the solution to (2) and (3) will be referred to as the **generalized method of moments** (GMM) estimator.

For example, the mean and variance of the exponential pdf can be shown to be

$$\beta \text{ and } \beta^2 \quad g(\beta) = \begin{pmatrix} \bar{Y} - \beta \\ s^2 - \beta^2 \end{pmatrix}$$

and there is not a method of

unless the sample variance is equal to the square of the sample mean. The GMM estimator would be obtained by solving equation (2) or (3) above for β which seeks to make the entries in $g(\beta)$ as small as possible, with small being measured by the value of the quadratic form.

c. Kernel Estimators

Kernel Estimators are referred to as non-parametric because no "distributional parameters" need be estimated. The kernel estimator might be thought of as "high-tech" histogram. However; the user still needs to make a decision about the *kernel* to use and about a *window width* to use. We will merely present some basic building blocks here and in class will show an example of the use of a kernel to describe stock returns.

Denote n observations on the random variable Z as $Z_i, i = 1, 2, \dots, n$. defined as :

The empirical cdf can then be

$$\hat{F}(z) = \Pr(Z \leq z) \left(\frac{\# \text{observations} \leq z}{\text{Total number of observations}} \right)$$

The corresponding empirical pdf (histogram) with "window width" or "band width" h is defined by

$$\hat{f}_h(z) = \frac{\hat{F}(z+h) - \hat{F}(z-h)}{2h}$$

- the empirical pdf gives the fraction of observations which lie within a neighborhood of width h of z . h is referred to as the bandwidth.
- Increasing h makes the pdf smoother and decreasing h makes the pdf choppier

An alternative way of expressing the histogram is as follows:

$$\hat{f}_h(z) = \left(\frac{1}{nh}\right) \sum_{i=1}^n I\left(-\frac{1}{2} \leq \frac{z - z_i}{h} \leq \frac{1}{2}\right)$$

where $I()$ denotes an indicator function which equals 1 if the inequality is valid and zero otherwise, (Pagan and Ullah).

If the indicator function is replaced by a pdf, say $K()$, the corresponding function is referred to as a kernel estimator

$$\hat{f}_h(z) = \left(\frac{1}{nh}\right) \sum_{i=1}^n K\left(\frac{z - z_i}{h}\right)$$

with bandwidth h and kernel $K()$.

An alternative development of the kernel density estimator can be structured in terms of a new variable W defined by $W = Z + hU$ where

- U is continuous with pdf $K(u)$
- h is the bandwidth
 - Small h , W behaves like the data
 - Large h , W behaves like U

The CDF of W (you can skip this and go straight to the kernel pdf) is given by

$$\begin{aligned} F_W(w) &= Pr(W \leq w) \\ &= Pr(Z + hU \leq w) = Pr\left(U \leq \frac{w - Z}{h}\right) \\ &= \frac{1}{n} \sum_{i=1}^n \int_{-\infty}^{\frac{w - z_i}{h}} K(u) du \end{aligned}$$

How would we estimate the pdf of Z ? (Use Leibnitz Rule).

The kernel estimator of the pdf is given by the following equation:

$$\hat{f}_h(z) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{z - z_i}{h}\right)$$

- $K(u)$ is the Kernel (often the standard normal, but there are other choices)
- It is easily verified by integrating over z . that the empirical pdf is in fact a pdf.
- h is the bandwidth. As h increases, the pdf gets smoother, variance decreases, but the bias increases. Silverman suggests: the following rule of thumb:

$$h = 1.06 \hat{\sigma} n^{-1/5}$$

as the optimal bandwidth.

STATA facilitates kernel estimation of density functions with a number of different kernels and window widths. The default kernel density uses a Epanechnikov kernel and optimal window width.

The command is

kdensity "variable name"

kdensity "variable name", normal overlays the kernel density and a fitted normal
histogram "variable name", bin(b) normal kdensity overlays the normal and kernel on the histogram
with b bins

4. Large Sample Theory

Reference: Amemiya [1985].

Let $\{X_n\}$ denote a sequence of random variables. An example of a sequence of random variables is the sample mean based upon a random sample of size n . Another example is the estimator of a regression coefficient based on a sample of size n . We will discuss (a) four different types of convergence of a sequence of random variables and the relationships between them and then consider some (b) central limit theorems which help us identify approximate distribution functions for the sample mean drawn from different populations.

a. Modes of convergence

- (1) Convergence in probability.

A sequence of random variables $\{X_n\}$ is said to converge to a random variable X (could be a constant) in probability if

$$\lim_{n \rightarrow \infty} P_r(X_n - X > \epsilon) = 0$$

for any $\epsilon > 0$. This is frequently written as

$$\boxed{\begin{matrix} P \\ X_n \rightarrow X \end{matrix}}$$

or

$$\boxed{\text{plim}_{n \rightarrow \infty} X_n = X}$$

- (2) Convergence in mean square.

$\{X_n\}$ is said to converge to X in mean square if $\lim_{n \rightarrow \infty} E(X_n - x)^2 = 0$
which is frequently denoted

$$\boxed{\begin{matrix} M \\ X_n \rightarrow X. \end{matrix}}$$

Recall that $E(X_n - X)^2 = E(X_n - E(X))^2 + (E(X_n) - X)^2 = \text{Var}(X_n) + \text{Bias}^2(X_n)$. Hence, convergence in mean square requires that the variance and bias approach zero as n grows indefinitely larger.

- (3) Convergence in distribution.

$\{X_n\}$ is said to converge to X in distribution, $X_n \rightarrow X$, if

$\lim_{n \rightarrow \infty} F_n(X) = F(X)$ for every continuity point of $F(X)$ where

$F_n(\cdot)$ and $F(\cdot)$, respectively, denote the distributions of X_n and X .

(4) Convergence almost everywhere.

$\{X_n\}$ is said to converge to X almost everywhere,

$$\boxed{\begin{array}{ccc} & \text{a.e.} \\ X_n & \rightarrow & X, \end{array}}$$

if $\Pr \left\{ \lim_{n \rightarrow \infty} X_n(w) = X(w) \right\} = 1$

where $X_n(w) = X(w)$ denotes the events in the sample space which are associated with both random variables having the same value.

(5) Relationship between the modes of convergence.

There is a hierarchical relationship between the modes of convergence which Amemiya depicts as

a.e

↓

$M \longrightarrow P \longrightarrow d$

Consequently, convergence in mean square or almost everywhere will imply convergence in probability and in distribution.

(6) Some Theorems

1. If $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{D} Y$, then

$$\begin{array}{c} D \\ X_n + Y_n \xrightarrow{D} X + Y \end{array}$$

$$\begin{array}{c} D \\ X_n Y_n \xrightarrow{D} X \bullet Y \end{array}$$

$$\begin{array}{c} X_n \xrightarrow{D} X \\ \frac{Y_n}{Y} \xrightarrow{P} 1 \end{array}$$

provided $\Pr(Y = 0) \neq 0$.

2. Slutsky's Theorem. Let X_n denote a vector of random variables and $g(\cdot)$ a real valued continuous function,

then $X_n \xrightarrow{P} \alpha$ implies $g(X_n) \xrightarrow{P} g(\alpha)$, where α is a constant vector.

$$\boxed{\begin{array}{c} \text{plim}_{n \rightarrow \infty} g(X_n) = g(\text{plim}_{n \rightarrow \infty} X_n) \end{array}}$$

D P

3. If $X_n \rightarrow X$ and $Y_n \rightarrow \alpha$, then

$$X_n + Y_n \xrightarrow{D} X + \alpha$$

$$X_n \cdot Y_n \xrightarrow{D} \alpha X$$

$$\frac{X_n}{Y_n} \xrightarrow{D} \frac{X}{\alpha} \text{ if } \alpha \neq 0.$$

b. Law of Large Numbers (LLN) and Central Limit Theorems (CLT)

- (1) Law of Large Numbers (LLN)

A Law of Large Numbers specifies conditions under which difference between the sample mean (\bar{X}) and its expected value will converge to zero.

- a. Kolmogorov LLN1

Let $\{X_t\}$ be independent with finite variance, $\text{Var}(X_t) = \sigma_t^2$. If

$$\sum_{t=1}^{\infty} \sigma_t^2 / t^2 < \infty,$$

$$\text{then } \bar{X}_n - E\bar{X}_n \xrightarrow{a.e.} 0.$$

- b. Kolmogorov LLN2

Let $\{X_t\}$ be independently distributed. Then a necessary and sufficient

condition that $\bar{X}_n \xrightarrow{a.e.} \mu$ is

that $E(X_t)$ exists and is equal to μ .

- (2) Central Limit Theorems

Suppose $\bar{X}_n - E(\bar{X}_n) \xrightarrow{P} 0$, then we can conclude that

It is more interesting to investigate distributions which approximate that of

$$Z_n = \frac{\bar{X}_n - E(\bar{X}_n)}{\sqrt{\text{Var}(\bar{X}_n)}}$$

Central limit theorems (CLT) specifies conditions under which Z_n will converge to $N(0,1)$, a standard normal.

Three central limit theorems will be stated and the reader is referred to Amemiya for additional discussion and references.

a. Lindeberg-Levy CLT

Let $\{X_t\}$ be independently and identically distributed

with $E(X_t) = \mu$ and $\text{Var}(X_t) = \sigma^2$, then

$$\begin{matrix} D \\ Z_n \rightarrow N(0,1). \end{matrix}$$

b. Liapounov CLT

Let $\{X_t\}$ be independent with $E(X_t) = \mu_t$, $\text{Var}(X_t) = \sigma_t^2$

and $E|X_t - \mu_t|^3 = \rho_t^3$.

If

$$\lim_{n \rightarrow \infty} \frac{\left(\sum_{t=1}^n \rho_t^3 \right)^{1/3}}{\left(\sum_{t=1}^n \sigma_t^2 \right)^{1/2}} = 0,$$

then

$$Z_n \xrightarrow{P} N(0, 1).$$

c. Lindeberg-Feller CLT

Let $\{X_t\}$ be independent with distribution F_t , $E(X_t) = \mu_t$, $\text{Var}(X_t) = \sigma_t^2$. Define

$$C_n = \left(\sum_{t=1}^n \sigma_t^2 \right)^{1/2}.$$

$$\text{If } \lim_{n \rightarrow \infty} \left(\frac{1}{C_n^2} \right) \sum_{t=1}^n \int_{A_n} (x - \mu_t)^2 dF_t(x) = 0.$$

with $A_n = \{X | |X - \mu_t| > \varepsilon C_n\}$

for all $\varepsilon > 0$,

then

$$\begin{matrix} D \\ Z_n \rightarrow N(0,1). \end{matrix}$$

References

- Amemiya, T. Advanced Econometrics, Cambridge: Harvard University, 1985.
- Arnold, B. C. (1983). Pareto Distributions, Bartonsville: International Cooperative Publishing House.
- Bookstaber, R. M. and J. B. McDonald, A general Distribution for Describing Security Price Returns," Journal of Business, 60(1987), 401-424.
- Dutta, Kabir K. and D. F. Babbel, "Extracting Probabilistic Information from the Prices of Interest Rate Options: Tests of Distribuitonal Assumptions," Journal of Business, 78(2005), 841-870.
- Glaser, Ronald E. "Bathtub and Related Failure Rate Characterizations," Journal of the American Statistical Association, 75(1980), 667-672.
- Goldberger, A. S. (1964). Econometric Theory, New York: Wiley.
- Graybill, F. A. (1961). An Introduction to Linear Statistical Models, New York: McGraw-Hill.
- Huber, P. J. (1981). Robust Statistics, New York: Wiley.
- Intriligator, M. (1977). Econometric Models, Techniques, and Applications, Englewood Cliffs, N.J.: Prentice-Hall.
- Johnson, N. L. (1949). "Systems of frequency curves generated by methods of translation," Biometrika 36, 149-176.
- Johnson, N. L. and S. Kotz. (1970). Continuous Univariate Distributions, Vols. 1, 2, New York: John Wiley & Sons.
- Johnston, J. (1972). Econometric Methods, New York: McGraw-Hill.
- Kalbfleisch, J. D. and R. L. Prentice. (1980). The Statistical Analysis of Failure Time Data, New York: John Wiley & Sons.
- Kmenta, J. (1971). Elements of Econometrics, New York: Macmillan.
- Lindgren, B. W. (1976). Statistical Theory, 3rd ed., New York: Macmillan.
- Maddala, G. S. (1977). Econometrics, New York: McGraw-Hill.
- Mauler,D. and J. B. McDonald, "Option Pricing and Distribution Characteristics," working paper, 2012.
- McDonald, J. B. "Some Generalized Functions for the Size Distribution of Income," Econometrica, 52(1984), 647-663.
- McDonald, J. B. and R. M. Bookstaber, "Optin Pricing for Generalized Distributions," Communications in Statistics--Theory and Method, 20(1991), 4053-4068.
- McDonald, J. B. and R. J. Butler. "Some Generalized Mixture Distributions with an Application to Unemployment Duration," Review of Economics and Statistics, 69(1987), 232-240.
- McDonald, J. B. and D. O. Richards. "Hazard Functions and Generalized Beta Distributions," IEEE Transactions on Reliability, 36(1987b), 463-466.

- McDonald, J. B. and D. O. Richards. "Model Selection: Some Generalized Distributions," Communications in Statistics, 16(1987), 1049-1074.
- McDonald, J. B. and W. K. Newey. "Partially Adaptive Estimation of Regression Models Via the Generalized T Distribution," Econometric Theory, 4(1988), 428-457.
- McDonald, J.B. and Y. Xu, "A Generalization of the Beta Distribution with Applications," Journal of Econometrics, 66(1995), 133-152.
- Patil, G. P., M. T. Boswell and M. V. Ratnaparkhi. Dictionary and Classified Biography of Statistical Distributions in Scientific Work: Continuous Univariate Models, Bartonsville: International Cooperative Clearing House, 1984.
- Rao, C. R. (1973). Linear Statistical Inference and Its Applications, New York: Wiley, 2nd edition.
- Theil, H. (1971). Principles of Econometrics, New York: Wiley.
- Tukey, J. W. (1977). Exploratory Data Analysis, Reading. MA: Addison-Wesley.
- White, H. "A Heteroskedasticity Consistent Covariance Matrix Estimator and a Direct Test of Heteroskedasticity," Econometrica, 48(1980), 817-838.
- White, H. "Maximum Likelihood Estimation of Misspecified Models," Econometrica, 50(1982), 1-25.
- 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.

APPENDIX A

Proofs of Some Theorems Associated with the Distribution of Quadratic Forms of Normally Distributed Variables

1. If $Y \sim N(\mu, \Sigma)$ where Σ is positive definite, then
 $n \times 1$

$$Q = (Y - \mu)' \Sigma^{-1} (Y - \mu) \sim \chi^2(n).$$

Proof:

Since Σ is positive definite, it can be factored

$$\Sigma = \Sigma^{1/2} (\Sigma^{1/2})' \text{ (e.g., Cholesky Decomposition).}$$

This implies that

$$\Sigma^{-1} = ((\Sigma^{1/2}))^{-1} (\Sigma^{1/2})^{-1}.$$

Making this substitution into the quadratic form yields

$$\begin{aligned} Q &= (Y - \mu)' \Sigma^{-1} (Y - \mu) = (Y - \mu)' (\Sigma^{-1/2})' \Sigma^{1/2} (Y - \mu) \\ &= [\Sigma^{-1/2} (Y - \mu)]' [\Sigma^{-1/2} (Y - \mu)] \\ &= Z'Z \quad \text{where } Z = \Sigma^{-1/2} (Y - \mu) \sim N[0, I_n] \end{aligned}$$

because

$$E(Z) = \Sigma^{-1/2} (E(Y) - \mu) = 0$$

$$Var(Z) = \Sigma^{-1/2} \Sigma (\Sigma^{-1/2})' = I;$$

therefore

$$Q \sim \chi^2(n).$$

2. If $Y \sim N[0, I]$ and A is a real symmetric idempotent matrix, then $Q = Y'AY \sim \chi^2(\text{trace}(A))$.

Proof:

The characteristic roots of an idempotent matrix are zero or one and the trace of the idempotent matrix will equal the number of unitary characteristic roots.

Let C denote the orthogonal matrix of characteristic vectors associated with A . In the previous section on matrix results it was shown that

$$C'AC = \Lambda$$

where Λ is a diagonal matrix with the characteristic roots on the diagonal.

The quadratic form Q can be rewritten as

$$\begin{aligned} Q &= Y'AY = Y'C'C'ACC'Y \\ &= (C'Y)' \Lambda (C'Y) \quad (C \text{ is orthogonal}) \end{aligned}$$

$$= Z' \Lambda Z \text{ where } Z = C'Y \sim N[C'0, C'IC] = N[0, I]$$

Therefore,

$$Q = Z' \Lambda Z \sim \chi^2(\text{Rank}(A) = \text{trace}(A)).$$

This follows from Z being a vector of "standard normal" variables and Λ having 0's and one's on the main diagonal. The number of ones on the diagonal of Λ is equal to the trace of A .

3. Let $Y_{nx1} \sim N(0, I)$ and let A denote an $n \times n$ real symmetric idempotent matrix. The vector of random variables LY and the quadratic form $Q = Y'AY$ will be independent if $LA = 0$.

Proof: The symmetry and idempotent property of A permits rewriting Q as

$$Q = Y'AY = Y'A'AY = (AY)'(AY)$$

The covariance between LY and AY is given by

$$\begin{aligned} \text{COV}(LY, AY) &= E(LY - E(LY))(AY - E(AY))' \\ &= E(LY)(AY)' = E(LYY'A') \\ &= L(E(YY'))A' \\ &= LIA' \quad \text{Because } Y \sim N[0, I] \\ &= LA \quad \text{Because } A \text{ is symmetric} \end{aligned}$$

It follows that $LA = 0$ implies that LY and AY are stochastically independent; hence Q and LY will be stochastically independent.

APPENDIX B

Cramer-Rao Inequality--Information Matrix
Theil, pp. 384-86

1. The equivalence of two important matrices

Let $L(Y_1, \dots, Y_n; \theta) = \prod_t f(Y_t; \theta) = L(Y; \theta)$ and let

$$\ell = \ln(L) = \sum_t \ln f(Y_t; \theta) = \ell(Y; \theta)$$

Integrating the likelihood function over all values of the Y 's is unity,

i.e.,

$$(B.1) \quad \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} L(Y_1, \dots, Y_n; \theta) dY_1 \dots dY_n = \int L(Y; \theta) dY = 1.$$

Differentiating both sides of (B.1) with respect to θ yields

$$(B.2) \quad \int \frac{dL(Y; \theta)}{d\theta} dY = 0, \quad \text{but this is the same as}$$

$$(B.2)' \quad \int \frac{\frac{dL(Y; \theta)}{d\theta}}{L(Y; \theta)} L(Y; \theta) dY = \int \frac{d\ell(Y; \theta)}{d\theta} L(Y; \theta) dY = 0$$

or
$$E\left(\frac{d\ell(Y; \theta)}{d\theta}\right) = 0$$

Differentiating the second expression in (B.2)' with respect to θ yields

$$(B.3) \quad \begin{aligned} & \int \frac{d^2\ell}{d\theta^2} L(Y; \theta) + \frac{d\ell}{d\theta} \frac{dL(Y; \theta)}{d\theta} dY \\ &= \int \left\{ \frac{d^2\ell(Y; \theta)}{d\theta^2} + \left(\frac{d\ell}{d\theta} \right) \left(\frac{dL}{d\theta} \right) \right\} L(Y; \theta) dY = 0. \end{aligned}$$

This implies that

$$(B.4) \quad E\left(\left(\frac{d\ell}{d\theta} \right) \left(\frac{d\ell}{d\theta} \right)\right) = -E\left\{ \frac{d^2\ell(Y; \theta)}{d\theta^2} \right\} = \text{var}\left(\frac{d\ell}{d\theta}\right)$$

This is a very important theoretical and practical result which can be used in estimating variances of MLE of θ as well as in testing for correct specifications.

2. The asymptotic distribution of the MLE of θ

Let $\hat{\theta} = \hat{\theta}(Y)$ denote an estimator of θ . The expected value of $\hat{\theta}(Y)$ is given by

$$(B.5) \quad E(\hat{\theta}(Y)) = \int \hat{\theta}(Y) L(Y; \theta) dY.$$

Differentiating (B.5) with respect to θ yields

$$(B.6) \quad \frac{\partial E(\hat{\theta}(Y))}{\partial \theta} = \int \hat{\theta}(Y) \frac{dL(Y; \theta)}{d\theta} dY = \int \hat{\theta}(Y) \frac{d\ell(Y; \theta)}{d\theta} L(Y; \theta) dY \\ = \text{cov}(\hat{\theta}(Y), \partial \ell(Y; \theta)/\partial \theta).$$

It will be useful to make a couple of observations at this point. First, for unbiased estimators $E(\hat{\theta}(Y)) = \theta$

$$\text{and } \frac{dE(\hat{\theta}(Y))}{d\theta} = \frac{d\theta}{d\theta} = 1$$

and then equation (B.6) implies $\text{Cov}(\hat{\theta}(Y), d\ell/d\theta) = 1$. Second, recall

correlation²(Z_1, Z_2) ≤ 1 implies that

$$\text{COVAR}^2(Z_1, Z_2) \leq \text{Var}(Z_1)\text{Var}(Z_2).$$

Combining these results yields

$$(B.7) \quad 1 = \left(\frac{\partial E(\hat{\theta}(Y))}{\partial \theta} \right)^2 = \text{Cov}^2(\hat{\theta}(Y)) \leq \text{Var}(\hat{\theta}(Y)) \text{Var}\left(\frac{\partial \ell(Y; \theta)}{\partial \theta} \right);$$

i.e.,

$$1 \leq \text{Var}(\hat{\theta}(Y)) \left\{ -E\left(\frac{\partial^2 \ell}{\partial \theta^2} \right) \right\}$$

Therefore

$$(B.8) \quad \boxed{\begin{aligned} \text{var}(\hat{\theta}(Y)) &\geq \left\{ -E\left(\frac{\partial^2 \ell}{\partial \theta^2} \right) \right\}^{-1} \\ &\geq \left\{ E\left(\frac{d\ell}{d\theta} \frac{d\ell}{d\theta} \right) \right\}^{-1} \end{aligned}}$$

which is the Cramer-Rao lower bound for unbiased estimators.

5. Problem sets

Consider the exponential distribution defined by the probability density function

$$f(X; \beta) = \frac{e^{-x/\beta}}{\beta} \quad \text{for } X \geq 0$$

= 0 otherwise.

1. Show that the cumulative distribution function for the exponential is given by

$$F(X; \theta) = 1 - e^{-X/\beta}$$

2. Derive an expression for the corresponding hazard function.

3. Show that the moment generating function for the exponential distribution is given by $\left(\frac{1}{1 - \beta t} \right)$

4. Write an expression for the cumulant generating function for the exponential distribution.

5. Using the cumulant generating function for the exponential evaluate

$$E(X) = \mu =$$

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

$$E(X - \mu)^3 =$$

$$E(X - \mu)^4 =$$

Note that these results imply that

$$\text{Skewness (normalized)} = \frac{\text{E} (X - \mu)^3}{\sigma^3} = 2 > 0$$

$$\text{Kurtosis (normalized)} = \frac{\text{E} (X - \mu)^4}{\sigma^4} = 9$$

6. Show that the mode and median for the exponential distribution are given by 0 and $(\beta \ln 2)$, respectively.

Hint: $f(\text{mode}) \geq f(x) \quad \int_{-\infty}^{\text{median}} f(s) ds = .5$

7. Demonstrate that the MLE of β in the exponential distribution is given by $\tilde{\beta} = \bar{X} = \sum_{i=1}^N \left(\frac{X_i}{N} \right)$

8. The mgf for the gamma probability density,

$$GA(x; p, b) = \frac{x^{p-1} e^{-x/b}}{b^p \Gamma(p)},$$

is given by

$$\frac{1}{(1-\beta t)^p}$$

Show that the sum of n independent and identically distributed exponential variables is distributed as $GA(\ ; \beta, p=n)$.

9. Using your results in question (8), demonstrate that the exact distribution of the sample mean (\bar{Y}) is given by:

$$\bar{Y} \sim GA(\ ; \beta/n, p=n).$$

10. Consider a random sample

$\{y_1, y_2, \dots, y_n\}$ where the density of y_i is given by

$$\text{GG}(y_i; a, \beta, p).$$

- (a) Form the log likelihood function.
- (b) Obtain the MLE of β for the case where $a = p = 1$, i.e., the exponential. (same as # 7)
- (c) Evaluate the second derivative of the log likelihood function with respect to β for the exponential?
- (d) What is the asymptotic distribution of the MLE of β for the exponential?

11. The exact distribution of the sample mean of exponentially distributed variables is a gamma, $\text{GA}(z; \beta/n, p=n)$, (see problem 9) and the approximating "asymptotic" normal distribution is $N(z; \beta, \beta^2/n)$. The corresponding moment generating functions are given by:

$$\text{Gamma: } M_{\text{GA}}(t) = \frac{1}{(1 - \beta t/n)^n}$$

$$\text{Normal: } M_N(t) = e^{\beta t + (\beta t)^2/2n}$$

Compare the mean, variance, skewness, and kurtosis of the *exact* distribution and the approximating or asymptotic normal distribution.

	$\text{GA}(; \beta/n , p=n)$	$N(\mu = \beta, \sigma^2 = \beta^2/n)$
Mean		
Variance		
Skewness*		
Kurtosis*		

Hint: Use the cumulant generating functions. * Note you are only asked to "normalize" the skewness and kurtosis coefficients. This is done by dividing the skewness and kurtosis by σ^3 and σ^4 , respectively.

12. The Burr distribution and hazard functions

- a. Demonstrate that

$$GB2(y; a, b, p = 1, q) = \frac{|a|q(y/b)^{a-1}}{b(1 + (y/b)^a)^{q+1}}$$

Hint:

$$B(1, q) = \frac{\Gamma(1)\Gamma(q)}{\Gamma(q+1)} = \frac{\Gamma(1)\Gamma(q)}{q\Gamma(q)} = \left\{ \frac{1}{q} \right\}$$

$$F_{Burr12}(y; a, b, p) = 1 - \left(\frac{1}{1 + (y/b)^a} \right)^q$$

- b. Use these results to obtain an expression for the hazard function for the BR12. Recall that the hazard function is defined to be pdf/(1-cdf)
- c. The hazard function for the Burr 12 with $a > 1$ is upside down "U" shaped. Investigate the shape of the hazard function for the two cases (1) $0 < a < 1$ and (2) $a = 1$. Assume that all parameter values are positive. Hint: What happens to the value of the hazard function as y increases for arbitrary but fixed values of a ?
13. What are the possible shapes of the hazard function for the gamma, Weibull and exponential and distributions? (Hint: This problem is not requesting a derivation. Refer to figure 4 for the generalized gamma function).
14. If $Y \sim LN(\mu, \sigma^2)$, verify that $Z = Y^2 \sim LN(2\mu, 4\sigma^2)$. Don't forget the Jacobian in your analysis of the transformation.
15. If $X \sim GA(x; \beta = 1, p)$, verify that $Y = \beta X^{1/\alpha} \sim GG(y; a, \beta, p)$. Don't forget the Jacobian in your analysis of the transformation.
16. Use the cumulant generating function for the normal, $N(\mu, \sigma^2)$, to show that the normalized or scaled kurtosis for a normally distributed random variable is given by 3.
17. Given that the $n \times 1$ vector Y is distributed as $N(\mu, \Sigma)$ where Σ is positive definite, demonstrate that the mode of the multivariate density $N(\mu, \Sigma)$ occurs at $Y = \mu$, i.e.,

demonstrate $\frac{df(Y)}{dY} = 0$ at $Y = \mu$ and note that $\frac{d^2f(Y)}{dY^2}$ is negative definite.

of the expression of the multivariate normal density function.

18. Given $g(X) = 2x_1^2 + (5/2)x_2^2 + 4x_1x_2 + x_1 + 3x_2$:

- a. Demonstrate that

$$g(X) = (x_1, x_2) \begin{pmatrix} 2 & 2 \\ 2 & 5/2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + (x_1, x_2) \begin{pmatrix} 1 \\ 3 \end{pmatrix}$$

- b. Using the technique of differentiating $g(X)$ with respect to the vector $X = (x_1, x_2)'$, determine the vector X which is associated with an optimum of $g(X)$.

- c. Evaluate $\frac{d^2g(X)}{dx^2}$ and determine whether your answer to (b) corresponds to a maximum or minimum.

19. Let $\mathbf{i} = (1, 1, \dots, 1)'$

- a. Evaluate $\mathbf{i}'\mathbf{i}$
 b. Evaluate $\mathbf{i}\mathbf{i}'$
 c. Demonstrate that the matrix

$$\left(\frac{1}{n} \mathbf{i} \mathbf{i}' \right)$$

is symmetric and idempotent.

20. Let y_t be independently distributed as $N(O, \sigma^2)$, i.e.,

$$\mathbf{Y} = (y_1, y_2, \dots, y_n)' \sim N[O, \sigma^2 I_n].$$

- a. Determine the distribution of

$$\bar{Y} = \sum_{t=1}^n y_t / n.$$

$$\text{Hint: } \bar{Y} = \left(\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n} \right) \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \left(\frac{1}{n} \right) \mathbf{i}' \mathbf{y}$$

$$= BY \text{ where } B = \left(\frac{1}{n} \right) \mathbf{i}' .$$

Consider the first "useful" theorem in the section on multivariate statistics.

$$\text{Now consider } s^2 = (\sum (y_i - \bar{Y})^2 / (n - 1))$$

$$= (y_1, \dots, y_n) \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & & \vdots \\ 0 & \dots & & 1 \end{pmatrix} - \frac{1}{n} \begin{pmatrix} 1 & 1 & \dots & 1 \\ \vdots & \vdots & & \vdots \\ 1 & 1 & & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} / n-1$$

$$= Y' (I - \frac{1}{n} \mathbf{i} \mathbf{i}') Y / (n-1)$$

$$= Y' A Y / (n-1)$$

$$\text{where } A = I - \frac{1}{n} \begin{pmatrix} 1 & \dots & 1 \\ \vdots & & \vdots \\ 1 & \dots & 1 \end{pmatrix} = I - \frac{1}{n} \mathbf{i} \mathbf{i}'$$

- b. Demonstrate that A is idempotent.
- c. Verify that trace (A) = n - 1.
- d. Demonstrate that the distribution of $s^2(n-1)/\sigma^2$ is $\chi^2(n-1)$.
- e. Verify that \bar{Y} and s^2 are independently distributed.
Hint: Is BA = 0?

f. What is the distribution of $\left(\frac{\bar{Y} - 0}{\sigma/n} \right) = \frac{\bar{Y} - 0}{s/n}$

$$\left(\frac{s^2(n-1)}{\sigma^2} / (n-1) \right)^{1/2}$$

21. Method of Moments

The mean and variance of a Gamma are given by βp $\beta^2 p$, respectively. Let the corresponding sample moments be denoted by \bar{Y} and s^2 .

- a. Derive the method of moments estimators of the parameters β p .
- b. Derive the method of moments estimator of the parameter β in an exponential distribution.
- c. Using two sample moments discuss how you could use generalized method of moments estimation to estimate the parameter β in an exponential distribution.

22. Optional problem. Verify

a. $EGB(\varepsilon; \delta, \sigma, c, p, q) = GB(e^\varepsilon; a=1/\sigma, b=e^\delta, c, p, q)e^\varepsilon$

Why is the expression for $GB()$ multiplied by e^ε ?

b. $EGB2(\varepsilon; \delta, \sigma, p, q) = (EGB(\varepsilon; \delta, \sigma, c=1, p, q)$

$$= \frac{e^{p(\varepsilon-\delta)/\sigma}}{\sigma B(p, q) (1 + e^{(\varepsilon-\delta)/\sigma})^{p+q}}$$

for $-\infty < \varepsilon < \infty$

c. $M_{EGB2}(t) = \frac{e^{\delta t} \Gamma(p + t\sigma) \Gamma(q - t\sigma)}{\Gamma(p) \Gamma(q)}$

Hint: Consider the moments for the GB2 reported in Table 1.

- d. Obtain the cumulant generating function for the EGB2 and show that the mean of the EGB2 is zero if $\delta = \sigma[\psi(q) - \psi(p)]$ where $\psi(x) = d \ln \Gamma(x)/dx$.

III. The Generalized Regression Model and Related Topics

- 1. The Model**
- 2. Estimators of the coefficient vector (β)**
 - a. Least squares**
 - (1) Sufficient condition
 - (2) Distribution of OLS estimator
 - b. Maximum likelihood**
 - (1) Sufficient conditions for a maximum
 - (2) Relationship to BLUE
 - (3) Distribution of the MLE of β
 - (4) Cramer-Rao matrix
 - (5) Efficiency
 - (6) MLE using OLS
 - (7) Likelihood Ratio tests
 - c. Alternative estimators**
- 3. Forecasting**
- 4. A review of some important special cases**
 - a. No autocorrelation with homoskedastic disturbances**
 - (1) Coefficient estimators
 - (2) Variance estimators
 - (3) Hypothesis testing
 - (a) Simple t-tests
 - (b) F-tests
 - (c) Chow tests
 - (d) Likelihood ratio tests
 - (4) Forecasting
 - (5) Consequences of using least squares when $\Sigma \neq \sigma^2 I$
 - b. Heteroskedasticity**
 - (1) Definition
 - (2) Statistical tests
 - (3) Estimation
 - c. Autocorrelation**
 - (1) Definition
 - (2) Statistical tests
 - (3) Estimation
 - (4) Forecasting
- 5. Autoregressive conditional heteroskedasticity (ARCH) models**
 - a. Introduction -- an example**
 - b. Estimation**
 - c. Test for ARCH disturbances**

6. Stochastic Regressors

7. Instrumental variables

- a. Background
- b. Instrumental variables estimator
- c. Some special cases
 - (1) Least squares
 - (2) Generalized least squares
 - (3) Projection of X on Z
 - (4) Stochastic regressors
- d. Selection of instrumental variables
- e. Instrumental variables in quantile regression models
- f. Some other issues—weak instruments, application references

8. Specification tests

- 1. Background
- 2. Hausmann test
- 3. Applications

9. Seemingly unrelated regression models (Sure-Zellner)

10. Models for Panel Data: cross sectional and time series

- a. OLS and GLS
- b. Random and fixed effects models
- c. Differences in differences
- d. Statistical inference

11. Regression discontinuities

12. Exercise set

The Generalized Regression Model

1. The Model

Let Y_t , X_{ti} denote endogenous (dependent) and exogenous (independent) variables.

Consider the relationship

$$Y_t = \beta_1 + \beta_2 X_{t2} + \dots + \beta_K X_{tK} + \varepsilon_t \quad (1.1)$$

$$t = 1, 2, \dots, N, N > K;$$

(1.1) can be represented in terms of matrices as

$$Y = X\beta + \varepsilon \quad (1.1)'$$

where

$$\begin{aligned} Y &= \begin{bmatrix} Y_1 \\ \vdots \\ Y_N \end{bmatrix}, & X &= \begin{bmatrix} 1 & X_{12} & \cdots & X_{1K} \\ \vdots & \vdots & & \vdots \\ 1 & X_{N2} & \cdots & X_{NK} \end{bmatrix}, \\ \beta &= \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_K \end{bmatrix} \text{ and } \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{bmatrix}. \end{aligned}$$

It will also be assumed that

$$(A.1) \quad \varepsilon \sim N[0, \Sigma],$$

(A.2) the X's are nonstochastic and

$$\lim_{N \rightarrow \infty} \frac{(X' \Sigma^{-1} X)}{N} = \Sigma_x$$

which is nonsingular.

(A.1) implies the random disturbances are distributed normally, each with mean zero and

$$\begin{aligned} \text{Var}(\boldsymbol{\varepsilon}) &= \begin{bmatrix} \text{Var}(\varepsilon_1) & \text{Cov}(\varepsilon_1, \varepsilon_2) & \cdots & \text{Cov}(\varepsilon_1, \varepsilon_N) \\ & \text{Var}(\varepsilon_2) & \cdots & \text{Cov}(\varepsilon_2, \varepsilon_N) \\ & \ddots & & \vdots \\ & & & \text{Var}(\varepsilon_N) \end{bmatrix} \\ &= \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1N} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2N} \\ \vdots & \vdots & & \vdots \\ \sigma_{N1} & \sigma_{N2} & \cdots & \sigma_{NN} \end{bmatrix} \\ &= \Sigma. \end{aligned}$$

Note that this specification doesn't require that the random disturbances are independent of each other (autocorrelation is possible) nor that the variances of the random disturbances are the same (heteroskedasticity is possible); however this specification allows for homoskedastic and nonautocorrelated errors if $\Sigma = \sigma^2 I$.

The unknown parameters in the generalized regression model are

- (1) the coefficients β
- (2) the variance-covariance matrix Σ .

2. Estimators of the Coefficient Vector (β)

- a. Least Squares. The least squares technique is based upon the principle of selecting an estimator of β , which minimizes the associated sum of squared errors--vertical deviations between the observed dependent variables and the estimated regression line (plane).

$$\begin{array}{ccc}
 Y_t & & Y_N \\
 | & & | \\
 Y_1 & & e_N \\
 | & & | \\
 e_1 & & Y = X\beta + \epsilon \\
 | & & | \\
 e_2 & & = \hat{X}\hat{\beta} + \epsilon \\
 | & & | \\
 Y_2 & & X_t
 \end{array}$$

$$\begin{array}{c}
 SSE(\hat{\beta}_1, \hat{\beta}_2) \\
 | \\
 | \\
 | \\
 | \\
 \beta_2
 \end{array}$$

$$(\hat{\beta}_1, \hat{\beta}_2)$$

$$\hat{\beta}_1$$

The sum of squared errors can be expressed as

$$\begin{aligned}
 \text{SSE}(\hat{\beta}) &= \sum e_t^2 \\
 &= (\mathbf{e}_1, \dots, \mathbf{e}_N) \begin{pmatrix} \mathbf{e}_1 \\ \vdots \\ \mathbf{e}_N \end{pmatrix} \\
 &= \mathbf{e}'\mathbf{e} \\
 &= (\mathbf{Y} - \mathbf{X}\hat{\beta})' (\mathbf{Y} - \mathbf{X}\hat{\beta})
 \end{aligned} \tag{2.1 a-c}$$

depends upon the estimators of β . The least squares estimators are obtained by

minimizing $\text{SSE}(\hat{\beta})$ with respect to the vector $\hat{\beta}$, i.e.,

$$\begin{aligned}
 &\underset{\hat{\beta}}{\text{minimize}} \text{ SSE}(\hat{\beta}) \\
 &= \underset{\hat{\beta}}{\min} (\mathbf{Y} - \mathbf{X}\hat{\beta})' (\mathbf{Y} - \mathbf{X}\hat{\beta}) .
 \end{aligned}$$

After expanding the expression (2.1d) for the sum of squared errors, we obtain

$$\text{SSE}(\hat{\beta}) = \mathbf{Y}'\mathbf{Y} - 2\hat{\beta}'\mathbf{X}'\mathbf{Y} + \hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta} .$$

Differentiating this expression with respect to $\hat{\beta}$ yields the necessary conditions

for a solution to (2.2):

$$\frac{d\text{SSE}(\hat{\beta})}{d\hat{\beta}} = -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\hat{\beta} = 0$$

or

(2.3)'

$$(X'X)\hat{\beta} = X'Y$$

(2.3)' is also referred to as the system of normal equations.

If $|X'X| \neq 0$, (2.3)' can be solved to yield the least squares estimator

(2.4)

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

Note: (1) The sufficient condition for (2.4) to be a solution to (2.2) is that

$$\frac{d^2SSE}{d\hat{\beta}d\hat{\beta}'} = 2(X'X)$$

is positive definite.

(2) The least squares estimator is distributed normally as

$$\hat{\beta} \sim N(\beta; (X'X)^{-1}X'\Sigma X(X'X)^{-1})$$

Proof. (a) See I.B.2 for a "useful" theorem

(b) Alternatively,

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

Therefore, $E(\hat{\beta}) = \beta$

$$\begin{aligned} \text{Var}(\hat{\beta}) &= E(\hat{\beta} - \beta)(\hat{\beta} - \beta)' \\ &= E(X'X)^{-1} X' \epsilon \epsilon' X (X'X)^{-1} \\ &= (X'X)^{-1} X' \Sigma X (X'X)^{-1}. \end{aligned}$$

b. Maximum Likelihood Estimators

The likelihood function associated with (1.1)' is given by

$$L(Y; \beta, \Sigma) = \frac{e^{(Y-X\beta)'\Sigma^{-1}(Y-X\beta)/2}}{(2\pi)^{N/2} |\Sigma|^{1/2}}$$

and the log likelihood function is given by

$$\begin{aligned} \ell(Y; \beta, \Sigma) &= \ln(L(Y; \beta, \Sigma)) \\ &= (-1/2)(Y - X\beta)' \Sigma^{-1} (Y - X\beta) - (\frac{1}{2})[N \ln(2\pi) + \ln |\Sigma|] \\ &= (-1/2)(Y' \Sigma^{-1} Y - 2\beta' X' \Sigma^{-1} Y + \beta' X' \Sigma^{-1} X\beta) - (\frac{1}{2})[N \ln(2\pi) + \ln |\Sigma|]. \end{aligned} \tag{2.7}$$

The maximum likelihood estimator (MLE) of β is obtained by maximizing (2.6) or (2.7) with respect to β , i.e.,

$$\underset{\beta}{\operatorname{Max}} \ell(Y; \beta, \Sigma). \tag{2.8}$$

The necessary conditions are given by

$$\frac{d\ell}{d\beta} = -[-X'\Sigma^{-1}Y + X'\Sigma^{-1}X\tilde{\beta}] = 0$$

or

$X'\Sigma^{-1}X\tilde{\beta} = X'\Sigma^{-1}Y$

(2.10) is sometimes referred to as the system of modified normal equations.

If $|X'\Sigma^{-1}X| \neq 0$, the maximum likelihood estimator is the solution to (2.10),

$$\tilde{\beta} = (\mathbf{X}' \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}' \Sigma^{-1} \mathbf{Y}$$

This estimator is also referred to as Aitken's estimator and also as the generalized or weighted least squares (GLS) estimator.

Note: (1) The sufficient condition for β to yield a maximum of (2.7) is for

$$\frac{d^2\ell}{d\beta d\beta'} = -\mathbf{X}' \Sigma^{-1} \mathbf{X}$$

to be negative definite.

(2) $\tilde{\beta}$ can also be shown to be the best linear unbiased estimator (BLUE) of β .

(3) $\tilde{\beta}$ is distributed normally and

$$\tilde{\beta} \sim N(\beta; (\mathbf{X}' \Sigma^{-1} \mathbf{X})^{-1});$$

hence, the least squares and MLE are both unbiased estimators.

The variances of $\hat{\beta}$ and $\tilde{\beta}$ may be different if $\Sigma \neq \sigma^2 I$.

(4) The Cramer-Rao matrix is given by

$$\Sigma_{\tilde{\beta}}^* = - \left[E \frac{\partial^2 \ell}{\partial \beta \partial \beta'} \right]^{-1}$$

$$= (X' \Sigma^{-1} X)^{-1}$$

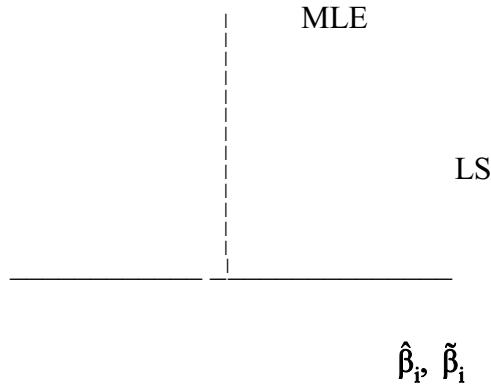
which is the variance covariance matrix of $\tilde{\beta}$.

(5) $\text{Var}(\hat{\beta}) - \text{Var}(\tilde{\beta})$

$$= (X'X)^{-1} X' \Sigma X (X'X)^{-1} - (X' \Sigma^{-1} X)^{-1}$$

is a positive semi-definite matrix. This implies that the variance of

$$\text{Var}(\hat{\beta}_i) \geq \text{Var}(\tilde{\beta}_i).$$



(6) MLE Using OLS

If a matrix T can be found such that $T \Sigma T' = \sigma^2 I$ ($T'T = \Sigma^{-1}$), then

the MLE of β can be obtained by estimating

$$TY = TX\beta + T\epsilon$$

using least squares.

(7) Likelihood Ratio (LR) test

MLE lends itself to testing hypotheses of the form $q(\beta) = 0$

where $q(\beta)$ denotes an $r \times 1$ vector of continuous functional constraints on the vector β , e.g.

$$q(\beta) = \begin{pmatrix} \beta_2\beta_3 - 1 \\ \beta_4 \\ \beta_1 + \beta_2 - 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

This hypothesis can be tested by estimating the model with and without the constraints imposed. Let the corresponding log likelihood values be denoted by ℓ^* and ℓ .

$$LR = 2(\ell - \ell^*)$$

provides the basis for testing $H_0: q(\beta) = 0$ and is asymptotically distributed as a chi-square,

$$LR = 2(\ell - \ell^*) \stackrel{a}{\sim} \chi^2(r)$$

c. Alternative estimators

There are many alternatives to least squares and MLE. These include, among others:

Least absolute deviation (LAD) estimators

$$\underset{\beta}{\text{Min}} \sum_{t=1}^n |\epsilon_t|;$$

L_p estimators

$$\underset{\beta}{\text{Min}} \sum_{t=1}^n |\varepsilon_t|^p;$$

M estimators

$$\underset{\beta}{\text{Min}} \sum \rho(\varepsilon_t);$$

among others. Each of these estimators works well for errors from a particular distribution, but not necessarily for others. There is considerable interest in estimators which are "robust" over many distributions. See the Appendix to this section for additional discussion of this material and some examples.

3. Forecasting

Recall that the model under consideration is given by (1.1).

$$\begin{aligned} Y_t &= \beta_1 + \beta_2 X_{t2} + \dots + \beta_k X_{tk} + \varepsilon_t \\ &= (1, X_{t2}, \dots, X_{tk}) \beta + \varepsilon_t \\ &= X_t \beta + \varepsilon_t. \end{aligned} \tag{3.1}$$

Goldberger (1962, JASA) demonstrated that the minimum variance h -period ahead unbiased predictor of Y is given by

$$Y_N(h) = \tilde{Y}_{N+h} = X_{N+h} \tilde{\beta} + W' \Sigma^{-1} e$$

(3.2)

where

- $\tilde{\beta}$ denotes the MLE of β

- $e = Y - X\tilde{\beta}$, the estimated residual vector and

$$\begin{aligned} \bullet W' &= E(\varepsilon_{N+h} \varepsilon) \\ &= E \begin{bmatrix} \varepsilon_{N+h} & \varepsilon_1 \\ \vdots & \\ \varepsilon_{N+h} & \varepsilon_N \end{bmatrix} = \begin{bmatrix} \text{cov}(\varepsilon_1, \varepsilon_{N+h}) \\ \vdots \\ \text{cov}(\varepsilon_N, \varepsilon_{N+h}) \end{bmatrix} \end{aligned}$$

- N = Sample size associated with the sample used in estimating β .

Note that if the error terms are uncorrelated, $E(\varepsilon_t \varepsilon_s) = 0$ for $t \neq s$, $W = 0$; hence (3.2)

simplifies to

$$Y_N(h) = \tilde{Y}_{N+h} = X_{N+h}\tilde{\beta}$$

(3.3)

for random disturbances which are not autocorrelated.

4. Some Important Special Cases

a. No autocorrelation with homoskedastic random disturbances

The model can then be written as

$$Y = X\beta + \varepsilon \quad (4.1)$$

$$(A.1) \quad \varepsilon \sim N(0, \sigma^2 I)$$

(1) Coefficient Estimators

Exercise: Demonstrate that the least squares and MLE of β are identical in this case and

$$\hat{\beta} = \tilde{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \sim N(\beta; \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$$

This model is generally referred to as the classical normal linear regression

model. The least squares estimators of β , $\hat{\beta}$, will be

- unbiased
- minimum variance of all unbiased estimators (hence BLUE)
- consistent
- asymptotically efficient
- normally distributed

(2) Variance Estimators

- An unbiased estimator of σ^2 is given by

$$s^2 = \sum e_t^2 / (N - K) \quad (4.4)$$

$$= e'e / (N - K)$$

$$\text{Exercise: } \frac{(N - K)s^2}{\sigma^2} \sim \chi^2(N - K).$$

- An unbiased estimator of $\text{var}(\hat{\beta})$ = $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ is given by

$$s^2(\mathbf{X}'\mathbf{X})^{-1} \quad (4.5)$$

(3) Hypothesis Testing

This form of the model facilitates testing numerous hypotheses of interest, e.g.

- (a) $H_0: \beta_i = \beta_i^o$

This hypothesis can be tested using a t-test

$$\frac{\hat{\beta}_i - \beta_i^0}{s_{\hat{\beta}_i}} \sim t(N - K)$$

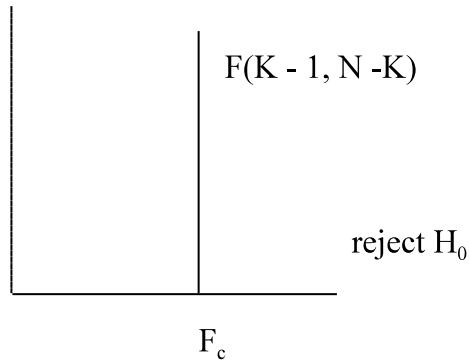
where the pdf of the t-statistic and corresponding critical values (2-tailed test) appear as in the following figure:



(b) $H_0: \beta_2 = \beta_3 = \dots = \beta_K = 0$

This test of the overall "explanatory" power of the model can be performed with an F test.

$$F = \frac{SSR/K - 1}{SEE/N - K} = \left(\frac{R^2}{1 - R^2} \right) \left(\frac{N - K}{K - 1} \right)$$

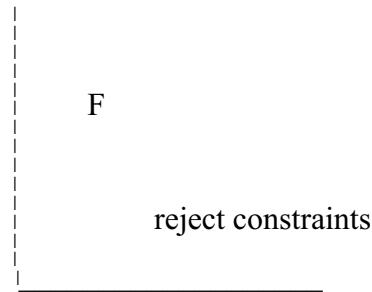


(c) The Chow test can be used to test these hypotheses as well as more general constraints on the coefficient vector. The test statistic is based

upon the results of estimating the model without the constraints imposed and again with the constraints imposed.

The test statistic is defined by

$$\frac{\frac{\text{SSE}^* - \text{SSE}}{r}}{\frac{\text{SSE}}{N-K}} \sim F(r, N - K)$$



where the * denotes the results from estimating the constrained model, and $r = (N-K)^* - (N-K)$ is the number of independent restrictions imposed by the hypotheses.

(d) Likelihood Ratio Test

The log likelihood value for the special case of homoskedastic and independent (not autocorrelated) residuals is given by

$$\ell(\beta, \sigma^2) = -\text{SSE}/2\sigma^2 - \frac{N}{2} \ln(2\pi) - \frac{N}{2} \ln(\sigma^2)$$

The concentrated (over σ^2) likelihood function is obtained from (4.9) by replacing σ^2 by its MLE, SSE/N, to obtain

$$\ell_c(\beta) = -\frac{N}{2} - \frac{N}{2} \ln(2\pi) - \frac{N}{2} \ln(\text{SSE}/N)$$

$$\ell_t(\beta) = -\frac{N}{2} \left(1 + \ln(2\pi) + \ln\left(\frac{\text{SSE}}{N}\right) \right).$$

The likelihood ratio (LR) test is obtained by estimating the model with and without the constraints imposed to yield ℓ^* and ℓ , respectively. The LR test statistic is constructed by taking twice the difference of ℓ and ℓ^* and has an asymptotic χ^2 distribution with degrees of freedom equal to the number of independent constraints, i.e.,

$$\text{LR} = 2(\ell - \ell^*) \stackrel{a}{\approx} \chi^2(r)$$

For normally distributed errors and

$$\bullet \underline{\sigma \text{ known}} \quad \text{LR} = \frac{\text{SSE}^* - \text{SSE}}{\sigma^2}$$

$$\bullet \sigma^2 \text{ unknown}$$

$$\text{LR} = N \ln(\text{SSE}^*/\text{SSE}). \quad (4.10'')$$

(4.10') corresponds to the case of known variance and is quite similar in structure to the Chow test in (4.8) (recall SSE/N-K is an unbiased estimator of σ^2).

(4) Forecasting (conditional on given X's)

From (3.2) and (3.3) we see that the best linear unbiased forecasts of Y, given X's, are given by

$$\tilde{Y}_{N+h} = X_{N+h}\tilde{\beta}.$$

\tilde{Y}_{N+h} is distributed normally with mean $X_{N+h}\beta$ and variance

$$\text{var}(\hat{Y}_{N+h}) = \sigma_{\hat{Y}}^2 = X_{N+h} \sigma^2 (X'X)^{-1} X_{N+h}'.$$

which can be estimated by

$$s_{\hat{Y}}^2 = X_{N+h} s^2 (X'X)^{-1} X_{N+h}'.$$

The Forecast error is defined to be the difference between the predicted and observed values for Y, i.e.,

$$FE = Y_{N+h} - \hat{Y}_{N+h}$$

The forecast error is distributed normally with mean 0 and variance

$$\sigma_{FE}^2 = \sigma^2 + \sigma_{\hat{Y}}^2 \quad (4.15)$$

measure of uncertainty of ϵ_t	+	measure of uncertainty of $X_{N+h}\beta$
----------------------------------------	---	------------------------------------------

σ_{FE}^2 can be estimated by

$$s_{FE}^2 = s^2 + s_{\hat{Y}}^2. \quad (4.16)$$

In summary,

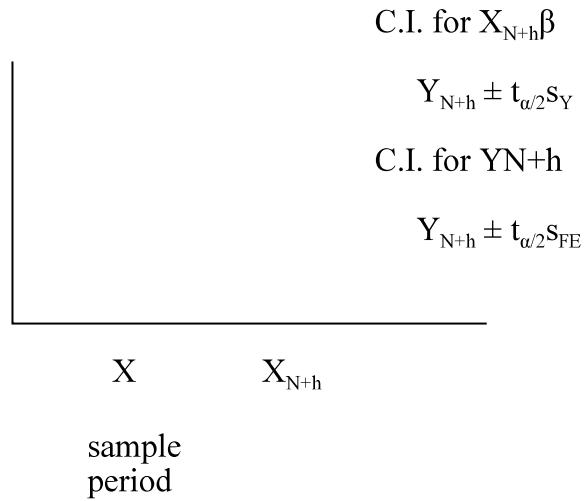
$$\begin{aligned} \tilde{Y}_{N+h} &\sim N(X_{N+h}\beta, \sigma_{\hat{Y}}^2) \\ FE &= Y_{N+h} - \hat{Y}_{N+h} \\ &\sim N(0, \sigma_{FE}^2 = \sigma^2 + \sigma_{\hat{Y}}^2) \end{aligned} \quad (4.18)$$

Confidence intervals for the regression line, $X_{N+h}\beta$, and for the actual value of Y, Y_{N+h} respectively are given by

$$X_{N+h}\hat{\beta} \quad \pm t_{\alpha/2} s_Y$$

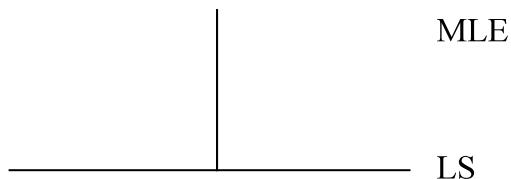
$$Y_{N+h}: X_{N+h}\hat{\beta} \quad \pm t_{\alpha/2} s_{FE}$$

where $t_{\alpha/2}$ denotes the critical value for a t-statistic with $N-K$ degrees of freedom at the α -level of significance. This can be graphically depicted as follows:



(5) Consequences of using least squares estimation when $\Sigma \neq \sigma^2 I$.

- (a) Least squares estimators will still be unbiased, consistent, and normally distributed, but will not be minimum variance estimators (Σ -known).



(b) The t and F statistics reported will not be distributed as $t(N - K)$

and $F(K - 1, N - K)$ because the wrong formulas $(s^2(X'X)^{-1})$

will be used to estimate the variance $(X'X)^{-1} X' \Sigma X (X'X)^{-1}$

$\hat{\beta}$. Hal White has proposed a consistent estimator of the var-

covariance matrix which will yield asymptotically appropriate t-statistics in this case. The command is operational in Shazam by listing HETCOV as an option to OLS.

(c) $\text{Var}(\hat{Y}_{\text{OLS}}) \geq \text{Var}(\tilde{Y}_{\text{MLE}})$

*For these reasons and others it is very important to perform tests of the assumptions of the model.

b. Heteroskedastic Random Disturbances

(1) Definition:

The heteroskedastic regression model corresponds to the situation in which the variances of the random disturbances are not constant. This situation frequently arises when working with cross-sectional data.

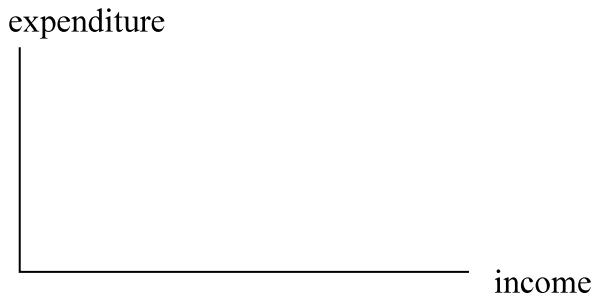
The model can be written as

$$Y = X\beta + \epsilon$$

where $\epsilon \sim N[0, \Sigma]$

$$\text{and } \Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & & \vdots \\ \vdots & \ddots & & \vdots \\ 0 & \cdots & & \sigma_n^2 \end{bmatrix}.$$

As an example, cross sectional analysis of consumption patterns often appear in the form:



Larger income levels are frequently seen to be associated with greater variation in expenditure levels.

(2) Statistical tests

There are many tests of the hypothesis of homoskedastic random disturbances,

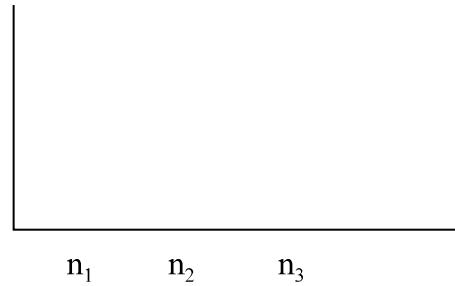
$$H_0: \sigma_1^2 = \sigma_2^2 = \dots = \sigma_N^2$$

These are based upon an analysis of estimated random disturbances (see exercise set number 1.b for a precautionary note).

The tests check for systematic behavior in the magnitudes of the random disturbances (or variances) and include the Goldfeld Quandt test, the Bartlet

test, graphical analysis, Park test, rank correlation tests, Glejser test and White test.

The Goldfeld Quandt test is an exact test (many of the others are asymptotic tests). The data is divided into three groups of approximately equal size.



Separate regressions are run on the first and third groups, yielding s_1^2 and s_3^2 .

Under the hypothesis of constant variances (homoskedasticity)

$$\frac{s_3^2}{s_1^2} \sim F(n_3 - K, n_1 - K)$$

$$F(n_3 - K, n_1 - K)$$

reject H_0 .

The second group isn't used for statistical reasons (power of the test) and separate regressions are used so that s_3^2 and s_1^2 will be distributed independently and the test statistic will be distributed as an F statistic. The larger of s_1^2 and s_3^2 is placed in the numerator of the test statistic.

Other tests for heteroskedasticity are based on an attempt to estimate relationships of the form $\sigma_t^2 = f(X_t)$. Recall that homoskedasticity implies that $\sigma_t^2 = f(X_t) = \sigma^2$ and σ_t^2 (even if observed) will not depend on X_t . Recall that $\sigma_t^2 = E(\varepsilon_t^2)$ and neither ε_t , σ_t^2 nor $f(\cdot)$ are observed.

- $e_t^2 = (Y_t - X_t \hat{\beta})^2$ is often used as a proxy for σ_t^2 (see eqn. 1).

for the notation used here).

- Alternative function forms for $f(X_t)$ have been considered in the literature.

The White test for heteroskedasticity [Econometrica, 1980, pp. 817-38] is based on using a second order Taylor Series approximation for the unknown function $f(X_t)$. This test is performed by regressing the squares of the OLS residuals (e_t^2) on an intercept, each of the X 's, the squares and cross products of the X 's. A Lagrangian multiplier (LM) test is used to test for the collective explanatory power of the X 's, i.e.,

$$\boxed{LM = nR^2 \approx \chi^2 \left(\frac{(K-1)(K+2)}{2} \right)}$$

For example, if the original model was

$$Y_t = \beta_1 + \beta_2 X_{t2} + \beta_3 X_{t3} + \varepsilon_t$$

the White test would involve using the OLS residuals

$$e_t^2 = (Y_t - \hat{\beta}_1 - \hat{\beta}_2 X_{t2} - \hat{\beta}_3 X_{t3})^2$$

in the regression

$$e_t^2 = \delta_1 + \delta_2 X_{t2} + \delta_3 X_{t3} + \delta_4 X_{t2}^2 + \delta_5 X_{t3}^2 + \delta_6 X_{t2} X_{t3}$$

and using

$$LM = nR^2 \approx \chi^2(5).$$

A lack of statistical significance of LM is consistent with the null hypothesis of homoskedasticity.

The Modified White Test is similar to the White Test except that the squares of the estimated residuals are regressed on the predicted Y's and the squares of the predicted Y's. The corresponding LM test is distributed as a Chi-square with two degrees of freedom. An obvious advantage of the Modified White Test, over the White Test, is in applications characterized by many explanatory variables.

The Park and Glejser tests correspond to using $|e_t|$ as a proxy for σ_t (or e_t^2 for σ_t^2) and then relationships of the form $|e_t| = f(X_t)$ or $e_t^2 = g(X_t)$ are estimated for various functions $f(\cdot)$ or $g(\cdot)$.

Rank correlation tests are based upon correlation between the magnitude of $|e_t|$ and the explanatory variables.

In Shazam, the command "DIAGNOS/HET CHOWTEST", immediately following the OLS estimation command, performs several tests for heteroskedasticity.

(3) Estimation.

Least squares estimation attributes equal weights to each observation and does not yield minimum variance estimators in the case of heteroskedastic errors. MLE attributes less "weight" to observations associated with large variances and can be obtained by applying least squares to

$$Y_t/\sigma_t = \beta_1(1/\sigma_t) + \beta_2(X_{t2}/\sigma_t) + \dots + \beta_K(X_{tK}/\sigma_t) + \varepsilon_t/\sigma_t$$

whose errors have constant variances.

The most difficult task is generally involved with the determination of the form of σ_t^2 . Note the variance of the transformed error term is constant.

Weighted least squares in STATA and Shazam perform this estimation.

STATA:

vwls dep_var indep_vars,std (where std is the name of the standard error estimated before running vwls which stands for variance weighted least squares)

c. Autocorrelation(1) Definition.

Autocorrelation exists if the random disturbances corresponding to different observations are correlated, i.e., the variance covariance matrix (Σ) is not diagonal. This situation frequently arises when working with time series. Autocorrelation can also arise from the use of an incorrect functional form or from the deletion of a relevant variable. It will be

assumed that an "appropriate" functional form has been selected and no relevant variables have been deleted.

The model can be written as

$$Y = X\beta + \varepsilon$$

where

$$\varepsilon \sim N[0, \Sigma]$$

and

$$\Sigma = \begin{bmatrix} \sigma^2 & \sigma_{12} & \dots & \sigma_{1N} \\ \sigma_{21} & \sigma^2 & \dots & \sigma_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{N1} & \sigma_{N2} & \dots & \sigma^2 \end{bmatrix}$$

The form of the σ_{ij} will depend upon the nature of the correlations between ε_i and ε_j , $\sigma_{ij} = \text{corr}(\varepsilon_i, \varepsilon_j)\sigma^2$.

One of the most commonly adopted models for autocorrelation is

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t$$

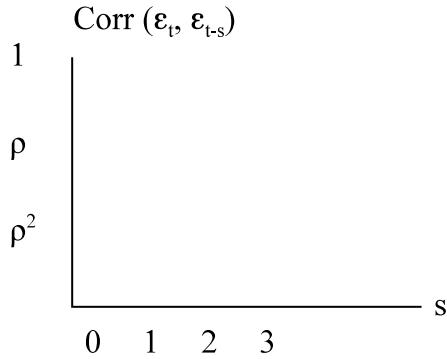
where $-1 < \rho < 1$, $u = (u_1, \dots, u_n)' \sim N(0, \sigma_u^2 I)$

This model is referred to as a first order autoregressive process, AR(1).

The corresponding Σ matrix can be shown to be

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{N-1} \\ \rho & 1 & \rho & \dots & \rho^{N-2} \\ \vdots & \vdots & & & \vdots \\ \rho^{N-1} & \rho^{N-2} & \dots & & 1 \end{bmatrix}$$

where $\sigma^2 = \sigma_u^2/(1-\rho^2)$. The correlation between ε_i and ε_j is given by $\rho^{|i-j|}$ which approaches zero as the observations get further apart. This might be graphically depicted with a correlogram.



The STATA command to plot the correlogram is

corrgram e

where “e” denotes the estimated errors from the regression model.

The reasons for the dominance of the AR(1) model in applications may be due to the availability of computer software and the widespread usage of annual data. It is no longer uncommon to have quarterly, monthly, or daily observations on some series. Consequently, other models for the behavior of autocorrelated error terms have been considered. These include:

Autoregressive models of order p, AR(p)

$$\varepsilon_t = \varphi_1 \varepsilon_{t-1} + \dots + \varphi_p \varepsilon_{t-p} + u_t;$$

Moving average model of order q, MA(q)

$$\varepsilon_t = u_t - \theta_1 u_{t-1} - \dots - \theta_q u_{t-q}; \text{ and}$$

Autoregressive moving average process or order (p,q)
ARMA (p,q)

$$\varepsilon_t = \varphi_1 \varepsilon_{t-1} + \dots + \varphi_p \varepsilon_{t-p} + u_t - \theta_1 u_{t-1} - \dots - \theta_q u_{t-q}.$$

These models will be considered in more detail in the section on time series analysis.

(2) Tests for Autocorrelation

Many tests for autocorrelation have been developed. These include a "signs test" which is based upon the number of times that the sign of the random disturbances (estimated) changes.

STATA reg y x's; predict e, resid; runtest e

Wooldridge proposes regressing the OLS residuals on the lagged estimated residuals and then using a "t-test" to determine the statistical significant of the estimated coefficient. This has problems if the estimated value of ρ is near one.

Probably the most common test for autocorrelated error terms of the AR(1) form is the Durbin Watson test statistic.

The Durbin Watson test statistic is defined by

$$D.W. = \sum_{t=2}^N (e_t - e_{t-1})^2 / \sum_{t=1}^N e_t^2$$

and can be rewritten as

$$\begin{aligned} D.W. &= \frac{\sum_{t=1}^N e_t^2 - 2\sum_{t=2}^N e_t e_{t-1}}{\sum_{t=1}^N e_t^2} - \frac{e_1^2 + e_N^2}{\sum_{t=1}^N e_t^2} \\ &= 2(1 - \hat{\rho}) - \frac{e_1^2 + e_N^2}{\sum_{t=1}^N e_t^2} \\ &\doteq 2(1 - \hat{\rho}) \end{aligned}$$

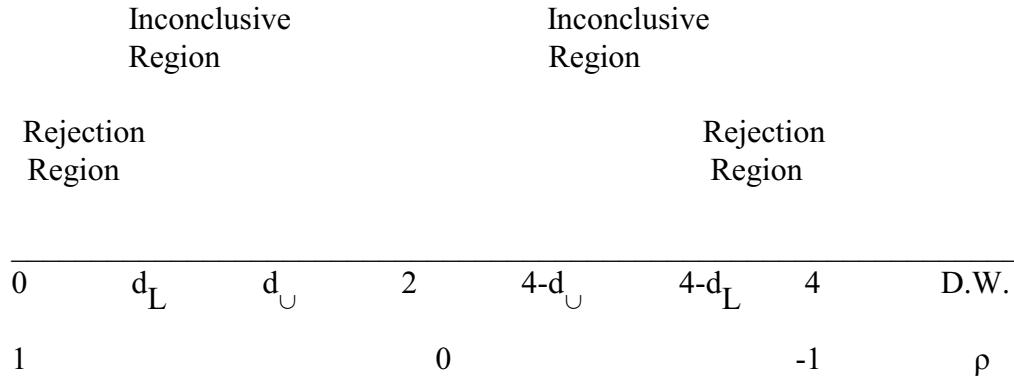
where $\hat{\rho} = \sum e_t e_{t-1} / \sum e_t^2$ is an estimator of the correlation between e_t and e_{t-1} . "Critical values" for D.W. are available and depend upon the " α level," sample size N and the number of slope coefficients ($K-1 = k'$).

The model is assumed to include an intercept and not include any lagged

dependent variables. For a given N, K-1 and α -level we obtain (d_L, d_U) .

The following figure is useful in testing for autocorrelation.

Fail to reject ($H_0: \rho = 0$)



This test statistic is not appropriate if the model includes a lagged dependent variable. In that case Durbin has proposed the h-test

$$h = \hat{\rho} \sqrt{\frac{N}{1 - N s_{\hat{\beta}_1}^2}}$$

where $s_{\hat{\beta}_1}^2$ denotes the least squares estimate of the variance of the coefficient of Y_{t-1} on the right hand side of the equation. The asymptotic distribution of h is $N(0,1)$.

The Breusch-Godfrey test can be used to test for higher autocorrelation.

The command, estat dwatson, following a STATA regression command will calculate the value of the Durbin Watson test statistic. The statistical significance of an exact D.W. Test (independent of the X matrix) is available in some other programs, but is not a Stata option.

(3) Estimation

Least squares estimators do not take account of the correlations between the random disturbances in the estimation process and will not be minimum variance; however, least squares estimators will still be unbiased and consistent.

If Σ is known the MLE, BLUE, GLS estimator of β is given by

$$\tilde{\beta} = (\mathbf{X}' \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}' \Sigma^{-1} \mathbf{Y} .$$

For random disturbances which are AR(1)

$$\Sigma = \frac{\sigma_u^2}{1-\rho^2} \begin{bmatrix} 1 & \rho & \dots & \rho^{N-1} \\ \rho & 1 & \dots & \rho^{N-2} \\ \vdots & \vdots & & \vdots \\ \rho^{N-1} & \dots & & 1 \end{bmatrix}$$

and it can be shown that Σ^{-1} is equal to

$$\Sigma^{-1} = \left(\frac{1}{\sigma_u^2} \right) \begin{bmatrix} 1 & -\rho & 0 & \dots & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1+\rho^2 & -\rho \\ 0 & 0 & 0 & \dots & -\rho & 1 \end{bmatrix} .$$

Let

$$T = \begin{bmatrix} \sqrt{1-\rho^2} & 0 & 0 & \cdots & 0 & 0 \\ -\rho & 1 & 0 & \cdots & 0 & 0 \\ 0 & -\rho & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\rho & 1 \end{bmatrix}.$$

Since

$$T'T = \sigma_u^2 \Sigma^{-1},$$

an application of least squares to the transformed model

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

$$TY = TX\beta + T\varepsilon$$

$$\begin{bmatrix} \sqrt{1-\rho^2} & Y_1 \\ Y_2 & -\rho Y_1 \\ \vdots & \\ Y_n & -\rho Y_{n-1} \end{bmatrix} = \begin{bmatrix} \sqrt{1-\rho^2} & \sqrt{1-\rho^2}X_{12} & \cdots & \sqrt{1-\rho^2}X_{1K} \\ 1-\rho & X_{22}-\rho X_{12} & \cdots & X_{2K}-\rho X_{1K} \\ \vdots & \vdots & \ddots & \vdots \\ 1-\rho & X_{N2}-\rho X_{N-12} & \cdots & X_{NK}-\rho X_{N-1K} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix} + \begin{bmatrix} \sqrt{1-\rho^2} & \varepsilon_1 \\ \varepsilon_2-\rho\varepsilon_1 & \\ \vdots & \\ \varepsilon_N-\rho\varepsilon_{N-1} & \end{bmatrix}$$

$$\text{yields } \hat{\beta}_T = (X^{*'}X)^{-1}X^{*'}Y^* = (X'T'TX)^{-1}X'T'TY$$

$\hat{\beta}_T$ is the maximum likelihood or generalized least squares estimator,

$$\hat{\beta}_T = (X'\Sigma^{-1}X'\Sigma^{-1}Y \quad \text{if } T'T \text{ is constructed to be } \Sigma^{-1}.$$

The Prais-Winsten approach is based upon using the N observations in Y^* , X^* . The Cochrane-Orcutt approach is based upon all but the first observation on Y^* and X^* and yields an approximation to

the MLE. When Σ (or ρ) is unknown (almost always) various iterative techniques are available to estimate ρ and β . For more general models for the random disturbances alternative estimation techniques are available. **STATA** will perform the Prais-Winsten or Cochran-Orcutt estimation, respectively, using the commands

```
tsset time_var
prais dep_var indep_vars
prais dep_var indep_vars, corc
```

(4) Forecasting (AR(1))

The interdependence between the random disturbances can be used in forming predictions to reduce the variance of conditional forecasts.

Recall

$$Y_t + X_t\beta + \varepsilon_t$$

$$\text{where } \varepsilon_t = \rho\varepsilon_{t-1} + u_t.$$

From (3.2) the minimum variance unbiased predictor is given by

$$Y_N(h) = X_{N+h} = X_{N+h}\tilde{\beta} + W'\Sigma^{-1}e$$

where

$$\begin{aligned}
 W &= E(\epsilon_{N+h} \epsilon) = E \begin{bmatrix} \epsilon_{N+h} \epsilon_1 \\ \vdots \\ \epsilon_{N+h} \epsilon_N \end{bmatrix} \\
 &= \sigma^2 \begin{bmatrix} \rho^{N+h-1} \\ \vdots \\ \rho^h \end{bmatrix}
 \end{aligned}$$

for an AR(1) model. Consequently, this simplifies to

$$\tilde{Y}_{N+h} = X_{N+h} \tilde{\beta} + \rho^h e_N$$

The second term takes account of the dependency of the error terms and approaches zero as h increases.

5. Auto Regressive Conditional Heteroskedasticity (ARCH) Models

a. Introduction and an example

Some times series are characterized by "clumps," "clusters," or groups of large residuals and groups of small residuals. This pattern of residuals has led to the development of autoregressive conditional heteroskedasticity (ARCH) models.

Graph:



ARCH models are extremely popular. A Google search for “ARCH models” on 2/3/2011 reported 7,240,000 results.

ARCH models attempt to model the behavior of the mean and variance. Many of the basic properties of ARCH models can be introduced by considering a simple ARCH model. Assume that

$$Y_t = X_t \beta + \varepsilon_t \quad (5.1)$$

$$\text{where } \varepsilon_t = u_t [\alpha_0 + \alpha_1 \varepsilon_{t-1}^2]^{\frac{1}{2}}, \text{ } u_t \text{ is iid } N[0,1]. \quad (5.2)$$

It follows that $E[\varepsilon_t | \varepsilon_{t-1}] = 0$.

$$\begin{aligned} E[\varepsilon_t] &= \int E[\varepsilon_t | \varepsilon_{t-1}] f(\varepsilon_{t-1}) d\varepsilon_{t-1} \\ &= 0 \end{aligned} \quad (5.3)$$

from the relationship of the joint and conditional pdf's.

Similarly,

$$\sigma_t^2 = \text{Var}[\varepsilon_t | \varepsilon_{t-1}] = E[\varepsilon_t^2 | \varepsilon_{t-1}] \quad (5.4)$$

$$\begin{aligned}
&= E(u_t^2) [\alpha_0 + \alpha_1 \varepsilon_{t-1}^2] \\
&= [\alpha_0 + \alpha_1 \varepsilon_{t-1}^2] \\
&= \text{Var}[Y_t | Y_{t-1}];
\end{aligned}$$

hence, conditional on ε_{t-1} , ε_t is heteroskedastic. However, the unconditional variance of ε_t is

$$\begin{aligned}
\text{Var}[\varepsilon_t] &= E[\text{Var}[\varepsilon_t | \varepsilon_{t-1}]] \\
&= \alpha_0 + \alpha_1 E[\varepsilon_{t-1}^2] \\
&= \alpha_0 / [1 - \alpha_1]
\end{aligned} \tag{5.5}$$

if the underlying process is variance stationary.

(5.4) can be written as

$$\begin{aligned}
\text{ARCH}(1): \quad \sigma_t^2 &= \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 && \text{and more generally} \\
\text{ARCH}(q): \quad \sigma_t^2 &= \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \dots + \alpha_q \varepsilon_{t-q}^2 \\
\text{GARCH}(q, p): \quad \sigma_t^2 - \delta_1 \sigma_{t-1}^2 - \dots - \delta_p \sigma_{t-p}^2 &= \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \dots + \alpha_q \varepsilon_{t-q}^2
\end{aligned}$$

Engle (1982) and Bollerslev (1986).

b. Estimation

Since assumptions A.1 - A.5 still hold, OLS will still be the minimum variance linear unbiased estimator. However, there will be a more efficient nonlinear estimation-the maximum likelihood estimator. The log-likelihood function for this model, given by

$$\ell = -\frac{1}{2} \sum_{t=1}^n \ln[(2\pi)(\sigma_t^2)] - \frac{1}{2} \sum_{t=0}^n \frac{\varepsilon_t^2}{\sigma_t^2} \tag{5.6}$$

$$= -\frac{1}{2} \sum_{t=1}^n \ln[(2\pi)(\alpha_0 + \alpha_1 \varepsilon_{t-1}^2)] - \frac{1}{2} \sum_{t=0}^n \frac{(Y_t - X_t \beta)^2}{(\alpha_0 + \alpha_1 \varepsilon_{t-1}^2)}$$

is maximized over the regression parameters (β) and the ARCH parameters $(\alpha_i's)$

STATA will perform this estimation with the commands

tset time_variable

arch depvar indep_vars, arch(order of ARCH, number of lagged errors squared) garch(order of GARCH process, number of lagged conditional variances), e.g., arch y x's, arch(1) garch(1)

Additonal model flexibility is obtained by selecting alternative distributions for the standardized error, u_t . The normal is the default, the student t and generalized exponential distributions can be used. The corresponding commands are

arch y x's, arch(p) garch(q) dist(t)

arch y x's, arch(p) garch(q) dist(ged)

- c. A test for ARCH disturbances can be performed by regressing the square of the OLS residuals on lagged values of the same variables and using and using a LM test

(nR^2) to test the null hypothesis of no ARCH effects

$(H_0: \alpha_1 = \alpha_2 = \dots = \alpha_p = 0)$ $\sigma_t^2 = \alpha_0 u_t$ which implies that

asymptotic chi square distribution ($\chi^2(df=p)$).

STATA can perform this test using the commands we have discussed earlier, “archlm”, “archlm, lags(p)”, or “archlm, lags(1/p)” following the reg y x's command.

d. Generalizations of the (G)ARCH models

Numerous generalizations of (G)ARCH models have been proposed in the literature which allow for asymmetries, nonlinearities, and other variations of the basic model.

Some of these specifications have catchy acronyms such as AARCH, SAARCH, TARCH, NARCH, PARCH, ABARCH, EGARCH, ATARCHE, among others.

6. Stochastic Regressors

This classical normal linear regression model is given by

$$Y = X\beta + \varepsilon \quad (6.1)$$

where

$$(A.1)' \varepsilon \sim N(0, \sigma^2 I)$$

(A.2)' X_t (rows of X) are nonstochastic and

$$\text{Limit}_{N \rightarrow \infty} \left(\frac{X'X}{N} \right) = \Sigma_{xx} \quad \text{is nonsingular}$$

We have already discussed variations in the first assumption (A.1)'.

This assumption provides a convenient foundation to develop the classical model. However, situations in which we can assume the X 's to be fixed in repeated samples are rare in economic modeling. This assumption is particularly problematic if any of the X 's are correlated with the error, as might be the case with endogenous regressors. We will look at a simple macro model containing an endogenous regressor and then formally outline the consequences of relaxing/violating A.5.

a. A simple macro model .

$$C_t = \alpha + \beta Y_t + \varepsilon_t$$

$$Y_t = C_t + Z_t$$

simple macro model

Dependent variables: C, Y

Independent variable(s): $Z=I+G+X$

Note: Y on the right handside of the consumption function is an endogenous variable and is referred to as an endogenous regressor

The corresponding reduced form equations, expressing each endogenous variable in terms of the exogenous or independent variables, are given by:

$$C_t = \frac{\alpha}{1-\beta} + \frac{\beta}{1-\beta}Z_t + \frac{\varepsilon_t}{1-\beta}$$

$$Y_t = \frac{\alpha}{1-\beta} + \frac{\beta}{1-\beta}Z_t + \frac{\varepsilon_t}{1-\beta}$$

Note: The endogenous regressor (Y_t) in the consumption function and ε_t are not independent since

$$\text{cov}(Y_t, \varepsilon_t) = \frac{\sigma^2}{1-\beta}$$

$$\text{Therefore, } \text{plim } \hat{\beta}_{OLS} = \beta + \frac{(1-\beta)\sigma^2}{\sigma_Z^2 + \sigma^2}.$$

This is also an example of the simultaneous equation problem where least squares estimators are biased and inconsistent.

One of the main lessons to be learned from this example is that if any of regressors (variables on the right hand side of the equation) have nonzero correlation with the random disturbances, then least squares estimators can be biased and inconsistent.

b. Formal analysis of the consequences of having stochastic X's.

We will consider two cases, (1) where the X's are stochastic, but uncorrelated with the disturbances and (2) where the X's are correlated with the disturbances.

(1) Case 1 of relaxing (A.2)'

(A.2)* The X_t 's are stochastic.

X_t and ε_t are stochastically independent.

$\text{plim } (X'X/N) = \sum_{XX}$ is nonsingular
 $N \rightarrow \infty$

The least squares estimator can be written as

$$\begin{aligned}\hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \boldsymbol{\varepsilon}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} \\ &= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\end{aligned}$$

Taking the expected value of $\hat{\beta}$ yields

$$E(\hat{\beta}) = \beta + E(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\boldsymbol{\varepsilon})$$

$$= \beta, \text{ hence } \hat{\beta} \text{ is unbiased.}$$

The variance of $\hat{\beta}$ is given by

$$\begin{aligned}\text{Var}(\hat{\beta}|\mathbf{X}) &= E(\hat{\beta} - \beta)(\hat{\beta} - \beta)' \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= ((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\end{aligned}$$

The consistency of $\hat{\beta}$ can be proven as follows:

$$\begin{aligned}\text{plim}_{N \rightarrow \infty} \hat{\beta} &= \beta + \text{plim}_{N \rightarrow \infty} (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} \\ &= \beta + \text{plim}_{N \rightarrow \infty} \left(\frac{\mathbf{X}'\mathbf{X}}{N} \right)^{-1} \frac{\mathbf{X}'\boldsymbol{\varepsilon}}{N} \\ &= \beta + \text{plim}_{N \rightarrow \infty} \left(\frac{\mathbf{X}'\mathbf{X}}{N} \right)^{-1} \text{plim}_{N \rightarrow \infty} \frac{\mathbf{X}'\boldsymbol{\varepsilon}}{N}\end{aligned}$$

Slutsky's Theorem

$$= \beta + \Sigma_{xx}^{-1} 0$$

$$= \beta$$

Consequently, if the regressors are stochastic but distributed independently of the random disturbances, the least squares estimators are still unbiased and consistent. However, the t and F statistics need not be valid for small samples since $\hat{\beta}$ will no longer be distributed normally due to the X's being stochastic.

(2) Case 2 of relaxing (A.2)

(A.2)** (a) X_t is stochastic

(b) X_t and ε_t are stochastically dependent and $\text{cov}(X_t, \varepsilon_t) \neq 0$.

(c) $\underset{N \rightarrow \infty}{\text{plim}} [(X'X)/N] = \Sigma_{xx}$ is nonsingular.

Thus,

$$E(\hat{\beta}) = \beta + E\{(X'X)^{-1}X'\varepsilon\}$$

$$\neq \beta$$

$$\underset{N \rightarrow \infty}{\text{plim}} (\hat{\beta}) = \beta + \underset{N \rightarrow \infty}{\text{plim}} (X'X)^{-1}X'\varepsilon$$

$$= \beta + \Sigma_{xx}^{-1} \text{Cov}(X\varepsilon)$$

$$\neq \beta$$

- **Thus, if the regressors and errors have nonzero correlation then the least squares estimators will be biased and inconsistent.**

- **We now discuss one of the most common methods of obtaining consistent estimators in the presence of stochastic X's.**

7. Instrumental Variables

a. Some background

Consider the generalized regression model

$$Y = X\beta + \varepsilon \quad (7.1)$$

where

$$(A.1) \quad \varepsilon \sim N(0, \Sigma)$$

Suppose that there exists a set of variables (N observations on K instrumental variables) $Z_t, Z = (Z'_1, \dots, Z'_N)'$ such that

$$\operatorname{plim}_{N \rightarrow \infty} (Z'X/N) = \Sigma_{ZX}$$

is nonsingular and

$$(Z'\varepsilon/N) \xrightarrow{p} N(0, \psi)$$

where \xrightarrow{p} means converges in probability to a normally distributed vector with

mean zero (null vector) and variance ψ .

This implies that

$$\operatorname{plim}(Z'\varepsilon/N) = 0.$$

b. The instrumental variables estimator (same number of Z's as X's, Z and X are NxK))

Now consider the estimator defined by the modified normal equations:

$$Z'Y = Z'X\beta_{IV} \quad \text{or } Z'\varepsilon = 0$$

$$\beta_{IV} = (Z'X)^{-1}Z'Y = \beta + (Z'X/n)^{-1}(Z'\varepsilon/n)$$

is referred to as the instrumental variables estimator of β based upon the instruments Z .

The motivation for the modified normal equations can be seen by multiplying equation (7.1) by $Z'/_N$, i.e.,

$$\frac{Z'Y}{N} = \frac{Z'X\beta}{N} + \frac{Z'\epsilon}{N}$$

with the last term converging to zero as N increases, leaving (7.4).

The asymptotic distribution of $\tilde{\beta}_{IV}$ is

$$\tilde{\beta}_Z \xrightarrow{a} N(\beta; \Sigma_{ZX}^{-1} \psi (\Sigma_{ZX})^{-1}/N)$$

(7.7)

If $Var(\epsilon) = \sigma^2 I$, then the variance of β_{IV} can be written as

$$\sigma^2 (Z'X)^{-1} (Z'Z) (X'Z)^{-1}$$

(7.8)

The bias of the IV estimator is given by

$$E((Z'X/n)^{-1} (Z'\epsilon/n))$$

The consistency of $\tilde{\beta}_Z$ follows from

$$\begin{aligned} \text{plim } \tilde{\beta}_Z &= \text{plim } (\beta) + \text{plim} \left(\frac{Z'X}{N} \right)^{-1} \text{plim} \left(\frac{Z'\epsilon}{N} \right) \\ &= \beta + \Sigma_{ZX}^{-1} 0 \\ &= \beta. \end{aligned}$$

c. Some special cases

Instrumental variables can be used to circumvent a number of problems in econometrics as well as providing a unified approach to many estimation problems.

(1) Least squares

If $Z=X$, then the corresponding estimator is the least squares estimator

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

(2) Generalized least squares

If Z is selected to be equal to $\Sigma^{-1}X$, then the associated estimator is given

by

$$\tilde{\beta}_Z = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y$$

which is the generalized least squares estimator.

estimator.

(3) “Projections of X on Z”

(where we have possible more instruments (m) than X's, $m \geq K$)

If the selected instruments are the projections of X on the variables Z ,

$$(Z(Z'Z)^{-1}Z')X , \text{ then the IV estimator will be given by}$$

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

which is asymptotically normal with variance-covariance matrix

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

- Not surprisingly, if $Z=X$ this estimator simplifies to the regular OLS estimator. This estimator can also be looked at as having resulted from a generalized least squares estimator of the parameters in equation (7.6).
- Also, if Z is $N \times K$, then $\hat{\beta}_{IV}$ simplifies to $(Z' X)^{-1} Z' Y$

with the result from the previous section, but differs from the projection result when $m > K$.

- IV estimators of the parameters in the model

$$y_1 = \beta_1 y_2 + \beta_2 y_3 + \gamma_1 x_1 + \gamma_2 x_2 + \varepsilon$$

can be obtained in STATA (11) using the command:

`ivregress 2sls y1 (y2 y3 = z1 z2 z3) x1 x2`

where y2 and y3 are endogenous regressors and z1, z2, and z3 are instrumental variables. There must be at least as many instrumental variables as endogenous regressors

Generalized method of moments (gmm) and limited information maximum likelihood (liml) are alternative estimators and may be used instead of 2sls.

The following postestimation commands may be useful in your analysis

estat firststage is used to explore the correlation of the instruments

with the endogenous regressors

estat overid **is used to test whether extra instruments are correlated with the error term.**

This particular estimator is an instrumental variables estimator and is also referred to as the two-stage least squares estimator which will be discussed in greater detail in another section.

d. Selection of instrumental variables. A valid instrumental variable (Z) is not included in the equation being estimated and should satisfy two conditions: relevance and exogenous

(1) Relevance: $\text{corr}(X, Z) \neq 0$

(2) Exogenous: $\text{corr}(\varepsilon, Z) = 0$

A **necessary condition** to perform IV estimation is that there must be at least as many instrumental variables (m) as there are endogenous regressors (k), $m \geq k$. The regression coefficients are said to be exactly identified if the $m=k$. IV estimators can't be obtained if $m < k$ (underidentified). The model is said to be overidentified if $m > k$. The validity of the overidentifying assumptions is tested using the “estat overid” command.

It is common practice to use an F-test to test for the “relevance” of the instruments. The endogenous regressor(s) is (are) regressed on the independent and instrumental variables. An F-test is then used to test for statistical significance of the instruments. The instruments are said to be “weak” if the F statistic (\tilde{F}) is

statistically small. Weak instruments can result in instrumental variables being worse than OLS. For one endogenous regressor, a rule of thumb is that the maximal bias of the IV estimator (2SLS) will not be greater than 10% of the OLS bias if $10 \leq \tilde{F}$.

A Hausman test (discussed in the next section) compares the OLS and instrumental (or other consistent estimator) variables estimator to test the endogeneity of the endogenous regressor.

- e. Instrumental variables estimation in quantile regression models. Chernozhukov and Hansen (*Econometrica*, 2005, 245-261) outline an approach for Instrumental Variables to be applied to quantile regression models which include endogenous regressors. Recall that quantile regression models attempt to model the impact that different variables will have on the distribution of a variable of interest such as what is the impact of the number of years of schooling on the distribution of income.

f. Some other issues

- Structural vs. experimental camps in econometrics. M. Keane (2010) wrote an interesting article in the Journal of Econometrics (“Structural vs. atheoretic approaches to econometrics,” 156 (1), pp. 3-20) making a case for the importance of making the underlying assumptions explicit. Insightful comments on this survey paper are provided by J. Rust, R. Blundell, and J. Heckman and Urzua.
- Numerous examples of instrumental variables can be found in recent literature. One example in modeling housing values with violent crime as an explanatory variable, some authors have used the number of murders as an instrumental variable. Some other references are given.
- Angrist, J.D., K. Graddy, and G.D. Imbens (2000). “The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish.” *Review of Economic Studies* 67, 499-527.
- Levitt, S. (1997). “Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime.” *American Economic Review*, June 1997.
- If the instruments are *weak* (not highly correlated with the rhs endogenous variables), problems can arise. In the case of endogenous regressors, OLS is inconsistent. IV estimation (or a variant) is often suggested in this case. However if the instrument is weakly correlated with the rhs endogenous variable, then even a small correlation between the instrument and the error can produce a larger inconsistency in the IV estimate of beta than in the OLS

estimator. An instrumental variable is often said to be *invalid* if it is correlated with the error term, regardless of the its correlation with the endogenous regressor. Bound, Jaeger, and Baker (1996, JASA, 443-450). This can be seen from an inspection of the bias of the IV estimator given after equation (7.8).

- Bekker (1994, Econometrica, 657-681) provides some approximations to the distributions of IV estimators which are useful in exploring the properties of alternative IV estimators.
- Donald, S.G and W. K. Newey, “Choosing the Number of Instruments,” Econometrica, 2001, 1161-1191. They use asymptotic expansions of the MSE of linear combinations of the coefficient estimators to explore the question of the optimal number of valid instruments which will minimize the MSE. They apply their methodology to the widely explored Angrist and Krueger data (1991, QJE, 979-1014) to estimate returns to schooling in the model

$$\ln(w) = \beta_0 + \beta_1(\text{years of school}) + \gamma's(9 - \text{year of birth}, 50 - \text{state of birth}) + \varepsilon$$

With instruments equaling subsets of

$$Z = (\text{quarter of birth}, QOB * SOB, QOB * YOB)$$

$\beta_1 \doteq .10$, for 2SLS, LIML with various combinations of instruments whereas the OLS estimates are between .6 and .7. The use of quarter of birth and interactions with year of birth is an interesting twist of this paper. Angrist

(1990, AER, 313-335) used lottery numbers as an instrument for military service.

- If the instruments are weak (not highly correlated with the rhs endogenous variables, then GMM and IV statistics are non-normal and hypothesis tests are unreliable. Stock, Wright, and Yogo (2002, JBES, 518-529). This paper also provides an excellent survey of related issues and alternatives.
- Kleibergen (September 2002, Econometrica, 1781-1803). Pivotal statistics for testing structural parameters in instrumental variables regression. Bad instruments not only lead to imprecise estimates of the structural parameters but also imply that the standard statistics we use to assess these estimates are unreliable.
- Angrist, Imbens, and Krueger (Journal of Applied Econometrics, 1999, 57-67) propose using Jackknife instrumental variables which will be, by construction, independent of the error terms in finite samples.

(g) Stochastic Regressors and IV estimation

The consumption function from the section on stochastic regressors can be written as

$$C = X\beta + \epsilon$$

or $\begin{pmatrix} C_1 \\ \vdots \\ C_N \end{pmatrix} = \begin{pmatrix} 1 & Y_1 \\ \vdots & \vdots \\ 1 & Y_N \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_N \end{pmatrix}.$

Recall that the Y_t are correlated with the random disturbances; hence, the least squares estimators will be biased and inconsistent.

It can be shown that selecting

$$Z = \begin{pmatrix} 1 & \hat{Y}_1 \\ \vdots & \vdots \\ 1 & \hat{Y}_N \end{pmatrix}$$

where $\hat{Y}_t = \hat{\Pi}_1 + \hat{\Pi}_2 Z_t$ (from macro example), the least squares predictions will yield consistent and asymptotically normal estimators of β_1 and β_2 . These estimators can be thought of as IV estimators and are frequently referred to as two stage least squares (2SLS) estimators.

8. Specification Tests in Econometrics

Jerry Hausman, *Econometrica*, 46(1978), pp. 1251-1270.

a. Background

Consider the model

$$Y = X\beta + \epsilon$$

where

$$(A.1)' \quad \text{var}(\epsilon | x) = \sigma^2 I \quad (8.1)$$

$$(A.2)' \quad E(\epsilon | x) = 0 \text{ or } \text{plim}_N \left(\frac{X\epsilon}{N} \right) = 0 \quad (8.1)$$

Violations of (A.1)' result in least squares estimators being unbiased, consistent, but not minimum variance. Violations of (A.2)' result in least squares being biased and inconsistent.

(A.2)' is very important and a few tests have been developed to investigate its validity: Wu, *Econometrica* (1973); Ramsey, *Frontiers in Econometrics*; Hausman, *Econometrica* (1978).

b. Outline of the Hausman test

Let $\hat{\beta}_0$ denote a consistent, asymptotically normal, asymptotically efficient

estimator of β if (A.1)' and (A.2)' are satisfied, but biased and inconsistent if (A.2)' is violated.

Let $\hat{\beta}_1$ denote an alternative estimator which is consistent under both the null and alternative hypothesis. Under these conditions the Hausmann test is based upon the difference,

$$\hat{q} = \hat{\beta}_1 - \hat{\beta}_0 .$$

If (A.1)' and (A.2)' are satisfied, then $\text{plim } \hat{q} = 0$. If (A.2)' is not satisfied, then $\text{plim } \hat{q} \neq 0$ and its (\hat{q}) asymptotic distribution provides the basis for rejecting (testing) (A.2)'. This is formalized by the theorem

(1) Let $\hat{\beta}_0$ and $\hat{\beta}_1$ denote consistent asymptotically normally estimators $\hat{\beta}_0$ with

attaining the Cramer Rao bound

$$\sqrt{n}(\hat{\beta}_0 - \beta) \xrightarrow{a} N(0, V_0)$$

$$\sqrt{n}(\hat{\beta}_1 - \beta) \xrightarrow{a} N(0, V_1)$$

The variance of the asymptotic distribution of $\hat{\beta}_1$ is V_1 / n , which goes to zero as n increases because of being a consistent estimator if the null hypothesis is valid.

The asymptotic distribution of \hat{q} ($\hat{q} = \hat{\beta}_1 - \hat{\beta}_0$) is given by

$$\sqrt{N}(q - 0) \xrightarrow{a} N(0, V(\hat{q}))$$

where $V(\hat{q}) = V_1 - V_0$.

The statistic

$$N\hat{q}' V(\hat{q})^{-1} \hat{q} \xrightarrow{a} \chi^2(\# \text{ par})$$

provides the basis for testing for specification error. **Perhaps an easier form for the Hausman Test is as follows:**

$$(\hat{\beta}_1 - \hat{\beta}_0)' \left(\text{Var}(\hat{\beta}_1) - \text{Var}(\hat{\beta}_0) \right)^{-1} (\hat{\beta}_1 - \hat{\beta}_0) \stackrel{a}{\sim} \chi^2 (\# \text{ slope pars})$$

In this form you don't need to worry about the sample size being in the formula.

Proof.

It can be shown that

- $[\hat{q} - 0]' \text{Var}^{-1}(\hat{q}) [\hat{q} - 0] \sim \chi (\# \text{ parameters})$
 - \hat{q} and $\hat{\beta}_0$ are uncorrelated.
 - $\hat{\beta}_1 - \hat{\beta}_0 = + \hat{q}$
 - $\text{Var}(\hat{\beta}_1) = \text{Var}(\hat{\beta}_0) + \text{Var}(\hat{q})$
 - $\text{Var}(\hat{q}) = \text{Var}(\hat{\beta}_1) - \text{Var}(\hat{\beta}_0)$
- $$= N [V_1 - V_0]$$

c. Applications: simultaneous equations, measurement error, panel data

(1) - Simultaneous equations

Simultaneous equations: Let $\beta_0 = \beta_{OLS}$ and $\beta_1 = \beta_{2SLS \text{ or } 3SLS}$.

β_{OLS} - efficient if A.1 - A.5 are valid

Inconsistent - if A.5 is violated

β_{2SLS} - consistent even if A.5 is violated

(2) Measurement error

Let $\beta_0 = \beta_{OLS}$ and $\beta_1 = \beta_{IV}$. Under the assumption of no measurement error, both estimators are consistent estimators of β with the OLS estimator being efficient. However in the presence of measurement error (in the X's) the OLS estimator is inconsistent and an appropriate IV estimator is consistent.

9. Seemingly Unrelated Regression Models (SURE-Zellner)

Consider the problem of estimating the coefficients in G separate regression equations.

$$\begin{aligned} Y_1 &= X_1\beta_1 + \varepsilon_1 \\ Y_2 &= X_2\beta_2 + \varepsilon_2 \\ &\vdots &&\ddots \\ &\vdots &&\ddots \\ Y_G &= X_G\beta_G + \varepsilon_G \end{aligned} \tag{9.1}$$

where Y_i denotes an $N \times 1$ vector of observations on the dependent variables in the i^{th} regression equation. X_i denotes the $N \times K_i$ matrix of observations on the explanatory variables and β_i denotes the $K_i \times 1$ vector of associated coefficients.

Assume that

$$\text{Var}(\varepsilon_i) = \sigma_{ii}^2 I_N \tag{9.2}$$

i.e., the random disturbances in each equation are uncorrelated over time and characterized by homoskedasticity.

If the random disturbances in each equation are independent of the random disturbances in all other equations, then least squares estimators of the β_i in each equation in (9.1) will yield MLE and BLUE of the β_i if

$$\varepsilon_i \sim N[0, \sigma_{ii}^2 I_N] .$$

$$\hat{\beta}_i = (X_i' X_i)^{-1} X_i' Y_i$$

$$\sim N[\beta_i; \sigma_{ii} (X_i' X_i)^{-1}] .$$

If the covariances between contemporaneous random disturbances in the i^{th} and j^{th} equation are given by σ_{ij} , i.e.,

$$\begin{aligned} E(\varepsilon_i \varepsilon_j') &= E \begin{bmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{iN} \end{bmatrix} \begin{bmatrix} \varepsilon_{j1} & \dots & \varepsilon_{jN} \end{bmatrix}' \\ &= E \begin{bmatrix} \varepsilon_{i1} \varepsilon_{j1} & \dots & \varepsilon_{i1} \varepsilon_{jN} \\ \vdots & & \vdots \\ \varepsilon_{iN} \varepsilon_{j1} & \dots & \varepsilon_{iN} \varepsilon_{jN} \end{bmatrix} \\ &= \begin{bmatrix} \sigma_{ij} & 0 & \dots & 0 \\ 0 & \sigma_{ij} & \dots & 0 \\ 0 & & & \sigma_{ij} \end{bmatrix} \\ &= \sigma_{ij} I. \end{aligned}$$

THEN, the Zellner Seemingly Unrelated Estimator (SURE) can yield unbiased estimators with smaller variances than the least squares estimators.

The SURE estimators can be defined in terms of an alternative representation of (9.1):

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_G \end{bmatrix} = \begin{bmatrix} X_1 & 0 & \cdots & 0 \\ 0 & X_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & X_G \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_G \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_G \end{bmatrix}$$

or

$$Y = X\beta + \epsilon \quad (9.7)$$

where Y , X , β and ϵ are implicitly defined in (9.6) and have dimensions $NGx1$, $NGx (K_1 + \dots + K_G)$, $(K_1 + \dots + K_G) \times 1$ and $NGx1$.

From (9.3) and (9.5),

$$\epsilon \sim N(0, \Omega = \sum \otimes I_N),$$

i.e., $\text{var}(\epsilon) = \sum \otimes I_N = \Omega$

$$\begin{aligned} &= \begin{bmatrix} \sigma_{11}I_N & \sigma_{12}I_N & \cdots & \sigma_{1N}I_N \\ \sigma_{21}I_N & \sigma_{22}I_N & \cdots & \sigma_{2N}I_N \\ \vdots & \vdots & & \vdots \\ \sigma_{N1}I_N & \sigma_{N2}I_N & \cdots & \sigma_{NN}I_N \end{bmatrix} \\ &= \begin{bmatrix} \text{var}(\epsilon_1) & \text{cov}(\epsilon_1\epsilon_2) & \cdots & \text{cov}(\epsilon_1\epsilon_N) \\ \text{cov}(\epsilon_2\epsilon_1) & \text{var}(\epsilon_2) & \cdots & \text{cov}(\epsilon_2\epsilon_N) \\ \vdots & \vdots & & \vdots \\ \text{cov}(\epsilon_N\epsilon_1) & \text{cov}(\epsilon_N\epsilon_2) & \cdots & \text{var}(\epsilon_N) \end{bmatrix}. \end{aligned}$$

The generalized least squares estimator (Zellner SURE) is given by

$$\tilde{\beta} = (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} Y$$

$$\tilde{\beta} \sim N(\beta, (X' \Omega^{-1} X)^{-1})$$

(9.10)

The least squares estimator (9.4) can be rewritten in the notation (9.7) by

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

$$\hat{\beta} \sim N(\beta, (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Omega\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}).$$

The least squares estimator has larger variances than $\tilde{\beta}$ unless $\sigma_{ij} = 0$ for $i \neq j$ or $X_1 = X_2 = \dots = X_G$, i.e., the random disturbances in different equations are mutually independent or the independent variables and related observations are identical in each equation.

ESTIMATION:

The STATA commands for Zellner's Seemingly Unrelated Regression Estimation is

sureg (dep_var1 ind_vars1) (dep_var2 ind_vars2) or

sureg (dep_var1 ind_vars1) (dep_var2 ind_vars2), isure

where the option “isure” iterates the estimation process until convergence is achieved.

Comments:

1. If Σ is unknown, then consistent estimators of σ_{ij} can be obtained using

$$s_{ij} = \frac{1}{N - K} e_i' e_j$$

where $e_i = Y_i - X_i \hat{\beta}_i$

denotes the least squares estimates of the residuals.

2. The seemingly unrelated regression estimator will be the same as the least squares estimator if either

(a) $X_i = X_j$ for all i, j or

- (b) $\sigma_{ij} = 0$ for $i \neq j$
3. If $\sigma_{ij} \neq 0$ and the X_i matrices are not identical, then the variances of the least squares estimators will have variances which are at least as large as Zellner's SURE of β .
 4. A similar approach can be used to combine cross sectional and time series data.
 5. This approach can be modified to handle autocorrelation.

10. Models for independent cross sectional data and for panel data

Independent cross sectional data over time is obtained by conducting random samples at different points in time. For example, random samples of demographic data (age, household size, income, employment status, and educational attainment) over time would be referred to as being an independent cross sectional data set. The sample sizes might be the same or different.

Panel data refers observational data on individuals ($i, i= 1, 2, \dots m$) over time($t=1,2,\dots, T_i$)

(two dimensions) and might be denoted as (Y_{it}) . The panel data set is referred to as balanced if

every individual is observed for every point of time ($T_1 = T_2 = \dots = T_m = T$). Otherwise

panel data set is referred to as unbalanced. Observations for a given individual over time are time series; whereas, cross sectional data are observations for different individuals at a given point in time. In many applications, the data are for short periods of time, but include many individuals.

a. Models for independent cross sections over time

Models for these data sets can take a number of different forms. Perhaps the simplest representation is given by

$$Y_{it} = X_{it}\beta + \varepsilon_{it} \quad (1)$$

where X_{it} denotes a $1 \times k$ vector of observations on k -exogenous variables for i^{th} individual in

the t^{th} time period and where the marginal impact of the X 's on Y is assumed constant over individuals and time (including the intercept). This specification is sometimes called the **pooled**

model and may include binary variables to represent the time period. Let the model be rewritten in matrix form as

$$\begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ y_m \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \\ \cdot \\ \cdot \\ X_m \end{bmatrix} \beta + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \varepsilon_m \end{bmatrix}$$

or

$$Y = X\beta + \varepsilon$$

OLS estimates of β $\hat{\beta} = (X'X)^{-1}X'Y$, can be obtained with the command
reg y x's

Recall, that in the presence of heteroskedasticity OLS estimators are inefficient and have invalid t-statistics. The same tests (White, modified White, Breusch-Pagan, etc.) used in the regular regression model can be used to test for the present of heteroskedasticity in the pooled regression model. Robust standard errors can be used to obtain appropriate t-statistics using the command:

reg y x's, vce(robust, bootstrap, or jackknife)

b. Models for Panel data

With panel data the same individuals (persons, firms, countries, etc.) are followed over time. Thus in the model

$$Y_{it} = X_{it}\beta + \varepsilon_{it}$$

X_{it} and Y_{it} , for a given value of “I”, correspond to the observations on the dependent and independent variables over time for a given person.

Regular least squares can be used to estimate the coefficients which are assumed to be invariant over time and for different individuals. As we showed in a previous chapter, OLS will not be efficient if either autorcorrelation or heteroskedasticity exists. Generalized least squares (GLS), BLUE, or MLE can provide more efficient estimators in this case. We may also want to allow for different intercepts for individuals or time periods and this leads to random or fixed effects specifications.

- (1) GLS (generalized least squares estimators) can provide more efficient estimators than OLS. The formulas for the GLS estimators and corresponding variance-covariance matrix are given by

$$\begin{aligned}\tilde{\beta} &= (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} Y \\ Var(\tilde{\beta}) &= (X' \Omega^{-1} X)^{-1}\end{aligned}$$

where $\text{Var}(\varepsilon) = \Omega = \sum_{mon} \otimes I_{T_i \times T_i} \quad T_i \geq m$,

In order to obtain GLS (generalized least squares) estimators, simplifying assumptions about the variance of ε, Ω , need to be made and the nature of the longitudinal/panel data must be provided to STATA with the “**xtset**” command as follows:

xtset panel_var or
xtset panel_var time_var.

where *panel_var* denotes the individual identification code or group variable name and *time_var* is an index which represents the time variable which defines the panels being used.

This is similar to using the `tsset time_variable` command to alert Stata that time series are being used. To clear the `xt` settings, use the command **`xtset, clear`**

An alternative approach is to list $i(panel_var)$ $t(time_var)$ after the first `xt` commands that will be discussed next.

Different intercepts for the different panels and or time periods can be obtained using the command **`xtreg y x's i.panel i.time`**.

Various generalized least squares estimators of β , depending on the form of the variance-covariance of the error term, can be obtained with the “**`xtgls`**” command.

If there is heteroskedasticity across panels,

$$\Omega = \begin{bmatrix} \sigma_1^2 I & 0 & \dots & \dots & 0 \\ 0 & \sigma_2^2 I & \dots & \dots & 0 \\ \vdots & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & \dots & \sigma_m^2 I \end{bmatrix},$$

corresponding GLS estimators can be obtained using the command

`xtgls y x's, panels(hetero)`

If there is correlation across panels (cross-sectional correlation) of the form

$$\Omega = \begin{bmatrix} \sigma_1^2 I & \sigma_{1,2} I & \dots & \dots & \sigma_{1,m} I \\ \sigma_{2,1} I & \sigma_2^2 I & \dots & \dots & \sigma_{2,m} I \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \sigma_{m,1} I & \sigma_{m,2} I & \dots & \dots & \sigma_m^2 I \end{bmatrix},$$

the GLS estimator is obtained with the command (this can only be applied to balanced panels)

xtgls y x's, panels(correlated)

The command

xtgls y x's, igls

iterates the generalized least squares procedure until convergence is obtained.

STATA allows for autocorrelation within the panels. The STATA manual, (Longitudinal/Panel Data, version 10, p. 150) states that three options are allowed: "corr(independent) or no autocorrelation, corr(ar1) (serial correlation where the correlation parameter is common for all panels), or corr(psar1) (serial correlation where the correlation parameter is unique for each panel)." A couple of observations are in order: (1) xtgls y X's, panels(iid) corr(independent) is equivalent to regress y X's; (2) when corr(ar1) or corr(psar1) are specified the iterated GLS estimator does not converge to the MLE.

Some examples and variations include:

xtgls y x's, panel(hetero) **fit panel-data model with hetero across panels**

xtgls y x's, panels(correlated) **correlation and hetero across panels**

xtgls y x's, panels(correlated) igls **uses iterative gls**

xtgls y x's, panels(hetero) corr(ar1) hetero across panels and auto within panels

xtgls y x's, panels(iid) corr(psar1)

Tests for heteroskedasticity and autocorrelation

An LR test can be used to test for heteroskedasticity across panels.

The unrestricted (with hetero) model is estimated using the command

xtgls y X's, igls panels (hetero)

estimates store hetero

The restricted model (assuming homoskedasticity) is estimated using the command

xtgls Y X's

The degrees of freedom is $e(N-g)-1$, so LR test is obtained by typing

local df = e(N_g)-1

lrtest hetero ., df(`df')

Testing for autocorrelation.

A user-written program to test for autocorrelation was written by David Drukker.

This program, called xtserial, can be downloaded using the program

findit xtserial

net sj 3-2 st0039

net install st0039

The program is then executed by typing

xtserial y X's

A significant value of the test statistic suggests serial correlation

(2). Fixed and random effects specifications

The fixed and random effects representations are a little different than the form just considered, in particular, they can be represented as:

$$Y_{it} = X_{it}\beta + \alpha_i + \varepsilon_{it}$$

where the marginal impact of changes in the X 's are still assumed to be constant across individuals, i.e. the β 's are the same for each individual. The only difference in the

relationship across firms is in the intercept term. In fixed effects (fe) models the α_i are

unknown constants and in random effects models (re) models the α_i are random. OLS can

be used to estimate the unknown parameters in this form where binary variables are added to the set of exogenous variables.

STATA uses a slight variation on this formulation in estimation

$$\alpha_i = \alpha + v_i \quad v_i \text{ where the } v_i \text{ are estimated such that } \sum_i v_i = 0 \text{ that ;}$$

hence,
$$Y_{it} = \alpha + X_{it}\beta + v_i + \varepsilon_{it}$$

Consider taking the following averages of (3):

$$\bar{y}_i = \alpha + \bar{x}_i\beta + v_i + \bar{\varepsilon}_i \quad (4) \text{ (average over } i)$$

$$\bar{\bar{y}} = \alpha + \bar{\bar{x}}\beta + \bar{v} + \bar{\bar{\varepsilon}} \quad (5) \text{ (average over } i \text{ & } t)$$

Combining equations (3) and (4), and (3), (4), and (5), respectively, enables us to write

$$Y_{it} - \bar{y}_i = (X_{it} - \bar{x}_i)\beta + (\varepsilon_{it} - \bar{\varepsilon}_i) \quad (6)$$

$$Y_{it} - \bar{y}_i + \bar{\bar{y}} = \alpha + (X_{it} - \bar{x}_i + \bar{\bar{x}})\beta + v_i + (\varepsilon_{it} - \bar{\varepsilon}_i + \bar{v}) + \bar{\bar{\varepsilon}} \quad (7)$$

STATA's **fixed effects (within)** estimation procedure, `xtreg y x's, fe`, correspond to

estimating β in equation (6) or equation (7), as adding in the overall mean of y has no

impact on the estimates of β . Note, from (6) or (7), that coefficients of explanatory

variables which do not vary across time can't be estimated. Estimates of v_i , $\hat{v}_i = u_i$ are

obtained, but not reported, from (4) as $u_i = \bar{y}_i - \hat{\alpha} - \bar{x}_i \hat{\beta}$ with $\bar{x}_i \hat{\beta}$ correlation with

being reported.

Three R^2 's are reported:

$$R_{within}^2 = \text{corr}^2(y_{it} - \bar{y}_i, (X_{it} - \bar{x}_i)\hat{\beta}) \quad R^2, \quad \text{from}$$

$$R_{between}^2 = \text{corr}^2(\bar{y}_i, \bar{x}_i \hat{\beta}) \quad R^2 \quad \bar{y}_i, \bar{x}_i \quad \text{from regressing} \quad \text{on}$$

$$R_{Overall}^2 = \text{corr}^2(y_{it}, X_{it}\hat{\beta}) \quad R^2 \quad y_{it}, X_{it} \quad \text{from regressing} \quad \text{on}$$

regression

Least squares estimation with a dummy variable (LSDV) for the different intercepts is equivalent to running a fixed effects regression and yields estimates of the different intercepts in the fixed effects specification. This estimation is facilitated with the command

`xi:reg y x's i.firm or xi: reg y x's i.firm i.time`

where “firm” and “time” denote the cross sectional and time variables, respectively.

The hypothesis that there is no heterogeneity in the fixed effects or that the grouped effects are all the same ($\nu_i = 0$, for all i) , can be tested using a Chow Test by comparing the pooled and LSDV regressions as follows:

$$F(m-1, mT-m-K) = \left[\frac{(R_{LSDV}^2 - R_{Pooled}^2)/(m-1)}{(1-R_{LSDV}^2)/(mT-m-K)} \right]$$

where m = number of groups and T = length of time series. The results of testing this hypothesis are also reported Stata when estimating the fixed effects model.

STATA’s **between effects** estimators can be obtained by estimating equation (4) using the Stata command,

`xtreg y x's, be`

The same R^2 reported with the fixed effects methods are reported for the between effects printouts, with the $R_{Between}^2$ corresponding to the fitted model with this estimation procedure.

In the **random effects** model the ν_i in the regression model

$$Y_{it} = \alpha + X_{it}\beta + \nu_i + \varepsilon_{it}$$

are assumed to be distributed identically and independently with mean zero and constant variance. The term $(\nu_i + \varepsilon_{it})$ can be thought of as a composite error term with

$$Var(\alpha_i + \varepsilon_{it}) = \sigma_s^2 I_T + \sigma_u^2 i_T i_T' = \Sigma \quad \varepsilon \quad \Omega = I_m \otimes \Sigma \quad \text{and var(}) =$$

obtain the desired estimators using the command, **xtreg y x's, re**. A maximum likelihood estimator could be used, using **xtreg y x's, mle**, and will generally give results similar to those obtained from **xtreg y x's, re** unless $\sum_i T_i$ is small. **f the** are uncorrelated with

the explanatory variables, random effects estimators will be efficient; however, they will be inconsistent if the ν_i are correlated with the explanatory variables. Random

effects estimation can estimate coefficients of explanatory variables which do not change over time for individuals. The output associated with random effects estimation includes the same three R^2 's as with fixed effects (within groups) and between groups estimation.

The fixed effects estimator is consistent whether the data are generated by a fixed effects model or a random effects model; however, it is less efficient than the random effects estimator if the data generating process is a random effects model. A Hausman test can be used to test the null hypothesis that the data are generated by a random effects model (there is no correlation between the intercepts and the explanatory variables). The Stata commands for performing a Hausman test are given by

xtreg y x's, fe

estimates store fe

xtreg y x's, re

estimates store re

hasuman fe re

In summary, the STATA commands for estimating fixed (within), between, and random effects models, respectively, are given by

xtset panel_var or xtset panel_var time_var

xtreg y x's, fe

xtreg y x's, be

xtreg y x's, re

A few general comments (some taken from Kennedy, 6th edition):

- (1) Since most panel data sets have a time dimension, time series problems such as unit roots or cointegration may arise and should be tested for.
- (2) Fixed effects estimation is OLS estimation when using the fixed effects model and the random effects model is actually GLS applied when using the random effects model.
- (3) Fixed and random effects estimators assume that the slopes are equal across-sectional units.
- (4) When there are intercept differences across individuals and time periods, the model is referred to as a two-way effects model to distinguish it from a one-way effects model where the intercepts only differ across individuals (or time). Reminder: xi: reg y x's i.firm i.time

(5) A Chow test can be used to test whether the intercepts are the same by using OLS on the pooled data and comparing the results to those obtained from fixed effects estimation. If the intercepts do not differ, OLS on the pooled data is preferred. An alternative approach is to use a Lagrange multiplier (LM) test to see if the variance of the intercept component of the composite error term is zero.

(6) Random effects is recommended when the composite error term is uncorrelated with the explanatory variables. The Hausman test can be used to explore this issue.

(7) The estimated standard errors can be sensitive to underlying assumptions with obvious implications relative to tests of hypotheses and statistical significance. Hence, you may want to consider correcting for clustered standard errors using the commands

xtreg y x's, fe cluster(cluster variable)

reg y x, cluster(cluster variable)

where the cluster variable might be the firm code.

Some additional Stata options:

(1) vce (variance covariance estimator) options:

vce(oim) observed information matrix

vce(opg) outer product of the gradient (OPPG) vectors

vce(robust) Huber/White sandwich estimator

vce(bootstrap [, bootstrap_options]) bootstrap estimation

vce(jackknife {,jackknife_options}) jackknife estimation

- (2) The command “**xtregar y x’s, re or fe**” can be used to estimate fixed effects or random effects models when the error term is characterized by a first order autoregressive process.

(3) **xtpcse. panel-corrected standard errors**

Format: xtpcse depvar [indepvars] [if] [in] [weight] [, options]

Description: xtpcse calculates panel-corrected standard error (PCSE) estimates for linear cross-sectional time-series models where the parameters are estimated by OLS or Prais-Winsten regression. When computing the standard errors and the variance-covariance estimates, xtpcse assumes that the disturbances are, by default, heteroskedastic and contemporaneously correlated across panels.

Options	description
noconstant	suppress constant term
correlation(independent)	use independent autocorrelation structure
correlation(ar1)	use AR1 autocorrelation structure
correlation(psar1)	use panel-specific AR1 autocorrelation structure
rhotype(calc)	specify method to compute autocorrelation parameter; see Options for details; seldom used
np1	weight panel-specific autocorrelations by panel sizes
hetonly	assume panel-level heteroskedastic errors
independent	assume independent errors across panels

(4) **xtivreg**

This command is used for fitting panel-data models in which some of the right-hand side variables are endogenous.

(5) xtlogit, xtpoisson, xtprobit, xttobit are available options

(6) Numerous variations are possible, e.g., consider

$$Y_{it} = \alpha + X_{it}\beta + v_i + \gamma_t + \varepsilon_{it}$$

which allows for cross-sectional effects and time contrasts.

(7) xtsum [varlist] [if] [, i(varname_i)] xtsum, is a generalization of summarize, reports means and standard deviations for cross-sectional time-series (xt) data; it differs from summarize in that it decomposes the standard deviation into between and within components.

(8) areg y x's, absorb(var_name)

Stata command which runs a linear regression with many binary variables which are absorbed into the designated var_name.

(9) A special edition of the Journal of Econometrics (editited by Baltagi, Kelejian, and Prucha(140, 2007) focuses on an analysis of spatially dependent data discusses related issues of identification, estimation, and testing.

An example (also see exercise in the problem set)

Consider the data set (STATATEST.txt) (from the STATA website)

t	code	x	y	d1	d2	d3	d4
1	1	0	-5	1	0	0	0
2	1	8	23	1	0	0	0
3	1	14	44	1	0	0	0
4	2	10	29	0	1	0	0
5	2	16	26	0	1	0	0
6	3	4	17	0	0	1	0
7	3	11	17	0	0	1	0
8	3	5	31	0	0	1	0
9	4	18	50	0	0	0	1
10	4	5	26	0	0	0	1
11	4	2	17	0	0	0	1

This data set is an unbalanced panel.

Consider the output corresponding to the following commands:

`reg y x d1 d2 d3`

`xtreg y x,fe`

Note that (1) the estimated coefficient for x and standard errors are the same

`xtreg y x,be`

`xtreg y x,re`

c. Difference in differences

Consider the impact of an experiment in a particular market. In setting up the experiment there will be treatment and control groups and observations before and after the experiment. A simple model to estimate the impact of the experiment might be written as

$$y_{it} = \beta_1 + \beta_2 treat_{it} + \beta_3 after_{it} + \beta_4 (treat_{it} after_{it}) + \varepsilon_{it}$$

where $treat = 1$ if in the experimental group, 0 otherwise; $after=1$ if after the experiment and 0 if before the experiment. The following table summarizes the expected levels for different combinations of experiment/control groups and before/after, along with the corresponding marginal impacts.

	Treatment Group	Control Group	Difference
Before the experiment	$\beta_1 + \beta_2$	β_1	β_2
After the experiment	$\beta_1 + \beta_2 + \beta_3 + \beta_4$	$\beta_1 + \beta_3$	$\beta_2 + \beta_4$
Difference	$\beta_3 + \beta_4$	β_3	β_4

Thus, the coefficient of the interaction term (β_4) captures the before/after effect of the experiment and is equal to the differences of before and after experiment and treatment and control group levels. Obviously, other controls could be added to the regression model. Card and Krueger (1994, AER) use this approach to investigate the impact of minimum wage on employment in the fast food industry in New Jersey and Pennsylvania. The before/after takes account of a national

recession. Another example might be the impact of Craig's list on the housing in the BYU/UVU rental market. Let CL=1 if listed on Craig's list, 0 otherwise and BYU=1 if the rental unit qualifies for BYU student rental and 0 otherwise. A possible model might be

$$\text{Re}nt_{it} = \beta_1 + \beta_2 CL_{it} + \beta_3 BYU_{it} + \beta_4 (CL_{it}BYU_{it}) + \varepsilon_{it}$$

where β_4 denotes the impact of Craig's list on the expected rental price of BYU housing.

These formulations need to be adjusted if an endogenous variable is included as a regressor and also if the data are serially correlated. Instrumental variables provide one approach to the endogeneity problem. See Bertrand, Duflo, and Mullainathan (2004, QJE) for a discussion of the inconsistency of standard errors in the presence of serial correlation.

d. Statistical Inference

- See Chris Hansen ("GLS inference in panel and multilevel models with serial correlation and fixed effects," Journal of Econometrics, 140(2007), 670-694) for a summary of some of the issues, literature review, and suggested solutions

11. Regression Discontinuity Formulations

Regression Discontinuity (RD) were introduced by Thistlethwaite and Campbell (1960) as a way of estimating treatment effects in a non-experimental setting where a treatment is determined depending on whether a variable (forcing variable) is above a threshold or cutoff point. The RD design has become an increasingly popular method of exploring a number of empirical problems, including questions dealing with education and health issues.

For pedagogical purposes assume that all students who score above 1000 on their SAT's attend an IVY league school and all students scoring below 1000 attend Slippery Rock State. To determine the value of an IVY league degree in the market place one might be tempted to compare average salaries of graduates; however, doing so would fail to account for differences in intelligence, family background, and a host of other factors.

An approach which could account for inherent differences in the educational programs is to select individuals who scored just above and just below 1000. These individuals are likely similar across other dimensions, only differing slightly in their SAT performance. The essential features of the RD design is (1) a continuous forcing or sorting variable, (2) all persons on one side of the cutoff are assigned to one group, and (3) all persons on the other side are assigned to another group.

This model could be formulated as follows:

$$Y_i = \alpha + X_i\beta + D_i\gamma + \varepsilon_i$$

where Y denotes salary, X represents a vector of other explanatory factors, and $D_i = 1$ if the forcing variable (SAT score in this example) is greater than 1000 and 0 otherwise. The coefficient of D would represent the value of an IVY league degree.

Lee and Lemieux (2009, NBER, working paper 14723) provide an excellent users guide to Regression Discontinuity models. A couple of interesting observations emerge from the work of Lee and Lemieux: (1) “When optimizing agents do not have precise control over the forcing variable, then the variation in the treatment will be as good as randomized in a neighborhood around the discontinuity threshold.” and (2) “The distribution of observed baseline covariates should not change discontinuously at the threshold.”

How could a RD model be used to determine the impact of a degree from BYU upon different measures of output such as salary and church activity?

12. Exercises: Generalized Regression Model

1. Consider the classical normal linear regression model

$$Y = X\beta + \varepsilon$$

where $\varepsilon \sim N(0, \sigma^2 I)$

and the X's satisfy (A.2).

Unbiased estimators of β and σ^2 are given by

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

$$s^2 = SSE/N-K$$

$$\text{where } SSE = (Y - \hat{Y})'(Y - \hat{Y}) = e'e$$

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

- a. Demonstrate that $R^2 = 1$ if $K=N$.

Hint: $R^2 = 1 - (SSE/SST)$, demonstrate that $Y = X\hat{\beta}$; hence $SSE=0$.

- b. Demonstrate that e can be written as

$$e = (I - X(X'X)^{-1}X')Y.$$

The distribution of the estimated residuals e is

$$N(0, \sigma^2 (I - X(X'X)^{-1}X'))$$

$$\begin{aligned} \text{because } E(e) &= (I - X(X'X)^{-1}X') E(Y) \\ &= (I - X(X'X)^{-1}X') X\beta \\ &= 0 \end{aligned}$$

$$\text{Var}(e) = (I - X(X'X)^{-1}X') \text{Var}(Y)(I - X(X'X)^{-1}X')$$

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

What can be said about the elements on the main diagonal of $\sigma^2(I - X(X'X)^{-1}X')$, are they equal? Will the off diagonal elements be zero? What implications does this result have for determining whether the unobserved random disturbances (ϵ) are homoskedastic or nonautocorrelated, based upon an analysis of the estimated random disturbances?

- c. Demonstrate that s^2 can be expressed as

$$\begin{aligned}
s^2 &= Y'(I - X(X'X)^{-1}X')Y/N-K \\
&= \epsilon'(I - X(X'X)^{-1}X')\epsilon/N-K.
\end{aligned}$$

- d. Recall that $I - X(X'X)^{-1}X'$ is symmetric and idempotent. Demonstrate that the rank of $I - X(X'X)^{-1}X'$ is $N - K$.

- e. Verify that

$$\frac{\epsilon'(I - X(X'X)^{-1}X')\epsilon}{\sigma^2} = \frac{Y'(I - X(X'X)^{-1}X')Y}{\sigma^2} = \frac{(N - K)s^2}{\sigma^2}$$

is distributed as a $\chi^2(N - K)$.

Note that this implies

$$\frac{(N - K)s^2}{\sigma^2} = \frac{(N - K)s^2 a_{ii}}{\sigma^2 a_{ii}} = \frac{(N - K)s_{\beta_i}^2}{\sigma_{\beta_i}^2} \sim \chi^2(N - K)$$

where a_{ii} denotes the i^{th} diagonal element of $(X' X)^{-1}$ and

f. Using the results in (e), demonstrate that s^2 is an unbiased estimator of σ^2 . Hint: Take the expected value of the first equation in problem (e)

g. Demonstrate that $\hat{\beta}$ and s^2 are independent. This will impl $\hat{\beta}$ and s^2 are

independent.

$$\text{Hint: } \hat{\beta} = (X' X)^{-1} X' Y = A Y$$

$$(N-K) s^2 = Y'(I - X(X' X)^{-1} X')Y = Y'BY.$$

(See Statistics B.2—Hint: $AB=0$)

h. Indicate what the distribution of the following statistics is and provide support for your answer:

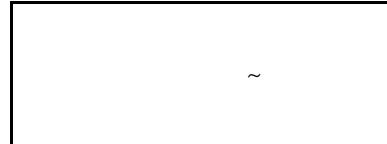
$$(1) \frac{\hat{\beta}_i - \beta_i}{s_{\hat{\beta}_i}} = \frac{(\hat{\beta}_i - \beta_i)/\sigma_{\hat{\beta}_i}}{\sqrt{\left(\frac{(N-K)s_{\hat{\beta}_i}^2}{\sigma^2} \right)/(N-K)}}$$



~

$$(2) \frac{R^2/(K-1)}{(1-R^2)/(N-K)} = \frac{SSR/(K-1)}{SSE/(N-K)}$$

$$= \frac{\left(\frac{SSR}{\sigma^2} \right)/(K-1)}{\left(\frac{SSE}{\sigma^2} \right)/(N-K)}$$



~

Hint: $SSR/\sigma^2 \sim \chi^2(K-1)$. Also see Statistics Notes (2. Multivariate Distributions, c.

Distribution Theory, (4) and (5))

2. In the generalized normal regression model

$$Y = X\beta + \epsilon$$

$$\epsilon \sim N(0, \Sigma)$$

where the X's satisfy A.2, the least squares and MLE of β are given by

$$\hat{\beta} = (X'X)^{-1}X'Y \sim N[\beta, (X'X)^{-1}X'\Sigma X(X'X)^{-1}]$$

$$\tilde{\beta} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y \sim N[\beta, (X'\Sigma^{-1}X)^{-1}].$$

Demonstrate that the estimators yield identical results and hence have the same distribution if the random disturbances are independent and have constant variance, i.e., $\Sigma = \sigma^2 I$.

3. Verify the following for the classical normal linear regression model ($\Sigma = \sigma^2 I$).

$$(a) \quad \frac{d\ell}{d\beta} = X'\epsilon/\sigma^2$$

$$(b) \quad E\left(\frac{d\ell}{d\beta} \frac{d\ell}{d\beta'}\right) = X'X/\sigma^2$$

$$(c) \quad -E\left[\frac{d^2\ell}{d\beta d\beta'}\right] =$$

(d) What is the relationship between (b), (c), the Cramer-Rao matrix and the variance of $\tilde{\beta}$?

(e) Obtain an expression for the sandwich estimator of the variance. The sandwich estimator is given by

$$\left(E \frac{d^2 \ell}{d\beta d\beta'} \right)^{-1} \left(E \frac{d\ell}{d\beta} \frac{d\ell}{d\beta'} \right) \left(E \frac{d^2 \ell}{d\beta d\beta'} \right)^{-1}$$

4. Consider the case of heteroskedasticity,

$$Y_t = \beta_1 + \beta_2 X_{t2} + \dots + \beta_K X_{tK} + \varepsilon_t$$

$$= (1, X_{t2}, \dots, X_{tK}) \beta + \varepsilon_t$$

$$= X_t \beta + \varepsilon_t$$

$$\text{where } \text{Var}(\varepsilon) = \begin{pmatrix} \text{Var}(\varepsilon_1) & 0 & \dots & 0 \\ 0 & \text{Var}(\varepsilon_2) & \dots & 0 \\ \vdots & & \ddots & \\ 0 & \dots & & \text{Var}(\varepsilon_N) \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & & 0 \\ \vdots & \vdots & & \vdots \\ 0 & \dots & & \sigma_N^2 \end{pmatrix}$$

Demonstrate that an application of least squares to the transformed model

$$Y_t/\sigma_t = \beta_1(1/\sigma_t) + \beta_2(X_{t2}/\sigma_t) + \dots + \beta_K(X_{tK}/\sigma_t) + \varepsilon_t/\sigma_t$$

$$= (X_t/\sigma_t)\beta + \varepsilon_t/\sigma_t$$

or

$$\begin{pmatrix} \frac{1}{\sigma_1} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_2} & 0 & \dots & 0 \\ \vdots & & \vdots & & \\ 0 & 0 & \dots & \frac{1}{\sigma_N} \end{pmatrix} \begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma_1} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_2} & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{\sigma_N} \end{pmatrix} \begin{pmatrix} 1 & X_{12} & \dots & X_{1K} \\ 1 & X_{22} & \dots & X_{2K} \\ \vdots & & & \vdots \\ 1 & X_{N2} & \dots & X_{NK} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{pmatrix}$$

$$+ \begin{pmatrix} \frac{1}{\sigma_1} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_2} & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & & \frac{1}{\sigma_N} \end{pmatrix} \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{pmatrix}$$

or $TY = TX\beta + T\varepsilon$

yields maximum likelihood estimators

$$\tilde{\beta} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y$$

Hint: $T'T$ is proportional to Σ^{-1} .

5. Verify that $T'T$ is proportional to Σ^{-1} for the case of random disturbances of the form

$$\text{AR}(1), \text{ hence } \hat{\beta}_T = \tilde{\beta}_{MLE}.$$

Optional problem. Now, delete the first row from T to form T^* and evaluate T^*T^* and compare this result with Σ^{-1} . What implications do these results have with respect to the relationship between the Prais-Winsten, Cochrane Orcutt and maximum likelihood estimators?

6. Consider the returns data in S&P500.dat

- a. Investigate the distributional characteristics (mean, variance, skewness, and kurtosis) of the returns data. You might consider using the command, "sum y,detail."

b. Use the reg command, reg y, to estimate the sample mean and to calculate deviations from the sample mean, with predict e, resid.

(1) Test for the presence of autocorrelation using

(a) the Durbin Watson test statistic and

(b) a t-statistic based on regressing the OLS estimated errors on their lags

(2) Reconcile your answers to (1) (a-b). You might use a non-parametric test, like the runs test, runtest e.

(3) Test for ARCH behavior using

(a) the estat command with one lag

(b) the estat command with fourteen lags

c. Taking account of possible ARCH/GARCH behavior, estimate the mean return.

7. Instrumental variables

Consider the model defined by

$$\log(wage) = \beta_1 + \beta_2 educ + \beta_3 exper + \beta_4 exper^2 + \varepsilon_t$$

In models like this it is often argued that education (educ) can be thought of as an endogenous regressor. One approach to circumventing this problem is to implement the method of instrumental variables. Two instrumental variables which have been considered are mother's education and father's education.

- a. Using the mroz data set (mroz.dta), estimate this regression model using
 - (1) OLS
 - (2) Instrumental variables estimation with Z=fatheduc
 - (3) Instrumental variables estimation with Z={motheduc, fatheduc}
- b. Compare the coefficients obtained from (1), (2), and (3) above and comment on the relative merits of each. Which estimates would you feel most comfortable with (why)?
- c. Comparing the results from (1) and (3), perform a Hausman test of the assumption that education is uncorrelated with the error term.
- d. An alternative to the Hausman Test is to (1) calculate the OLS residuals obtained from regressing education on experience, experience², mother's education, and father's education; (2) include these residuals as a potential regressor in the equation of interest (the log wage equation), and (3) perform a test of the statistical significance of the corresponding coefficient. Perform this test and compare the results with those obtained from the Hausman Test.

8. Verify that the least squares and SURE yield the same results if either

$$(a) \quad \Sigma = \sigma^2 I \text{ or}$$

$$(b) \quad X_i = X_j \text{ for all } i \text{ and } j.$$

Hint: (a) and (b) are equivalent to

$$(a)' \quad \Omega = \sigma^2 I \otimes I = \sigma^2 I$$

or

$$(b)' \quad X = \begin{pmatrix} X_1 & 0 & \cdots & 0 \\ 0 & & & 0 \\ \vdots & X_2 & & \vdots \\ 0 & 0 & \cdots & X_G \end{pmatrix} = I \otimes X_*$$

where $X_* = X_1 = \dots = X_G$

Hint: Use the properties of Kronecker products.

9. Use the data (statatext.txt) in the class notes for Generalized Regression.

It might be helpful to learn some STATA conversion procedures to create a STATA data file. Go onto Blackboard and copy the table on page 62 of the pdf to the Clipboard. Go into Excel and past the data. With the data blocked, go to Data? Text to Columns? Delimited? Space? Finish. Then the data will appear in columns. Copy it to the clipboard, open Stata, go to the Data Editor, paste in your data and close the data editor. As a former TA, Christopher Palmer would say, “Voila, you showed that text data whose boss.”

Using this data estimate and contrast the results obtained from the following commands:

`reg y x d1 d2 d3`

`xtgls y x, i(code)`

`xtgls y x d1 d2 d3, i(code)`

`xtreg y x,fe i(code)`

`xtreg y x,be i(code)`

`xtreg y x,re i(code)`

10. Another panel data problem

Consider the following data:

Year	Firm	Cost	Output	D1	D2	D3	D4	D5	D6
1955	1	3.154	214	1	0	0	0	0	0
1960	1	4.271	419	1	0	0	0	0	0
1965	1	4.584	588	1	0	0	0	0	0
1970	1	5.849	1025	1	0	0	0	0	0
1955	2	3.859	696	0	1	0	0	0	0
1960	2	5.535	811	0	1	0	0	0	0
1965	2	8.127	1640	0	1	0	0	0	0
1970	2	10.966	2506	0	1	0	0	0	0
1955	3	19.035	3202	0	0	1	0	0	0
1960	3	26.041	4802	0	0	1	0	0	0
1965	3	32.444	5821	0	0	1	0	0	0
1970	3	41.180	9275	0	0	1	0	0	0
1955	4	35.229	5668	0	0	0	1	0	0
1960	4	51.111	7612	0	0	0	1	0	0
1965	4	61.045	10206	0	0	0	1	0	0
1970	4	77.885	13702	0	0	0	1	0	0
1955	5	33.154	6000	0	0	0	0	1	0
1960	5	40.044	8222	0	0	0	0	1	0
1965	5	43.125	8484	0	0	0	0	1	0
1970	5	57.727	10004	0	0	0	0	1	0
1955	6	73.050	11796	0	0	0	0	0	1
1960	6	98.846	15551	0	0	0	0	0	1
1965	6	138.880	27218	0	0	0	0	0	1
1970	6	191.560	30958	0	0	0	0	0	1

- a. Generate $y = \log(\text{cost})$ and $x = \log(\text{output})$ and perform *pooled OLS (POLS)* by estimating

$$y_{it} = \beta_1 + \beta_2 x + \varepsilon_{it}$$

using the Stata command `reg y x`. This formulation assumes that the coefficients are the same over time and across firms, uncorrelated for a given individual, with the ε_{it} being homoskedastic over time and across individuals.

- b. Now relax the assumption of identical coefficients and allow different firms to have the same slopes, but with possibly different intercepts,

$$y_{it} = \beta_i + \beta_2 x + \varepsilon_{it} = \alpha + \beta_2 x + \alpha_i + \varepsilon_{it}$$

(where $\sum_i \alpha_i = 0$) .

- (1) Perform the following Stata estimations and explain the results

xtset firm

reg y x D1 D2 D3 D4 D5 or reg y x D1 D2 D3 D4 D5 D6, noconstant
xtreg y x, fe

This estimator is referred to as the *fixed effects estimator* or the *within groups estimator*. It can be obtained by estimating a regular regression with binary variables for the intercept or by regression the deviations of the dependent variable from their individual means on deviations of the explanatory variables from their individual means.

- (2) Test the hypothesis that the intercepts are the same for each firm. Use a Chow test and look for related information on the fixed effects printout.

- c. Another estimator that is sometimes considered is obtained by regressing the average y for each firm on the average x's for each firm. This estimator is called the *between-groups estimator* and can be obtained in Stata by using the command

xtreg y x, be

Perform this estimation and compare the results with those obtained in b(1).

- d. Consider the case where the α_i 's are not constant for each firm, but may vary. Then $Var(\alpha_i + \varepsilon_i) = \sigma_s^2 I_T + \sigma_u^2 i_T i_T' = \Sigma$ and $\varepsilon \sim \Omega = I_m \otimes \Sigma$ and var() assumes that the α_i 's ε 's, and x's are all independent. The corresponding GLS estimator is called the random effects estimator and can be obtained using the Stata command,

xtreg y x, re

Obtain the random effects estimators of the cost function.

- e. Use the Hausman test to explore the use of fixed vs. random effects estimators.

11. (Just for fun, but not required)

Consider the model defined by

$$Y = X\beta + U$$

where

$$\begin{aligned} U &= \sigma \operatorname{sign}(\mu) [\sinh(\theta v) - F(\theta, \mu)] / \theta \\ V &\sim N[\mu, 1] \\ E(\sinh(\theta v)) &= F(\theta, \mu) \\ &= e^{\theta^2/2} \sinh(\mu \theta). \end{aligned}$$

$\text{Sin } h(s)$ denotes the hyperbolic sin $[e^s + e^{-s}]/2$. This formulation allows for skewness and thick-tailed error distributions.

It can be shown that

$$\begin{aligned} E(U) &= 0 \\ \operatorname{Var}(U) &= \frac{\sigma^2 [e^{\theta^2} - 1][e^{\theta^2} \cosh(2\mu\theta) + 1]}{2\theta^2} \\ \operatorname{Skew}(U) &= E(\mu^3) = -\frac{\operatorname{sign}(\mu)\sigma^3}{4\theta^3} \{e^{\theta^2/2} (e^{\theta^2} - 1)^{1/2}\} \\ &\quad \cdot \frac{e^{\theta^2} (e^{\theta^2} + 2) \sinh(3\mu\theta) + 3 \sin(\mu\theta)}{(e^{\theta^2} \cosh(2\mu\theta) + 1)^{3/2}} \end{aligned}$$

This information is given to help place you in the top .001% of undergraduate economics majors in the world. The previous results can be obtained, with patience, from the result

$$E(e^{\theta v}) = e^{h^2\theta^2/2 + h\mu\theta}$$

where $V \sim N[\mu, \sigma^2 = 1]$.

generating function.

Derive this result. Hint: Consider the n

IV. Estimation Frameworks in Econometrics (GR: Ch. 14-18)

- 1. Parametric, semiparametric, and nonparametric specifications**
- 2. Maximum likelihood**
- 3. M-estimators and partially adaptive methods of estimation**
- 4. Generalized method of moments**
- 5. Extremum estimation**
- 6. Kernel estimation**
- 7. Empirical likelihood**
- 8. Homework exercises**

IV. Estimation Frameworks in Econometrics (GR: Ch. 14-18)

1. Parametric, semiparametric, nonparametric specifications

The literature on econometric modeling and methodologies refer to parametric, semi-parametric, and non-parametric methods. The following table (modified from the table in Mittlehammer, Judge, and Miller) summarizes attributes of these methods.

	Parametric	Partially adaptive (QMLE)	Semiparametric	Nonparametric
Model: Y	$Y_t = h(X_t, \beta) + \varepsilon_t$	$Y_t = h(X_t, \beta) + \varepsilon_t$	$Y_t = h(X_t, \beta) + \varepsilon_t$	$Y_t = h(X_t) + \varepsilon_t$
Systematic component	$h(X_t, \beta) = X_t\beta$ X_t known and fixed	$h(X_t, \beta) = X_t\beta$ X_t known and fixed	$h(X_t, \beta) = X_t\beta$ X_t known and fixed	* $h(X_t)$ is unspecified X_t known and fixed
Error term	ε_t 's are iid $N[0, \sigma^2]$	ε_t 's are iid $\sim f(\varepsilon_t; \theta)$	ε_t 's are independent with zero mean and variance σ^2	ε_t 's are independent with zero mean and variance σ^2
Parameters	$\beta \in R^k, \sigma^2 > 0$	$\beta \in R^k$, θ (dist. parameters) capture skew & kurt	$\beta \in R^k, \sigma^2 > 0$	$\sigma^2 > 0$

* If the function is not linear, then it is of a known functional form with possible unknown parameters in the parametric, QMLE, and semiparametric formulations; however, no functional form is specified in the nonparametric formulation.

The differences between the approaches depends upon the detail in the model specification. Parametric methods correspond to a complete specification of the functional form of the systematic component as well as a rather restrictive error distribution. The other extreme is a nonparametric specification in which neither the functional form of the systematic component nor the error distribution are specified. In the parametric and semiparametric specifications, the systematic component need not be linear as shown above, but only needs to be a specified functional form with possibly unknown parameters. The partially adaptive formulation is similar to the parametric specification, except that the error distribution may allow for flexibility in the skewness and kurtosis of the error distribution.

The parametric specification is often estimated using methods of maximum likelihood . Kernel methods can be used to estimate the functional form in nonparametric models. Kernel methods can also be used to approximate the unknown error distribution in semiparametric models. An intermediate step between the parametric and semiparametric specifications is that of partially adaptive or quasi maximum likelihood specifications in which the assumption of normally distributed errors is replaced with the assumption of of a more general flexible error distribution.

2. Maximum likelihood estimation

a. Theoretical development and main results

Assume that the underlying model is given by

$$y_i = h(X_i, \beta) + \varepsilon_i \quad , i = 1, 2, \dots, n ,$$

where the disturbances are independently and identically distributed with the probability density function $f(\varepsilon, \theta)$ and θ denotes a vector of unknown distributional parameters. We will review and contrast several alternative estimators. Recall that such a comparison is often based upon their relative bias and variance and whether they are consistent.¹ We now turn to the development of maximum likelihood estimation and then we briefly summarize the notion of extremum estimators.

The likelihood function associated with the random sample $\{y_t | t = 1, 2, \dots, n\}$ is given by

$$L(\psi; Y) = \prod_{i=1}^n f(\varepsilon_i = y_i - h(X_i, \beta); \theta)$$

with corresponding log likelihood function

$$\begin{aligned} \ell(\psi; Y) &= \ln L(\psi; Y) = \sum_{i=1}^n \ln f(y_i - h(X_i, \beta), \theta) \\ &= \sum_{i=1}^n \ell(\varepsilon_i; \Psi) \end{aligned}$$

¹ Recall that $\hat{\theta}$ is an *unbiased estimator* $E(\hat{\theta}) = \theta$.

Consistency requires that $p \lim_{n \rightarrow \infty} \hat{\theta} = \theta$ or $\lim_{n \rightarrow \infty} \Pr(|\hat{\theta} - \theta| \geq \varepsilon) = 0$ for any $\varepsilon > 0$

If the variance and bias of $\hat{\theta}$ approach zero, then $\hat{\theta}$ will be consistent; however, there are cases in which the variance/bias may not be defined and the estimator is still consistent. An unbiased estimator is said to be efficient if its variance is less than the variance of any other unbiased estimator. In the case of a vector of estimators, one would investigate whether the difference of the variance covariance matrices is positive definite. In the case of biased estimators, comparisons can be made using the mean squared error which is equal to sum of the variance and square of the bias, i.e. $MSE(\hat{\theta}) = VAR(\hat{\theta}) + BIAS^2(\hat{\theta})$.

where $\ell(\varepsilon; \psi)$, $\ell(\varepsilon_i; \psi) = \ln f(\varepsilon_i = y_i - h(X_i, \beta); \theta)$

(multivariate pdf) as a product of independent pdf's assumes that the observations are independent. If there is a correlation between the observations, then the likelihood function will be a multivariate pdf such as a multivariate normal, multivariate t, or some other multivariate pdf.

The maximum likelihood estimators (MLE) of the parameters ψ are implicitly defined by the equation

$$\frac{d\ell(Y; \psi)}{d\psi} = \begin{pmatrix} \frac{\partial \ell}{\partial \beta} \\ \vdots \\ \frac{\partial \ell}{\partial \theta} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} = 0.$$

These estimators will optimize the log likelihood function $\ell(y; \theta)$. It is possible that multiple solutions to $\frac{d\ell(Y; \psi)}{d\psi} = 0$ will exist; however, a unique maximum will exist if $\ell(Y; \theta)$ is concave in θ . The concavity can be investigated by determining whether the matrix

$$\frac{d^2\ell(Y; \psi)}{d\psi^2} = \begin{pmatrix} \ell_{11} & \dots & \ell_{1k} \\ \vdots & & \vdots \\ \ell_{k1} & \dots & \ell_{kk} \end{pmatrix}$$

is negative definite where $\ell_{ij} = \frac{\partial^2 \ell(Y; \psi)}{\partial \psi_i \partial \psi_j}$.

Under rather general regularity conditions the MLE of

ψ (*the distributional and model parameters*), $\hat{\psi}$

distribution. The regularity conditions will be given and the main results cited. The literature contains a number of alternative formulations of "regularity conditions." The following conditions are fairly typical.

Regularity Conditions:

Let Ψ denote the set of permissible values ψ

(R.1) For almost all y_i

$$\frac{d\ell(y_i; \psi)}{d\psi}, \frac{d^2\ell(y_i; \psi)}{d\psi^2}, \frac{d^3\ell(y_i; \psi)}{d\psi^3}$$

exist for all $\psi \in \Psi$

(R.2) There exist integrable functions $F_1(y)$, $F_2(y)$ and $F_3(y)$ such that

$$\frac{d\ell(y_i; \psi)}{d\psi} < F_1(y_i)$$

$$\frac{d^2\ell(y_i; \psi)}{d\psi^2} < F_2(y_i)$$

$$\frac{d^3\ell(y_i; \theta)}{d\psi^3} < F_3(y_i)$$

for all $\psi \in \Psi$

and

$$E(F_3(y)) = \int_{-\infty}^{\infty} F_3(y) f(y; \theta) dy < M$$

where M is independent of θ .

$$(R.3) E\left[\frac{\partial \ell(y_i; \psi)}{\partial \psi}\right]^2 = \int_{-\infty}^{\infty} \left[\frac{\partial \ln f(y; \theta)}{\partial \psi}\right]^2 f(y; \theta) dy < +\infty.$$

These assumptions are important in demonstrating the asymptotic normality of MLE of ψ . R.1

insures the existence of necessary Taylor Series expansions while R.2 permits differentiation

under an integral and (R.3) implies that the random variable(s) $\left\{ \frac{\partial \ln f(y_t; \theta)}{\partial \psi} \right\}$

will have a finite variance.

Theorem. Under the regularity conditions (R.1), (R.2) and (R.3) the MLE of ψ will be consistent and asymptotically normal.

$$\hat{\psi}_{MLE} \xrightarrow{a} N[\theta, \Sigma_{\hat{\psi}_{MLE}}]$$

where

$$\begin{aligned} \Sigma_{\hat{\psi}_{MLE}} &= - \left(\frac{E d^2 \ell}{d\psi d\psi'} \right)^{-1} \\ &= \left(E \left(\frac{d\ell}{d\psi} \right) \left(\frac{d\ell}{d\psi} \right)' \right)^{-1}. \end{aligned}$$

This distributional result can be alternatively written

$$\sqrt{n}(\hat{\psi}_{MLE} - \psi) \xrightarrow{a} N[0, n\Sigma_{\hat{\psi}_{MLE}}].$$

2. A simple example

Recall from the statistics section what an asymptotic distribution is. Example from previous homework. MGF for sample mean (exponential elements) $(1 - \beta t/n)^{-n}$

	Exact	Asymptotic
	$GA(z; \beta/n, p=n)$	$N(\beta, \beta^2/n)$
Mean	β	β
Variance	β^2/n	β^2/n
Skewness	$2\beta^3/n^2$	0
Kurtosis	$6\beta^4/n^3 + 3\beta^4/n^2$	$3\beta^4/n^2$

3. M-estimators and partially adaptive methods of estimation

A large body of statistics and econometric literature, summarized in Koenker (1982), discusses robust regression estimation. Huber (1981), Hampel (1986) and Rey (1983) overview many topics related to robust statistics. Much of the material in this literature can be structured in terms of influence functions and M-estimators.

In order to consider robust regression, first define the simple regression model as

$$y_i = X_i \beta + \varepsilon_i \quad , \quad i = 1, 2, \dots, n ,$$

where the y_i and X_i denote the i -th observation on the dependent variable and a $1 \times K$ vector of independent variables. The β represents a $K \times 1$ parameter vector of regressor coefficients. The ε_i are independently, identically distributed random variables and are also independent of X_i .

a. M-Estimators.

As commonly described in the literature, the M-estimation technique selects an estimator of β which minimizes a function of the residuals $\rho(\varepsilon_i)$, i.e.,

$$\min_{\beta} \sum \rho(Y_i - X_i \beta - \varepsilon_i)$$

The solution to this minimization problem yields an influence function which follows the Hampel (1974) definition. In the notation of the above linear model, the influence function is given by:

$$\psi(\varepsilon_i) = \rho'(\varepsilon_i)$$

The ψ function measures the "influence" that a residual will have in the estimation process.

M-estimators include

Estimator	$\rho(\cdot)$	$\psi(\cdot)$
OLS	$\rho(\varepsilon) = \varepsilon^2$	$\psi(\varepsilon) = 2\varepsilon$
LAD	$\rho(\varepsilon) = \varepsilon $	$\psi(\varepsilon) = \text{sign}(\varepsilon)$
L_p	$\rho(\varepsilon) = \varepsilon ^p$	$\psi(\varepsilon) = p\varepsilon^{p-1} \text{ sign}(\varepsilon)$

$$\text{sign}(\varepsilon) = 1 \text{ if } (\varepsilon > 0) \text{ and } -1 \text{ if } (\varepsilon < 0)$$

Consider the shapes of these influence functions. Note that large disturbances or outliers will have a large impact on the OLS estimators; whereas, LAD is not as sensitive to outliers. The L_p estimators obviously include the OLS and LAD estimators as special cases, corresponding to $p=2$ and $p=1$. OLS, LAD, and L_p correspond to MLE with a normal, Laplace, and GED, respectively. Maximum likelihood estimators will be obtained if the p -function is the negative of the natural logarithm of the density of the residuals. These estimators are also referred to as quasi-maximum likelihood estimators (QMLE). They may offer some advantages over arbitrarily selected OLS, LAD, L_p or other estimators. However, the optimality properties of partially adaptive estimators is conditional on the degree of agreement between the assumed and actual pdf for the errors.

STATA allows for various forms of M-estimation. The general form of the *robust regression* command is

rreg y x's, options

A common variation is the command

qreg y x's

which yields the LAD, LAE, or MAD estimator obtained by selecting the regression estimates to minimize the sum of absolute values of the residuals.

Options: `tune(#)` use # as the biweight tuning constant; default is `tune(7)` meaning 7 times the median absolute deviation from the median residual (MAD). Lower tuning constants downweight outliers rapidly but may lead to unstable estimates (below 6 is not recommended). Higher tuning constants produce milder downweighting.

See **rreg postestimation** for additional capabilities of estimation commands.

The section of the course, “Nonlinear models”, will outline how STATA can be used to obtain additional M-estimators.

b. Partially adaptive estimators.

Partially adaptive estimators allow some additional flexibility to distributional characteristics of the errors if the estimators are defined by

$$\min_{\beta, \varphi} \sum \rho(\varepsilon_i = y_i - X_i \beta; \theta)$$

$$= -\min_{\beta, \varphi} \sum \ln f(\varepsilon_i; \theta)$$

where θ denotes the distributional parameters for the error distribution. Thus the estimation procedure may be able to accommodate diverse distributional assumptions. Optimizing the objective function over the regression and distributional parameters allows the influence function to *adapt* to the distributional characteristics. For example, selecting the pdf to be the GED not only includes the OLS, LAD, and L_p estimators, but endogenizes the selection of the parameter p if the objective function is optimized over p .

Partially adaptive estimation of regression models requires the use of nonlinear optimization algorithms which will be covered in the section on nonlinear models.

4. Generalized Method of Moments

a. Basic Theory

Recall that the **method of moments** (MOM) estimators of the parameters in the data generating process

$$\text{DGP: } f(Y, \theta)$$

are obtained by solving

$$g(\hat{\theta}) = \hat{m} - m(\hat{\theta}) = 0$$

where $\hat{m} = (\hat{m}_1, \hat{m}_2, \dots, \hat{m}_m)'$ & $m(\theta) = (m_1(\theta), m_2(\theta), \dots, m_m(\theta))'$

with $\hat{m}_h = \frac{1}{n} \sum_{i=1}^n y_i^h$, (sample moments), and $m_h(\theta) = E(Y^h)$, (theoretical moments)

$h = 1, 2, \dots, m$, the number of moments (m) used in estimation (number of equations) is the same as the number of parameters (p). If a solution exists, then not only will $g(\hat{\theta}) = \hat{m} - m(\hat{\theta}) = 0$

$$g(\hat{\theta})' g(\hat{\theta}) = (\hat{m} - m(\hat{\theta}))' (\hat{m} - m(\hat{\theta})) = 0$$

$$g(\hat{\theta})' W g(\hat{\theta}) = (\hat{m} - m(\hat{\theta}))' W (\hat{m} - m(\hat{\theta})) = 0$$

where W is any positive definite weighting matrix.

Generalized method of moments estimation arises when there are more moment conditions (m) imposed than parameters (p) in the model and there is not a solution to

$$g(\hat{\theta}) = \hat{m} - m(\hat{\theta}) = 0.$$

How can you take account of more sample moments than parameters? Generalized method of moments estimators (GMM) estimators are defined by

$$\hat{\theta} = \arg \min_{\theta} g(\theta)' g(\theta) = (\hat{m} - m(\theta))' (\hat{m} - m(\theta))$$

or more generally by

$$\hat{\theta} = \arg \min_{\theta} g(\theta)' W g(\theta)$$

where W denotes a positive definite weighting matrix. Both of these quadratic forms will be nonnegative and their value can be used to perform a test of the consistency of the data with the hypothesized model in overidentified cases where there are more moment conditions than parameters ($m > p$). The GMM estimator of θ in the second equation depends upon the weighting matrix W . W is often selected to minimize the asymptotic variance of the GMM estimator. The W matrix which does this is given by $W = Var(g(\hat{\theta}))^{-1}$.

In this case, $g(\hat{\theta})'Var^{-1}(g(\hat{\theta}))g(\hat{\theta})$ has an asymptotic $\chi^2(m-p)$ distribution.

We will now consider a simple example and also demonstrate that GMM includes OLS, IV, and MLE as special cases.

b. Examples.

(1) Exponential

The pdf for the exponential is given by $f(y; \beta) = \frac{e^{-y/\beta}}{\beta}$ for $y > 0$ and has me-

and variance equal to β and β^2 , respectively.

Estimators:

MLE: $\hat{\beta} = \bar{Y}$

MOM: $\hat{\beta} = \bar{Y}$

What if the sample variance is not equal to the square of the sample mean?

Alternatively, what if the sample variance is not equal to the square of the sample mean and you want the estimator to take account of information about the sample mean and sample variance? GMM is one approach of doing this (two sample moments and one parameter)

GMM: minimize over beta sum of squares or weighted sum of squares

$$H(\beta) = (\bar{Y} - \beta, s^2 - \beta^2)' W (\bar{Y} - \beta, s^2 - \beta^2)'$$

$$= (\bar{Y} - \beta, s^2 - \beta^2)' \begin{bmatrix} w_1^2 & w_{12} \\ w_{21} & w_2^2 \end{bmatrix} \begin{pmatrix} \bar{Y} - \beta \\ s^2 - \beta^2 \end{pmatrix}$$

(2) OLS

$$\text{Let } g(\beta) = X'(Y - X\beta) = X'\varepsilon$$

, then

$$Var(g(\beta)) = Var(X'\varepsilon) = \sigma^2 (X'X) = W^{-1}$$

The corresponding minimization problem is to minimize

$$H(\theta) = (g(\theta))' W (g(\theta))$$

$$= (Y - X\beta)' X (X'X)^{-1} X'(Y - X\beta) / \sigma^2$$

The necessary conditions for a minimum of $H(\cdot)$ are

$$\frac{dH(\beta)}{d\beta} = \frac{-2}{\sigma^2} X'X (X'X)^{-1} X' (Y - X\hat{\beta}) = \frac{-2}{\sigma^2} X'(Y - X\beta) = 0$$

which yields the OLS estimator

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

which is based on the estimated covariances of the errors and the x's being zero.

(3) Instrumental variables

$$\text{Let } g(\beta) = Z'(Y - X\beta) = Z'\varepsilon$$

$Z_{n \times m}$ denotes a matrix including n observations on m instrumental variables which we would like to be uncorrelated with the estimated error terms. It can be shown that

$$Var(g(\beta)) = Var(Z'\varepsilon) = \sigma^2 (Z'Z)$$

The corresponding minimization problem is to minimize

$$\begin{aligned} H(\theta) &= (g(\theta))' D(g(\theta)) \\ &= (Y - X\beta)' Z (Z'Z)^{-1} Z' (Y - X\beta) / \sigma^2 \end{aligned}$$

The necessary conditions for a minimum are

$$\frac{dH(\beta)}{d\beta} = \frac{-2}{\sigma^2} X Z (Z'Z)^{-1} Z (Y - X\hat{\beta}) = 0$$

which yields the IV estimator

$$\tilde{\beta} = (X'Z (Z'Z)^{-1} Z'X)^{-1} X'Z (Z'Z)^{-1} Z'Y$$

which is based on the estimated covariances between the instrumental variables and the errors being zero.

(4) Maximum likelihood estimation

The assumed log-likelihood function is given by

$$\ell(\beta) = \sum_t \ln f(\varepsilon_t = Y - X_t\beta)$$

$$\text{Let } g(\beta) = \frac{d\ell}{d\beta} = \sum_t \left(\frac{d \ln f(\cdot)}{d\varepsilon_t} \right) X_t'$$

with corresponding variance estimated by

$$Var\left(\frac{d\ell}{d\beta}\right) = \sum_t (X_t X_t') \left(\frac{d \ln f(\cdot)}{d\varepsilon_t} \right)^2 / n$$

The objective function is given by:

$$H(\theta) = (g(\theta))' \left(Var\left(\frac{d\ell}{d\theta}\right) \right)^{-1} (g(\theta))$$

The variance matrix involves the unknown parameters; thus, this can be a very nonlinear optimization problem.

c. Comments

- (1) GMM encompasses most of the common estimation procedures such as MLE, OLS, IV, and 2SLS.
- (2) In the exactly identified case ($m=p$) we can often get exact answers
- (3) In the overidentified case, GMM requires the specification of a weighting matrix case. The estimators will depend upon the weighting matrix.

- (4) The optimal weighting matrix minimizes the asymptotic variance of the estimator
- (5) Tests of the overidentifying assumptions: The optimized objective function corresponding to the optimal weighting matrix is asymptotically distributed as a chi square with degrees of freedom $m-p$.
- (6) Chamberlain (1987, Journal of Econometrics, 305-334) showed that the class of GMM estimators achieve the semiparametric efficiency bound given the set of moment restrictions.
- (7) In the case of overidentification (number of moment restrictions exceed the number of parameters), in some cases GMM estimators exhibit substantial bias with related test statistics being problematical.
- (8) Woolridge (2001) survey article, “Applications of Generalized Method of Moments Estimation,” explores some different areas in which GMM can be productively applied, including cross-sectional applications, testing functional forms, for censored or truncated regression models, various time series problems, and with panel data sets. However, Woolridge notes that the “majority of empirical researchers in economics probably have never used a GMM procedure.”
- (9) The method of empirical likelihood appears to have less bias and yields test statistics similar in form to LR, W, and LM statistics. Imbens (2002, JBES, 493-506). Imbens’ paper includes a nice discussion of the method of empirical likelihood.

5. Extremum estimators (GR 6th: 14.5, 7th: 12.5)

a. Basic Theory and results

The book by Huber (Robust Statistics) discusses properties of more general extremum estimators which include the case of "MLE estimators." Consider the

extremum estimator $(\hat{\theta})$ defined by

$$\hat{\theta} = \arg \max_{\theta \in \Theta} H(\theta; data),$$

the estimator maximizes a function of the parameters, conditional on the data. Under fairly general regularity conditions the asymptotic distribution of the extremum estimator,

$$\hat{\theta}, \text{ is given by } \sqrt{n}(\hat{\theta} - \theta) \sim N[0, C]$$

where $C = A^{-1} B A^{-1}$, $A = \lim_{n \rightarrow \infty} E \left[\frac{1}{n} \frac{d^2 H(\cdot)}{d\theta d\theta'} \right]$, and

$$B = \lim_{n \rightarrow \infty} \left\{ \text{Var} \left(\frac{1}{\sqrt{n}} \frac{dh(\cdot)}{d\theta} \right) = E \left[\frac{1}{n} \frac{dh(\cdot)}{d\theta} \frac{dh(\cdot)}{d\theta'} \right] \right\}$$

The variance-covariance matrix C is often referred to as the *sandwich estimator*.

Proof: Consider the Taylor series expansion of $\frac{dH(\cdot)}{d\theta}$, the derivative of the objective

function,

$$\frac{dH(\cdot)}{d\theta}|_{\hat{\theta}} = 0 = \frac{dH(\cdot)}{d\theta}|_{\theta_0} + \frac{d^2 H(\cdot)}{d\theta d\theta'}|_{\theta_0} (\hat{\theta} - \theta_0)$$

where θ_* lies between $\hat{\theta}$ and θ_0 .

Since the extremum estimator is defined by $\frac{dH(\cdot)}{d\theta}|_{\hat{\theta}} = 0$,

we can express $(\hat{\theta} - \theta_0)$ as

$$\sqrt{n}(\hat{\theta} - \theta_0) = -\left(\frac{1}{n} \frac{d^2 H(\cdot)}{d\theta d\theta'}|_{\theta_*}\right)^{-1} \left(\frac{1}{\sqrt{n}} \frac{dH(\cdot)}{d\theta}|_{\theta_0} \right)$$

where the regularity conditions are sufficient to insure that the second term on the right hand side converges to

$$N[0, B]$$

and the desired result follows from an application of the *Useful Theorem*.

b. Special cases:

(1) OLS

$$\text{Let } H(\beta) = -SSE(\beta) = -(Y'Y - 2\beta' X'Y + \beta' X'X\beta)$$

$$\frac{dH}{d\beta} = 2X'\varepsilon,$$

$$Var\left(\frac{dH}{d\beta}\right) = 4\sigma^2(X'X) = B$$

$$\frac{d^2H}{d\beta d\beta'} = -2(X'X) = A$$

$$Var(\hat{\beta}) = A^{-1}BA^{-1} = (-2(X'X))^{-1} 4\sigma^2(X'X)(-2(X'X))^{-1}$$

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

(2) MLE

$$\text{Let } H(\hat{\theta}) = \ell(\hat{\theta})$$

If the model is correctly specified, $B=-A$, and the sandwich estimator gives the same results as MLE, the Cramer-Rao Matrix.

(3) QMLE

Quasi Maximum likelihood estimation (“MLE” with the wrong pdf) can be included in this framework. In the context of a QMLE framework a specification test (Hal White) can be performed by testing $A(\theta) + B(\theta) = 0$, using $A(\hat{\theta}) + B(\hat{\theta})$.

The expectations are taken with respect to the true distribution of Y . In the case of correct model specification $A(\theta) = -B(\theta)$, see Theil (pp. 384-6) and Appendix B of the Statistics section. In this case, $C(\theta)$ simplifies to the result for MLE. In this case $B(\theta)$ provides an alternative estimator of the variance covariance matrix which is based on the first derivatives of $\ell(Y; \theta)$. This estimator is often useful in empirical work.

(5) GMM

Extremum estimators also include GMM as special cases by defining

$$H(\theta) = -g(\theta)' W g(\theta) \quad (0)$$

6. Kernel or adaptive estimation

a. Semiparametric applications

Kernel estimators can be used to obtain semiparametric or nonparametric estimators of the relationship between variables. We first consider their use as semiparametric estimators of a known functional form for a regression relationship with an unknown error distribution. We will then consider the use of kernels to estimate an unknown relationship between variables with an unknown error distribution. These methods may require large samples for the estimators to reflect desirable large sample properties.

As discussed earlier, kernel estimators make use of an approximating pdf

$$f(\epsilon) = \left(\frac{1}{sn} \right) \sum_{t=1}^n K\left(\frac{\epsilon - e_t}{s} \right)$$

where K and e_t denote the a specified kernel pdf and least squares residuals,

$e_t = Y_t - X_t \hat{\beta}$, respectively. $K(\cdot)$ can be any pdf, but is often selected to be the standard normal. s is a smoothing parameter. Some variations of this procedure use of a trimming parameter. Silverman suggests selecting the smoothing parameter to be to $1.06 \hat{\sigma} n^{-1/2}$. The approximating pdf, $f(\cdot)$, can be thought of as a smoothed histogram.

"MLE" is then performed using the kernel in the likelihood function,

$$\ell(\beta; Y, s, K) = \ln L(\beta; Y, s, K) = \sum_{i=1}^n \ln f(y_i - X_i \beta, K, s)$$

where the estimators are conditional on the smoothing parameter (s) and the selected kernel (K). These estimators are also conditional on the OLS estimated parameters to obtain the estimated disturbances. An alternative approach is to iterate the procedure.

Start with the OLS residuals and obtain estimates of β_0 . Given the new estimate β_1 ,

calculate the corresponding residuals, then re-optimize over β_2 . This procedure until convergence is achieved. Kernel estimators are quite robust and, at least in this form, apply to the classical linear regression model with independently (no autocorrelation) and identically (homoskedasticity) error terms. Variations can be used for multiple regressors and to cover departures from the assumptions independence and identically distributed errors a. The interested reader is encouraged to review the paper by Hsieh and Manski [1987]. Pagan and Ullah [1999] provide an excellent introduction to many related techniques and application. An alternative adaptive estimator is discussed in Newey [1988].

b. Nonparametric application

(Reference: Wikipedia, the free encyclopedia and Greene (14.4.2))

Kernel regression methods are nonparametric procedures to estimate the conditional expectation of a random variable. These conditional expectations can involve nonlinear relations between random variables and can be written as

$$E(Y|X) = m(X).$$

The Nadaraya-Watson estimator of $m(X)$ is defined by

$$\hat{m}(x) = \frac{\sum_{i=1}^n K_h\left(\frac{x-X_i}{h}\right)Y_i}{\sum_{i=1}^n K_h\left(\frac{x-X_i}{h}\right)}$$

where $K_h(\cdot)$ denotes a kernel with window width h . While alternate regression estimators, such as the Priestly-Chao or Gasser-Muller kernel estimators have been considered, the Nadaraya-Watson estimator is one of the most commonly used.. The Nadaraya-Watson kernel estimator can be obtained by evaluating

$$E[Y|X=x] = \int y f(y|x) dy = \int y \frac{f(x,y)}{f(x)} dy$$

where kernel estimates are used for $f(x, y)$ and $f(x)$

$$\hat{f}(x, y) = \frac{\sum_{i=1}^n K\left(\frac{x-x_i}{h}\right)K\left(\frac{y-y_i}{h}\right)}{nh^2}$$

$$\hat{f}(x) = \frac{\sum_{i=1}^n K\left(\frac{x-x_i}{h}\right)}{nh}$$

The **STATA command** to implement kernel regression estimation is

```
kernreg yvar xvar [if exp] [in range], bwidth(#) kercode(#) npoint(#)  
[gen(mhvar gridvar) nograph graph_options]
```

Description: kernreg calculates the Nadaraya-Watson nonparametric regression. By default, kernreg draws the graph of the estimated conditional mean over the grid points used for calculation connected by a line without any symbol.

Options:

bwidth(#) specifies the smoothing parameter (bandwidth or halfwidth) of the kernel density estimation for xvar. This parameter defines the width of the weight function window around each grid point.

kercode(#) specifies the weight function (kernel) to calculate the required univariate densities according to the following numerical codes:

- 1 = Uniform
- 2 = Triangle
- 3 = Epanechnikov
- 4 = Quartic (Biweight)
- 5 = Triweight
- 6 = Gaussian
- 7 = Cosinus

npoint(#) specifies the number of equally spaced points (which define a grid) in the range of xvar used for the regression estimation.

gen(mhvar gridvar) creates two variables containing the estimated regression (conditional mean) values and the corresponding grid points, respectively.

nograph suppresses the graph.

graph_options are any of the options allowed with graph, twoway.

Remarks: bwidth, kercode, and nppoint are not optional. If the user does not provide them, the program halts and displays an error message on screen.

This program uses kernel density estimators modified from Salgado-Ugarte, et al. (1993) and based on the equations provided by Haerdle (1991) and Scott (1992). The smoothness of the resulting estimate can be regulated by changing the bandwidth: wide intervals produce smooth results; narrow intervals give noisier estimates.

Except for the Gaussian kernel, all the functions are supported on [-1,1].

While using the gen option, if the number of cases is less than nppoint then the program sets the number of the observations = nppoint to obtain the full set of estimations.

This procedure can be regarded as a descriptive smoother of scatterplots as well as a nonparametric regression estimator (Nadaraya-Watson).

Examples:

```
kernreg wait dura, bwidth(0.65) kercode(4) nppoint(100)
```

```
kernreg accel time, b(2.4) k(4) np(100) gen(m2p4 g2p4) nog
```

A recent reference: **The July 2010 issue of the *Journal of Econometrics* is entitled “Nonlinear and nonparametric methods in econometrics and includes a variety of papers dealing with applications of nonparametric methods in econometrics.

c. An application with parametric and nonparametric components

A researcher may have a particular interest in modeling the relationship between a dependent variable (y) and several independent variables (X 's and z), but has a particular interest in exploring the functional relationship between the dependent variable and one of the independent variables(z). This situation might be modeled as

$$y_t = X_t \beta + h(z_t) + \varepsilon_t$$

where the form of $h(z)$ is unspecified. Kennedy (2008), Robinson (1988), and Anglin and Gencay (1996) suggest methods which are similar to estimating the relationship between y and X using a traditional method such as least squares. The unknown form for $h(z)$ is estimated by using Kernel methods to estimate the relationship between the residuals from the first stage and the variable(s) z .

7. Empirical likelihood estimation

While GMM estimators have desirable large sample properties, their small sample properties can be improved upon by considering different estimation procedures. In particular, some applied studies have shown that GMM may be strongly biased in certain applications. Variations of GMM which update the weighting matrix, Continues updating estimators (CUE), appear to reduce the finite sample bias. Another family of estimation procedures, generalized empirical likelihood (GEL), also appears to reduce estimator bias. EL, GMM, CUE, and an estimator called exponential tilting (ET), share a common structure. We will use the notation used by Whitney and Smith (2004) to explore these interrelationships. .

Let z_i ($i=1,\dots,n$) be independent and identically distributed observations on a data vector.

Further let β denote a $px1$ parameter vector a $g(z, \beta)$ be an $mx1$ vector of functions where the number of functions is greater than or equal to the number of parameters, $m \geq p$. Further

assume that $E_z[g(z, \beta_0)] = 0$ for the β_0 parameter vector

The GMM estimator is defined by

$$\hat{\beta}_{GMM} = \arg \min_{\beta \in B} \hat{g}(\beta)' \Omega(\tilde{\beta}) \hat{g}(\beta)$$

where $\tilde{\beta}$ denotes a preliminary estimate β to calculate the weighting matrix. A variation of the GMM estimator is the continuous updating estimator (CUE) which updates the weighting matrix in the objective function and is defined by

$$\hat{\beta}_{CUE} = \arg \min_{\beta \in B} \hat{g}(\beta)' \Omega(\beta) \hat{g}(\beta)$$

To define empirical likelihood (EL) and exponential tilting (ET) estimators, consider a

concave function $\rho(\nu)$ on its domain, an open interval containing zero. Let

$$\hat{\Lambda}_n(\beta) = \left\{ \lambda : \lambda' g_i(\beta) \in \nu \right\} \lambda . \quad \text{The generalized empirical likelihood (GEL)}$$

defined by the solution to

$$\hat{\beta}_{GEL} = \arg \min_{\beta \in B} \sup_{\lambda \in \hat{\Lambda}} \sum_{i=1}^n \rho(\lambda' g_i(\beta))$$

The exponential tilting estimator corresponds to a special case of the GEL estimator where $\rho(\nu) = -e^\nu$, Kitamura and Stutzer (1997) and Smith (1997). The empirical likelihood estimator is also a special case of GEL where $\rho(\nu) = \ln(1-\nu)$, Qin and Lawless (1994). $\rho(\nu)$ is quadratic $\hat{\beta}_{GEL} = \hat{\beta}_{CUE}$

An interesting interpretation of the GEL estimators is available by considering

$$\hat{\pi}_i = \frac{\rho_1(\hat{\lambda}' \hat{g}_i)}{\sum_{i=1}^n \rho_1(\hat{\lambda}' \hat{g}_i)} \quad \rho_1(\) = \frac{\partial \rho(\nu)}{\partial \nu} \text{ here}$$

The $\hat{\pi}_i$ sum to one and satisfy $\sum_{i=1}^n \hat{\pi}_i \hat{g}_i = 0$ $\hat{\pi}_i$ Thus, the empirical probabilities for the observations.

Minimum discrepancy (MD) estimators are defined as the solution to the optimization problem

$$\bar{\beta} = \arg \min_{\beta \in B, \pi \in \Sigma} \sum_{i=1}^n h(\pi_i) \quad \text{and}$$

$$\text{subject to } \sum_{i=1}^n \pi_i g_i(\beta) = 0 \quad \sum_{i=1}^n \pi_i = 1 \quad \text{and}$$

Selecting $h(\pi) = -\ln(\pi) = -\pi \ln(\pi)$ or yields EL and ET, respectively

$$\text{selecting } h(\pi) = \frac{n^2 \pi^2 - 1}{2n} \quad \text{yields CUE.}$$

Thus the EL estimator can be seen to solve the following optimization problem:

$$\bar{\beta} = \arg \max_{\beta \in E, \pi_i > 0} \sum_{i=1}^n \ln(\pi_i)$$

subject to $\sum_{i=1}^n \pi_i g_i(\beta) = 0 \quad \sum_{i=1}^n \pi_i = 1 \quad \text{and}$

In a recent paper Chausse' presents an R program which can be used to obtain GMM and GEL estimators.

Guggenberger (2008) uses Monte Carlo methods to compare the finite sample properties of the GEL, CUE, and various k-class estimators based on sample median, mean, mean squared errors and coverage probabilities and length of confidence intervals. The results suggest that GEL and LIML behave similarly, having thick tails.

References:

- Chausse', P., "Computing Generalized Method of Moments and Generalized Empirical Likelihood with R," Rmetrics Workshop, 2009. (Listed in 588 resources)
- Guggenberger, P., "Finite Sample Evidence Suggesting a Heavy Tail Problem of the Generalized Empirical Likelihood Estimator," *Econometric Reviews*, 26 (2008), 526-541.
- Kitamura, Y. And M. Stutzer, "An Information-Theoretic Alternative to Generalized Method of Moments Estimation," *Econometrica*, 65 (1997), 861-874.
- Newey, W. K. And R. J. Smith, "Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators," *Econometrica*, 72 (2004), 219-255.
- Smith, R. J. , "Alternative Semi-parametric Likelihood Approaches to Generalized Method of Moments Estimation," *Economic Journal*, 107(1997), 503-519
- Qin, J. And J. Lawless, "Empirical Likelihood and General Estimating Equations," *Annals of Statistics*, 22(1994), 300-325.

8. Homework problems

1. Consider the model $Y = X\beta + \varepsilon$ Y, X, β , and ε denote $nx1, nxk, kx1$

matrices. Additionally, let Z denote a matrix of nxm observations on m instrumental variables.

- a. Derive the IV/GMM estimator which maximizes $H(\beta)$ where

$$\begin{aligned} H(\beta) &= -((Z'\varepsilon)'W(Z'\varepsilon)) = -((Y-X\beta)'ZWZ'(Y-X\beta)) \\ &= -(Y-X\beta)'ZWZ'(Y-X\beta) \end{aligned}$$

The estimator will depend upon the weighting matrix, W , which is assumed to be positive

definite. The optimal weighting matrix is given by $W = (Var(Z'\varepsilon))^{-1} = (Z'Var(\varepsilon)Z)^{-1}$

For parts (b) and (c) the following results will be helpful.

Note: $\frac{dH}{d\beta} = 2X'ZWZ'(Y-X\beta) = 2X'ZWZ'\varepsilon$

$$\frac{dH}{d\beta} \frac{dH}{d\beta'} = 4X'ZWZ'\varepsilon\varepsilon'ZWZ'X$$

$$\frac{d^2H}{d\beta d\beta'} = -X'ZWZ'X$$

- b. If $Var(\varepsilon) = \sigma^2 I$, demonstrate that the optimal weighting matrix is give by

$$W = (Z'Z)/\sigma^2$$

and evaluate the corresponding IV estimator and the Sandwich estimator of the variance.

c. . If $\text{Var}(\varepsilon) = \Sigma$, demonstrate that the optimal weighting matrix $\mathbf{W} = (\mathbf{Z}' \Sigma \mathbf{Z})^{-1}$

and evaluate the corresponding IV estimator and the sandwich estimator of its variance.

d. Derive the sandwich estimator of the variance of the OLS estimator when $\text{Var}(\varepsilon) = \Sigma$

Hint: Let $H(\beta) = -(Y - X\beta)'(Y - X\beta)$

Evaluate and take expected values of

$$\frac{dH(\beta)}{d\beta}, \frac{dH(\beta)}{d\beta'} \frac{dH(\beta)}{d\beta'}, \text{ and } \frac{d^2H(\beta)}{d\beta d\beta'}$$

This estimator should look familiar from Econ 388 and is what the reg y x's and newey y x's, lags(#) estimate when heteroskedasticity and/or autocorrelation of the errors are present.

e. Consider how you might estimate the sandwich estimator of the OLS estimator derived in 1(d).

(1) Heteroskedasticity

(2) Heteroskedasticity and autocorrelation

2. How can you obtain a sandwich estimator for MLE/QMLE estimators of β in the linear

regression model with flexible error distributions?

3. Consider a model for panel data,

$$y_1 = X_1\beta + \varepsilon_1$$

.

.

$$y_n = X_n\beta + \varepsilon_n$$

where y_i , X_i , β , and ε_i respectively denote $Tx1$, Txk , $kx1$, and tx matrices. Note that

the same for each panel. This system of equations can be written as

$$Y = X\beta + \varepsilon$$

$$\text{where } Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

where Y , X , and ε will be $Tnx1$ matrices and

$$Var(\varepsilon) = \begin{bmatrix} Var(\varepsilon_1) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & 0 & Var(\varepsilon_n) \end{bmatrix} = \begin{bmatrix} \Omega_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & 0 & \Omega_n \end{bmatrix} = \Omega$$

is $Tn \times Tn$.

Demonstrate that the OLS estimator and corresponding sandwich estimator can be written as

$$\hat{\beta} = \left(\sum_{i=1}^n X_i' X_i \right)^{-1} \left(\sum_{i=1}^n X_i' Y_i \right)$$

$$\text{Sandwich estimator} = \left(\sum_{i=1}^n X_i' X_i \right)^{-1} \left(\sum_{i=1}^n X_i' \Omega_i X_i \right) \left(\sum_{i=1}^n X_i' X_i \right)^{-1}$$

This estimator is related to what are referred to as clustered standard errors,

see C. B. Hansen, "Asymptotic Properties of a Robust Variance Matrix Estimator for Panel Data when T is Large," *Journal of Econometrics*, 141 (2007), 597-620.

Homework key:

1. Extremum, IV, and GMM estimation

a. Derive the IV/GMM estimator

(points)

$$(1) \text{ Min } H = (Y - X\beta)' Z W Z' (Y - X\beta)$$

$$\begin{aligned} \frac{dh}{d\beta} &= -2X' Z W Z' (Y - X\beta) = -2X' Z W Z' \epsilon = 0 \\ \rightarrow \hat{\beta}_{IV} &= (X' Z W Z' X)^{-1} X' Z W Z' Y \end{aligned}$$

b. Special case of $\text{Var}(\varepsilon) = \sigma^2 I$

Substitute $(\sigma^2(Z'Z))^{-1} = W$ into the equation for the IV estimator
 $\rightarrow \hat{\beta}_{IV} = (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'Y$

Sandwich estimator:

$$\frac{d^2H}{d\beta d\beta'} = 2X' Z W Z' X = \frac{2}{\sigma^2} X' Z (Z'Z)^{-1} Z' X$$

$$\left(\frac{dH}{d\beta} \right) \left(\frac{dH}{d\beta'} \right) = 4X' Z W Z' \epsilon \epsilon' Z W Z' X = \frac{4}{\sigma^4} X' Z (Z'Z)^{-1} Z' \epsilon \epsilon' Z (Z'Z)^{-1} Z' X$$

$$\rightarrow E \left[\left(\frac{dH}{d\beta} \right) \left(\frac{dH}{d\beta'} \right) \right] = \frac{4}{\sigma^4} X' Z (Z'Z)^{-1} Z' X$$

Sandwich:

$$\begin{aligned} &\left(\frac{d^2H}{d\beta d\beta'} \right)^{-1} E \left[\left(\frac{dH}{d\beta} \right) \left(\frac{dH}{d\beta'} \right) \right] \left(\frac{d^2H}{d\beta d\beta'} \right)^{-1} \\ &= \frac{\sigma^4}{4} (X' Z (Z'Z)^{-1} Z' X)^{-1} \left(\frac{4}{\sigma^2} \right) (X' Z (Z'Z)^{-1} Z' X) (X' Z (Z'Z)^{-1} Z' X)^{-1} \\ &= \sigma^2 (X' Z (Z'Z)^{-1} Z' X)^{-1}. \end{aligned}$$

1. c. Special case of $\text{Var}(\varepsilon) = \Sigma$

Substitute $(Z'\Sigma Z)^{-1}$ for W

$$(1) \hat{\beta}_{IV} = (X' Z W Z' X)^{-1} X' Z W Z' Y$$

$$(2) \hat{\beta}_{IV} == (X' Z (Z' \Sigma Z)^{-1} Z' X)^{-1} X' Z (Z' \Sigma Z)^{-1} Z' Y$$

(3) Sandwich estimator

$$\frac{d^2 H}{d\beta d\beta'} = 2X' Z W Z' X = 2X' Z (Z' \Sigma Z)^{-1} Z' X$$

$$\left(\frac{dH}{d\beta} \right) \left(\frac{dH}{d\beta'} \right) = 4X' Z W Z' \epsilon \epsilon' Z W Z' X = 4X' Z (Z' \Sigma Z)^{-1} Z' \epsilon \epsilon' Z (Z' \Sigma Z)^{-1} Z' X$$

$$\rightarrow E \left[\left(\frac{dH}{d\beta} \right) \left(\frac{dH}{d\beta'} \right) \right] = 4X' Z (Z' \Sigma Z)^{-1} Z' \Sigma Z (Z' \Sigma Z)^{-1} Z' X = 4X' Z (Z' \Sigma Z)^{-1} X$$

So sandwich is

$$\left(\frac{d^2 H}{d\beta d\beta'} \right)^{-1} E \left[\left(\frac{dH}{d\beta} \right) \left(\frac{dH}{d\beta'} \right) \right] \left(\frac{d^2 H}{d\beta d\beta'} \right)^{-1} = (X' Z (Z' \Sigma Z)^{-1} Z' X)^{-1}$$

1. d. Sandwich estimator for the OLS estimator

$$Max_{\beta} H(\beta) \{ -(Y - X\beta)'(Y - X\beta) \}$$

$$\frac{dH}{d\beta} = (-2)(-X')(Y - X\beta) = 2X'(Y - X\beta) = 2X'\epsilon$$

$$\frac{d^2 H}{d\beta d\beta'} = 2(X'X).$$

(1)

$$E \left[\frac{dH}{d\beta} \frac{dH}{d\beta'} \right] = E(2X'\epsilon)(2\epsilon'X') = 4X'E(\epsilon\epsilon')X = 4X'\Sigma X$$

(2)

Sandwich estimator:

$$\begin{aligned} \left(-E \frac{d^2 H}{d\beta d\beta'} \right)^{-1} \left(E \frac{dH}{d\beta} \frac{dH}{d\beta'} \right) \left(-E \frac{d^2 H}{d\beta d\beta'} \right)^{-1} &= \frac{1}{2}(X'X)^{-1}(4X'\Sigma X)\left(\frac{1}{2}\right)(X'X)^{-1} \\ &= (X'X)^{-1}X'\Sigma X(X'X)^{-1}. \end{aligned}$$

1.e. (1) Heteroskedasticity

Estimate

$$E(\epsilon\epsilon') = \begin{bmatrix} E(\epsilon_1^2) & E(\epsilon_1\epsilon_2) & \cdots & E(\epsilon_1\epsilon_n) \\ E(\epsilon_2\epsilon_1) & E(\epsilon_2^2) & \cdots & E(\epsilon_2\epsilon_n) \\ \vdots & \ddots & \ddots & \vdots \\ E(\epsilon_n\epsilon_1) & E(\epsilon_n\epsilon_2) & \cdots & E(\epsilon_n^2) \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

using $\begin{bmatrix} e_1^2 & 0 & \cdots & 0 \\ 0 & e_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e_n^2 \end{bmatrix}$

1.e. (2) Autocorrelation

Part (b): $E(\epsilon\epsilon') = \begin{bmatrix} E(\epsilon_1^2) & E(\epsilon_1\epsilon_2) & \cdots & E(\epsilon_1\epsilon_n) \\ E(\epsilon_2\epsilon_1) & E(\epsilon_2^2) & \cdots & E(\epsilon_2\epsilon_n) \\ \vdots & \vdots & \ddots & \vdots \\ E(\epsilon_n\epsilon_1) & E(\epsilon_n\epsilon_2) & \cdots & E(\epsilon_n^2) \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2 & \cdots & \sigma_1\sigma_n \\ \sigma_2\sigma_1 & \sigma_2^2 & \cdots & \sigma_2\sigma_n \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_n\sigma_1 & \sigma_n\sigma_2 & \cdots & \sigma_n^2 \end{bmatrix}$

A naive approach might be to use e_j^2 to estimate $E(\epsilon_j^2)$,

Use $e_i e_j$ to estimate $E(\epsilon_i \epsilon_j)$.

The Newey-West provides a variation on this method.

3. Sandwich estimator for the MLE

$$\text{Max}\{\sum_{i=1}^n \ln(y_i - X_i\beta; \theta)\}$$

$$(1) \quad \left(-E \frac{d^2 H}{d\psi d\psi'} \right)^{-1} \left(E \frac{dH}{d\psi} \frac{dH}{d\psi'} \right) \left(-E \frac{d^2 H}{d\psi d\psi'} \right)^{-1}.$$

$$(2) \quad \begin{aligned} E \frac{d^2 H}{d\psi d\psi'} &= \frac{1}{n} \sum \frac{d^2 H}{d\psi d\psi'} \\ E \frac{dH}{d\psi} \frac{dH}{d\psi'} &= \frac{1}{n} \sum \frac{dH}{d\psi} \frac{dH}{d\psi'} \end{aligned}$$

4. Panel data estimation

Coefficient estimator

$$\hat{\beta} = (X'X)^{-1}X'Y = \left\{ [X'_1 \quad \cdots \quad X'_n] \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} \right\}^{-1} \left\{ [X'_1 \quad \cdots \quad X'_n] \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \right\}$$

$$= \left(\sum_{i=1}^n X'_i X_i \right)^{-1} \sum_{i=1}^n X'_i Y_i$$

Sandwich estimator

$$(X'X)^{-1}X'\Omega X(X'X)^{-1} = \left(\sum_{i=1}^n X'_i X_i \right)^{-1} [X'_1 \quad \cdots \quad X'_n] \begin{bmatrix} \Omega_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Omega_n \end{bmatrix} \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} \left(\sum_{i=1}^n X'_i X_i \right)^{-1}$$

$$= \left(\sum_{i=1}^n X'_i X_i \right)^{-1} \left(\sum_{i=1}^n X'_i \Omega_i X_i \right) \left(\sum_{i=1}^n X'_i X_i \right)^{-1}.$$

Selected References

- Amemiya, Takeshi (1985), Advanced Econometrics, Cambridge: Harvard Press (especially Chapter 4, Asymptotic Properties of Extremum Estimators).
- Anglin, P. M., and G. Ramazan (1996), "Semiparametric Estimation of a Hedonic Price Function," *Journal of Applied Econometrics* 11, no. 6, 633-648.
- Anscombe, F. J. (1973), "Graphs in Statistical Analysis," *The American Statistician*, 27, 17-21.
- Boyer, B. H., J. B. McDonald, and W. K. Newey, (2003) "A Comparison of Partially Adaptive and Reweighted Least Squares Estimation," *Econometric Reviews*, 115-134.
- Hampel, F. R. (1974), "The Influence Curve and Its Role in Robust Estimation," *Journal of the American Statistical Association*, 69, 383-393.
- Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986), Robust Statistics: The Approach Based on Influence Functions, New York: John Wiley and Sons.
- Hsieh, D. A. and C. F. Manski, "Monte Carlo Evidence on Adaptive Maximum Likelihood Estimation of a Regression," *Annals of Statistics*, 15(1987), 541-551.
- Huber, P. J. (1981), Robust Statistics, New York: Wiley.
- Kennedy, P. (2008), A Guide to Econometrics, 6th edition, Malden, MA: Blackwell Publishing.
- Koenker, R. (1982), "Robust Methods in Econometrics," *Econometric Reviews*, 1, 213-255.
- Martin, R. D. and T. T. Simin (2003), "Outlier-Resistant Estimates of Beta," *Financial Analysts Journal*, 56-69.
- McDonald, J. B. and Newey, W. K. (1988), "Partially Adaptive Estimation of Regression Models via the Generalized T Distribution," *Econometric Theory*, 4(1988), 428-457.
- McDonald, J. B. and S. B. White, "A Comparison of Some Robust, Adaptive and Partially Adaptive Estimators of Regression Models," *Econometric Reviews*, 12(1993), 103-124.
- Newey, W. K., "Adaptive Estimation of Regression Models Via Moment Restrictions," *Journal of Econometrics*, 38(1988), 301-339.
- Pagan, A. And A. Ullah, Nonparametric Econometrics, Cambridge: Cambridge University Press, 1999.
- Rey, W. J. J. (1983), Introduction to Robust and Quasi-Robust Statistical Methods, New York: Springer-Verlag.
- Robinson, P. M. (1998), "Root-N consistent Semiparametric Regression," *Econometrica* 56, no. 4, 931-954.
- Rousseeuw, P. J. (1984), "Least Median Squares Regression," *JASA*, 871-880.
- Sharpe, W. F. (1964), "Capital Asset Prices: A Theory of Equilibrium Under Conditions of Risk," *Journal of Finance*, 19, 425-442.
- Silverman, B. W. 1986. Density estimation for statistics and data analysis. Chapman & Hall.

APPENDIX: DATA FOR CAPM EXAMPLES

DATE	RISKFREE RATE	VALUE- WEIGHTED MARKET RATE	AMERICAN CAN CO	MARTIN MARIETTA CORP
JAN 1982	0.0080	-0.022042	-0.05164	-0.12847
FEB 1982	0.0092	-0.049210	-0.16078	-0.06773
MAR 1982	0.0098	-0.008324	0.03738	-0.04769
APR 1982	0.0113	0.041886	0.01712	0.06393
MAY 1982	0.0106	-0.029107	0.00000	-0.03433
JUN 1982	0.0096	-0.019897	0.05455	-0.07627
JUL 1982	0.0105	-0.021088	-0.01379	-0.06373
AUG 1982	0.0076	0.125243	0.12108	0.69550
SEP 1982	0.0051	0.012631	-0.02800	-0.07187
OCT 1982	0.0059	0.115704	0.05267	0.09091
NOV 1982	0.0063	0.047138	-0.01600	0.00926
DEC 1982	0.0067	0.016163	0.00407	0.08208
JAN 1983	0.0069	0.037046	0.02348	-0.03429
FEB 1983	0.0062	0.028331	0.04049	-0.08284
MAR 1983	0.0063	0.033159	0.05058	0.23819
APR 1983	0.0071	0.072562	0.15481	0.11579
MAY 1983	0.0069	0.003940	0.17974	0.04443
JUN 1983	0.0067	0.039217	-0.03047	0.10251
JUL 1983	0.0074	0.030019	-0.03771	0.02479
AUG 1983	0.0076	0.012462	-0.01813	-0.06484
SEP 1983	0.0076	0.018097	0.05846	0.08261
OCT 1983	0.0076	0.018098	0.01105	-0.05120
NOV 1983	0.0070	0.025632	0.14035	-0.05498
DEC 1983	0.0073	0.008210	-0.03846	-0.03051
JAN 1984	0.0076	0.008873	0.12747	0.02448
FEB 1984	0.0071	0.036881	-0.06954	-0.07276
MAR 1984	0.0073	0.016679	-0.04381	-0.00743
APR 1984	0.0081	0.005258	-0.01132	0.01873
MAY 1984	0.0078	0.051286	-0.06094	-0.03426
JUN 1984	0.0075	0.023251	0.10030	0.00385
JUL 1984	0.0082	0.015648	-0.06220	0.09579
AUG 1984	0.0083	0.111437	0.09884	0.11077
SEP 1984	0.0086	0.002052	0.03175	-0.04127
OCT 1984	0.0100	0.003322	0.00462	0.20530
NOV 1984	0.0073	0.009374	0.01554	-0.06407
DEC 1984	0.0064	0.025158	0.03061	0.05325
JAN 1985	0.0065	0.079807	0.02178	0.13764
FEB 1985	0.0058	0.016262	0.00737	0.05185
MAR 1985	0.0062	0.000849	0.03415	-0.02657
APR 1985	0.0072	0.002710	0.01604	-0.00243
MAY 1985	0.0066	0.058722	0.08000	0.13382
JUN 1985	0.0055	0.017185	0.03704	-0.00216
JUL 1985	0.0062	0.003678	-0.02143	0.02273
AUG 1985	0.0055	0.004681	0.02174	-0.00952
SEP 1985	0.0060	0.036768	-0.09149	-0.14194
OCT 1985	0.0065	0.044130	0.06276	0.00000

APPENDIX:R CAPM EXAMPLES, CONT'D.

DATE	RISKFREE RATE	VALUE-WEIGHTED MARKET RATE	AMERICAN CAN CO	MARTIN MARIETTA CORP
NOV 1985	0.0061	0.068925	0.14286	0.04511
DEC 1985	0.0065	0.045591	-0.06250	0.02878
JAN 1986	0.0056	0.005765	0.11000	-0.06338
FEB 1986	0.0053	0.074143	0.16888	0.13910
MAR 1986	0.0060	0.054636	-0.01299	0.15182
APR 1986	0.0052	-0.012209	-0.06941	0.01153
MAY 1986	0.0049	0.050867	0.04821	0.07407
JUN 1986	0.0052	0.015193	0.03578	-0.01867
JUL 1986	0.0052	-0.054215	0.06875	-0.05163
AUG 1986	0.0046	0.072633	0.09627	0.08596
SEP 1986	0.0045	-0.079405	-0.07507	-0.08443
OCT 1986	0.0046	0.052730	0.08239	-0.08406
NOV 1986	0.0039	0.017544	-0.03138	0.10759
DEC 1986	0.0049	-0.027270	-0.00884	-0.11143

V. Nonlinear Models

- 1. Introduction**
- 2. Estimation**
 - a. Nonlinear least squares
 - b. Maximum likelihood
 - c. Estimation subject to constraints
 - d. Other methods of estimation
 - (1) Method of moments
 - (2) Generalized Method of Moments
 - (3) Methods for grouped data
 - (4) Extremum Estimators (QMLE)
- 3. Computational Algorithms**
- 4. Statistical Inference**
 - a. Individual parameters
 - b. More general tests
 - (1) Wald
 - (2) LM
 - (3) LR
- 5. Some Applications**
 - a. Production functions
 - b. Models for income distribution
 - c. A precautionary note
 - d. Binary or limited dependent variables
 - e. Ordered or categorical data
 - f. Poisson regression
 - g. Censored regression
 - h. Truncated and grouped regression
 - i. Interval regression
 - j. Quantile regression
 - k. Hazard functions
 - j. GMM
 - k. Bootstrapping
- 6. Some exercises**

V. Nonlinear Models

1. Introduction

Many nonlinear models can be transformed into a form such that linear techniques can be employed in estimating the unknown parameters. One such example is the Cobb-Douglas production function

$$Y_t = AL_t^{\beta_1}K_t^{\beta_2}\varepsilon_t.$$

Taking the logarithm of each side yields the equivalent expression

$$\ln(Y_t) = \ln A + \beta_1 \ln(L_t) + \beta_2 \ln(K_t) + \ln \varepsilon_t,$$

which is linear in the parameters β_1 and β_2 .

The second representation can be easily estimated using methods associated with linear models.

If the $(\ln \varepsilon_t)$'s are distributed normally, with zero mean, constant variance and independent of each other, then an application of least squares will yield maximum likelihood **estimators**.

Recall that $\ln(\varepsilon_t) \sim N[0, \sigma^2]$ implies that $\varepsilon_t \sim LN[\mu = 0, \sigma^2]^1$.

Many applied problems are associated with nonlinear models which can't be so easily estimated. For example, the constant elasticity of substitution production function (CES) is defined by

$$Y_t = A(\delta L_t^\rho + (1-\delta)K_t^\rho)^{M/\rho} \varepsilon_t$$

This is obviously a nonlinear relationship. Taking the logarithms of both sides yields

$$\ln Y_t = \ln A + (M/\rho) \ln(\delta L_t^\rho + (1-\delta)K_t^\rho) + \ln \varepsilon_t$$

which is nonlinear in the parameters and is of a form which can't be directly estimated using linear techniques.

¹ $E(\varepsilon_t) = e^{\mu + .5\sigma^2}$

The CES production function is an example of a nonlinear model which is of the general form (1.1) $Y_t = f(X_t; \beta) + \varepsilon_t$

where Y_t denotes an endogenous variable, X_t represents a $K \times 1$ vector of exogenous variables and ε_t is a random disturbance. $f(\cdot)$ represents a known function with unknown parameter values, denoted by β . Given N observations for Y_t and the vector X_t , we can also write

$$(1.1)' \quad Y = f(X; \beta) + \varepsilon$$

where $Y = (Y_1, \dots, Y_N)'$, $X = (X_1, \dots, X_N)'$, and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_N)'$. The vector function $f(X; \beta)$ is then defined to be $(f(X_1; \beta), \dots, f(X_N; \beta))'$. Also assume that β has K^* elements, $\beta = (\beta_1, \dots, \beta_{K^*})'$. The number of unknown parameters (K^*) need not equal the number of independent variables (K), e.g., the C.E.S. function production has four unknown parameters (A, δ, M, ρ) and two independent variables (K, L). For notational convenience, we will assume that the number of parameters is K . The reader is referred to the excellent survey paper, "Nonlinear Regression Models" by T. Amemiya in Handbook of Econometrics. Section 2 will discuss methods of estimating models which are nonlinear in the parameters. Section 3 will summarize algorithms used in obtaining estimators. Methods of statistical inference using the Wald, likelihood ratio and Lagrangian multiplier test will be presented in Section 4, and applications will be considered in the Section 5.

2. Estimation

The estimation problem is basically the problem of obtaining estimators of the vector of parameters β in

$$(2.1) \quad Y = f(X; \beta) + \varepsilon$$

from the sample data in Y and X . Many estimation techniques are available, but two of the most commonly used are nonlinear least squares and maximum likelihood estimation, although GMM, extremum, kernel, and other methods are discussed in the chapter on Estimation

Frameworks. Each of these methods generally involves solving a nonlinear system of equations. As previously noted, some computational methods of solving these equations will be surveyed in section three.

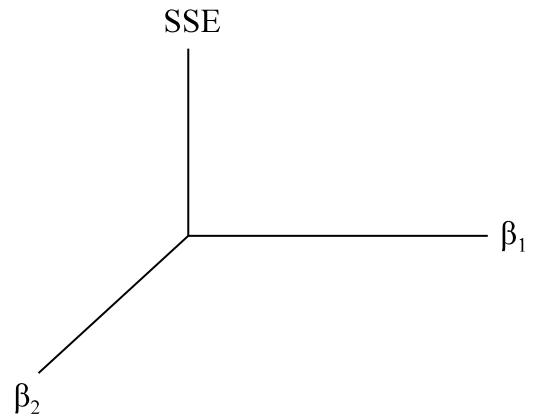
a. Nonlinear Least Squares Estimators (NLS):

The sum of squared errors associated with (2.1) is given by

$$(2.2) \quad \begin{aligned} SSE(\beta) &= \sum_t \varepsilon_t^2 \\ &= \varepsilon' \varepsilon \\ &= (Y - f(X; \beta))' (Y - f(X; \beta)). \end{aligned}$$

The least squares estimator of β , $\hat{\beta}_{NLS}$, satisfies the necessary conditions:

$$(2.3) \quad \frac{dSSE(\beta)}{d\beta} = 0$$



These equations are typically nonlinear in the β_i 's.

If the errors are independently and identically distributed with constant variance and a zero mean, $\varepsilon \sim N[0, \sigma^2 I]$, and the function $f(\cdot)$ satisfies certain regularity conditions, then the least squares estimator $\hat{\beta}_{NLS}$ will be consistent and asymptotically normal. σ^2 can be estimated by

$$(2.4) \quad \hat{\sigma}_{NLS}^2 = s_{NLS}^2 = SSE(\hat{\beta}_{NLS}) / (N - K)$$

and, the variance matrix of $\hat{\beta}_{NLS}$ can be estimated by

$$(2.5) \quad Var(\hat{\beta}_{NLS}) = \hat{\sigma}_{NLS}^2 [Z(\hat{\beta}_{NLS})' Z(\hat{\beta}_{NLS})]^{-1}$$

where $Z(\beta) = \left(\frac{\partial f(X; \beta)}{\partial \beta} \right)_{N \times K}$

$$= \begin{bmatrix} \frac{\partial f(X_1; \beta)}{\partial \beta_1} & \dots & \frac{\partial f(X_1; \beta)}{\partial \beta_K} \\ \vdots \\ \frac{\partial f(X_N; \beta)}{\partial \beta_1} & \dots & \frac{\partial f(X_N; \beta)}{\partial \beta_K} \end{bmatrix}$$

evaluated at β_{NLS} .

The **sandwich estimator** can also be used to provide an estimator of the variance.

In the linear model $Y = X\beta + \varepsilon$

$$Z(\beta) = \frac{\partial(X\beta)}{\partial\beta'} = X$$

and we obtain the regular least squares result

$$\beta_{NLS} = (X'X)^{-1}X'Y \stackrel{a}{\sim} N(\beta, \sigma^2(X'X)^{-1}).$$

Consider the nature of the matrix $Z(\beta)$ corresponding to the CES production function given on page 2.

Nonlinear least squares can be easily performed using popular software packages.

A simple example of using nonlinear least squares to estimate $y = \beta_1 + \beta_2/x + \varepsilon$

Obviously OLS could be used with a transformed variable.

STATA commands

nl(y=eqn), initial(coef value coef value) generic form

nl (y={b1}+{b2}/x), initial (b0 1 b1 3) with initial starting values or

nl (y={b1}+{b2}/x), nolog suppresses output and doesn't specify initial starting values

nl (y={b1}+{b2}/x), initial (b0 1 b1 3) robust reports robust standard errors

As another example, the CES production function just discussed could be estimated using the commands:

STATA:

```
gen ly=log(y)
```

```
nl (ly={g}+({M}/{rho})*log({d}*L^{{rho}}+(1-{d})*k^{{rho}})), initial (g 1 M 1 rho .5 d .5)
```

It is easy! Try it you will like it. To summarize, the “*nl*” option in STATA selects the parameter estimates to minimize the sum of squared errors.

b. Maximum Likelihood Estimation

In order to obtain MLE of β in (1.1), the distribution of the vector

$$\epsilon = Y - f(X; \beta)$$

must be given. Let this density be denoted by $g(\epsilon; \theta)$ where θ denotes the parameters associated with the density function $g(\cdot)$, e.g., σ^2 for the normal. The likelihood function $L(\cdot)$ and log likelihood functions, respectively, are defined by

$$(2.6a) \quad L(\beta, \psi) = g(\epsilon; \theta)$$

$$= g(Y - f(X; \beta); \theta)$$

$$(2.6b) \quad \ell(\beta, \psi) = \ln g(Y - f(X; \beta); \theta).$$

If the random disturbances in the model are identically and independently (homoskedastic and not autocorrelated) distributed, then the log likelihood function (2.6b) can be written as a sum

$$(2.7) \quad \ell = \sum_{t=1}^N \ln[g(Y_t - f(X_t; \beta); \theta)]$$

The maximum likelihood estimators of $\theta = (\beta, \psi)$ are defined by the solution of the equations

$$(2.7a,b) \quad \frac{\partial \ell}{\partial \beta} = 0$$

$$\frac{\partial \ell}{\partial \psi} = 0.$$

These equations are typically nonlinear in the parameters, and the solution again requires nonlinear optimization procedures. Under quite general regularity conditions:

(R.1) the range of the random variable Y_t is independent of the vector of parameters $\theta = (\beta, \psi)$;

(R.2) the density $g(\cdot)$ possesses derivatives of at least third order with respect to θ and these derivatives are bounded,

$$\frac{|\partial^3 L|}{|\partial \theta_i^3|} < M_i(y)$$

$E(M_i(Y)) < K$ constant; then

the maximum likelihood estimators are asymptotically normal

$$(2.9) \quad \tilde{\theta}_{ML} \stackrel{a}{\sim} N \left[\theta = \begin{pmatrix} \beta \\ \psi \end{pmatrix}; - \left(E \frac{\partial^2 \ell}{\partial \theta \partial \theta'} \right)^{-1} \right]$$

or

$$(2.10) \quad \sqrt{N}(\tilde{\theta}_{ML} - \theta) \stackrel{a}{\sim} N \left[0; N \left(-E \frac{\partial^2 \ell}{\partial \theta \partial \theta'} \right)^{-1} \right]$$

Note that the asymptotic variance covariance matrix is the *Cramer-Rao* lower bound.

One of the most common assumptions for the distribution of the random disturbances is that

$$(2.11) \quad \varepsilon \sim N[0, \Sigma]$$

in which case the log likelihood function (2.6b) is given by

$$(2.12) \quad \ell = \ln L = -(Y - f(X; \beta))' \Sigma^{-1} (Y - f(X; \beta)) / 2$$

$$-1/2 \ln \Sigma - \frac{N}{2} \ln(2\pi).$$

If the normal errors are also independently and identically distributed, i.e.,

$$\varepsilon \sim N[0, \Sigma = \sigma^2 I],$$

the corresponding loglikelihood function is given by

$$\begin{aligned}
 (2.13) \quad \ell &= \ln L = -\frac{1}{2\sigma^2} \sum_{t=1}^N (Y_t - f(X_t; \beta))^2 - \frac{N}{2}(\ln \sigma^2 + \ln 2\pi) \\
 &= -\frac{\text{SSE}(\beta)}{2\sigma^2} - \frac{N}{2}(\ln \sigma^2 + \ln 2\pi) \\
 &= -\frac{(Y - f(X; \beta))'(Y - f(X; \beta))}{2\sigma^2} - \frac{N}{2}(\ln \sigma^2 + \ln 2\pi)
 \end{aligned}$$

The maximum likelihood estimators in this case are defined by

$$\begin{aligned}
 \frac{\partial \ell}{\partial \beta} &= -\frac{1}{2\sigma^2} \frac{\partial \text{SSE}(\beta)}{\partial \beta} = \left(\frac{1}{\sigma^2} \right) \left(\frac{\partial f}{\partial \beta} \right) (f(X; \beta) - Y) = \frac{\partial f(\epsilon)}{\partial \beta} \left(\frac{\epsilon}{\sigma^2} \right) \\
 \frac{\partial \ell}{\partial \sigma^2} &= \frac{\text{SSE}(\beta)}{2\sigma^4} - \frac{N}{2\sigma^2} \\
 &= \frac{\epsilon' \epsilon}{2\sigma^4} - \frac{N}{2\sigma^2}
 \end{aligned}$$

Setting the derivatives to zero and solving implies

$$\hat{\beta}_{\text{NLS}} = \tilde{\beta}_{\text{ML}}$$

$$\tilde{\sigma}_{\text{ML}} = \text{SSE}(\beta_{\text{ML}})/N.$$

Substituting $\text{SSE}(\beta)/N$ into (2.13) for σ^2 yields what is called the *concentrated likelihood function*

$$(2.13)' \quad \ell = -\frac{N}{2} \left(1 + \ln(2\pi) + \ln \left(\frac{\text{SSE}}{N} \right) \right)$$

Recall from our previous discussion of MLE estimators that the asymptotic variance matrix (Cramer Rao) can also be obtained from

$$\left[-E \left(\frac{d^2 \ell}{d\theta^2} \right) \right]^{-1} = \left[E \left[\frac{d\ell}{d\theta} \frac{d\ell}{d\theta'} \right] \right]^{-1}$$

Given the previous results, the second expression for the variance covariance matrix, i.e.,

$$\begin{aligned}
(2.14) \quad \text{Var}(\beta_{\text{ML}}) &= \left[E \frac{dl}{d\beta} \frac{dl}{d\beta'} \right]^{-1} \\
&= \left[E \frac{df}{d\beta} \left(\frac{\epsilon}{\sigma^2} \right) \frac{\epsilon'}{\sigma^2} \left(\frac{df}{d\beta} \right), \right]^{-1} \\
&= \left[\frac{1}{\sigma^4} \frac{df}{d\beta} E(\epsilon \epsilon') \left(\frac{df}{d\beta} \right), \right]^{-1} \\
&= \sigma^2 \left[\frac{df}{d\beta} \left(\frac{df}{d\beta} \right), \right]^{-1}.
\end{aligned}$$

Note that (2.5) corresponds to the results for nonlinear least squares, as it should. Thus we see that least squares and MLE will be equivalent for the case of normally distributed residuals which are uncorrelated and homoskedastic. In summary, NLS and MLE will be equivalent if $\epsilon \sim N[0, \sigma^2 I]$.

Exercise:

Consider the case of the generalized error distribution in which the density of the random disturbances is given by

$$GED(\varepsilon_t; \psi=(p,s)) = \frac{pe^{-\left(|\varepsilon_t|^p/s^p\right)}}{2\Gamma(1/p)}$$

$$= \frac{pe^{-\left(|Y_t - f(X_t;\beta)|^p/\sigma^p\right)}}{2\sigma\Gamma(1/p)}$$

- (a) Write out an expression for the log likelihood function.
- (b) Interpret the sum in the log likelihood function corresponding to
 - (1) $p = 1$
 - (2) $p = 2$
- (c) What relationship does this density function have to the normal density? Hint: $\Gamma(1/2)$
- = $\sqrt{\pi}$
- (d) How could MLE of $\beta = (\beta_1, \dots, \beta_K)'$, σ and "p" be obtained (conceptually)?

The **STATA commands** for obtaining estimates of the unknown regression and distributional parameters with the GED distribution are given as follows:

```

clear
cap prog drop GED
program define GED
version 1.0
args lInf xb s p
quietly replace `lInf'=ln(abs(`p'))-((abs($ML_y1-'xb')/`s')^abs(`p'))-ln(2*`s')-lngamma(abs(1/^`p')))
end

clear
infile y x using g:\american.dat

ml model lf GED (y=x) (s:) (p:), technique(dfp)
ml search
ml maximize

```

For comparative purposes, the commands for MLE with a normal pdf are given by (STATA 10)

STATA 10

```
STATA commands for a normal pdf in a linear regression model
cap prog drop normalpdf
program define normalpdf
version 1.0
args lnf mu sigma (lnf, mu, and sigma are the parameters of estimation)
quietly replace `lnf'=ln(normalden($ML_y1, `mu', `sigma'))
end
ml model lf normalpdf (y=x's) (sigma: ), technique(dfp)
ml search
ml maximize, difficult
```

The fifth problem in the homework section involves an application of MLE using STATA with a GED pdf.

c. Estimation subject to constraints

Assume that we believe that the parameters θ satisfy the constraint

$$(2.15) \quad q(\theta)=0$$

where $q(\theta)$ can denote an $rx1$ vector of univariate real valued functions if multiple constraints exist (r constraints or restrictions). For example, if the CES production function is characterized by constant returns to scale, it can be shown that

$$M = 1$$

or

$$M - 1 = 0.$$

There are several approaches to estimation subject to constraints on the parameters. One is to reduce the number of parameters by substitution. Another approach provides some necessary groundwork for testing hypotheses of the form

$$q(\theta)=0$$

where $q(\theta)$ is differentiable and will be briefly discussed.

Assume that the objective function to be maximized is $H(\theta)$, e.g., if nonlinear least squares estimation is used then $H(\theta) = -SSE(\beta)$; if MLE is used then

$$H(\theta)=\ell(\beta, \sigma^2).$$

A second estimation procedure which takes account of constraints on parameters is based on Lagrangian multipliers.

Define

$$\mathcal{L}(\theta, \lambda) = H(\theta) + \lambda' q(\theta)$$

where $\lambda' = (\lambda_1, \dots, \lambda_r)$ and $q(\theta) = (q_1(\theta), \dots, q_r(\theta))'$ correspond to r constraints imposed on the parameters. λ_i is generally referred to as the i^{th} Lagrangian multiplier which corresponds to the i constraint $q_i(\theta) = 0$.

The necessary conditions associated with this constrained optimization problem,

$$(2.16) \quad \max_{\theta} H(\theta)$$

subject to $q(\theta) = 0$,

are given by

$$(2.17) \quad \frac{\partial \mathcal{L}(\theta, \lambda)}{\partial \theta} = 0$$

$$\begin{aligned} \frac{\partial \mathcal{L}(\theta, \lambda)}{\partial \lambda} &= q(\theta) \\ &= 0. \end{aligned}$$

The Lagrangian multipliers in this formulation can be used to test the validity of the constraints.

d. Other methods of estimation (generally require nonlinear optimization procedures—covered in another section of the notes in more detail)

- (1) Method of moments. Sample moments of the data are equated to the theoretical moments which are functions of the parameters. These are equated and the resulting estimators are usually consistent, but not necessarily efficient.
- (2) Generalized Method of Moments. See the discussion in the Regression Section which describes estimation where more moment restrictions are considered than the number of parameters.
- (3) Methods for grouped data. Data is in the form of cell frequencies. Predicted frequencies depend upon unknown parameters. Estimation procedures are based on "matching" observed

and predicted frequencies. Possible criteria: Least Squares, MLE, minimum or modified minimum chi-square.

(4) Extremum estimators. As discussed earlier, extremum estimators are defined by the equation

$$\tilde{\theta} = \arg \max_{\theta} H(\theta)$$

Where under appropriate regularity conditions,

$$\tilde{\theta} \sim^a N\left[\theta; C = B^{-1}AB^{-1}\right]$$

with $B = \left(E \frac{d^2H}{d\theta d\theta'}\right)$ and $A = \left(E \frac{dH}{d\theta} \frac{dH}{d\theta'}\right)$

This estimator includes the nonlinear least squares and maximum likelihood estimators as special cases. Additionally if MLE is performed with a questionable specification for the pdf, the corresponding estimator is frequently referred to as a Quasi Maximum Likelihood estimator (QMLE).

3. Computation

The derivation of either nonlinear least squares or maximum likelihood estimators requires the solution of a system of nonlinear equations. For example if the estimation problem has been written as a maximization problem

$$(3.1) \quad \max_{\theta} H(\theta), \quad H(\theta) = \begin{cases} - \text{SSE}(\beta) \text{ for L.S.} \\ \ell(\beta, \sigma^2) \text{ for M.L.E} \\ -g(\beta)'Var^{-1}(g(\beta))g(\beta) \end{cases}$$

estimation requires solving the equations:

$$(3.2) \quad H_{\theta}(\theta) = \frac{dH(\theta)}{d\theta} = 0$$

Before discussing numerical methods of computing solutions to (3.2) it should be mentioned that

- (1) Round off error can accumulate.
- (2) The objective function may have a "plateau" or several local optima.
- (3) Many techniques are based upon gradient techniques.

a. Generalized Gradient Methods

The procedure for maximizing (3.1) generally involves selecting an initial guess for θ , and then determining whether this corresponds to a local maximum. If not, determine a direction " δ " in which to adjust and how far (step size) "t" to move in that direction. This procedure is continued until the desired convergence is achieved.

The gradient vector $\left(\frac{dH}{d\theta} \right)$

indicates the direction of movement or adjustment in

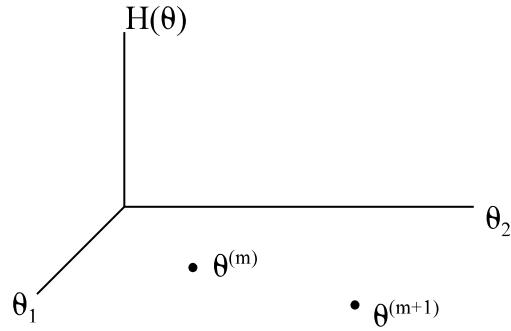
the θ 's which will result in the "most rapid" increase in $H(\theta)$. Some algorithms modify the direction of movement by multiplying the gradient vector by a matrix P to yield $P \left(\frac{dH}{d\theta} \right)$.

Let "t" denote a scale factor which determines the distance moved. Thus, an iterative search procedure can be implemented as follows:

$$(3.3) \quad \boxed{\theta^{(n+1)} = \theta^{(n)} + t_n \left\{ P_n \left(\frac{dH(\theta^{(n)})}{d\theta} \right) \right\}} \quad \text{or}$$

$$(3.4) \quad \boxed{\Delta\theta = t_n \left\{ P_n \left(\frac{dH(\theta^{(n)})}{d\theta} \right) \right\} \\ = t_n \delta_n}$$

This can be visualized as in the following figure



What type of matrices P_n will lead to an improvement, i.e. $H(\theta^{(n+1)}) > H(\theta^{(n)})$?

Consider the Taylor Series expansion of $H(\theta)$ about $\theta^{(n)}$

$$(3.5) \quad H(\theta^{(n+1)}) \doteq H(\theta^{(n)}) + \frac{dH(\theta^{(n)})}{d\theta} \Delta\theta^{(n)}$$

$$+ \frac{1}{2} \Delta\theta^{(n)} \frac{dH^2(\theta^{(n)})}{d\theta d\theta} \Delta\theta^{(n)}$$

Substituting (3.4) into (3.5) yields

$$(3.5)' \quad H(\theta^{(n+1)}) = H(\theta^{(n)}) + \frac{dH(\theta^{(n)})}{d\theta} t_n P_n \frac{dH}{d\theta}$$

$$+ \frac{t_n^2}{2} \frac{dH(\theta^{(n)})}{d\theta} P_n \cdot \frac{d^2H(\theta^{(n)})}{d\theta d\theta} P_n \frac{dH(\theta^{(n)})}{d\theta}$$

or

$$(3.6) \quad H(\theta^{(n+1)}) - H(\theta^{(n)})$$

$$\doteq \frac{dH(\theta^{(n)})}{d\theta} \left\{ t_n P_n + \frac{t_n^2}{2} P_n \cdot \frac{[d^2H(\theta^{(n)})]}{d\theta d\theta} P_n \right\} \frac{dH(\theta^{(n)})}{d\theta}$$

The iterative adjustment will correspond to an increase in $H(\theta)$, i.e.,

$$H(\theta^{(n+1)}) > H(\theta^{(n)})$$

if the bracketed expression is positive definite.

Special cases

(1.) Simple Gradient. ($P_n = I$).

$$\theta^{(n+1)} = \theta^{(n)} + t_n P_n \left(\frac{dH}{d\theta} \right)$$

$$= \theta^{(n)} + \left(\frac{dH}{d\theta} \right), \quad P_n = I, \quad t_n = 1$$

This method will converge, but may do so very slowly.

$$(2.) \text{ Newton Raphson} \quad \left(P_N = \left(\frac{d^2H}{d\theta d\theta} \right)^{-1} \right) \quad \text{_____}$$

$$\theta^{(n+1)} = \theta^{(n)} + t_n P_n \left(\frac{dH}{d\theta} \right)$$

The Newton Raphson works very well near an optimum. If we are a long way from an optimum, Newton-Raphson need not work well and may even send us in the wrong direction. If the original objective function $H(\theta)$ is quadratic, the Newton-Raphson algorithm will yield the solution in one step. The motivation for this selection of P_n is that (3.5)' is maximized with respect to $\theta^{(n+1)}$ to yield

$$P_n = \left(\frac{d^2H}{d\theta d\theta'} \right)^{-1}.$$

- (3.) Fletcher-Powell. Similar to Newton-Raphson except numerical derivatives are used to evaluate $d^2H/d\theta d\theta'$ in the previous equation.
- (4.) The Berndt, Hall, Hall, Hausman (BHHH) algorithm corresponds to using.

$$P_n = \left[\frac{dH}{d\theta} \frac{dH}{d\theta'} \right]^{-1}$$

- (5) other variations exist and many are characterized by different selections for P_n and t_n .

b. Direct Search and Grid Search Techniques.

See Goldfeld and Quandt--Nonlinear Methods in Econometrics, 1972 Amsterdam: North Holland.

4. Statistical Inference

The exact distribution of parameter estimators in nonlinear models is generally unknown. However, approximating asymptotic distributions can be determined in some important leading cases. There is no generally valid rule of thumb relating sample size for a desired accuracy of the asymptotic distributions. Consequently, the only justification for statistical inference in nonlinear models must be based on asymptotic considerations. One approach to estimating the unknown distribution is using what is called the bootstrap which will be briefly discussed in another section.

The asymptotic distribution of θ_{MLE} is

$$N(\theta, \Sigma_{\theta_{MLE}})$$

where the asymptotic variance of θ_{MLE} is given by

$$\Sigma_{\theta_{MLE}} = - \left(E \frac{\partial^2 \ell}{\partial \theta \partial \theta'} \right)^{-1} \quad \left(E \frac{d\ell}{d\theta} \frac{d\ell}{d\theta'} \right)^{-1}$$

the Cramer-Rao matrix, and also by

$$\sigma_{MLE}^2 \left[\frac{\partial f(X; \beta_{MLE})}{\partial \beta} \frac{\partial f(X; \beta_{MLE})}{\partial \beta'} \right]^{-1}$$

for the case of $\varepsilon \sim N[0; \sigma^2 I]$.

a. Confidence Intervals (asymptotic) for Individual θ_i s.

$$(4.1) \quad H_0: \theta_i = \theta_i^0$$

Confidence intervals for θ_i based on the asymptotic distribution, are given by

$$(4.2) \quad \theta_{MLE,i} \pm Z_{\alpha/2} \sigma_{MLE,i}$$

where $Z_{\alpha/2}$ = appropriate critical level for a standard normal,

$$\begin{aligned} \sigma_{MLE,i}^2 &= \text{MLE of asymptotic variance of } \theta_{MLE,i} \\ &\equiv \text{i}^{\text{th}} \underline{\text{diagonal element of }} \Sigma_{\theta_{MLE}} \end{aligned}$$

b. More General Tests, e.g.,

$$(4.3) \quad H_0: q(\theta) = 0,$$

against the alternative $H_a: q(\theta) \neq 0$. $q(\theta)$ is a $(rx1)$ vector of r -functional constraints on the individual parameters θ , $q(\theta)$ must be continuously differentiable and $(\partial q / \partial \theta')$ must be of full rank, r . If we want to test the hypothesis that the parameters in the CES production function satisfied the constraints

$$M = 1 \text{ and}$$

$$\delta = 1/2,$$

the constraints could be written in the form (4.3) as

$$q(\theta) = \begin{pmatrix} M & - & 1 \\ \delta & - & 1/2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

The Wald (W) test, Lagrangian Multiplier (LM) or Rao test and Likelihood ratio (LR) test can be used to test hypotheses of the form (4.3). They differ in ease of application, but share the same asymptotic distribution.

(1) The Wald test.

To perform a Wald test, the *unconstrained* (without the constraints imposed) model is estimated and the test statistic is defined by

$$(4.4) \quad \boxed{\begin{aligned} W &= q(\theta_{ML})' (\text{Var}(q(\theta_{ML}))^{-1}) q(\theta_{ML}) \\ W &= q(\theta_{ML})' \left(\frac{\partial q}{\partial \theta}, \sum_{\theta_{ML}} \frac{\partial q}{\partial \theta} \right)^{-1} q(\theta_{ML}) \end{aligned}}$$

(evaluated at θ_{ML}). The asymptotic distribution of W is a chi-square with degrees of freedom equal to r (the number of independent constraints on the θ_i 's),

$$W \stackrel{a}{\sim} \chi^2(r)$$

Note: (a) The construction of the Wald test only requires the estimation of the unconstrained model.

(b) Equivalence of the two forms of the Wald test.

Given that

$$\theta_{ML} \xrightarrow{a} N(\theta, \Sigma_{\theta_{ML}}),$$

it follows that $q(\theta_{MLE})$ has the following asymptotic distribution,

$$q(\theta_{ML}) \xrightarrow{a} N\left[q(\theta) = 0, \Sigma_{q(\theta_{ML})} = \frac{dq}{d\theta}, \Sigma_{\theta_{ML}} \frac{dq}{d\theta}\right]$$

Demonstration(this methodology is very important for PhD bound students):

Consider the Taylor series expansion of $q(\theta_{ML})$ about the true value of θ :

$$q(\theta_{ML}) \doteq q(\theta) + \frac{dq}{d\theta} (\theta_{ML} - \theta)$$

Consequently,

$$\text{Asympt mean } q(\theta_{ML}) = q(\theta) = 0$$

$$\text{Asympt var } q(\theta_{ML}) = \frac{dq}{d\theta}, \Sigma_{\theta_{ML}} \frac{dq}{d\theta}$$

Therefore,

$$q(\theta_{ML}) \xrightarrow{a} N\left[q(\theta) = 0, \Sigma_{q(\theta_{ML})} = \frac{dq}{d\theta}, \Sigma_{\theta_{ML}} \frac{dq}{d\theta}\right].$$

(c) It follows that

$$(q(\theta_{MLE}) - 0) \left(\frac{dq}{d\theta}, \Sigma_{\theta_{MLE}} \frac{dq}{d\theta} \right)^{-1} (q(\theta_{MLE}) - 0)$$

is asymptotically distributed as a $\chi^2(r)$.

(2) Lagrangian Multiplier (LM) or Rao Test.

The test statistic is defined by

$$(4.5) \quad \begin{aligned} LM &= \left(\frac{\partial \ell}{\partial \theta} \right)' \left(\text{Var} \left(\frac{\partial \ell}{\partial \theta} \right) \right)^{-1} \left(\frac{\partial \ell}{\partial \theta} \right) \\ LM &= \left(\frac{\partial \ell}{\partial \theta} \right)' \Sigma_{\theta_R} \left(\frac{\partial \ell}{\partial \theta} \right) \end{aligned}$$

where the statistic $\frac{\partial \ell}{\partial \theta}$ is evaluated at the restricted maximum likelihood estimator (θ_R - the solution to 2.16). The motivation for the form and distribution of

LM follows from the results established I.B Appendix B (Cramer-Rao Inequality):

The relevant results are

- $E \left(\frac{d\ell}{d\theta} \right) = 0$
- $\text{Var} \left(\frac{d\ell}{d\theta} \right) = -E \left(\frac{d^2\ell}{d\theta d\theta'} \right) = (\text{Var}(\theta_R))^{-1} = E \left(\frac{\partial \ell}{\partial \theta} \frac{\partial \ell}{\partial \theta'} \right)$

for the case where $q(\theta) = 0$. Therefore the quadratic form in (4.5) has an asymptotic chi-square distribution,

$$LM \xrightarrow{a} \chi^2(r)$$

- Note:
- (1) The evaluation of LM using (4.5) requires ML estimates of θ subject to the restriction, (θ_R) , and also the associated $(\partial \ell / \partial \theta)$.
 - (2) Unconstrained estimates of θ are not required.
 - (2) An alternative form for (4.5) can be expressed in terms of λ_{ML} , the vector of Lagrangian multipliers associated with $q(\theta) = 0$. See (2.16) and (2.17).

Recall that the restricted MLE of θ , θ_R , satisfies

$$\begin{aligned} & \text{Max } l(\theta) \\ & \ell \\ & \text{s.t. } q(\theta) = 0. \end{aligned}$$

The solution can be structured in terms of

$$l(\theta) + \lambda' q(\theta).$$

Maximizing this with respect to θ yields the condition

$$\frac{\partial l(\theta)}{\partial \theta} + \frac{\partial q}{\partial \theta} \lambda_{ML} = 0$$

$$\lambda_{ML}' \frac{\partial q}{\partial \theta} = -\frac{\partial l(\theta)}{\partial \theta}.$$

Therefore,

$$(4.5)' \quad LM = \lambda_{ML}' \frac{\partial q}{\partial \theta} \Sigma_{\theta_R} \frac{\partial q}{\partial \theta} \lambda_{ML}$$

(4.5) and (4.5)' can be further simplified if the residuals are $N[0, \sigma^2 I]$. In this case

$$\frac{-\partial l}{\partial \beta} = (Y - f(X; \beta_R))' \frac{df}{d\beta} / \sigma_R^2$$

from (2.13)

Therefore,

$$\begin{aligned} LM &= \left(\frac{(Y - f(X; \beta_R))'}{\sigma_R^2} \right) \frac{df}{d\beta} \left[\sigma_R^2 \left(\frac{df}{d\beta} \frac{df}{d\beta} \right)^{-1} \right] \\ &\cdot \frac{df}{d\beta} \frac{(Y - f(X; \beta_R))}{\sigma_R^2} \end{aligned}$$

or

$$(4.5)'' \quad LM = \frac{(Y - f(X; \beta_R))' \frac{df}{d\beta} \left(\left(\frac{df}{d\beta} \frac{df}{d\beta} \right)^{-1} \frac{df}{d\beta} (Y - f(X; \beta_R)) \right)}{SSE(\beta_R)}$$

This form doesn't require estimates of λ_{ML} .

(3) The Likelihood Ratio Test.

This test involves obtaining MLE estimates of $\theta = (\beta', \psi')$ with the constraints imposed and without the constraints imposed, denoted (β_R, ψ_R) and (β_{ML}, ψ_{ML}) .

The test statistic is defined by

$$(4.6) \quad \boxed{\begin{aligned} LR &= 2[\ell(\beta_{ML}, \psi_{ML}) - \ell(\beta_R, \psi_{ML})] \\ &= 2(\ell - \ell*) \end{aligned}}$$

and is asymptotically distributed as a chi square with the same degrees of freedom as for the Wald and LM test statistics. For the case in which

$$\varepsilon \sim N(0, \sigma^2 I),$$

the reader can use (2.13) to demonstrate that

$$(4.6)' \quad LR = \frac{[SSE(\beta_R) - SSE(\beta_{ML})]}{\sigma^2} \quad \text{or}$$

$$(4.6)'' \quad \boxed{LR = (N) \left(\ln \left(\frac{SSE_R}{SSE} \right) \right) = N[\ln SSE_R - \ln SSE]}$$

depending on whether σ^2 is known or not. Consider the relationship between (4.6) and the familiar Chow test used to test hypotheses of the form $q(\beta) = 0$ for linear regression models.

It can also be shown that for the special case of errors

$$\varepsilon \sim N[0, \sigma^2 I]$$

in the linear regression model, i.e.,

$$Y = f(X; \beta) + \varepsilon$$

$$= X\beta + \varepsilon,$$

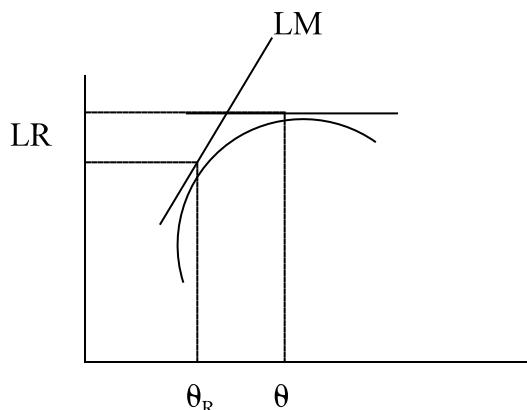
the W, LM and LR tests for testing overall explanatory power $H_0: \beta_2 = \beta_3 = \dots = \beta_K = 0$ can be expressed in terms of R^2 as

$$\boxed{\begin{aligned} W &= NR^2/(1-R^2) \\ LM &= NR^2 \\ LR &= -N \ln(1-R^2). \end{aligned}}$$

We also note in this case that $W \geq LR \geq LM$. A proof of this result uses the inequality $X \geq \log_e(1 + X) \geq X/(1 + X)$.

The relationship between the LR, W and LM test can be easily visualized as follows:

The Wald, LR, and LM



tests are appropriate for nested

hypotheses, where one model includes another as a special case. Several approaches to testing nonnested hypotheses have been developed. One approach to testing nonnested hypotheses is to nest both hypotheses (models) as special cases of a more general model.

5. Some Applications

This section includes a brief discussion of a number of nonlinear econometric models including (a) the variable elasticity of substitution production function, (b) models for the distribution of income, (c) a precautionary note pertaining to computations associated with nonlinear models, (d) qualitative response models, (e) ordered or categorical data, (f) Poisson regression, (g) censored, (g) truncated or grouped regression models, and (h) quantile regression.

a. Variable Elasticity of Substitution (VES) Production Function

The VES production function is defined by

$$(5.1) \quad Y_t = A[(\rho - 1) K_t + L_t]^{\alpha \delta \rho} K_t^{\alpha(1-\delta\rho)} \varepsilon_t \\ = A[\rho - 1 + L_t/K_t]^{\alpha \delta \rho} K_t^\alpha \varepsilon_t.$$

Taking the logarithm of both sides of (5.1) yields

$$(5.1)' \quad \ln Y_t = \ln A + \alpha \delta \rho \ln[(\rho - 1) K_t + L_t] + \alpha(1 - \delta \rho) \ln K_t + \ln \varepsilon_t \\ = \ln A + \alpha \ln K_t + \alpha \delta \rho \ln[(\rho - 1) + L_t/K_t] + \ln \varepsilon_t.$$

The elasticity of substitution of this production function is given by

$$(5.2) \quad \sigma = \frac{\% \Delta(K/L)}{\% \Delta(\omega_L/\omega_K)} \\ = 1 + \left(\frac{\rho - 1}{1 - \delta \rho} \right) \frac{K}{L}.$$

The parameter α presents the returns to scale.

The VES production function includes the Cobb-Douglas production as a special case, i.e., if $\rho=1$ then (5.1) simplifies to

$$(5.3) \quad Y_t = A L_t^{\alpha \delta} K_t^{\alpha(1-\delta)} \\ = A L_t^{\alpha \delta} K_t^{\alpha(1-\delta)} \varepsilon_t$$

Note that the corresponding elasticity of substitution for the Cobb-Douglas production function is one. See (5.2) with $\rho=1$.

Estimation:

If the $\ln\epsilon_t$ are identically and independently distributed as $N(0, \sigma^2)$, then nonlinear least squares estimators are the same as maximum likelihood estimators. The sum of squared errors is given by

$$(5.4) \quad SSE(A, \alpha, \delta, \rho) = \sum_{t=1}^N (\ln Y_t - \ln A - \alpha \delta \rho \ln[(\rho-1)K_t + L_t] - \alpha (1-\delta\rho) \ln K_t)^2$$

and the log likelihood function is given by (see equation (2.13))

$$(5.5) \quad \ell(A, \alpha, \delta, \rho, \sigma^2) = -SSE(\cdot)/2\sigma^2 - \frac{N}{2}(\ln \sigma^2 + \ln 2\pi).$$

The parameter σ^2 can be estimated by

$$(5.6) \quad \hat{\sigma}_{NLS}^2 = SSE(\cdot)/N-4$$

$$\text{or} \quad \hat{\sigma}_{MLE}^2 = SSE(\cdot)/N$$

where the least squares estimates of A, α, δ, ρ , are used in evaluating SSE. Fortunately econometric packages not only calculate the parameter estimates, but also estimate standard errors. The asymptotic variance covariance matrix of $A_{NLS}, \alpha_{NLS}, \delta_{NLS}, \rho_{NLS}$ (same as maximum likelihood estimators) is given by

$$(5.7) \quad \sigma_{NLS}^2 \left(\frac{\partial f}{\partial \beta} \frac{\partial f}{\partial \beta'} \right)^{-1}$$

where $f(A, \alpha, \delta, \rho) = \begin{pmatrix} \ln A + \alpha \ln K_1 + \alpha \delta \rho \ln[(\rho-1) + L_1/K_1] \\ \vdots \\ \ln A + \alpha \ln K_N + \alpha \delta \rho \ln[(\rho-1) + L_N/K_N] \end{pmatrix}$

(5.8)

$$\frac{\partial f}{\partial \beta} = \begin{pmatrix} \frac{1}{A} & \cdot & \frac{1}{A} \\ \delta\rho \ln[\rho-1 + L_1/K_1] + \ln K_1 & \cdot & \delta\rho \ln[\rho-1 + L_N/K_N] + \ln K_N \\ \alpha \ln[\rho-1 + L_1/K_1] & \cdot & \alpha \ln[\rho-1 + L_N/K_N] \\ \frac{\alpha \delta \rho}{\rho-1 + \frac{L_1}{K_1} + \alpha \delta \ln[\rho-1 + \frac{L_1}{K_1}]} & \cdot & \frac{\alpha \delta \rho}{\rho-1 + \frac{L_N}{K_N} + \ln[\rho-1 + \frac{L_N}{K_N}]} \end{pmatrix}$$

The columns correspond to observations and the rows correspond to derivatives of $f(\cdot)$ with respect to the parameters A , α , δ and ρ , respectively. The derivatives are evaluated at the non linear least squares estimates. The Cramer-Rao matrix could also be used. Several optimization algorithms use numerical approximations to (5.8). Numerical approximations of this matrix are standard output from many software packages.

Testing Hypotheses

In order to test hypotheses about individual parameters, the asymptotic standard errors in (5.7) can be used. For example, in order to test whether (5.1) is characterized by constant returns to scale, we test the hypothesis

$$(5.9) \quad H_0: \alpha = 1.$$

The corresponding test statistic is given by

$$(5.10) \quad \left(\frac{\alpha_{NLS} - 1}{\hat{\sigma}_{\alpha_{NLS}}} \right) \stackrel{a}{\sim} N(0,1)$$

where

$\hat{\sigma}_{NLS}$ = square root of the element in the 2nd row and 2nd column of (5.7).

In order to test whether (5.1) can be modeled by a Cobb-Douglas production function with constant returns to scale, we test the hypothesis

$$(5.11) \quad H_0: \rho=1 \text{ and } \alpha=1,$$

and the Wald test can be used, or the Lagrange or Likelihood ratio test could be applied.

Define

$$(5.12) \quad q(\theta) = \begin{pmatrix} \rho - 1 \\ \alpha - 1 \end{pmatrix}.$$

The Wald test of $q(\beta) = 0$ is constructed as

$$(5.13) \quad W = (q(\theta_{ML}))' (\text{Var}(q(\theta_{ML})))^{-1} q(\theta_{ML})$$

$$= (\rho_{ML} - 1, \alpha_{ML} - 1) \begin{bmatrix} \sigma_{\rho\rho} & \sigma_{\rho\alpha} \\ \sigma_{\alpha\rho} & \sigma_{\alpha\alpha} \end{bmatrix}^{-1} \begin{pmatrix} \rho_{ML} - 1 \\ \alpha_{ML} - 1 \end{pmatrix} \stackrel{a}{\sim} \chi^2(2).$$

Note that this test statistic only requires that the unconstrained model (5.1) be estimated.

In this case $\text{Var}(q(\theta_{ML}))$ is easily constructed from the simple form of the hypotheses being tested. More generally, note that

$$(5.14) \quad \text{Var}(q(\theta_{ML})) = \left(\frac{\partial q}{\partial \theta'} \right) \Sigma_{\theta_{ML}} \left(\frac{\partial q}{\partial \theta} \right)$$

$$= \begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \sigma_{AA} & \sigma_{A\alpha} & \sigma_{A\delta} & \sigma_{Ap} \\ \sigma_{\alpha A} & \sigma_{\alpha\alpha} & \sigma_{\alpha\delta} & \sigma_{\alpha p} \\ \sigma_{\delta A} & \sigma_{\delta\alpha} & \sigma_{\delta\delta} & \sigma_{\delta p} \\ \sigma_{pA} & \sigma_{p\alpha} & \sigma_{p\delta} & \sigma_{pp} \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 0 & 0 \\ 1 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_{\rho\rho} & \sigma_{\rho\alpha} \\ \sigma_{\alpha\rho} & \sigma_{\alpha\alpha} \end{bmatrix}$$

where

$$(5.15) \quad \frac{\partial q}{\partial \theta'} = \begin{bmatrix} \frac{\partial(\rho-1)}{\partial A} & \frac{\partial(\rho-1)}{\partial \alpha} & \frac{\partial(\rho-1)}{\partial \delta} & \frac{\partial(\rho-1)}{\partial p} \\ \frac{\partial(\alpha-1)}{\partial A} & \frac{\partial(\alpha-1)}{\partial \alpha} & \frac{\partial(\alpha-1)}{\partial \delta} & \frac{\partial(\alpha-1)}{\partial p} \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

The Lagrangian and Likelihood Ratio tests require that the constrained model be estimated. The constrained model associated with (5.11) is given by

$$(5.16) \quad Y_t = AL_t^\delta K_t^{1-\delta} \epsilon_t$$

or

$$(5.16)' \quad \ln Y_t = \ln A + \delta \ln L_t + (1-\delta) \ln K_t + \ln \varepsilon_t$$

or

$$(5.16)'' \quad \ln(Y_t/K_t) = \ln A + \delta \ln(L_t/K_t) + \ln \varepsilon_t.$$

Equation (4.5)'' can be used to evaluate LM.

$$(5.17) \quad \frac{\partial f}{\partial \beta} = \begin{pmatrix} \frac{1}{A} & \dots & \frac{1}{A} \\ \ln\left(\frac{L_1}{K_1}\right) & \dots & \ln\left(\frac{L_N}{K_N}\right) \end{pmatrix}$$

$$(5.18) \quad \frac{\partial f}{\partial \beta} \frac{\partial f}{\partial \beta'} = \begin{pmatrix} \frac{N}{A} & \sum_t \ln(L_t/K_t)/A \\ \sum_t \ln(L_t/K_t)/A & \sum_t \ln^2(L_t/K_t) \end{pmatrix}$$

$$(5.19) \quad (Y - f(x; \beta_R))' = (\ln Y_1 - \ln A - \delta \ln L_1 - (1-\delta) \ln K_1, \dots, \ln Y_N - \ln A - \delta \ln L_N - (1-\delta) \ln K_N).$$

SSE(β_R) is the sum of squared errors associated with the constrained model, (5.16)'. Substituting these expressions into (4.5)'' yields the LM test statistic which has a $\chi^2(2)$ asymptotic distribution.

(4.6) and (4.6)'' are easily implemented forms for the likelihood ratio test statistic. This form only requires that the sum of squared errors for the constrained model (5.16)'' and the sum of squared errors for the unconstrained model (5.1)' be obtained. This test statistic is also asymptotically distributed as $\chi^2(2)$. If normality is not assumed, then equation (4.6) can be used.

b. Models for the Distribution of Income

Many of the models for the distribution of income are special cases of the *generalized beta of the second kind* (GB2) and *generalized beta of the first kind* (GB1).

$$(5.20a) \quad GB2(y; a, b, p, q) = \frac{|a| y^{(ap-1)}}{b^{ap} B(p, q) (1 + (y/b)^a)^{p+q}}$$

$$(5.20b) \quad GB1(y; a, b, p, q) = \frac{|a| y^{(ap-1)} (1 - (y/b)^a)^{q-1}}{b^{ap} B(p, q)}$$

These include the *generalized gamma* (GG)

$$(5.21) \quad GG(y; a, \beta, p) = \frac{|a| y^{ap-1} e^{-(y/\beta)^a}}{\beta^{ap} \Gamma(p)}$$

which corresponds to the limit of (5.20a-b) as $q \rightarrow \infty$, $b = \beta q^{1/a}$;

the *gamma* (G)

$$(5.22) \quad G(y; \beta, p) = GG(y; a, 1, \beta, p) = \frac{y^{p-1} e^{-y/\beta}}{\beta^p \Gamma(p)};$$

the *Singh Maddala* (SM)

$$(5.23) \quad SM(y; a, b, q) = GB2(y; a, b, p=1, q) = \frac{|a| y^{a-1}}{b^a B(1, q) (1 + (y/b)^a)^{q+1}};$$

the Fisk (F)

$$(5.24) \quad FISK(y; a, b) = GB2(y; a, b, p=q=1) = \frac{|a| y^{a-1}}{b^a (1 + (y/b)^a)^2};$$

the Weibull (W)

$$(5.25) \quad W(y; a, \beta) = GG(y; a, \beta, p=1) = \frac{|a| y^{a-1} e^{-(y/\beta)^a}}{\beta^a}$$

and the Lognormal (LN)

$$(5.26) \quad \text{LN}(y; \mu, \sigma^2) = \frac{e^{\frac{-1}{2\sigma^2}(\ln y - \mu)^2}}{y\sqrt{2\pi}\sqrt{\sigma^2}}$$

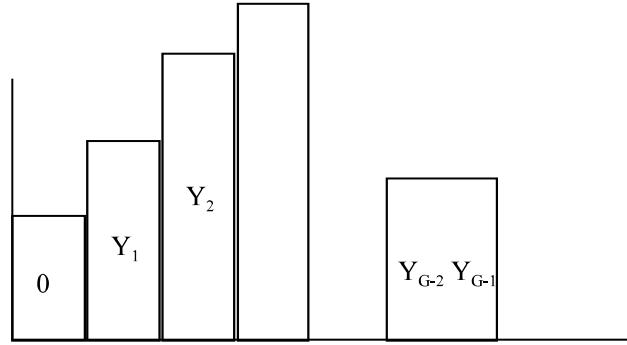
which is the limit of (5.21) as $a \rightarrow 0$, $b^a = \sigma^2 a^2$, $p = (a\mu + 1)/b^a$.

These distributions can be fit to individual observations using maximum likelihood procedures in the usual manner by maximizing

$$\ell = \sum_{t=1}^n \ln f(y_t; \theta)$$

with respect to the underlying parameters θ , where $f(y; \theta)$ denotes the assumed distribution.

If the data are in a grouped format, with G groups, n_i observations will be reported in the i^{th} group, (Y_{i-1}, Y_i) , for $i = 1, 2, \dots, G$ and alternative procedures must be employed.



The corresponding histogram might appear as above where

$0 < Y_1 < Y_2 < \dots < Y_{G-1} < Y_G = \infty$ partitions all possible incomes into G groups. Further, let $F(Y; \theta)$ denote the cumulative distribution function

$$F(Y; \theta) = \int_{-\infty}^Y f(y; \theta) dy.$$

The observed relative frequencies in the i^{th} interval are given by n_i/n where $(n = \sum_{i=1}^G n_i)$

and the predicted probabilities for the i^{th} interval is given by

$$p_i(\theta) = \int_{Y_{i-1}}^{Y_i} f(y; \theta) dy = F(Y_i; \theta) - F(Y_{i-1}; \theta).$$

$p_i(\theta)$ is a fu

Estimators of the parameters θ can be obtained for grouped data by solving any of the following problems:

$$\min_{\theta} \left\{ \text{SSE} = \sum_{i=1}^G \left(\frac{n_i}{n} - p_i(\theta) \right)^2 \right\}$$

$$\min_{\theta} \left\{ \text{SAE} = \sum_{i=1}^G \left| \frac{n_i}{n} - p_i(\theta) \right| \right\}$$

$$\min_{\theta} \left\{ \chi^2 = \sum_{i=1}^G \left(\frac{n_i}{n} - p_i(\theta) \right)^2 / p_i(\theta) \right\}$$

$$\max_{\theta} \left\{ \ell = n! \prod_{i=1}^G (p_i(\theta))^{n_i} / n_i! \right\}.$$

The first two estimators might be thought of as least squares and least absolute deviations estimators.

The third is a minimum chi-square estimator and the fourth is a maximum likelihood estimator. The third and fourth estimators are asymptotically efficient.

An example of fitting statistical distributions to the following grouped data from Census Population Reports for 1980 family income by maximizing the multinomial loglikelihood function is reported in table 1. The corresponding 1980 data are given by

(thousands)	(%)
2.5	2.1
5.0	4.1
7.5	6.2
10.0	6.5
12.5	7.3
15.0	6.9
20.0	14.0
25.0	13.7
35.0	19.8
50.0	12.8
∞	6.7
ENDPOINTS	(n = 40,000)

TABLE 1. ESTIMATED DISTRIBUTION FUNCTIONS
1980 FAMILY INCOME

	GB1	GB2	B1	GG	B2	SM	LN	G	W	F	E
a	1.4008	2.5373	1.0000	1.4008	1.0000	1.6971	$\mu=2.9372$	1.0000	1.6057	2.2768	1.0000
b(β)	273102.	40.7667	163.757	(21.8145)	3535660.	87.6981	$\sigma=0.7797$	(11.0473)	(26.3368)	19.7450	(24.5944)
p	1.2454	0.6117	1.9173	1.2454	2.1555	1.0000		2.1557	1.0000	1.0000	1.0000
q	549517.	2.1329	11.3828		320081.	8.3679				1.0000	
Mean*	23.644	23.931	23.604	23.646	23.810	23.730	25.564	23.815	23.065	27.749	24.59
Gini**	.353	.359	.353	.353	.363	.355	.419	.363	.351	.439	.50
SAE	.0004	0.0002	0.0006	0.0004	0.0008	0.0003	0.0070	0.0008	0.0005	0.0053	0.02
SSE	.0545	0.0385	0.609	0.0545	0.0775	0.0495	0.2326	0.0775	0.0561	0.2038	0.41
χ^2	143.0	84.3	188.4	143.0	353.5	114.2	4620.7	353.6	183.6	2438.9	9107
$-\ell$	129.1	99.1	151.6	129.1	228.3	114.5	1859.4	228.3	150.0	1219.5	5108

*Census Estimate: 23.974

**Census Estimate: .365

Exercises:

- (1) Compare the relative fits of the various distributions. Which distributions provide the best fits for 2, 3 and 4 parameter models.
- (2) Indicate how a statistical test can be performed to check whether the SM model is consistent with the generalized beta of the second kind, $H_0: GB2 = SM$. Hint: consider the likelihood ratio test. Wald and LM tests could also be used, but would require more information than is reported in the table. A comparison of the sum of absolute errors (SAE), sum of squared errors, χ^2 value and log likelihood estimates reveals considerable difference in "goodness of fit" between the various models.

c. A Precautionary Note: Nonlinear Optimization

Numerous estimation problems can arise in nonlinear estimation which may yield questionable results. An indirect check of the validity of the parameter estimates obtained from a nonlinear optimization routine is provided by comparing estimated population characteristics such as the mean with independently obtained results where available. The following example provides an example of the importance of this.

Thurow's widely cited [AER, 1970] paper provides an example of estimation problems. The underlying distribution of income was assumed to be modeled by a beta density function of the first kind (B1).

$$B1(y; b, p, q) = \frac{y^{p-1}(b-y)^{q-1}}{B(p, q)b^{p+q}} \quad 0 < y < b, \quad p, q > 0$$

Thurow assumed that the maximum income (b) was equal to \$15,000 and obtained separate estimates of p and q for the distribution of income (1959 dollars) of families and unrelated individuals for whites and nonwhites for the period 1949-1966. Income characteristics associated with the estimated parameter values for (p, q, b) are inferred and their relationship with hypothesized explanatory variables considered. Thurow's results raise questions as to whether economic growth is associated with a more egalitarian distribution as well as

suggesting that inflation may lead to a more equal distribution of income for whites. The accuracy of the estimated (p,q)'s is a critical element in the validity of the analysis of the estimated relationship between the hypothesized explanatory variables and the distribution of income. Thurow's estimates of (p,q) were not reported in his paper, but were provided on request and are given in Table 2. The mean and Gini coefficient associated with the beta function are given by

$$E(Y) = \frac{bp}{p+q}.$$

$$G = \frac{\Gamma(p+q)\Gamma(p+1/2)\Gamma(q+1/2)}{\Gamma(p+q+1/2)\Gamma(p+1)\Gamma(q)\Gamma(1/2)},$$

(McDonald [1984]). The mean income level and Gini coefficients implied by Thurow's estimates can be readily obtained by substituting parameter estimates into these equations. . These corresponding estimates are reported in Table 2. Independent estimates obtained in various census publications are also given in this table to provide a useful comparison.

An analysis of the entries in this table suggests that the distribution of income for whites is more egalitarian with a higher mean than for nonwhites. This qualitative result is consistent with Thurow's estimates; however, other implications of these two sets of estimates are not. For example, all of the associated estimated density functions are either "U" shaped ($p < 1, q < 1$) or "L" shaped ($p < 1, q > 1$) rather than "U" shaped. The magnitude and intertemporal behavior (reductions in excess of 30 percent) of the associated Gini coefficients implied by the estimated parameters (p,q) for the period under consideration are inconsistent with the census estimates and provide additional evidence of an estimation problem. The agreement between the implied and census estimates of the mean is much closer than for the Gini coefficients.

Thus there is relatively close agreement between the two estimates of mean income, but very poor agreement between the measures of inequality. The estimation procedure appears

to have roughly preserved the mean characteristic, but implicitly modeled intra and/or intergroup variation incorrectly. The results could also have been partially due to the conjunction of the nature of the income groups and treatment of the maximum income.

The previous section included examples in which estimated distributions (B1) of the form considered by Thurow provide relatively accurate estimates of population characteristics.

THUROW'S ESTIMATES OF p, q^a

Whites				Nonwhites				
Year	p	q	mean ^b	Gini ^b	p	q	mean ^b	Gini ^b

1949	.258	.666	4.188	(4.052)	.615	(.404)	.160	.930	2.202	(2.044)	.752	(.443)
1950	.279	.687	4.332	(4.294)	.600	(.407)	.172	.930	2.341	(2.245)	.738	(.438)
1951	.269	.625	4.513	(4.441)	.596	(.387)	.182	.908	2.505	(2.334)	.725	(.433)
1952	.298	.649	4.720	(4.579)	.576	(.398)	.205	.921	2.731	(2.518)	.702	(.407)
1953	.327	.667	4.935	(4.798)	.556	(.395)	.228	.920	2.979	(2.638)	.679	(.428)
1954	.334	.697	4.859	(4.741)	.557	(.401)	.217	.929	2.840	(2.579)	.691	(.456)
1955	.368	.718	5.083	(5.034)	.536	(.397)	.225	.913	2.966	(2.679)	.681	(.431)
1956	.411	.731	5.398	(5.311)	.511	(.391)	.249	.978	3.044	(2.875)	.666	(.427)
1957	.406	.728	5.370	(5.229)	.513	(.385)	.269	1.025	3.118	(2.860)	.652	(.435)
1958	.411	.750	5.310	(5.291)	.514	(.388)	.276	1.075	3.064	(2.893)	.651	(.448)
1959	.460	.765	5.633	(5.571)	.488	(.396)	.286	1.051	3.209	(2.977)	.641	(.452)
1960	.504	.815	5.732	(5.646)	.473	(.398)	.330	1.061	3.559	(3.276)	.608	(.459)
1961	.622	.979	5.828	(5.817)	.443	(.408)	.346	1.173	3.417	(3.268)	.607	(.462)
1962	.663	.971	6.086	(5.987)	.426	(.395)	.338	1.107	3.509	(3.278)	.607	(.443)
1963	.712	.985	6.294	(6.167)	.411	(.396)	.356	1.073	3.737	(3.513)	.591	(.440)
1964	.785	1.017	6.534	(6.333)	.391	(.400)	.406	1.095	4.057	(3.788)	.562	(.444)
1965	.842	1.029	6.750	(6.552)	.376	(.393)	.452	1.124	4.302	(3.859)	.538	(.427)
1966	.955	1.044	7.166	(6.912)	.348	(.390)	.514	1.104	4.765	(4.192)	.504	(.426)

^aThurow did not estimate the parameter b, but rather assumed it to be 15 (\$15,000) and included any higher incomes in the group with an upper bound of \$15,000.

^bThe mean and Gini coefficients were evaluated using the given equations. The numbers in parentheses are the corresponding census estimates reported in current populations reports (P60). The nominal figures for mean income were adjusted by the CPI to obtain the figures in 1959 dollars.

d. Models with binary dependent variables or limited dependent variables

(1) Introduction

Consider models in which one might want to explain:

- (a) when there will be a default on a loan ($Y = 1$) or no default ($Y = 0$);
- (b) whether a tax return has been filed by someone who has misrepresented their financial position ($Y = 1$) or accurately reflects their financial situation ($Y = 0$);
- (c) The market share of a firm ($0 \leq Y \leq 1$)

These are known as limited dependent variable problems. Amemiya (1981) has an excellent survey paper on these models in the [Journal of Economic Literature](#).

In each case the dependent variable (Y) in the function

$$Y = f(X; \beta) + \varepsilon$$

is constrained in value.

Numerous approaches have been adopted to solve this problem and these include regression analysis, linear probability models, discriminant analysis and limited dependent models.

(2) Linear Probability Model (LPM)

$$\text{Let } Y_t = X_t\beta + \varepsilon_t$$

where

$$\begin{aligned} Y_t &= 1 \text{ if first option chosen} \\ &= 0 \end{aligned}$$

X_t vector of values of attributes
 (independent variables)

β corresponding vector of unknown coefficients

ε_t independently distributed random variable
 with zero mean

The specification of the model implies:

$$E(Y_t) = X_t\beta.$$

Now let $P_t = \text{Prob}(Y_t = 1)$

$$Q_t = 1 - P_t = \text{Prob}(Y_t = 0)$$

So that

$$\begin{aligned} E(Y_t) &= 1 \cdot \text{Prob}(Y_t = 1) + 0 \cdot \text{Prob}(Y_t = 0) \\ &= 1 \cdot P_t + 0 \cdot Q_t \\ &= P_t = X_t \beta. \end{aligned}$$

Thus the regression equation can be interpreted as describing, given the vector X_t , the probability that the first choice is made. The vector β measures the effect on the probability of choosing the first alternative corresponding to unit changes in

explanatory variables, $\beta_i = \frac{\partial Y_t}{\partial X_{ti}}$. If we estimate the equation using OLS, we may

obtain estimates of β . However, there is some question about the applicability of OLS in this model. To explore this issue, note the following:

$$\epsilon_t = Y_t - X_t \beta$$

$$E(\epsilon_t) = P_t(1 - X_t \beta) + (1 - P_t)(-X_t \beta) = P_t - X_t \beta$$

$$E(\epsilon_t) = 0 \text{ implies } P_t = X_t \beta \text{ and } (1 - P_t) = 1 - X_t \beta.$$

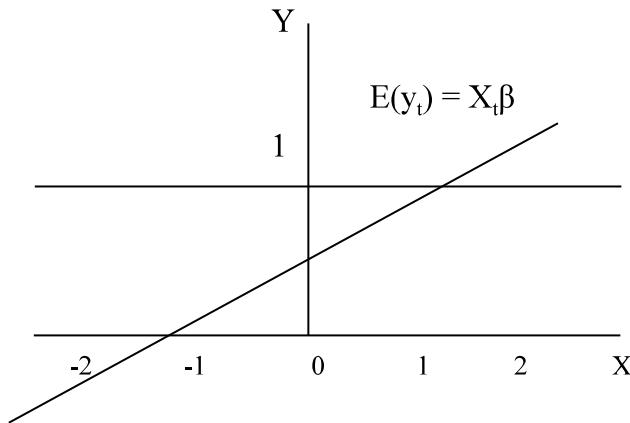
To find the variance of the error term ϵ_t , we evaluate

$$\begin{aligned} \text{Var}(\epsilon_t) &= E(\epsilon_t^2) = (1 - X_t \beta)^2 P_t + (-X_t \beta)^2 (1 - P_t) \\ &= (1 - X_t \beta)^2 (X_t \beta) + (X_t \beta)^2 (1 - X_t \beta) \\ &= (1 - X_t \beta)(X_t \beta) \\ &= P_t(1 - P_t) \\ &= E(Y_t)(1 - E(Y_t)) = (X_t \beta)(1 - X_t \beta), \end{aligned}$$

which clearly shows the error is heteroskedastic. One possible solution is to use weighted least squares.

Another problem with the LPM model is that of prediction:

Note that with the linear probability model there is a chance that predicted values for Y_t may lie outside the interval $[0, 1]$.

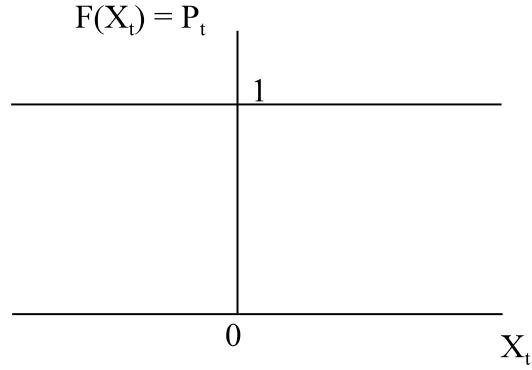


One possible solution is to set all predictions greater than 1 equal to 1 and all predictions less than 0 equal to zero. However, this approach presents a problem in running weighted least squares.

(3) Limited Dependent Variable Models

(a) Introduction

Multiple discriminant analysis (MDA) (Multivariate Statistics classes) is still another approach to problems of this type. MDA is closely related to the LPM model. Another possibility for binary or limited dependent variables is to use constrained estimation. Since Y_t or a transformation is constrained to the interval $(0,1)$, functional forms $F(x_t)$ which are constrained to the interval $(0,1)$ can be selected. This quite naturally suggests the use of cumulative distributions for $F(X_t)$.



This possibility admits many alternative models:

$$P_t = \Pr(Y_t = 1 | X_t) = F(X_t\beta; \theta) = \int_{-\infty}^{X_t\beta} f(s; \theta) ds$$

where $f(s; \theta)$ denotes a "well behaved" density function with distributional parameters θ . $X_t\beta$ are often referred to as the "scores." $F(x_t\beta; \theta)$ is then in the form of a cumulative distribution function. Two models which have been widely used are the standard normal and logistic models:

	$f(s; \theta)$	$F(z) = \int_{-\infty}^z f(s; \theta) ds$	
Normal	$\frac{e^{-s^2/2}}{\sqrt{2\pi}}$	$\int_{-\infty}^z \frac{e^{-s^2/2}}{\sqrt{2\pi}} ds$	Probit
Logistic	$\frac{e^{-s}}{(1 + e^{-s})^2}$	$\frac{1}{1 + e^{-z}}$	Logit

These two distributions are only two of many which could have been used; however, this literature has been dominated by the use of the probit (based on the normal) and logit (based on the log-logistic) models.

(b) Interpretation of Coefficients

Using Leibnitz's Rule (also the chain rule in this case)

$$\frac{\partial P_t}{\partial X_{ti}} = \frac{\partial F(X_i\beta)}{\partial X_{ti}} = \frac{\partial F(X_i\beta)}{\partial (X_i\beta)} \frac{\partial (X_i\beta)}{\partial X_{ti}} = f(X_i\beta) \beta_i$$

Thus the β_i 's, alone, do not yield the change in P_t corresponding to a change in X_{ti} , it is necessary to estimate $f(X_i\beta)$, the pdf evaluated at the score. However, the relative values of the beta's to all estimate the relative impact of changes in the exogenous variables on P_t .

(c) Estimation

The estimation of limited dependent models depends upon the model or density selected and the nature of the data: (i) $Y_t = 0$ or 1 or (ii) $0 < Y_t < 1$.

(i) $Y_t = 0$ or 1.

If we have a micro data based on discrete choices, then $Y = 0$ or 1.

The likelihood function in this case is given by

$$\begin{aligned} L(\beta; \theta; Y_t) &= \prod_{t=1}^n P_t^{Y_t} (1 - P_t)^{1-Y_t} \\ &= \prod_{t=1}^n F(X_t\beta; \theta)^{Y_t} (1 - F(X_t\beta; \theta))^{1-Y_t} \end{aligned}$$

and the log likelihood function is

$$\ell(\beta, \theta) = \sum_{t=1}^n \{Y_t \ln F(X_t\beta; \theta) + (1 - Y_t) \ln(1 - F(X_t\beta; \theta))\}.$$

which is optimized over the parameters β and θ to obtain maximum likelihood estimators. This procedure can be quite involved if the expression for the cumulative distribution is complicated. Recall that

$$F(X_t\beta; \theta) = \Pr(s \leq X_t\beta) = \int_{-\infty}^{X_t\beta} f(s; \theta) ds$$

where θ denotes unknown distributional parameters.

The STATA commands for fitting the probit and logit models are as follows:

probit y x's
logit y x's

Using the command **dprobit** reports the marginal effects. **dlogit** is not functional in version 10 of STATA; however, the **margins** command available in Stata 11 and 12 facilitates the evaluation of the marginal impact of changes in the independent variables on the dependent variable in a many econometric models: **margins, dydx(*) atmeans** or **margins, dydx(selected x's) atmeans**

The command

estat class, cutoff(#) reports the prediction matrix corresponding to the selected threshold value.

Qualitative response models and heteroskedasticity

There is a literature on solutions to treating heteroskedasticity in qualitative response models (Green 6th ed. , 788-790). The STATA command for one model of heteroskedasticity in the probit model is

hetprob y x's, het (all or selected X;s).

This command uses the normal cdf where the σ parameter is expressed as a function of the explanatory variables $\sigma = e^{x\delta}$. If the vector $\delta = 0$, then $\sigma = 1$ which is the regular probit model. Thus, the heteroskedasticity specification includes the regular probit model as a special case. The likelihood ratio test statistic can be used to check for statistically significant differences.

Qualitative response models and specification error

If the true qualitative response model corresponds to a pdf other than the normal or loglogistics, a nonlinear index (rather than $(X_t\beta)$), or involves heteroskedasticity, then the probit/logit estimators will be inconsistent.

Alternatives to the normal or log-logistics include using more general pdf's such as the EGB2, SGT, or IHS. Klein-Spadey (1993, An efficient semi-parametric estimator for binary response models, *Econometrica*, 387-421) propose the use of a kernel estimator in the log-likelihood function. The Klein-Spadey methodology can be implemented in Stata by downloading and installing a subroutine with the command, ***findit ST0144***.

Qualitative response models and endogenous regressors

If the right hand side variables (X) includes any endogenous regressors, you will probably want to consider using the “ivprobit” estimator. The form for this command is

ivprobit depvar X1 (Y1=X2 or Y1=X1 X2) where X2 denotes the instruments

(ii) Bounded data $0 < Y_t < 1$.

If we have a discrete choice model with grouped data or a model with the dependent variable strictly between 0 and 1, alternative estimation techniques are available. For the discrete choice problem, the MLE approach just considered can be adopted.

One approach is to use

$$\hat{p}_t = \frac{v_t}{m_t} \quad v_t = \text{number choosing the first response in the } t^{\text{th}} \text{ group}$$

m_t = number in the t^{th} group

$$F^{-1}(\hat{P}_t) = X_t\beta \text{ or } F^{-1}(Y_t) = X_t\beta .$$

If F is known, then regression techniques can be employed to estimate the vector β .

Recall that the probit model is based upon the normal cumulative distribution function and

$$F(X_t\beta) = \int_{-\infty}^{X_t\beta} \frac{s^{-s^2/2} ds}{\sqrt{2\pi}}.$$

The logit model is based upon the logistic distribution function

$$F(X_t\beta) = \frac{1}{1 + e^{-X_t\beta - \epsilon_t}}.$$

The probit model involves rather complicated estimation and there is no compelling reason that the normal should be used. The logit approximates the probit model, but has thicker tails. The logit model is particularly well suited for grouped data or other situations in which $0 < Y_t = F(x_t\beta) < 1$.

Note that

$$F^{-1}(Y_t) = \ln\left(\frac{Y_t}{1 - Y_t}\right) = X_t\beta + \epsilon_t$$

and regression techniques can be used. Further, note that $Y_t \neq 0$ or 1 in this representation.

e. Ordered or categorical data for Y .

For situations in which more than two ordered outcomes are possible, the user may want to consider the ordered logit or the ordered probit model using the STATA commands

ologit y x's

oprobit y x's

For multiple categorical dependent variables, not necessarily ordered, the multinomial logit model is a possibility

mlogit y x's

f. Poisson regression

For models in which the dependent variable represents count data, the Poisson regression model may be appropriate. The basic form for the Poisson regression model is as follows:

$$\Pr[Y = y_i | X_i] = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} \quad \lambda_i = e^{X_i \beta}, \quad y_i = 0, 1, 2, \dots$$

$$E(Y_i | X_i) = \lambda_i \beta$$

The MLE of the unknown parameters are obtained using the STATA command: **poisson y x's**

The Poisson model is characterized by the mean being equal to the variance. The negative binomial model relaxes the assumption of the mean and variance being equal.

g. Censored and Grouped or interval Regression Models

(1) Censored Regression

Consider the model

$$Y_t^* = X_t \beta + \varepsilon_t$$

where

$$Y_t = 0 \text{ if } Y_t^* \leq 0$$

$$Y_t = X_t \beta + \varepsilon_t \quad Y_t^* > 0.$$

This model has been used to describe the purchase of consumer durables, hours worked by women in the labor force, arrests after prison, etc.

Note $Y_t = 0$ if $X_t \beta + \varepsilon_t \leq 0$ or $\varepsilon_t \leq -X_t \beta$ and $Y_t > 0$ if $\varepsilon_t > -X_t \beta$. In the censored regression model observations are available for (X_t, Y_t) for $Y_t = 0$ **and** $Y_t > 0$. In truncated regression models observations are only available for $Y_t > 0$.

It should be noted that

$$E(Y_t | X_t) = X_t\beta + E(\varepsilon_t | \varepsilon_t < -X_t\beta) \neq X_t\beta.$$

This result suggests one approach to estimation based on estimating the last term involving the conditional expected value of the error term. This approach has been discussed by Heckman and involves what is called the Mill's ratio to represent $E(\varepsilon_t | \varepsilon_t < -X_t\beta)$.

Rather than using the Heckman correction, we consider regular maximum likelihood estimation of the censored regression model. If $F(\varepsilon; \theta)$ denotes the cdf for the random disturbance, then

$$\begin{aligned} \Pr(Y_t = 0 | X_t) &= \Pr(\varepsilon_t \leq -X_t\beta) \\ &= F(-X_t\beta; \theta). \end{aligned}$$

Let

$$d_i = \begin{cases} 1 & Y_t > 0 \\ 0 & Y_t = 0 \end{cases}$$

The log likelihood function for a random sample is given by

$$\ell = \sum_t \{ d_t \ln f(Y_t - X_t\beta; \theta) + (1 - d_t) \ln F(-X_t\beta; \theta) \}$$

ℓ is maximized over β and θ . In cases where the density of the random disturbance is assumed to be symmetric about zero, the likelihood function can be rewritten as

$$\ell = \sum_t \{ d_t \ln f(Y_t - X_t\beta; \theta) + (1 - d_t) \ln(1 - F(X_t\beta; \theta)) \}$$

If $f(\cdot)$ is assumed to be the normal, then the corresponding censored regression model is known as the Tobit model.

The STATA command for the Tobit model is

tobit y x's

If the Tobit models is estimated, but the true residuals are not normal, then the estimators may be **inconsistent**. Ignoring the observations for $Y=0$ and just running least squares on the observations ($Y>0$) yields inconsistent estimators. Flexible parametric and kernel based formulations have been considered for the censored regression model. Nonlinearities in the index as well as the presence of heteroskedasticity can lead to inconsistent estimates.

Amemiya, T. (1979) considers estimation of simultaneous equation models involving Tobit-like variables and Smith and Blundell (1986) outline an exogeneity test for simultaneous equation Tobit models. For Tobit models with endogenous regressors, the reader may want to consider using the **ivtobit** in STATA.

Powell (1984, 1986) introduces the censored least absolute deviations (CLAD) and symmetrically trimmed least squares (SCLS) estimators of the censored regression model which are robust to heteroskedasticity and distributional assumptions, except for assuming the error distribution is symmetric. Newey (1984) outlines how the SCLS and Tobit estimators can be combined, using a Hausman test, to check the homoskedasticity and normality assumptions. Cosslett (2004) outlines some semi-parametric approaches to estimating censored regression models, including the use of the Kaplan-Meier estimator of the cdf , to circumvent possible problems with the Tobit model. Lewbell and Linton (2002) outline other semi-parametric estimation methods for censored regression models.

Partially adaptive estimation can be modified to model different forms of heteroskedasticity and make adjustments for non-normal heteroskedastic errors. The following table summarizes the results of a simulation study with various estimators with normal, thick tailed, and skewed error terms along with two types of heteroskedasticity. Note that heteroskedasticity as well as non-normality results in Tobit estimators performing poorly.

Table 2 – Simulation Results for 25% Censoring with Homoskedastic Errors

25% Censoring									
$y = a + b*x$	Normal			Mixed Normal			ST		
	Slope: $b=1$	bias	std. dev.	RMSE	bias	std. dev.	RMSE	bias	std. dev.
OLS	-0.238	0.102	0.258	-0.194	0.096	0.216	-0.209	0.070	0.220
SCLS	0.036	0.199	0.202	0.014	0.103	0.104	0.101	0.136	0.170
CLAD	0.025	0.189	0.191	0.008	0.082	0.082	0.016	0.141	0.141
Tobit	-0.002	0.128	0.128	0.175	0.139	0.223	0.050	0.086	0.100
GED - hom	-0.001	0.131	0.131	-0.003	0.077	0.077	0.014	0.091	0.093
SGED - hom	-0.001	0.134	0.134	-0.002	0.078	0.078	0.006	0.078	0.078
IHS - hom	-0.006	0.129	0.129	-0.002	0.059	0.059	0.002	0.073	0.073
Tobit-het	0.001	0.132	0.132	0.184	0.120	0.220	0.090	0.080	0.120
GED - het	0.002	0.135	0.135	-0.003	0.078	0.078	0.047	0.093	0.104
SGED - het	0.005	0.140	0.140	-0.001	0.079	0.079	0.007	0.080	0.080
IHS - het	-0.007	0.133	0.133	0.000	0.060	0.060	-0.003	0.078	0.078

N = 200, T = 1000 Simulations

Table 3 – Simulation Results for 25% censoring with Heteroskedastic Errors (Type I)

25% Censoring									
$y = a + b*x$	Normal			Mixed Normal			ST		
	Slope: $b=1$	bias	std. dev.	RMSE	bias	std. dev.	RMSE	bias	std. dev.
OLS	0.173	0.243	0.298	0.043	0.213	0.218	-0.050	0.164	0.171
SCLS	0.001	0.353	0.353	0.039	0.213	0.217	-0.061	0.289	0.295
CLAD	-0.011	0.324	0.324	0.022	0.168	0.169	-0.119	0.221	0.251
Tobit	0.277	0.303	0.411	0.463	0.269	0.535	0.221	0.220	0.312
GED - hom	0.060	0.318	0.323	-0.012	0.164	0.164	-0.105	0.225	0.249
SGED - hom	0.281	0.404	0.492	-0.008	0.167	0.167	-0.163	0.256	0.303
IHS - hom	0.109	0.349	0.365	0.025	0.135	0.137	-0.152	0.242	0.286
Tobit-het	-0.005	0.260	0.260	0.260	0.216	0.338	0.014	0.173	0.173
GED - het	0.000	0.264	0.264	-0.016	0.161	0.161	-0.080	0.181	0.198
SGED - het	0.001	0.275	0.275	-0.023	0.159	0.161	-0.004	0.164	0.164
IHS - het	-0.003	0.258	0.258	0.004	0.115	0.115	-0.008	0.159	0.160

N = 200, T = 1000 Simulations

Table 4 – Simulation Results for 25% censoring with Heteroskedastic Errors (Type 2)

25% Censoring									
$y = a + b*x$	Normal			Mixed Normal			ST		
	Slope: $b=1$	bias	std. dev.	RMSE	bias	std. dev.	RMSE	bias	std. dev.
OLS	0.219	0.264	0.343	0.074	0.232	0.243	-0.029	0.177	0.180
SCLS	-0.007	0.358	0.358	0.038	0.215	0.218	-0.081	0.291	0.303
CLAD	-0.002	0.343	0.343	0.020	0.165	0.166	-0.148	0.225	0.270
Tobit	0.289	0.329	0.438	0.464	0.271	0.538	0.205	0.236	0.312
GED - hom	0.071	0.352	0.359	-0.020	0.175	0.176	-0.141	0.225	0.265
SGED - hom	0.253	0.432	0.501	-0.014	0.181	0.182	-0.194	0.277	0.338
IHS - hom	0.110	0.381	0.396	0.024	0.137	0.139	-0.200	0.262	0.329
Tobit-het	0.000	0.289	0.289	0.265	0.224	0.347	-0.008	0.186	0.186
GED - het	0.005	0.293	0.293	-0.025	0.171	0.173	-0.106	0.190	0.218
SGED - het	0.007	0.310	0.310	-0.032	0.171	0.174	-0.025	0.183	0.185
IHS - het	0.000	0.286	0.286	0.011	0.119	0.120	-0.027	0.172	0.174

N = 200, T = 1000 Simulations

(2) Grouped Regression

A regression model where Y is reported as being in different groups, e.g., in income intervals, (a_i, b_i) . The corresponding log-likelihood function is given by

$$\sum_i \ln(F(b_i - X_i\beta; \theta) - F(a_i - X_i\beta; \theta))$$

The stata command

```
intreg depvar1 depvar2 x's
```

fits a model of $y=[\text{depvar1}, \text{depvar2}]$ on independent variables, where y for each observation is point data, interval data, left-censored data, or right-censored data.

depvar1 and depvar2 should have the following form:

type of data	depvar1	depvar2
point data	a = [a,a]	a a
interval data	[a,b]	a b
left-censored data	(-inf,b]	. b
right-censored data	[a,inf)	a .

Additional options available with the intreg command include het() and vce, among others.

The “intreg” command assumes homoskedasticity and normally distributed error terms. These estimators will be inconsistent if the errors are not normally distributed or are

characterized by heteroskedasticity. Semi-parametric or partially adaptive estimators may be useful if either of these assumptions are violated.

References:

- Amemiya, T. (1979). "The Estimation of a Simultaneous-Equation Tobit Model." International Economic Review, 20, 1969-181
- Cosslett, S.R. (2004). "Efficient Semiparametric Estimaion of Censored and Truncated Regressions via a Smoothed Self-consistency Equation." Econometrica, 72, 1277-1293.
- Lewbell, A. And O. Linton (2002). "Nonparametric censored and truncated regression." Econometrica 70, 765-779.
- Newey, W. K. (1984). "Specification Tests for Distributional Assumptions in the Tobit Model." Journal of Econometrics, 34: 125-145.
- Powell, J.L. (1984). "Least Absolute Deviations Estimation for the Censored Regression Model." Journal of Econometrics. 25: 303-325.
- Powell, J.L. (1986). "Symmetrically Trimmed Least Squares Estimation for Tobit Models." Econometrica, 54: 1435-1460.
- Smith, R.J. and R.W. Blundell (1986). "An Exogeniety Test for a Simultaneous Equation Tobit Model with an Application to Labor Supply." Econometrica, 54, 679-85.

h. Truncated Regression Models

In truncated regression models (X_t, Y_t) is only observed (utilized) if $Y_t > a$,

$$Y_t = X_t\beta + \varepsilon_t \geq a, \text{ or}$$

$$\varepsilon_t \geq a - X_t\beta.$$

The conditional pdf of ε_t is given by

$$\frac{f(Y_t - X_t\beta; \theta)}{1 - F(a - X_t\beta; \theta)}$$

It is important to note that

$$E(Y_t | X_t) = X_t\beta + E(\varepsilon_t | \varepsilon_t \geq a - X_t\beta)$$

The log likelihood function for a random sample is given by

$$\ell(\beta, \theta) = \sum_{i=1}^n \left\{ \ln(f_\varepsilon(y_i - X_i\beta; \theta)) - \ln[1 - F_\varepsilon(a - X_i\beta; \theta)] \right\}$$

which is maximized over β and θ . In the case of upper (b) and lower threshold values (a), the log-likelihood function can be written as

$$\ell(\beta, \theta) = \sum_{i=1}^n \left\{ \ln(f_\varepsilon(y_i - X_i\beta; \theta)) - \ln[F_\varepsilon(b - X_i\beta; \theta) - F_\varepsilon(a - X_i\beta; \theta)] \right\}$$

The STATA command for normal truncation is

truncreg y x's, ll(#l) ul(#u)

where the $a = \#l$ and $b = \#u$ denote the lower and upper censoring values.

Incorrectly specifying the error distribution can lead to inconsistent estimators. Flexible parametric pdf/cdf's or non-parametric estimators of the pdf/cdf can be used

Cosslett's (2004) approach using the Kaplan-Meier non-parametric kernel estimator of the cdf can be applied to the truncated regression model.

Cosslett, S.R. (2004). "Efficient Semiparametric Estimation of Censored and Truncated Regressions via a Smoothed Self-consistency Equation." *Econometrica*, 72, 1277-1293.

Lewbel, A. And O. Linton (2002). "Nonparametric censored and truncated regression." *Econometrica* 70, 765-779.

Summary of qualitative response, censored, truncated, and interval regression model specifications

	Data	Model	ℓ (Log-likelihood)	Stata Commands
Qualitative response	$Y = 0, 1$	$\Pr(Y_i = 1 X_i) = F(X_i\beta; \theta)$	$\sum_i \left(y_i \ln(F(X_i\beta; \theta)) + (1-y_i) \ln(1-F(X_i\beta; \theta)) \right)$	probit y x's probit y x's, vce(robust) hetprob y x's logit y x's margins, dydx(*) atmeans
Censored regression	$0 \leq Y$	$Y = X\beta + \varepsilon \geq 0$	$\sum_i \left(d_i \ln(f(\varepsilon_i = y_i - X_i\beta; \theta)) + (1-d_i) \ln(F(-X_i\beta; \theta)) \right)$ $d_i = 1 \text{ if } y_i > 0$	tobit y x's tobit y x's, vce(robust) tobit y x's, vce(boot)
Truncated regression	$L < Y < U$	$L < Y = X\beta + \varepsilon < U$ $f_\varepsilon(\varepsilon; \theta) = \frac{f(\varepsilon; \theta)}{F(U - X\beta; \theta) - F(L - X\beta; \theta)}$	$\sum_i \ln(f_\varepsilon(y_i - X_i\beta; \theta))$	truncreg y x's, ll(L) UL(U) margins dydx(*) atmeans
Interval regression	$Y_i \in (a_i, b_i)$ only know endpoints	$a_i < Y_i = X_i\beta + \varepsilon_i < b_i$	$\sum_i \ln \left(\frac{F(b_i - X_i\beta; \theta)}{F(a_i - X_i\beta; \theta)} \right)$	intreg depvar1 depvar2 x's, options depvar1=a's, depvar2=b's

Estimators of β are inconsistent if the error pdf is misspecified, heteroskedasticity exist, or if the $\text{ind}(X\beta)$

Alternative estimators can be obtained using flexible or kernel specifications for the error distribution.

should be nonlinear

i. Quantile regression

A popular estimation procedure when outliers may be present is that of least absolute deviations (LAD) estimation defined by

$$\hat{\beta}_{LAD} = \arg \min_{\beta} \left\{ \sum_{t=1}^n |y_t - X_t \beta| \right\}$$

the solution of which involves what is referred to as a linear programming problem.

The corresponding estimated relationship is given by $\hat{y}_t = X_t \hat{\beta}_{LAD}$. This actually estimates median corresponding to X_t ; the protus, for e: X

If the relationship between the dependent variable and the explanatory variables is desired corresponding to different quantiles (e.g. top 75%), the quantile regression methods can be used. The quantile regression model is a generaliztion of the LAD and is defined by

$$\begin{aligned} \Pr(y_t \leq \tau | X_t) &= F_{\varepsilon_\theta}(X_t \beta_\theta + \varepsilon_t \leq \tau | X_t) = F(\varepsilon_t \leq \tau - X_t \beta_\theta | X_t) \\ &= F_\varepsilon(\tau - X_t \beta_\theta) = \theta \end{aligned}$$

The θ^{th} quantile is defined $Quantile_\theta(y_t | X_t) = X_t \beta_\theta$

The corresponding β_θ can be obtained by

$$\begin{aligned} \hat{\beta}_\theta &= \arg \min_{\beta} \left\{ \sum_{t:y_t \geq X_t \beta} \theta |y_t - X_t \beta| + \sum_{t:y_t < X_t \beta} (1-\theta) |y_t - X_t \beta| \right\} \\ &= \arg \min_{\beta} \sum_{t=1}^n \left\{ \left(\theta - \frac{1}{2} + \frac{1}{2} sign(y_t - X_t \beta) \right) (y_t - X_t \beta) \right\} \end{aligned}$$

Stata's *qreg* command facilitates the evaluation of quantiles with the command

qreg y x's, quantile(θ)

- *qreg y x's, quantile(.5)* is the equivalent to *qreg y x's — LAD*
- *qreg y x's, quantile(.9)* reports the 90% quantile

MLE using the Laplace pdf yields the LAD results

The skewed Laplace pdf with the scale parameter expressed in terms of the regressors could accommodate skewed error distributions with heteroskedasticity.

j. Hazard Functions

Hazard functions are important characterizations of random variables in labor, quality control, and in demography. For example, let $f(s; \theta)$ denote the probability density function for a completed spell of unemployment. Let the cumulative distribution function be denoted by

$$F(s; \theta) = \Pr(S \leq s)$$

$$= \int_0^s f(t; \theta) dt$$

and the probability that an unemployment spell will have length less than s .

The mean length of unemployment spells is given by

$$\mu_s = E(S) = \int_0^\infty sf(s; \theta) ds.$$

The probability of a spell of length greater than s is called the survivor function

$$S(s; \theta) = 1 - F(s; \theta).$$

The hazard function gives the conditional probability that a person unemployed for length s will find employment at time s ,

$$h(s; \theta) = \Pr(S=s | S>s)$$

$$= \frac{f(s; \theta)}{1 - F(s; \theta)}$$

The parameters θ can be estimated from the pdf, $f(s; \theta)$, or from the hazard functions, $h(s; \theta)$. To see this, consider a random sample which includes completed spells and interrupted spells. The pdf for completed spells is $f(s_i; \theta)$ and $1-F(s_i; \theta)$ for interrupted spells. Let $d_i = 1$ for a complete spell and 0 for an uninterrupted spell.

The likelihood function for the sample can be written as

$$L = \prod_i (f(s_i; \theta))^{d_i} (1 - F(s_i; \theta))^{1-d_i};$$

hence, the log likelihood function is given by

$$\ell = \ln L = \sum_i \{d_i \ln f(s_i; \theta) + (1 - d_i) \ln(1 - F(s_i; \theta))\}$$

Note the relationship between this log likelihood function and that for the qualitative responsive model. Regrouping terms in the log likelihood function yields

$$\ell = \sum_i d_i \ln \left(\frac{f(s_i; \theta)}{1 - F(s_i; \theta)} \right) + \sum_i \ln(1 - F(s_i; \theta))$$

Note the first sum is structured in terms of the hazard functions.

k. Generalized Method of Moments (GMM)

The Regression Section of the notes outlines the formulation of the generalized method of moments (GMM) as minimizing a quadratic form

$$q(\hat{\theta})' VAR^{-1}(q(\hat{\theta})) q(\hat{\theta})$$

where $q(\hat{\theta})$ is a vector whose elements are the difference between sample moments and theoretical

moments. In many applications of GMM this optimization involves solving nonlinear optimization

problems. The $VAR(q(\hat{\theta}))$ matrix may depend on unknown parameters. Frequently these are

estimated and then the corresponding quadratic form is minimized to obtain GMM. If the minimization is over the parameters in $q(\cdot)$ and in the variance matrix, the corresponding estimators are referred to as continuous updating estimators (CUE) and tend to have smaller bias than without updating the variance matrix.

I. The Bootstrap

The bootstrap attempts to approximate the distribution of the actual estimator. The bootstrap is performed by generating repeated samples (with replacement) from the original sample and estimating the unknown parameters. This gives us a “bootstrapped” sample of estimates of the desired parameters or distributional characteristics. The multiple parameter estimates, or distributional characteristics, are used to approximate the exact distribution and construct confidence intervals.

As a simple example, consider using the bootstrap to estimate the standard errors of the OLS regression coefficient estimator in the presence of heteroskedastic errors. For pedagogical purposes the estimated standard errors of the OLS could be compared to the robust standard errors and bootstrap standard errors. These could be obtained using the STATA commands

reg y x's

reg y x's, robust

reg y x's, vce(boot)

The estimated standard errors obtained from the resampling with the bootstrap methodology could be obtained using the commands:

bootstrap _b _se, reps(#r) size(#s): reg y x's

where **reps(#r)** indicates the number of random samples (with replacement) and **#s** denotes the sample size ($\#s \leq n$); hence, **#r** is the number of sample estimates of the desired entity that are generated.

6. SOME EXERCISES USING NONLINEAR MODELS

Production functions:

Three commonly used production functions are the:

COBB DOUGLAS:

$$(1) \quad Y(t) = A(t) L_t^{\beta_1} K_t^{\beta_2} \cdot \varepsilon_t$$

CONSTANT ELASTICITY OF SUBSTITUTION (CES):

$$(2) \quad Y(t) = A(t) [\delta L^\rho + (1-\delta)K^\rho]^{M/\rho} \cdot \varepsilon_t$$

VARIABLE ELASTICITY OF SUBSTITUTION (VES):

$$(3) \quad Y(t) = A(t) [L_t + (\rho-1)K_t]^{\alpha\rho} K_t^{\alpha(1-\rho)} \cdot \varepsilon_t$$

The $A(t)$ in (1), (2), and (3) can be used as a proxy for technological change allowing output (Y_t) to change without changes in labor (L_t) and capital (K_t) inputs. Some characteristics and interrelationships between these functions are given by

	CD	CES	VES
Returns to scale	$\beta_1 + \beta_2$	M	α
Elasticity of substitution $\% \Delta(K/L) / \% \Delta(W_L/W_K)$	1	$\frac{1}{1 - \rho}$	$1 + \left(\frac{\rho-1}{1-\delta} \right) \left(\frac{K}{L} \right)$
Labor's share of output (if constant returns to scale)	β_1	$\frac{\delta}{\delta + (1-\delta) \left(\frac{K}{L} \right)^\rho}$	$\frac{\delta P}{1 + (\rho-1) \frac{K}{L}}$

The Cobb Douglas is a special case of the CES and the VES production functions, in particular

The VES production function with $\rho = 1$ is of a Cobb Douglas form.

The CES production function with $\rho \rightarrow 0$ is also of a Cobb Douglas form.

Consider the following data set:

B.	t	Output (Y_t)	Labor (L_t)	Capital (K_t)
	1	40.26	64.63	133.14
	2	40.84	66.30	139.24
	3	42.83	65.27	141.64
	4	43.89	67.32	148.77
	5	46.10	67.20	151.02
	6	44.45	65.18	143.38
	7	43.87	65.57	148.19
	8	49.99	71.42	167.12
	9	52.64	77.52	171.33
	10	57.93	79.46	176.41

This set consists of the first ten observations used by Solow in his study dealing with estimation of technological change.

Exercises

1. Use Solow's data and assume the possibility of technological change as modeled by

$$A(t) = e^{\beta_1 + \beta_2 t}$$

$$\ln \varepsilon_t \sim N[0, \sigma^2],$$

in the production functions, obtain nonlinear least squares estimates of the parameters in the CD, CES, and VES production functions. Also estimate the Cobb Douglas production function under the assumption of constant returns to scale.

Test the following hypotheses:

- (a) The CES production function has constant returns to scale

- (b) The Cobb Douglas production function has constant returns to scale

- (c) CES production function is a Cobb Douglas production function

- (d) The VES production function is a Cobb Douglas production function with constant returns to scale. Perform the test using the WALD Test and the likelihood ratio test.

2. Many economists feel that the assumption of normally distributed random disturbances is too restrictive and that in particular, the tails of the normal density are "too thin" to describe many economic series. One alternative is provided by the following:

$$GED(\varepsilon; \sigma, p) = \frac{p e^{-|\varepsilon/\sigma|^p}}{2\sigma\Gamma(1/p)}$$

- (a) Assume that no autocorrelation exists. Obtain the log likelihood functions corresponding to $y_t = x_t\beta + \varepsilon_t$.
- (b) Discuss how one might obtain MLE estimators of the parameters in each case.
- (c) How could you test the assumption of Normality of the error distribution?

3. Consider the model

$$Y_t = X_t\beta + \varepsilon_t$$

where $X_t = (1, X_{t1}, \dots, X_{tk})$, and $\beta = (\beta_1, \beta_2, \dots, \beta_k)$. Assume that $\varepsilon = (\varepsilon_1, \dots, \varepsilon_N) \sim N[0, \sigma^2 I]$. Show that the Wald, LR and LM tests for $H_0: \beta_2 = \dots = \beta_k = 0$ can be expressed in terms of the coefficient of determination as

$$W = N R^2 / (1 - R^2)$$

$$LR = -N \ln(1 - R^2)$$

$$LM = N R^2$$

4. Fourteen applicants to a graduate program had the following quantitative and verbal scores on the GRE examination. Six students were admitted to the program.

GRE Scores

Student number	Quantitative Q	Verbal V	Admitted Yes = 1
1	760	550	1
2	600	350	0
3	720	320	0
4	710	630	1
5	530	430	0
6	650	570	0
7	800	500	1
8	650	680	1
9	520	660	0
10	800	250	0
11	670	480	0
12	670	520	1
13	780	710	1

Source: Donald F. Morrison, Applied Linear Statistical Methods, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1983, p. 279.

- (a) Use a linear probability model to "predict" admissions into the graduate program based upon the quantitative and verbal scores.
- (b) Repeat (4a) using a probit model.
- (c) Repeat (4a) using a logit model.

5. Consider the following data [Hampel, Robust Statistics]:

Y	X
---	---

15.7	17.6
------	------

44.9	20.0
------	------

18.0	20.9
------	------

19.9	21.6
------	------

23.4	26.0
------	------

19.7	27.1
------	------

23.1	27.6
------	------

23.8	27.8
------	------

24.9	32.6
------	------

26.1	33.4
------	------

27.6	35.1
------	------

26.1	37.0
------	------

31.3	38.7
------	------

- (a) Plot the data.

- (b) Estimate $Y_t = \beta_1 + \beta_2 X_t + \varepsilon_t$ using least squares with
 - (1) the entire data set

 - (2) the second observation deleted.

- (c) Use the “qreg y x’s” in STATA or the “robust y x’s/LAE” command in SHAZAM and compare these results with your answer in (b).

- (d) Estimate this model ($Y_t = \beta_1 + \beta_2 X_t + \varepsilon_t$) under the assumption that the ε_t are distributed as a GED (generalized error distribution). .

STATA Commands for the normal and GED pdf's:

```

cap prog drop normal
program define normal
version 1.0
args lnf mu sigma (used in estimation)
quietly replace `lnf=ln(normalden($ML_y1,'mu','sigma')) STATA 10-calls normal density
end
use "e:\Econ Data\hbj.dta, clear or use infile command
ml search
ml maximize, difficult

cap prog drop GED
program define GED
version 1.0
args lnf xb s p
quietly replace `lnf=ln(`p')-((abs($ML_y1-'xb')/`s')^(abs(`p')))-ln(2*`s')-lngamma((abs(1/'p')))

This is the log-likelihood of the GED
end
clear
infile y x using g:\american.dat

ml model lf normal (y=x) (sigma:), technique(dfp)
ml search
ml maximize
ml model lf GED (y=x) (s:) (p:), technique(dfp)
ml search
ml maximize

```

- (e) Perform LR and W tests of the hypothesis $H_0: p=2$, i.e., the errors are normal.

7. Using the data in problem 6, calculate the estimated standard error for the OLS slope coefficient using (a) the OLS results, (b) the robust standard errors (using `reg y x,robust`), and (c) the estimated standard error based upon a bootstrap with 100 replications.

VI. Simultaneous Equation Models

1. An introduction to some problems associated with simultaneous equation models
2. Basic Notation and alternative representations
 - a. Structural representation
 - b. Reduced form representation
 - c. Final form or transfer function representation
3. Identification
 - a. Identification as a mapping (3 examples)
 - b. Examples
4. Estimation of reduced form parameters
 - a. Least squares no restrictions (LSNR)
 - b. Estimators derived from structural estimators
 - c. SURE—generally the same as LSNR
5. Estimation of Structural Models
 - a. Notation
 - b. Single equation methods: OLS, IV, 2SLS, LIML, k-class, GMM, IV, other
 - c. Simultaneous equation methods: FIML and 3SLS
6. Statistical inference
 - a. Identification
 - b. Reduced form coefficients
 - c. Structural parameters
7. Exercises

James B. McDonald
Brigham Young University
1/2013

VI. SIMULTANEOUS EQUATION MODELS

1. INTRODUCTION TO SIMULTANEOUS EQUATION MODELS

There are several problems encountered with structural equations in simultaneous equations models which are not generally associated with single equation models. These include (1) the identification problem, (2) inconsistency of ordinary least squares (OLS) estimators of structural parameters, (3) questions about the interpretation of structural parameters, and (4) the validity of the OLS "t statistics" associated with structural coefficients.

In order to introduce these problems, we review two important papers. The paper on identification by E. J. Working [1927, QJE] is considered in the first section. The work of Haavelmo [1947, JASA] dealing with alternative methods of estimating the marginal propensity to consume is described in the second section. The third section contains a brief summary.

a. STRUCTURAL AND REDUCED FORM REPRESENTATIONS,

IDENTIFICATION, AND INTERPRETATIONS OF COEFFICIENTS

Consider the problem of estimating the impact of an increase in the price of crude oil upon the equilibrium price and quantity of gasoline. The corresponding increase in the equilibrium price of gasoline will depend upon several factors including the slope of the demand curve.

This is illustrated in the following figure:

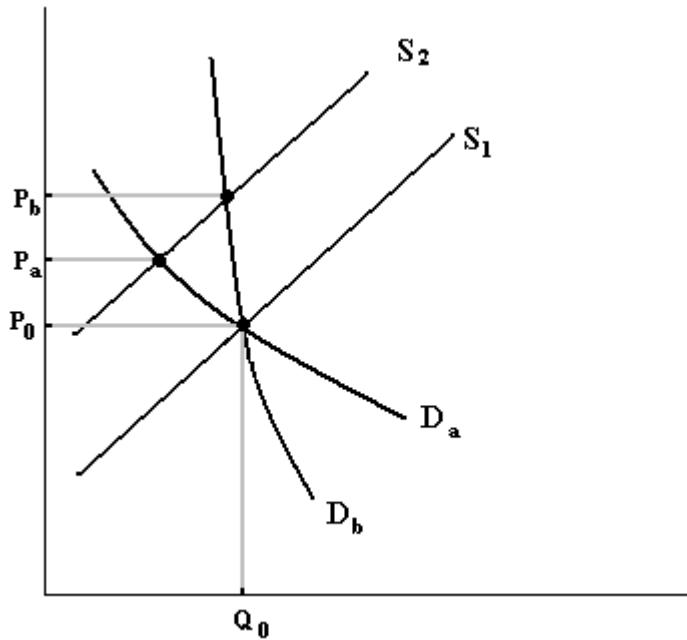


Figure 1

Assume that (Q_0, P_0) denotes the original equilibrium. Assume that the increase in the price of crude oil results in the supply curve shifting from S_1 to S_2 . The associated change in P depends upon the relevant demand schedule with the more inelastic curve being associated with the larger price increases. This example clearly indicates the importance of estimating the slope of the demand schedule in order to make predictions about the impact of changes in factor price upon equilibrium price.

Estimation of the slope of the demand curve might begin by collecting observations on (P, Q) which might appear as in Figure 2.

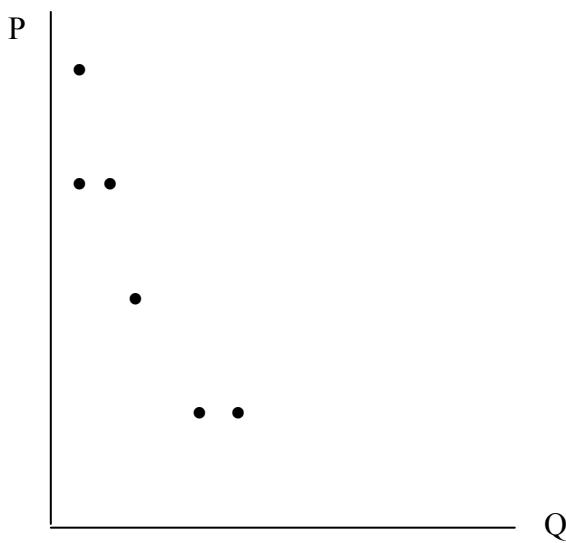


Figure 2

The reader would probably be tempted to draw a line through the points or perform a least squares estimation on $p = \beta_1 - \beta_2 Q$ in order to estimate the demand schedule. But how would we estimate the demand curve if a plot of P and Q appeared as in Figure 3 rather than as in Figure 2?

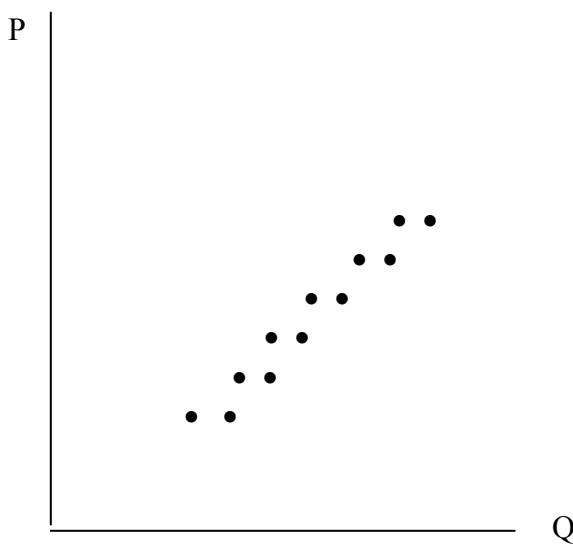


Figure 3

The data in Figure 3 appears to define a supply curve rather than a demand curve. Alternatively, how could a demand curve be estimated if the data appear as in Figure 4?

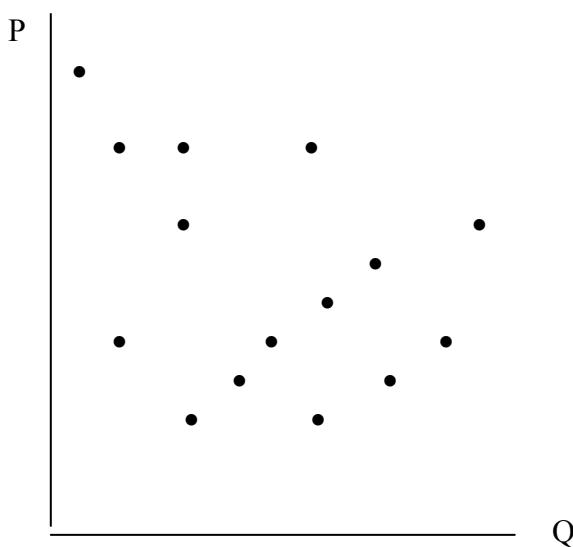


Figure 4

In order to answer this question, we need to recall that equilibrium price and quantity are determined by supply and demand factors and not by supply or demand alone. The observations depicted in Figure 2 could have been generated by either of the following scenarios:

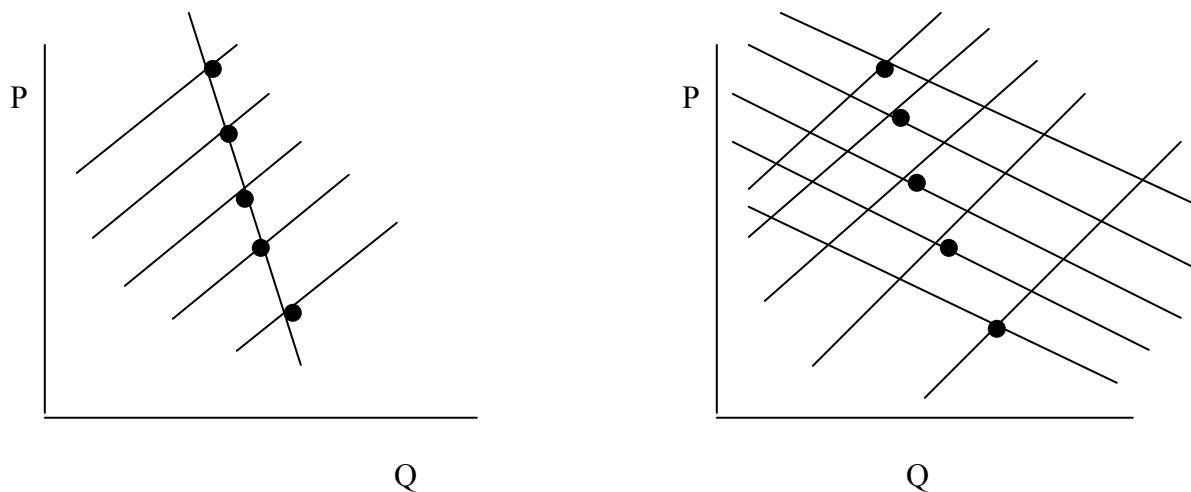


Figure 5

If the demand curve is stable and the supply curve shifts, then the demand curve is "traced out." If both curves shift, fitting a relationship to the observed (Q, P) would not correspond to the underlying demand curve(s). Similarly, Figure 3 could correspond to a relatively stable supply curve and a shifting demand curve or both curves shifting. Figure 4 would appear to correspond to both curves shifting.

Consider the following model:

$$(1.1) \text{Demand: } Q = \gamma_{11} - \beta_{12}P + \gamma_{12}Y + \varepsilon_{1t}$$

$$(1.2) \text{Supply: } Q = \gamma_{21} + \beta_{22}P - \gamma_{23}FC + \varepsilon_{2t}$$

or equivalently,

$$\begin{bmatrix} -1 & -\beta_{12} \\ -1 & \beta_{22} \end{bmatrix} \begin{bmatrix} Q_t \\ P_t \end{bmatrix} + \begin{bmatrix} \gamma_{11} & \gamma_{12} & 0 \\ \gamma_{21} & 0 & -\gamma_{23} \end{bmatrix} \begin{bmatrix} 1 \\ Y_t \\ FC_t \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix} = 0.$$

Equations (1.1) and (1.2) will be referred to as the structural model with Q and P as endogenous (dependent) variables and income (Y) and factor costs (crude oil, FC) as exogenous (independent) variables. In order to draw a demand curve or supply curve using (Q, P) as coordinates, Y and FC must be fixed at some arbitrary level.

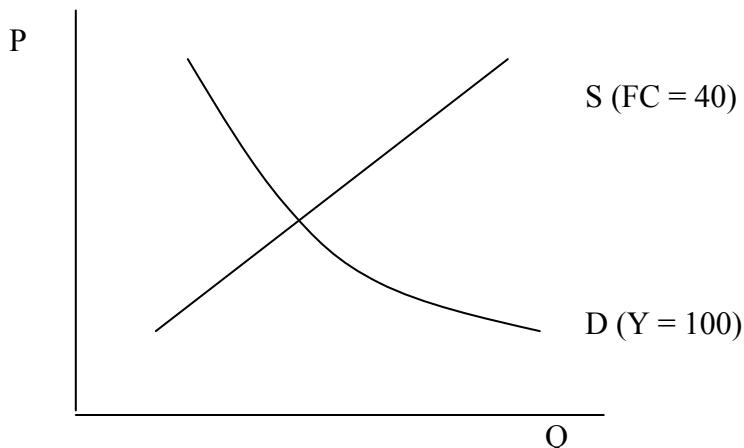


Figure 6

A change in factor costs (income fixed) will shift the supply curve and "trace" the depicted demand curve and a change in income (factor costs fixed) will shift the demand curve and "trace" the depicted supply curve, et cet. paribus. It is interesting to observe that by including factor costs (FC) in the supply equation and not the demand equation we are able to "identify" the demand equation. Similarly, by including income (Y) in the demand equation and not in the supply equation we are able to "identify" the supply equation. Hence, one way of "identifying" a structural equation is by excluding variables from the equation we want to estimate, which are included in other structural equations. These excluded variables are referred to as instrumental variables. This is the general approach to the identification problem developed by E. J. Working [1927]. A more formal development will be considered later.

We note from Figure 6 that for each level of factor costs and income there is a corresponding equilibrium price and quantity determined by the intersection of the supply and demand curves. If we solve the structural model for the explicit relationship between (P, Q) and FC and Y we obtain

$$\begin{aligned}
 \begin{bmatrix} Q_t \\ P_t \end{bmatrix} &= - \begin{bmatrix} -1 & -\beta_{12} \\ -1 & \beta_{22} \end{bmatrix}^{-1} \left\{ \begin{bmatrix} \gamma_{11} & \gamma_{12} & 0 \\ \gamma_{21} & 0 & -\gamma_{23} \end{bmatrix} \begin{bmatrix} 1 \\ Y_t \\ FC_t \end{bmatrix} + \begin{bmatrix} \varepsilon_{t1} \\ \varepsilon_{t2} \end{bmatrix} \right\} \\
 &= \left(\frac{1}{\beta_{12} + \beta_{22}} \right) \begin{bmatrix} \beta_{22} & \beta_{12} \\ 1 & -1 \end{bmatrix} \left\{ \begin{bmatrix} \gamma_{11} & \gamma_{12} & 0 \\ \gamma_{21} & 0 & -\gamma_{23} \end{bmatrix} \begin{bmatrix} 1 \\ Y_t \\ FC_t \end{bmatrix} + \begin{bmatrix} \varepsilon_{t1} \\ \varepsilon_{t2} \end{bmatrix} \right\} \\
 &= \left(\frac{1}{\beta_{12} + \beta_{22}} \right) \begin{bmatrix} \beta_{22}\gamma_{11} + \beta_{12}\gamma_{21} & \beta_{22}\gamma_{12} & -\beta_{12}\gamma_{23} \\ \gamma_{11} - \gamma_{21} & \gamma_{12} & \gamma_{23} \end{bmatrix} \begin{bmatrix} 1 \\ Y_t \\ FC_t \end{bmatrix} \\
 (1.3a-b) \quad &+ \begin{bmatrix} \frac{\beta_{22}\varepsilon_{t1} + \beta_{12}\varepsilon_{t2}}{\beta_{12} + \beta_{22}} \\ \frac{\varepsilon_{t1} - \varepsilon_{t2}}{\beta_{12} + \beta_{22}} \end{bmatrix}
 \end{aligned}$$

$$= \begin{bmatrix} \pi_{11} & \pi_{12} & \pi_{13} \\ \pi_{21} & \pi_{22} & \pi_{23} \end{bmatrix} \begin{bmatrix} 1 \\ Y_t \\ FC_t \end{bmatrix} + \begin{bmatrix} \eta_{t1} \\ \eta_{t2} \end{bmatrix}$$

Note: $\frac{\partial Q}{\partial Y} = \frac{\beta_{22}\gamma_{12}}{\beta_{12} + \beta_{22}} = \pi_{12} > 0$, $\frac{\partial Q}{\partial FC} = \frac{-\beta_{12}\gamma_{23}}{\beta_{12} + \beta_{22}} = \pi_{13} < 0$

$$\frac{\partial P}{\partial Y} = \frac{\gamma_{12}}{\beta_{12} + \beta_{22}} = \pi_{22} > 0$$
, $\frac{\partial P}{\partial FC} = \frac{\gamma_{23}}{\beta_{12} + \beta_{22}} = \pi_{23} > 0$

Equations (1.3a, b) are referred to as the reduced form equations for Q and P corresponding to the structural model defined by (1.1) and (1.2). Note that each reduced form equation expresses the equilibrium value (P or Q) as a function of the exogenous variables FC and Y.

In order to determine the impact of an increase in the price of crude oil upon the price of gasoline, we employ the reduced form representation, i.e.,

$$\frac{\partial P}{\partial FC} = \frac{\gamma_{23}}{\beta_{12} + \beta_{22}} = \pi_{23} > 0$$

which takes into account the slopes of the supply and demand curves as well as how far the supply curve shifts in response to an increase in the price of crude oil. The equilibrium quantity would also change according to

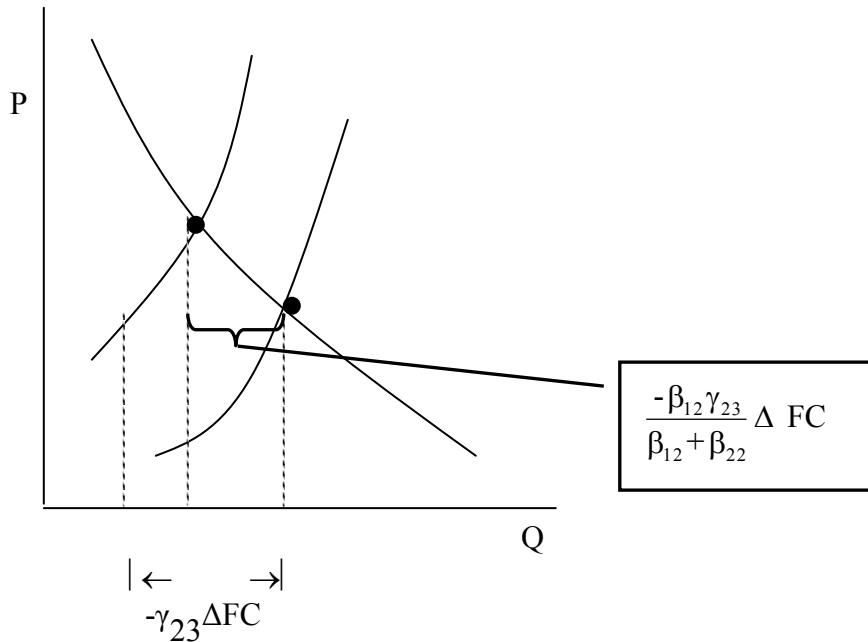
$$\frac{\partial Q}{\partial FC} = \frac{-\beta_{12}\gamma_{23}}{\beta_{12} + \beta_{22}} = \pi_{13} < 0.$$

The reader might wonder why

$$\frac{\partial Q^s}{\partial FC} = -\gamma_{23} < 0$$

doesn't characterize the change in equilibrium quantity.

The following figure will illustrate why the reduced form provides the necessary information.



Differentiating the supply equation with respect to FC yields $(-\gamma_{23})$ and assumes that P is fixed and hence merely represents the horizontal shift of the supply curve and not the change in equilibrium quantity. The reduced form equation for Q expresses the equilibrium quantity as a function of FC and Y and takes account of the increase in equilibrium price associated with an increase in factor costs.

To summarize, the reduced form coefficients represent the change in equilibrium values corresponding to changes in the predetermined or exogenous variables, i.e., the reduced form coefficients are the multipliers. The structural coefficients represent slopes or shifts of structural schedules in response to changes in predetermined or exogenous variables.

EXERCISE:

1. The Asymptotic Bias of the OLS estimator of the slope for the demand curve is given by

$$\frac{(\beta_{22} + \beta_{12}) \sigma_{\epsilon 1}^2}{\sigma_{\epsilon 1}^2 + \sigma_{\epsilon 2}^2 + \gamma_{23}^2 (1 - \text{COR}^2(Y, FC))}$$

where $\text{COR}(Y, FC)$ = correlation between Y and FC.

- (a) Mathematically analyze the impact of increases in $\sigma_{\epsilon 2}^2$, γ_{23}^2 , and $\text{COR}(Y, FC)$ upon the asymptotic bias of $\hat{\beta}_{12}$.
- (b) Graphically analyze the impact of increases in $\sigma_{\epsilon 2}^2$, γ_{23}^2 , and $\text{COR}(Y, FC)$ upon the "bias of β_{12} ".

b. INCONSISTENCY OF ORDINARY LEAST SQUARES, ALTERNATIVE ESTIMATORS, AND STATISTICAL INFERENCE

Haavelmo [1947] considered the following simple macro model:

$$(2.1) \quad C_t = \alpha + \beta Y_t + \varepsilon_t$$

$$(2.2) \quad Y_t = C_t + Z_t$$

where Y_t , C_t , and Z_t ($Z \equiv Y - C$) respectively denote income, consumption and nonconsumption expenditure.

The reduced form representation corresponding to (2.1) and (2.2) is given by

$$(2.3) \quad C_t = \pi_{11} + \pi_{12} Z_t + \eta_t$$

$$(2.4) \quad Y_t = \pi_{21} + \pi_{22} Z_t + \eta_t$$

where (2.5a-e) $\eta_t = \varepsilon_t / (1-\beta)$

$$\pi_{11} = \alpha / (1-\beta)$$

$$\pi_{12} = \beta / (1-\beta)$$

$$\pi_{21} = \alpha / (1-\beta)$$

$$\pi_{22} = 1 / (1-\beta)$$

Note that π_{12} and π_{22} correspond to the multipliers discussed in simple macroeconomics models. Haavelmo's analysis of the simple model defined by (2.1) and (2.2) pointed out many problems which are also associated with larger econometric models. For this reason we will consider this model in detail.

Estimation. Past experience might suggest that the OLS estimator of β would have desirable statistical properties if ε_t in (2.1) is not characterized by autocorrelation or heteroskedasticity.

The OLS estimator of β in (2.1) is defined by

$$(2.6) \hat{\beta} = \frac{\sum(Y - \bar{Y})(C - \bar{C})}{\sum(Y - \bar{Y})^2} = \frac{Cov(Y, C)}{Var(Y)}$$

but from (2.3) and (2.4), we see that

$$\begin{aligned} (2.7) C - \bar{C} &= \pi_{12}(Z - \bar{Z}) + \frac{\bar{\varepsilon} - \bar{\varepsilon}}{1 - \beta} \\ &= \frac{\beta}{1 - \beta}(Z - \bar{Z}) + \frac{\bar{\varepsilon} - \bar{\varepsilon}}{1 - \beta} \end{aligned}$$

and

$$\begin{aligned} (2.8) Y - \bar{Y} &= \pi_{22}(Z - \bar{Z}) + \frac{\bar{\varepsilon} - \bar{\varepsilon}}{1 - \beta} \\ &= \frac{1}{1 - \beta}(Z - \bar{Z}) + \frac{\bar{\varepsilon} - \bar{\varepsilon}}{1 - \beta}; \end{aligned}$$

hence, after substituting (2.7) and (2.8) into (2.6), we can write

$$\begin{aligned} (2.9) \hat{\beta} &= \frac{\sum \left\{ \frac{(Z - \bar{Z})}{(1 - \beta)} + \frac{(\bar{\varepsilon} - \bar{\varepsilon})}{1 - \beta} \right\} \left\{ \frac{\beta}{1 - \beta}(Z - \bar{Z}) + \frac{(\bar{\varepsilon} - \bar{\varepsilon})}{1 - \beta} \right\}}{\sum \left\{ \frac{(Z - \bar{Z})}{(1 - \beta)} + \frac{(\bar{\varepsilon} - \bar{\varepsilon})}{1 - \beta} \right\}^2} \\ \hat{\beta} &= \frac{\sum \left\{ \frac{\beta}{(1 - \beta)^2}(Z - \bar{Z})^2 + \frac{(1 + \beta)(Z - \bar{Z})(\bar{\varepsilon} - \bar{\varepsilon})}{(1 - \beta)^2} + \frac{(\bar{\varepsilon} - \bar{\varepsilon})^2}{(1 - \beta)^2} \right\}}{\sum \left\{ \frac{(Z - \bar{Z})^2}{(1 - \beta)^2} + 2 \frac{(\bar{\varepsilon} - \bar{\varepsilon})(Z - \bar{Z})}{(1 - \beta)^2} + \frac{(\bar{\varepsilon} - \bar{\varepsilon})^2}{(1 - \beta)^2} \right\}} \\ &= \frac{\beta \sum(Z - \bar{Z})^2 / N + (1 + \beta) \sum(Z - \bar{Z})(\bar{\varepsilon} - \bar{\varepsilon}) / N + \sum(\bar{\varepsilon} - \bar{\varepsilon})^2 / N}{\sum \left\{ (Z - \bar{Z})^2 / N + (\bar{\varepsilon} - \bar{\varepsilon})(Z - \bar{Z}) / N + (\bar{\varepsilon} - \bar{\varepsilon})^2 / N \right\}}. \end{aligned}$$

Assuming that:

$$\sum_{t=1}^N (Z - \bar{Z})^2 / N \rightarrow \sigma_Z^2 \quad \text{as } N \rightarrow \infty,$$

$$\sum_{t=1}^N (Z - \bar{Z})(\bar{\varepsilon} - \bar{\varepsilon}) / N \rightarrow 0 \quad \text{as } N \rightarrow \infty, \text{ and}$$

$$\sum_{t=1}^N (\varepsilon - \bar{\varepsilon})^2 / N \rightarrow \sigma^2 \quad \text{as } N \rightarrow \infty$$

$$(2.10) \quad \hat{\beta} \rightarrow \frac{\beta \sigma_z^2 + \sigma^2}{\sigma_z^2 + \sigma^2}.$$

$$= \beta + \frac{\sigma^2(1-\beta)}{\sigma_z^2 + \sigma^2}$$

as $N \rightarrow \infty$. Hence, we see from (2.10) that $\hat{\beta}$ is an inconsistent estimator of β with asymptotic bias equal to the second term in (2.10)

$$\frac{\sigma^2(1-\beta)}{\sigma_z^2 + \sigma^2}.$$

This may seem like a surprising result in light of the apparent simplicity of the consumption function. It may not be obvious which of the assumptions

$$(A.1) \quad \varepsilon_t \text{ distributed normally}$$

$$(A.2) \quad E(\varepsilon_t) = 0 \text{ for all } t$$

$$(A.3) \quad \text{Var}(\varepsilon_t) = \sigma^2 \text{ for all } t$$

$$(A.4) \quad E(\varepsilon_t \varepsilon_s) = 0 \text{ for } t \neq s$$

$$(A.5) \quad Y_t \text{ and } \varepsilon_t \text{ are independent.}$$

are violated. But upon closer inspection (hint: see (2.4)) we note that

$$\begin{aligned} E(Y_t \varepsilon_t) &= E\left[\left(\pi_{21} + \pi_{22} Z_t + \frac{\varepsilon_t}{1-\beta}\right)(\varepsilon_t)\right] \\ &= E(\varepsilon_t^2)/(1-\beta) \\ &= \sigma^2/(1-\beta) \neq 0; \end{aligned}$$

hence, (A.5) is violated and OLS estimators of the structural parameters α and β are biased and inconsistent. In fact, this is typically the case when OLS is used to estimate structural relationships which include endogenous variables on the right hand side of the structural equation. Right hand side endogenous variables are commonly referred to as ***endogenous regressors***.

As another example, the asymptotic bias of the OLS estimator of β_{12} in (1.1) is given by

$$(2.11) \quad \frac{(\beta_{22} + \beta_{12}) \sigma_{\varepsilon 1}^2}{\sigma_{\varepsilon 1}^2 + \sigma_{\varepsilon 2}^2 + \gamma_{23}^2 (1 - \text{Corr}^2(Y, FC))}.$$

How can we obtain consistent estimators of the unknown structural parameters?

Two stage least squares or an appropriate application of instrumental variables estimation provides a solution. It is instructive to consider an alternative estimator first. Recall that the ordinary least squares estimators of the reduced form equations (referred to as least squares no restrictions, LSNR) will yield unbiased and consistent estimators of the π_{ij} 's which will be denoted by $\hat{\pi}_{ij}$. This observation provides the basis for obtaining consistent estimators of α and β in the Haavelmo model. From (2.5 c,e) we note that

$$\beta = \pi_{12}/\pi_{22}$$

hence, a consistent estimator of β can be obtained from

$$(2.12) \quad \beta^* = \hat{\pi}_{12}/\hat{\pi}_{22}$$

where $\hat{\pi}_{12} = \frac{\sum(C - \bar{C})(Z - \bar{Z})}{\sum(Z - \bar{Z})^2}$

$$\hat{\pi}_{22} = \frac{\sum(Y - \bar{Y})(Z - \bar{Z})}{\sum(Z - \bar{Z})^2}$$

or

$$(2.13) \quad \beta^* = \frac{\sum(C - \bar{C})(Z - \bar{Z})}{\sum(Y - \bar{Y})(Z - \bar{Z})}$$

In order to verify the consistency of β^* in (2.13) we replace $(C - \bar{C})$ and $(Y - \bar{Y})$ in (2.12) by (2.7) and (2.8) to obtain

$$(2.14) \quad \beta^* = \frac{\sum \left\{ \left[\frac{\beta}{(1-\beta)}(Z - \bar{Z}) + \frac{\varepsilon - \bar{\varepsilon}}{1-\beta} \right] [Z - \bar{Z}] \right\}}{\sum \left\{ \left[\frac{1}{1-\beta}(Z - \bar{Z}) + \frac{(\varepsilon - \bar{\varepsilon})}{1-\beta} \right] [Z - \bar{Z}] \right\}}$$

$$= \frac{\beta \sum(Z - \bar{Z})^2 / N + \sum(\varepsilon - \bar{\varepsilon})(Z - \bar{Z}) / N}{\{\sum(Z - \bar{Z})^2 / N + \sum(\varepsilon - \bar{\varepsilon})(Z - \bar{Z}) / N\}}$$

Now as $N \rightarrow \infty$

$$\beta^* \rightarrow \beta;$$

hence, β^* is a consistent estimator and is obtained by obtaining consistent estimators of the reduced form (LSNR) and then deducing corresponding estimates of structural coefficients. This general method is referred to as indirect least squares (ILS), but is not applicable for all structural models.

The consistent estimator β^* can also be obtained by replacing the dependent variable on the right hand side of (2.1) by its predicted value (from the reduced form)

$$\hat{Y} = \hat{\pi}_{21} + \hat{\pi}_{22}Z$$

or $\hat{Y} - \bar{Y} = \hat{\pi}_{22}(Z - \bar{Z})$

and then applying least squares to the resultant expression. More explicitly,

$$\begin{aligned}
 (2.15 \text{ a-e}) \quad \beta^* &= \frac{\sum(\hat{Y} - \bar{Y})(C - \bar{C})}{\sum(\hat{Y} - \bar{Y})^2} \\
 &= \frac{\hat{\pi}_{22}}{\hat{\pi}_{22}^2} \frac{\sum(Z - \bar{Z})(C - \bar{C})}{\sum(Z - \bar{Z})^2} \\
 &= \frac{1}{\hat{\pi}_{22}} \frac{\sum(Z - \bar{Z})(C - \bar{C})}{\sum(Z - \bar{Z})^2} \\
 &= \left\{ \frac{\sum(Z - \bar{Z})^2}{\sum(Y - \bar{Y})(Z - \bar{Z})} \right\} \left\{ \frac{\sum(Z - \bar{Z})(C - \bar{C})}{\sum(Z - \bar{Z})^2} \right\} \\
 &= \frac{\sum(Z - \bar{Z})(C - \bar{C})}{\sum(Y - \bar{Y})(Z - \bar{Z})}
 \end{aligned}$$

which corresponds to (2.13). Compare (2.15 a) with (2.6) and note that the only difference is that

\hat{Y} (predicted value) replaces Y in (2.6). The structural estimator, obtained by applying least squares to the structural equation which has been modified by replacing the right hand dependent variables by their reduced form predictions is referred to as two stage least squares (2SLS).

2SLS yields consistent estimators, and is applicable even when indirect least squares is not.

Another way of looking at the alternative estimator is obtained by comparing (2.6) and (2.15e).

Here we see that the difference is that the right hand side dependent variable Y in (2.6) is replaced by Z (an instrumental variable) which is correlated with Y , but not with C ; hence, these estimators are sometimes referred to as instrumental variables estimators.

A numerical example: the Haavelmo data set, (**Haavelmo.dat**).

Using the data provided by Haavelmo, the regular OLS estimates of the consumption function given by

$$\begin{aligned}
 \hat{C}_{OLS} &= 84.01 + .732Y \\
 (s_{\hat{\beta}_i}) &\quad (14.55) \quad (.030)
 \end{aligned}$$

$$R^2 = .971$$

$$s^2 = 58.21.$$

The corresponding 2SLS estimates of the consumption function are given by

$$\hat{C}_{2SLS} = 113.1 + .672Y$$

(17.8) (.037)

$$s^2 = 71.29.$$

The LSNR estimates of the reduced form equations are given by

$$\hat{C} = 344.70 + 2.048Z$$

(16.48) (.341)

$$R^2 = .668$$

$$\hat{Y} = 344.70 + 3.048Z$$

(16.48) (.341)

$$R^2 = .668$$

The reader should verify that the indirect least squares estimators are equal to the 2SLS.

However, except for pedagogical examples, the reader will apply 2SLS or instrumental variables estimation directly and not use the two step procedure. Also, the two step procedure yields incorrect standard errors.

CONFIDENCE INTERVALS. In determining confidence intervals for structural parameters, the reader might be inclined to use the results associated with the OLS or 2SLS estimates of the structural equation under consideration. As an example of this we compute "95% confidence intervals for β (the MPC)."

(a) Based upon OLS: ($t = 2.101$)

$$\begin{aligned}\hat{\beta}_{OLS} &\pm ts_{\hat{\beta}} \\ &= (.732 \pm 2.101(.0299)) \\ &= (.669, .795)\end{aligned}$$

(b) Based upon 2SLS

$$\begin{aligned}\hat{\beta}_{2SLS} &\pm ts_{\hat{\beta}} \\ &= (.672 \pm 2.101(.0368)) \\ &= (.594, .748)\end{aligned}$$

These confidence intervals are very different and one might ask which if either is appropriate. As it turns out, neither is completely satisfactory since

$$\frac{\hat{\beta} - \beta}{s_{\hat{\beta}}}$$

is not exactly distributed as a t-statistic where $\hat{\beta}$ is obtained from the technique of OLS or 2SLS.

One way in which we can determine which (if either) of the previous confidence intervals is closest is to note that

$$\frac{\hat{\pi}_{ij} - \pi_{ij}}{s_{\hat{\pi}_{ij}}} \sim t(n-2);$$

hence,

$$\begin{aligned}1 - \alpha &= \Pr[-t_{\alpha/2} \leq \frac{\hat{\pi}_{22} - \pi_{22}}{s_{\hat{\pi}_{22}}} \leq t_{\alpha/2}] \\ &= \Pr[\hat{\pi}_{22} - t_{\alpha/2} s_{\hat{\pi}_{22}} \leq \pi_{22} \leq \hat{\pi}_{22} + t_{\alpha/2} s_{\hat{\pi}_{22}}] \\ &= \Pr[\hat{\pi}_{22} - t_{\alpha/2} s_{\hat{\pi}_{22}} \leq \frac{1}{1-\beta} \leq \hat{\pi}_{22} + t_{\alpha/2} s_{\hat{\pi}_{22}}] \\ &= \Pr[1 - \frac{1}{\hat{\pi}_{22} - t_{\alpha/2} s_{\hat{\pi}_{22}}} \leq \beta \leq 1 - \frac{1}{\hat{\pi}_{22} + t_{\alpha/2} s_{\hat{\pi}_{22}}}].\end{aligned}$$

Making the appropriate substitutions we obtain

$$(.57, .73)$$

which is much closer to the results obtained using two least squares than from OLS. One might be inclined to conjecture that a reason for the poor performance of OLS confidence intervals is due to the asymptotic bias of OLS estimator,

$$\frac{\sigma^2(1-\beta)}{\sigma^2 + \sigma^2}.$$

It might be instructive to estimate the asymptotic bias. Doing so we obtain for OLS estimates of $\sigma^2(s^2=58.2)$, $\beta(\hat{\beta}=.732)$, $\sigma_z^2(285.55)$; hence asymptotic bias ($\hat{\beta}_{OLS}$) = .0454; for 2SLS estimates of $\sigma^2(s^2=71.29)$, $\beta(\hat{\beta}=.672)$, $\sigma_z^2(285.55)$, asymptotic bias ($\hat{\beta}_{OLS}$) = .0655. Note that the difference between the OLS and 2SLS is (.732 - .672 = .06).

PREDICTIONS. In order to make predictions, one should use the reduced form representation.

c. A BRIEF OVERVIEW

The mathematical formulation of an economic model is generally referred to as the structural representation. The structural equations in the structural representation will often include endogenous regressors (endogenous variables on the right hand side) as well as exogenous variables.

The reduced form representation corresponding to the structural representation is characterized by separate equations which express each dependent variable as a function of the exogenous variables. The reduced form provides explicit expressions for the equilibrium values of the dependent variables in the model, conditional on an arbitrary, but given, set of values for the exogenous variables. The reduced form coefficients can be interpreted as "multipliers" and yield comparative static results. The reduced form representation is usually the form used for obtaining forecasts from econometric models.

After the econometrician is satisfied that a given econometric model is consistent with relevant economic theory, it is important that each structural equation be identified. Identification should be checked even before attempting to estimate the model. A necessary condition (order condition) for identification is that the number of exogenous (predetermined) variables excluded (K_2) from a structural equation is at least as large as the number of endogenous regressors (one less than the number of endogenous variables in the equation being checked (G_Δ)),

$$K_2 \geq G_\Delta - 1$$

Stated a little differently, the number of instrumental variables must be at least as large as the number of endogenous regressors. This condition must be satisfied for each structural

equation. The values for K_2 and G_Δ may vary from one equation to another. Identities do not contain unknown parameters and need not be checked for identification.

OLS estimates of parameters in **structural models** are typically biased and inconsistent and have unreliable t-statistics. This problem is due to non-zero correlation between the error term and the endogenous regressors on the right hand side of the equation. Two stage least squares estimators (2SLS) provide biased, but consistent estimators. They can also be viewed as instrumental variables estimators.

The **Stata command** for 2SLS is

```
ivregress 2sls y1 (Y2 Y3=X1 X2) X1
```

where Y = endogenous variables (y1 on lhs, y2 and y3 on the rhs),

X1 = exogenous variables in structural equation being estimated,

X2=Z = exogenous variables in the model, but excluded from the structural equation being estimated. The variables in X2 are often called instruments. An alternative form for the two stage estimators is given by

```
ivregress 2sls y1 (Y2 Y3=X2) X1
```

“ivreg” can be used in place of the command “ivregress”. 2sls is the default and the corresponding command can be written more compactly as “ivreg y1 (y2 y3=X2) X1.

Example 1: See the problem set for some sample data

Demand: $Q = \gamma_{11} - \beta_{12}P + \gamma_{12} Y + \varepsilon_{1t}$

Supply: $Q = \gamma_{21} + \beta_{22}P - \gamma_{23} FC + \varepsilon_{2t}$

ENDOGENOUS VARIABLES: Q, P

EXOGENOUS VARIABLES: Y, FC

(a) Identification(1) Demand $K_2 = 1$ FC is in the supply model, but
not in the demand equation

$$G_\Delta - 1 = 2 - 1 = 1 \quad \text{One endogenous regressor (P) in the demand equation}$$

(2) Supply $K_2 = 1$ Y is in the demand model, but
not in the supply equation

$$G_\Delta - 1 = 2 - 1 = 1 \quad \text{One endogenous regressor (P) in the supply equation}$$

Therefore $K_2 \geq G_\Delta - 1$ is satisfied for the supply and demand equation.(b) 2SLS estimation of the structural parameters (STATA commands)

(1) Demand

ivregress 2sls Q (P = FC) Y or ivregress 2sls Q (P=Y FC) Y

(2) Supply

Ivregress 2sls Q (P = Y) FC or ivregress 2sls Q (P=Y FC) FC

(c) Estimation of the reduced form (STATA commands)

(1) Q Equation

reg Q Y FC

(2) P Equation

reg P Y FC

Example 2. Consider the Haavelmo model and data:

$$C_t = \alpha + \beta Y_t + \epsilon_t$$

$$Y_t = C_t + Z_t$$

(a) Identification

The exogenous variable Z is not included in the consumption function, but is in the identity.

(b) 2SLS estimation of the structural parameters (STATA commands)

```
ivregress 2sls c (Y=Z)
```

(c) Estimation of the reduced form parameters (STATA commands)

```
reg c z
```

```
reg y z
```

The data used by Haavelmo is given

Y	C	Z
433	394	39
483	423	60
479	437	42
486	434	52
494	447	47
498	447	51
511	466	45
534	474	60
478	439	39
440	399	41
372	350	22
381	364	17
419	392	27
449	416	33
511	463	48
520	469	51
477	444	33
517	471	46
548	494	54
629	529	100

References

Haavelmo, T. "Methods of Measuring the Marginal Propensity to Consume," Journal of American Statistical Association, 42(1947):105-122.

Working, E. "What Do Statistical Demand Curves Show?" Quarterly Journal of Economics,
41(1926):212-235.

VI. Simultaneous Equation Models

- 1. An introduction to some problems associated with simultaneous equation models**
- 2. Basic Notation and alternative representations**
 - a. Structural representation**
 - b. Reduced form representation**
 - c. Transfer function representation**
- 3. Identification**
 - a. Identification as a mapping (3 examples)**
 - b. Necessary conditions for a structural equations to be identified**
- 4. Estimation of reduced form parameters**
 - a. Least squares no restrictions (LSNR)**
 - b. Estimators derived from structural estimators**
 - c. SURE—generally the same as LSNR**
- 5. Estimation of Structural Models**
 - a. Notation**
 - b. Single equation methods: OLS, IV, 2SLS, LIML, k-class, GMM, other**
 - c. Simultaneous equation methods: FIML and 3SLS**
- 6. Statistical inference**
 - a. Identification**
 - b. Reduced form coefficients**
 - c. Structural parameters**
- 7. Exercises**

Structural Models

	Structural representation	Reduced form representation
Identification		
Estimation		
Interpreting coefficients		
Statistical inference		
Forecasting		

VI. Simultaneous Equation Models

1. An introduction to some problems associated with simultaneous equation models

- a. A graphical presentation of the identification problem. See "What Do Statistical Demand Curves Show" by E. J. Working in the QJE, 1927, pp. 212-235.
- b. The inconsistency of OLS estimators of the marginal propensity to consume in a simple national income model. See the article by Haavelmo in the March 1947 issue of JASA.

2. Formulation of linear economic models - basic notation and concepts

Linear economic models can be represented as a system of linear equations using matrices as reviewed in this section. The general specification will allow lagged values of the variables.

a. Structural Representation (G equations, G dependent variables)

$$\mathbf{B}\mathbf{Y}'_t + \mathbf{B}_1\mathbf{Y}'_{t-1} + \dots + \mathbf{B}_s\mathbf{Y}'_{t-s} + \boldsymbol{\Gamma}\mathbf{X}'_t + \boldsymbol{\varepsilon}'_t = 0 \quad (2.1)$$

where

$\mathbf{Y}_t = (Y_{t1}, \dots, Y_{tG})$ is a row vector of observations on G endogenous variables at time t;

$\mathbf{X}_t = (X_{t1}, \dots, X_{tK})$ is a row vector of observations on K exogenous variables at time t;

\mathbf{B} is a $G \times G$ matrix of unknown coefficients and is assumed to be nonsingular;

$\boldsymbol{\Gamma}$ is a $G \times K$ matrix of unknown coefficients; and

$\boldsymbol{\varepsilon}_t = (\varepsilon_{t1}, \varepsilon_{t2}, \dots, \varepsilon_{tG})$ is a row vector of random disturbances which is

distributed $N(0, \Omega)$ where Ω is a $G \times G$ positive definite matrix.

Let L denote the “lag” operator, $LY_t = Y_{t-1}$ $L^i Y_t = Y_{t-i}$

Using the lag operator, we can rewrite equation (2.1) as

(2.1)'

$$(B + B_1 L + \dots + B_s L^s) Y_t' + \Gamma X_t' + \varepsilon_t' = 0$$

or as

(2.1)''

$$B(L) Y_t' + \Gamma X_t' + \varepsilon_t' = 0$$

where $B(L) = (B + B_1 L + \dots + B_s L^s)$

is a matrix polynomial in

“matrix coefficients.”

- b. **Reduced Form Representation.** If $|B| \neq 0$, then the previously discussed structural representation of an economic model is observationally equivalent to the *reduced form representation*

$$Y_t' = \Pi_1 Y_{t-1}' + \dots + \Pi_s Y_{t-s}' + \Pi X_t' + \eta_t'$$

(2.2)

where

$\Pi_i = -(B)^{-1} B_i$ ($i = 1, 2, \dots, s$) are $G \times G$

$\Pi = -(B)^{-1} \Gamma$ is $G \times K$, and

$\eta_t' = (B)^{-1} \varepsilon_t' \sim N(0, \Sigma = (B)^{-1} \Omega (B^{-1})')$.

Note:

- The reduced form representation, expresses the current value of each endogenous variable in terms of predetermined (exogenous and lagged endogenous) variables.
- (2.2) is obtained by multiplying (2.1) by the matrix B^{-1} and solving $\vec{Y}_t = \vec{r}$
- (2.1) is said to represent a dynamic model if at least one of the B_i matrices is not a null matrix.
- Static models contain no lagged endogenous variables ($B_i = 0$ for all i): whereas, dynamic models contain lagged endogenous variables.
- The matrix Π , in the reduced form representation, contains the impact multipliers $\left(\frac{d\vec{Y}_t}{d\vec{X}_t} = \Pi = B^{-1}\Gamma \right)$, the instantaneous response of \vec{Y} to changes in \vec{X} .

c. The Final Form or Transfer Function Representation

If the modulus of the roots of $|B(z)| = 0$ are greater than one, then

$B^{-1}(L)$ exists and the transfer function corresponding to (2.1), or (2.1)' is given by

$$(2.3) \quad \vec{Y}'_t = -B^{-1}(L)\Gamma\vec{X}'_t - B^{-1}(L)\vec{\varepsilon}'_t$$

which expresses \vec{Y}'_t in terms of current and lagged values \vec{X}'_t . (2.3) is obtained by multiplying (2.1)' by $B^{-1}(L)$ and so \vec{Y}'_t is for . An inspection of the coefficients of $\vec{X}_t, \vec{X}_{t-1}, \vec{X}_{t-2}, \dots$ in (2.3) yields the impact and interim multipliers. Two results are quite easily obtained. The impact multiplier can be obtained from the equation

$$\frac{d\vec{Y}_t}{d\vec{X}_t} = -B^{-1}(L=0)\Gamma = -B^{-1}\Gamma = \Pi$$

which is the same result obtained from the reduced form representation. The sum of

the impact and interim multipliers gives the long run cumulative multiplier which is

equal to $-B^{-1}(L=1)\Gamma = -(B + B_1 + \dots + B_s)^{-1}\Gamma$

obtained by expressing $B^{-1}(L)$ as a matrix polynomial of infinite order and

selecting the appropriate term in the expansion of $-B^{-1}(L)\Gamma\vec{X}'_t$. The exercises illustrate applications of these results.

VI. Simultaneous Equation Models

3. Identification logically precedes estimation

a. Identification as a Mapping Problem

Identification is a property of the mapping between the structural parameter

space $\mathcal{B} = \{(B, B_1, \dots, B_s, \Gamma, \Omega), \text{ where } B, \Omega, \text{ and } B_i \text{ are } G \times G \text{ matrices, } \Gamma \text{ is } G \times K,$

Ω is positive definite and symmetric, $|B| \neq 0\}$, and the corresponding reduced form

parameters space $\mathcal{A} = \{(\Pi_1, \dots, \Pi_s, \Pi, \Sigma) | \Pi_i = -B^{-1}B_i, i = 1, 2, \dots, s$

$\Pi = -B^{-1}\Gamma, \Sigma = (B)^{-1}\Omega(B')^{-1} \text{ and } (B, B_1, \dots, B_s, \Omega, \Gamma) \in \mathcal{B}\}$

A structural economic hypothesis can be represented as a definite proper subset

\mathcal{B}_I of the structural parameter space (\mathcal{B}). The set \mathcal{B}_I can be described by restrictions

on the elements of the matrices $B, B_1, \dots, B_s, \Gamma, \Omega$ such as exclusions of variables

from equations, functional dependencies among structural parameters, or restrictions

on the variance covariance matrices.

Define $\mathcal{A}_I = \{(\Pi_1, \dots, \Pi_s, \Pi, \Sigma) | \Pi_i = -B^{-1}B_i,$

$\Pi = -B^{-1}\Gamma, \Sigma = (B)^{-1}\Omega(B')^{-1} \text{ and}$

$(B, B_1, \dots, B_s, \Gamma, \Omega) \in \mathcal{B}_I\}$

\mathcal{A}_I is then said to be the reduced form hypothesis corresponding to the structural

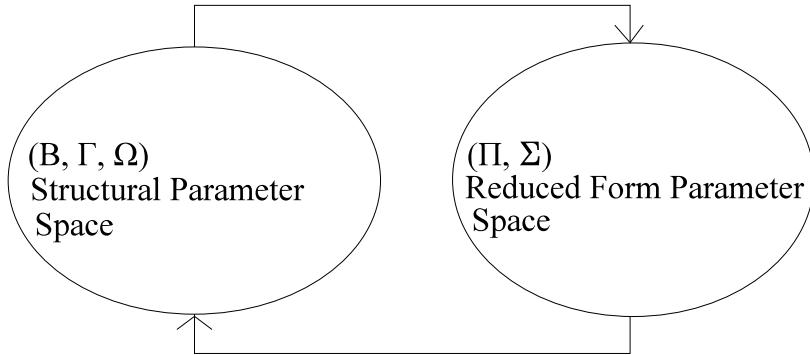
hypothesis \mathcal{B}_I .

The structural hypothesis (model) \mathcal{B}_I is identified if and only if the mapping

between \mathcal{B}_I and \mathcal{A}_I is one-to-one. .

For the static case, the identification problem can be graphically depicted as follows:

$$(1) \quad \Pi = -B^{-1}\Gamma \\ \Sigma = B^{-1}\Omega(B^{-1})^{-1}$$



(2) Can the equations
 $\Pi = -B^{-1}\Gamma$ and
 $\Sigma = B^{-1}\Omega(B^{-1})^{-1}$ or $\Omega = B\Sigma B^{-1}$
be solved for B , Γ and Ω in terms of Π and Σ ?

A **necessary condition** for the mapping between A_1 and B_1 to be one-to-one is that A_1 and B_1 have the same dimension (the same number of free parameters).

Dimension of B

(1) Coefficients

$$B: \quad G^2$$

$$\Gamma: \quad G \cdot K$$

Dimension of A

$$\Pi: \quad GK$$

(2) Variance-covariance parameters

$$\Omega \quad G + \frac{G(G-1)}{2}$$

$$\Sigma \quad G + \frac{G(G-1)}{2}$$

Total parameters:

$$G^2 + GK + G + \frac{(G-1)(G)}{2}$$

$$\text{Total: } GK + G + \frac{(G-1)(G)}{2}$$

The structural representation involves G^2 more parameters than the reduced form representation. A necessary condition for a one to one correspondence between B_1 and A_1 is for B_1 to be associated with at least G independent restriction on each structural equation. A common, but not a necessary, approach to this dimensionality problem is to impose restrictions of the following types on the coefficients of each of the structural equations. Ideally, economic theory provides the basis for these restrictions, where exclusions corresponds to specifying some coefficients are zero..

<u>Types of structural restrictions</u>	<u>Number of restrictions</u>
Normalize on one dependent variable	1
# of endogenous variables hypothetically excluded	$G_{\Delta\Delta}$
# of exogenous variables hypothetically excluded	K_2
Total restrictions	$1 + G_{\Delta\Delta} + K_2$

Necessary condition for exclusion restrictions to yield identification¹:

total number of restrictions $\geq G$ or

- $1 + G_{\Delta\Delta} + K_2 \geq G$, or
- $K_2 \geq G - G_{\Delta\Delta} - 1$, or
- $K_2 \geq G_{\Delta} - 1$ where G_{Δ} = the number of included endogenous variables

$G_{\Delta} - 1$ = the number of **endogenous regressors**

If $K_2 > G_{\Delta} - 1$, then $v = K_2 - G_{\Delta} + 1$ is said to be the number of overidentifying restrictions. In the terminology of the instrumental variables literature v is the number of extra instruments, $v = \text{number of instruments} - \text{number of endogenous regressors}$.

In summary, a necessary condition (sometimes referred to as the order condition) for each structural equation to be identified, by solving $B\Pi + \Gamma = 0$, based on coefficient restrictions is

- $K_2 \geq G_{\Delta} - 1$
- $K_2 - G_{\Delta} + 1 \geq 0$
- $\# \text{instruments} \geq \# \text{endogenous regressors}$

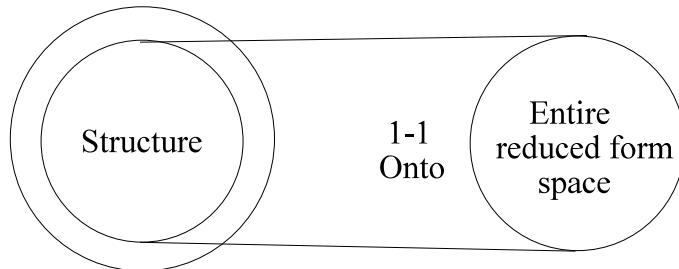
¹There are other types of restrictions which can be imposed on the structural coefficients to make the mapping one-to-one. For example, structural vector auto-regressive models (SVAR) use restrictions on the variance covariance matrix to achieve identification.

Another way of thinking about the condition $K_2 \geq G_\Delta - 1$ is that, for each structural equation, there be at least as many excluded explanatory (predetermined) variables as there are right hand side endogenous variables.

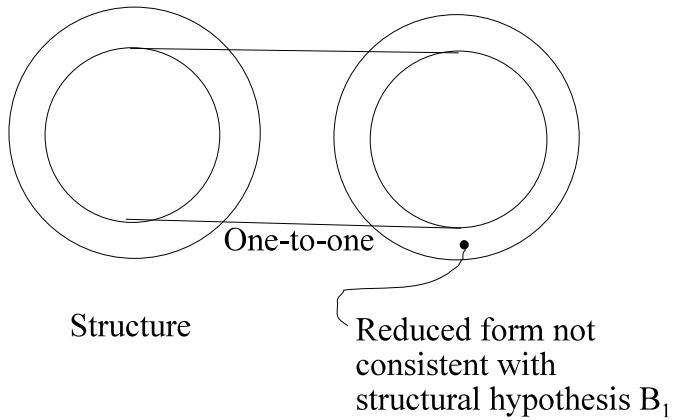
Alternatively, a necessary condition for a structural equation to be identified is that one predetermined variable must be excluded for each endogenous regressor, which is saying that you need at least one instrument for each endogenous regressor, $v = K_2 - (G_\Delta - 1) \geq 0$.

If the mapping between B_1 and A_1 is 1-1 and

- (1) $v = 0$, the structural equation is said to be exactly identified



- (2) If $v > 0$, those equations are said to be over identified



Note: overidentification is not bad, but it imposes constraints on the reduced form coefficients the validity of which can be tested. using the **estat overid** command in Stata.

(3) If $v < 0$ ($K_2 < G\Delta - 1$), the structural equation is said to be under identified and more than one structure can correspond to the same reduced form. The necessary condition for identification is not satisfied.

Note: $v = K_2 - G_\Delta + 1 \geq 0$

is a necessary condition, but is not sufficient.

Sufficient conditions (order conditions) are developed on a case-by-case basis. In the case where identifying restrictions are exclusion restrictions the sufficient condition is

$$\text{rank}[\Pi_{\Delta 2}] = G_\Delta - 1 \text{ or } \text{rank}[B_2 \Gamma_2] = G - 1 \quad \Pi_{\Delta 2}$$

reduced form coefficients corresponding to the endogenous variables in the structural equation being estimated and the excluded predetermined variables and

$[B_2 \Gamma_2]$ denotes the matrix of structural coefficients in all other structural equations

corresponding to the excluded dependent and independent or predetermined variables. Derivations of this condition can be found in older econometrics books, such as those by Johnston or Kmenta or in earlier versions of my class notes.

The following three examples will illustrate each of these situations: :

Example 1: Supply-Demand Model: exactly identified

Structural Model:

$$\text{Demand: } -Q_t - \beta_{12}P_t + \gamma_{11} + \gamma_{12}Y_t + \varepsilon_{t1} = 0$$

$$\text{Supply: } -Q_t + \beta_{22}P_t + \gamma_{21} - \gamma_{23}FC_t + \varepsilon_{t2} = 0$$

$$\begin{bmatrix} -1 & -\beta_{12} \\ -1 & \beta_{22} \end{bmatrix} \begin{bmatrix} Q_t \\ P_t \end{bmatrix} + \begin{bmatrix} \gamma_{11} & \gamma_{12} & 0 \\ \gamma_{21} & 0 & -\gamma_{23} \end{bmatrix} \begin{bmatrix} 1 \\ Y_t \\ FC_t \end{bmatrix} + \begin{bmatrix} \varepsilon_{t1} \\ \varepsilon_{t2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Where P, Q, Y and FC, respectively, denote price, quantity, income, and factor costs.

Reduced Form Representation:

$$\begin{aligned} \begin{bmatrix} Q_t \\ P_t \end{bmatrix} &= - \begin{bmatrix} -1 & -\beta_{12} \\ -1 & \beta_{22} \end{bmatrix}^{-1} \left\{ \begin{bmatrix} \gamma_{11} & \gamma_{12} & 0 \\ \gamma_{21} & 0 & -\gamma_{23} \end{bmatrix} \begin{bmatrix} 1 \\ Y_t \\ FC_t \end{bmatrix} + \begin{bmatrix} \varepsilon_{t1} \\ \varepsilon_{t2} \end{bmatrix} \right\} \\ &= \frac{1}{\beta_{12} + \beta_{22}} \begin{bmatrix} \beta_{22}\gamma_{11} + \beta_{12}\gamma_{21} & \beta_{22}\gamma_{12} - \beta_{12}\gamma_{23} \\ \gamma_{11} - \gamma_{21} & \gamma_{12} - \gamma_{23} \end{bmatrix} \begin{bmatrix} 1 \\ Y_t \\ FC_t \end{bmatrix} + \begin{bmatrix} \beta_{22}\varepsilon_{t1} + \beta_{12}\varepsilon_{t2} \\ \varepsilon_{t1} - \varepsilon_{t2} \end{bmatrix} \end{aligned}$$

$$= \begin{bmatrix} \pi_{11} & \pi_{12} & \pi_{13} \\ \pi_{21} & \pi_{22} & \pi_{23} \end{bmatrix} \begin{bmatrix} 1 \\ Y_t \\ FC_t \end{bmatrix} + \begin{bmatrix} \eta_{t1} \\ \eta_{t2} \end{bmatrix}$$

Note: There are six β 's and γ 's and six π_{ij} 's

Identification: Identification involves solving for the β_{ij} 's and γ_{ij} 's in terms of the π_{ij} 's. There is a one-to-one mapping between the structural and reduced form parameters. **Check each equation for identifiability (rank and order).**

Example 2. Supply-Demand Model: an over identified structural equation

Structural Model:

$$\begin{bmatrix} -1 & -\beta_{12} \\ -1 & \beta_{22} \end{bmatrix} \begin{bmatrix} Q_t \\ P_t \end{bmatrix} + \begin{bmatrix} \gamma_{11} & \gamma_{12} & \gamma_{13} & 0 \\ \gamma_{21} & 0 & 0 & -\gamma_{24} \end{bmatrix} \begin{bmatrix} Y_t \\ P_{st} \\ FC_t \end{bmatrix} + \begin{bmatrix} \varepsilon_{t1} \\ \varepsilon_{t2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

where P_{st} = price of substitutes.

Reduced Form Representation:

$$\begin{aligned} \begin{bmatrix} Q_t \\ P_t \end{bmatrix} &= -\begin{bmatrix} -1 & -\beta_{12} \\ -1 & \beta_{22} \end{bmatrix}^{-1} \left\{ \begin{bmatrix} \gamma_{11} & \gamma_{12} & \gamma_{13} & 0 \\ \gamma_{21} & 0 & 0 & -\gamma_{24} \end{bmatrix} \begin{bmatrix} Y_t \\ P_{st} \\ FC_t \end{bmatrix} + \begin{bmatrix} \varepsilon_{t1} \\ \varepsilon_{t2} \end{bmatrix} \right\} \\ &= \frac{1}{\beta_{12} + \beta_{22}} \left\{ \begin{bmatrix} \beta_{22}\gamma_{11} + \beta_{12}\gamma_{21} & \beta_{22}\gamma_{12} & \beta_{22}\gamma_{13} & -\beta_{12}\gamma_{24} \\ \gamma_{11} - \gamma_{21} & \gamma_{12} & \gamma_{13} & \gamma_{24} \end{bmatrix} \begin{bmatrix} Y_t \\ P_{st} \\ FC_t \end{bmatrix} + \begin{bmatrix} \eta_{t1} \\ \eta_{t2} \end{bmatrix} \right\} \\ &= \begin{bmatrix} \pi_{11} & \pi_{12} & \pi_{13} & \pi_{14} \\ \pi_{21} & \pi_{22} & \pi_{23} & \pi_{24} \end{bmatrix} \begin{bmatrix} Y_t \\ P_{st} \\ FC_t \end{bmatrix} + \begin{bmatrix} \eta_{t1} \\ \eta_{t2} \end{bmatrix} \end{aligned}$$

Note: There are seven β_{ij} 's and γ_{ij} 's and eight π_{ij} 's

Identification. Solve for the β_{ij} 's and γ_{ij} 's in terms of the π_{ij} 's. This example illustrates an overidentified model where restrictions are imposed on the reduced form parameters, e.g.

$$\beta_{22} = \frac{\pi_{12}}{\pi_{22}} = \frac{\pi_{13}}{\pi_{23}}$$

Check the necessary (order) condition for each structural equation. Check the sufficient condition (rank) condition for each structural equation.

Example 3. Supply-Demand Model: an under identified structural equation

Structural Model:

$$\begin{bmatrix} -1 & -\beta_{12} \\ -1 & \beta_{22} \end{bmatrix} \begin{bmatrix} Q_t \\ P_t \end{bmatrix} + \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & 0 \end{bmatrix} \begin{bmatrix} 1 \\ Y_t \end{bmatrix} + \begin{bmatrix} \epsilon_{t1} \\ \epsilon_{t2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Reduced Form Representation:

$$\begin{bmatrix} Q_t \\ P_t \end{bmatrix} = - \begin{bmatrix} -1 & -\beta_{12} \\ -1 & \beta_{22} \end{bmatrix}^{-1} \left\{ \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & 0 \end{bmatrix} \begin{bmatrix} 1 \\ Y_t \end{bmatrix} + \begin{bmatrix} \epsilon_{t1} \\ \epsilon_{t2} \end{bmatrix} \right\}$$

$$= \begin{bmatrix} 1 \\ \beta_{12} + \beta_{22} \end{bmatrix} \begin{bmatrix} \beta_{22}\gamma_{11} + \beta_{12}\gamma_{21} & \beta_{22}\gamma_{12} \\ \gamma_{11} - \gamma_{21} & \gamma_{12} \end{bmatrix} \begin{bmatrix} 1 \\ Y_t \end{bmatrix} + \begin{bmatrix} \eta_{t1} \\ \eta_{t2} \end{bmatrix}$$

$$= \begin{bmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{bmatrix} \begin{bmatrix} 1 \\ Y_t \end{bmatrix} + \begin{bmatrix} \eta_{t1} \\ \eta_{t2} \end{bmatrix}$$

Note: There are five β_{ij} 's and γ_{ij} 's and only four π_{ij} 's

Identification: Attempt to solve for the β_{ij} 's and γ_{ij} 's in terms of the π_{ij} 's. In this model β_{22} can be uniquely expressed in terms of the reduced form parameters, but β_{12} can not.

Check the necessary (order) condition for each structural equation. Check the sufficient condition (rank) condition for each structural equation.

VI. Simultaneous Equation Models

4. Estimation of Reduced Form Parameters

Given the Structural model

$$YB' + X\Gamma' + \varepsilon = 0 \text{ where } \varepsilon_t \sim N(0, \Omega) \text{ for all } t=1, 2, \dots, N \text{ and}$$

Y and X , respectively, have dimension $N \times G$ and $N \times K$, then

the associated reduced form can be written in the form

$$Y = XII' + \eta \quad \text{where } \eta_t \sim N(0, \Sigma).$$

Thus, each reduced form equation expresses the equilibrium value of an endogenous variable in terms of explanatory variables. There are two main methods of estimating the reduced form coefficients.

a. Reduced form estimators using Least Squares No Restrictions (LSNR)*

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

$$\text{where } \Pi = \begin{bmatrix} \pi_1 \\ \vdots \\ \pi_G \end{bmatrix}_{G \times K}$$

and π_i denotes the coefficients in the reduced form equation for Y_{ti} , $\pi_i = (\pi_{i1}, \pi_{i2}, \dots, \pi_{iK})$.

Let Π_{vec} denote the $(GK) \times 1$ column vector of reduced form coefficients obtained by stacking the Π 's in a column, i.e.,

$$\Pi'_{vec} = [\pi_1, \pi_2, \dots, \pi_G].$$

The distribution of $\hat{\Pi}_{vec}$, using LSNR, is given by

$$\hat{\Pi}_{vec} \sim N[\Pi_{vec}, \Sigma \otimes (X' X)^{-1}]$$

and Σ is estimated by

$$\hat{\Sigma} = \left(\frac{1}{N} \right) (Y - X\hat{\Pi})' (Y - X\hat{\Pi})$$

where $\hat{\Sigma}$ has a Wishart distribution. $(N-K)$ is often used as a divisor.

Note:

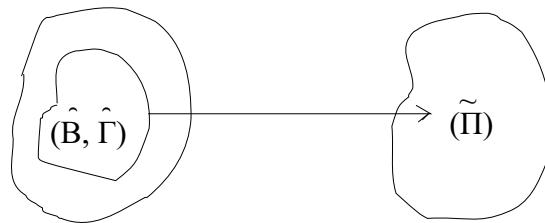
- (1) LSNR Estimators of Π are consistent, unbiased, and normally distributed; however, LSNR doesn't take account of overidentifying restrictions. (Schmidt (1976)) and need not be asymptotically efficient.
- (2) This approach merely amounts to applying least squares to each reduced form equation, regressing each of the dependent variables on the predetermined variables.

Stata commands:

`reg y x's`

b. Reduced form estimators obtained from estimators of (B, Γ) .

This approach might be visualized by the following figure:



and involves two steps:

- (1) obtain consistent estimates of B, Γ :

$$\hat{B}, \hat{\Gamma} .$$

- (2) calculate $\tilde{\Pi} = -\hat{B}^{-1} \hat{\Gamma}$

If each structural equation is exactly identified, then these estimators will be identical to least squares no restrictions estimators (LSNR). For models containing over identified structural equations, LSNR of II doesn't take account of the

overidentifying restrictions imposed on the population reduced form parameters; whereas

$$\tilde{\Pi} = -\hat{B}^{-1} \hat{\Gamma}$$

does. We will comment on the relative merits of LSNR of II and alternative estimators of the form

$$\tilde{\Pi} = -\hat{B}^{-1} \hat{\Gamma}$$

later. The reduced estimators obtained in this way are often named after the method used to estimate (B, Γ) . For example, if two stage least squares is used to estimate (B, Γ) , then corresponding estimator of Π would be referred to as two stage least squares estimation of Π . The asymptotic distribution of

$$\sqrt{N} (\tilde{\Pi}_{\text{vec}} - \Pi_{\text{vec}})$$

is given by

$$N [0, D' V D]$$

where

$$D = (B^{-1})' \otimes \begin{pmatrix} \Pi \\ I_K \end{pmatrix}$$

$$\sqrt{N} \text{vec} \left\{ \begin{bmatrix} \hat{\beta} \\ \hat{\Gamma} \end{bmatrix} - \begin{bmatrix} \beta \\ \Gamma \end{bmatrix} \right\} \rightarrow N [0, V]$$

See Schmidt (pp. 237-8) for a proof using a different notation.

Notes:

- (1) Restricted or derived reduced form estimators of the form

$$\tilde{\Pi} = -\hat{B}^{-1} \hat{\Gamma}$$

will be consistent if \hat{B} and $\hat{\Gamma}$ are consistent estimators

Proof: $\text{plim } \tilde{\Pi} = -\text{plim } \hat{B}^{-1} \hat{\Gamma}$

$$= -(\text{plim } \hat{B}^{-1})^{-1} \text{plim}$$

$$= -B^{-1} \Gamma = \Pi$$

- (2) 3SLS and FIML derived reduced form estimators of π are asymptotically efficient relative to LSNR; however, this doesn't imply that the corresponding estimators will have smaller (exact) variances than LSNR for any given sample size. (McCarthy, IER, 1971, pp. 757-751)
- (3) Moments (e.g. means and variances) of restricted reduced form estimators need not exist corresponding to 2SLS. This has implications for forecasts. (Dhrymes, (1973) Econometrica, pp. 119-134)
- (4) 2SLS and LIML restricted estimators of the reduced form coefficients are not necessarily asymptotically efficient relative to LSNR.
 - LSNR uses all sample information, not all restrictions.
 - Restricted estimation uses restrictions--not all sample information.
- (5) If all equations are exactly identified, then LSNR, 2SLS, LIML, 3SLS, FIML estimates of the reduced form will be identical.

References: Schmidt, Peter. Econometrics, Marcel Dekker; 1976.

c. SURE—(not so fast, SURE are generally the same as LSNR)

One might be tempted to use SURE to estimate the reduced form coefficients. If the random disturbances in the reduced form equations are correlated across equations, then SURE would yield estimators with smaller variance than LSNR if overidentifying restrictions imply that some of the exogenous variables should be deleted for some equations. Otherwise, SURE with identical regressors yields the same estimators as LSNR. This follows from the homework assignment which explored conditions under which SURE and OLS give identical results. One of these conditions is when the separate equations being estimated include identical regressors as is the case in estimating reduced form parameters without restrictions.

VI. Simultaneous Equation Models

5. Estimation of Structural Parameters (\mathbf{B} , Γ , Ω)

This section will introduce alternative methods of estimating structural parameters. These methods can be viewed as being of two types: (1) those in which the structural parameters are estimated one equation at a time and (2) those in which all structural parameters are estimated simultaneously.

a. Review of Notation

$$Y = (y_{ti}) \quad t = 1, 2, \dots, N; i = 1, 2, \dots, G;$$

$$X = (X_{tj}) \quad t = 1, 2, \dots, N; j = 1, 2, \dots, K;$$

$y_{\cdot i} = (y_{1i}, \dots, y_{Ni})'$ column vector, i^{th} column of Y matrix;

$y_{\cdot t} = (y_{t1}, \dots, y_{tG})'$ row vector, t^{th} row of Y matrix;

$$Y_1 = \begin{bmatrix} y_{12} & \dots & y_{1G_A} \\ y_{22} & \dots & y_{2G_A} \\ \vdots & & \\ y_{N2} & \dots & y_{NG_A} \end{bmatrix}$$

is an $N \times G_A - 1$ matrix of observations of the dependent variables (endogenous regressors) Y_{12}, \dots, Y_{tG_A} appearing on the right hand side of the structural equation under consideration;

$X_1 = (X_{tj}) \quad t = 1, 2, \dots, N, j = 1, 2, \dots, K_1$; X_1 is the $N \times K_1$ matrix of observations on the exogenous included in the structural equation under consideration.

$$X_2 = (X_{ti}) \quad t = 1, 2, \dots, N, j = K_{1+1}, \dots, K$$

X_2 is the $N \times K_2$ matrix of observations of the exogenous excluded from the particular structural equation under consideration.

β_i = coefficients of nonnormalized endogenous variables appearing in the i^{th} structural equation (endogenous regressors); and

γ_i = coefficients of exogenous variables which are hypothetically included in the i^{th} structural equation.

Let $y_{t1} = \beta_{12}y_{t2} + \dots + \beta_{1G_1}y_{tG_1} + \gamma_{11}X_{t1} + \dots + \gamma_{1K_1}X_{tK_1} + \varepsilon_{t1}$

denote the first structural equation from the system of structural equations,

$$\boxed{\begin{aligned} By'_{t1} + \Gamma X'_{t1} + \varepsilon'_{t1} &= 0 \\ YB' + X\Gamma' + \varepsilon &= 0 \end{aligned}} \quad (5.2)$$

Equation (5.1) can also be written in the following form in terms of the matrices

$$y_{11} = [y_{12} \dots y_{1G_1}] \begin{bmatrix} \beta_{12} \\ \vdots \\ \beta_{1G_1} \end{bmatrix} + [X_{11} \dots X_{1K_1}] \begin{bmatrix} \gamma_{11} \\ \vdots \\ \gamma_{1K_1} \end{bmatrix} + \varepsilon_{11}$$

•
•
•

$$y_{N1} = [y_{N2} \dots y_{NG_1}] \begin{bmatrix} \beta_{12} \\ \vdots \\ \beta_{1G_1} \end{bmatrix} + [X_{N1} \dots X_{NK_1}] \begin{bmatrix} \gamma_{11} \\ \vdots \\ \gamma_{1K_1} \end{bmatrix} + \varepsilon_{N1}$$

or, equivalently,

$$\boxed{y_{11} = Y_1 \beta'_{11} + X_1 \gamma'_{11} + \varepsilon_{11}} \quad (5.1)'$$

or

$$\boxed{y_{11} = (Y_1 X_1) \begin{pmatrix} \beta'_{11} \\ \gamma'_{11} \end{pmatrix} + \varepsilon_{11}} \quad (5.1)''$$

This notation will facilitate a formal presentation of estimation procedures. The *concentration parameter* associated with the structural equation (5.1) is defined by

$$\mu^2 = (X_2 \Pi_{22})' \left(I - X_1 (X_1' X_1)^{-1} X_1' \right) (X_2 \Pi_{22}) / (\text{Var}(\varepsilon_{t1}))$$

and provides a measure of the spread of the estimator around the true value and is related to the “strength” of the instruments (excluded exogenous variables). An unbiased estimator of this parameter is provided by

$$\hat{\mu}^2 = K_2 (\tilde{F} - 1)$$

where \tilde{F} denotes the Chow statistic associated testing the hypothesis that the instruments ($Z = X_2$) or exogenous variables excluded from structural equation have significant explanatory power for the endogenous regressor(s) in equation (5.1). Generally speaking, the larger the value of the concentration parameter, the smaller the variance (where defined) of the consistent structural estimators and the better the limiting t and F distributions will match the corresponding exact finite sample distributions.

b. Single Equation Techniques: OLS, 2SLS, LIML, k-class, Sawa's combined estimator.

(1) Ordinary Least Squares Estimators (OLS)

$$y_{1\cdot} = Y_1 \beta'_{1\cdot} + X_1 \gamma'_{1\cdot} + \varepsilon_{1\cdot} \quad (5.3)$$

$$= [Y_1 X_1] \begin{bmatrix} \beta'_{1\cdot} \\ \gamma'_{1\cdot} \end{bmatrix} + \varepsilon_{1\cdot}$$

The corresponding sum of squared errors can be written as

$$\begin{aligned} SSE &= (y_{1\cdot} - Y_1 \hat{\beta}_{1\cdot} - X_1 \hat{\gamma}_{1\cdot})' (y_{1\cdot} - Y_1 \hat{\beta}_{1\cdot} - X_1 \hat{\gamma}_{1\cdot}) \\ &= (-1, \hat{\beta}_{1\cdot}) Y_\Delta' \left(I - X_1 (X_1' X_1)^{-1} X_1' \right) Y_\Delta \begin{pmatrix} -1 \\ \hat{\beta}_{1\cdot} \end{pmatrix} \end{aligned}$$

where Y_Δ denotes the G_Δ matrix of observations on the endogenous variables included in the structural equation being estimated.

The ordinary least squares estimators (OLS) of $\beta_{1\cdot}$ and $\gamma_{1\cdot}$ in (5.1) minimize (5.4) and are given by

$$\begin{bmatrix} \beta_1' \\ \gamma_1' \end{bmatrix}_{OLS} = \left[(Y_1 X_1)' (Y_1 X_1) \right]^{-1} (Y_1 X_1)' y_1$$

$$= \begin{bmatrix} Y_1' Y_1 & Y_1' X_1 \\ X_1' Y_1 & X_1' X_1 \end{bmatrix} \begin{bmatrix} Y_1' y_1 \\ X_1' y_1 \end{bmatrix}$$

The OLS structural coefficient estimators are biased and inconsistent. Note that only observations on the variables appearing in the structural equation being estimated are required to obtain OLS estimators.

STATA commands—you usually don't want to estimate a structural equation with OLS
reg y1 Y1 X1

(2) Indirect Least Squares (ILS)–included for pedagogical purposes only .

For structural equations which are exactly identified we can obtain consistent estimators (biased) by using the technique of ILS which can be thought of as consisting of two steps

- (a) Obtain the least squares No restrictions estimates (LSNR) of Π in $\mathbf{Y} = \mathbf{X}\Pi' + \mathbf{V}$ $\hat{\Pi}$, say .
- (b) Solve $\hat{\Pi}\mathbf{B} + \Gamma = 0$ for the corresponding estimates of the structural parameters in the exactly identified structural equations. These estimators are referred to as the ILS estimates of B and Γ .

Example: Consider the simple macro structural model defined by

$$\begin{aligned} C_t &= \alpha + \beta Y_t + \varepsilon_t \\ Y_t &= C_t + Z_t \end{aligned}$$

The corresponding reduced form representation is given by

$$\begin{aligned} C_t &= \frac{\alpha}{1-\beta} + \frac{\beta}{1-\beta} Z_t + \frac{\gamma_t}{1-\beta} = \pi_{11} + \pi_{12} Z_t + v_{t1} \\ Y_t &= \frac{\alpha}{1-\beta} + \frac{\beta}{1-\beta} Z_t + \frac{\gamma_t}{1-\beta} = \pi_{11} + \pi_{12} Z_t + v_{t1} \end{aligned}$$

The consumption function is exactly identified. Using Haavelmo's data (1947, JASA, pp. 105-122) we obtain the LSNR estimates of Π :

$$\hat{\Pi} = \begin{bmatrix} \hat{\Pi}_{11} & \hat{\Pi}_{12} \\ \hat{\Pi}_{21} & \hat{\Pi}_{22} \end{bmatrix} = \begin{bmatrix} 344.70 & 2.048 \\ 344.70 & 3.048 \end{bmatrix}$$

and since $\beta = \frac{\Pi_{12}}{\Pi_{22}}$,

we obtain $\hat{\beta} = \frac{\hat{\Pi}_{12}}{\hat{\Pi}_{22}} = \frac{2.048}{3.048} = .672$ as the ILS of β .

Recall that the OLS estimator of β is .732.

(3) Two Stage Least Squares (2SLS). For overidentified structural equations the

technique of ILS is not applicable and the technique of OLS yields inconsistent estimators. Perhaps the most commonly used technique of consistent estimation in such cases is that of 2SLS. For the case of an exactly identified structural equation the 2SLS estimator is equal to the ILS estimator; hence, ILS need never be performed.

2SLS estimates can readily be obtained using STATA with the command

STATA commands

```
ivreg y1 (Y1=X1 X2) X1, options or
ivreg y1 (Y1=X2) X1 or
ivregress 2sls (Y1=X1 X2) X1 or
ivregress 2sls y1 (Y1=X2) X1
```

2SLS estimation was independently developed by H. Theil and R.L. Basmann. 2SLS estimators are asymptotically normally distributed.

(a) Theil's Development of 2SLS.

Theil's approach to two stage least squares involves running two separate regressions; hence, the name two stage least squares. Denote the reduced form representation corresponding to Y_1 (The dependent variables on the right side of the structural equation of interest) by

$$Y_1 = X\Pi_1' + V_1 \quad (NxG_\Delta - 1)$$

where Π_1' is $Kx(G_\Delta - 1)$.

The LSNR estimator of Π_1 is given by

$$\hat{\Pi}_1 = (X'X)^{-1}X'Y_1.$$

The 2SLS least squares estimators of $\beta_{1.}$ and $\gamma_{1.}$ in

$$y_{1.} = Y_1\beta_{1.}' + X_1\gamma_{1.}' + \epsilon_{1.}$$

can be thought of as being obtained by performing the following two-step process:

(1) replace Y_1 in (1.1)' by its least squares estimate

$$\hat{Y}_1 = X\hat{\Pi}_1 \quad Y_1 = X\hat{\Pi}_1 + \hat{V}_1 \quad)$$

$$Y_{1.} = \hat{Y}_1\beta'_{1.} + X_1\gamma'_{1.} + \varepsilon^*_{.1} \quad \varepsilon^* = \varepsilon_{.1} + \hat{V}_1\beta_{1.} \quad \text{where}$$

$$= (\hat{Y}_1, X_1) \begin{pmatrix} \beta'_{1.} \\ \gamma'_{1.} \end{pmatrix} + \varepsilon^*_{.1}$$

and then

(2) apply least squares to this result to yield, i.e.,

$$\begin{bmatrix} \beta'_{1.} \\ \gamma'_{1.} \end{bmatrix}_{2SLS} = ([\hat{Y}_1, X_1]' [\hat{Y}_1, X_1])^{-1} [\hat{Y}_1, X_1]' y_{.1}$$

$$= \begin{bmatrix} \hat{Y}_1 \hat{Y}_1 & \hat{Y}_1 X_1 \\ X_1 \hat{Y}_1 & X_1 X_1 \end{bmatrix}^{-1} \begin{bmatrix} \hat{Y}_1 & y_{.1} \\ X_1 & y_{.1} \end{bmatrix} \quad (5.6)$$

Compare this formula to that obtained for OLS, (5.5).

Exercises: Demonstrate that

$$\begin{bmatrix} \beta'_{1.} \\ \gamma'_{1.} \end{bmatrix}_{2SLS} = \begin{bmatrix} Y_1 Y_1 & -\hat{V}_1 \hat{V}_1 & Y_1 X_1 \\ X_1 Y_1 & & X_1 X_1 \end{bmatrix}^{-1} \begin{bmatrix} (Y_1 - \hat{V}_1) y_{.1} \\ X_1 y_{.1} \end{bmatrix}$$

$$= \begin{bmatrix} Y_1 X (X' X)^{-1} & X' Y_1 & Y_1 X_1 \\ X_1 Y_1 & & X_1 X_1 \end{bmatrix}^{-1} \begin{bmatrix} Y_1 X (X' X)^{-1} X' y_{.1} \\ X_1 y_{.1} \end{bmatrix}$$

Hint: $(\hat{Y}_1 = Y_1 - \hat{V}_1 = X\hat{\Pi}_1 = X(X'X)^{-1}X'Y_1)$
 $X_1 \hat{Y}_1 = (X_1 X (X' X)^{-1} X' Y_1) = X_1 Y_1$

Comment:

The second expression for $\begin{bmatrix} \beta_1 \\ \gamma_1 \end{bmatrix}_{2SLS}$ can be applied directly and doesn't require a two-stage process. It is this approach that is generally used in computer programs.

(b) Basmann's Development

Recall that the sum of squared errors associated with (5.1), can be written as

$$\begin{aligned} G_1(\beta_{1.}) &= (y_{1.} - Y_1 \beta_{1.} - X_1 \gamma_{1.})' (y_{1.} - Y_1 \beta_{1.} - X_1 \gamma_{1.}) \\ &= (-1, \beta_{1.}) Y_\Delta \left(I - X_1 (X_1' X_1)^{-1} X_1' \right) Y_\Delta \begin{pmatrix} -1 \\ \beta_{1.} \end{pmatrix} \end{aligned}$$

OLS estimates of $\beta_{1.}$ can be shown to be equivalent to minimizing $G_1(\beta_{1.})$ over $\beta_{1.}$ and using

$$\hat{\gamma}'_{1.} = \hat{\Pi}_{1\Delta} \begin{pmatrix} -1 \\ \beta'_{1.} \end{pmatrix}_{OLS}$$

to estimate $\gamma_{1.}$. Now consider

$$G_2(\beta_{1.}) = (-1, \beta_{1.}) Y_\Delta' (I - X(X'X)^{-1}X') Y_\Delta \begin{pmatrix} -1 \\ \beta'_{1.} \end{pmatrix} \quad (5.7)$$

$G_2(\beta_{1.})$ can be thought of as the sum of squared errors associated with

$$y_{1.} = Y_1 \beta_{1.} + X_1 \gamma'_{1.} + X_2 \gamma'_{2.} + \epsilon_{1.} \quad (5.8)$$

where all of the explanatory variables in the structural model are included. As discussed earlier, we hypothesize that, based on economic theory, $\gamma_{2.} = 0$ and this provides the basis for "identifying" the structural equation under consideration.

If the hypothesis $\gamma_{2.} = 0$ is "true," we might expect that the sum of squared errors associated with (5.1) and (5.8) would be very similar. Equivalently, the

reduction in sum of squared error associated with adding "X₂" to the structural equation (5.1) would be small.

It can be shown that 2SLS estimators of $\beta_{1.}$ and $\gamma_{1.}$ are defined by

$$\boxed{\min_{\beta_{1.}} [G_1(\beta_{1.}) - G_2(\beta_{1.})]} \quad (5.9)$$

$$\text{and } (\hat{\gamma}_{1.})_{\text{2SLS}} = \hat{\pi}_{1A} \begin{pmatrix} -1 \\ \beta'_{1.} \end{pmatrix}_{\text{2SLS}} \quad (5.10)$$

Two stage least squares are constructed so as to minimize the reduction in the sum of squared errors which would result from including the variables which are hypothesized to not be included in the structural equation of interest, kind of like minimizing the Chow test associated with the hypothesis $H_o: \gamma_2 = 0$.

$G_1(\beta_{1.})$ and $G_2(\beta_{1.})$ are sometimes used to define estimators of the variance of ε_{r1} . For this reason 2SLS estimators are also referred to as least variance difference estimators. Note that 2SLS requires observations on all of the independent variables in the structural model and the dependent variables included in the structural equation of interest. OLS only requires observations on the independent variables of the structural equation being estimated.

(c) $E(\beta_{2SLS}^k)$ is only valid for $\nu \leq h$. Hence, the 2SLS estimator does not have a finite mean in an exactly identified ($\nu = 0$) structural equation. The bias, where defined, is asymptotically proportional to $\left(\frac{\nu-1}{\mu^2}\right)$. Here μ^2 denotes the concentration parameter defined earlier. Thus, the bias of 2SLS can be substantial if $\left(\frac{\nu-1}{\mu^2}\right)$ is large. An unbiased estimator of the concentration parameter is given by $\hat{\mu}^2 = K_2(\tilde{F}-1)$, where \tilde{F} is the F-statistic associated with testing the hypothesis that the coefficients of the instruments (X_2) in the estimated reduced form for Y_1 are equal to zero. Thus weak instruments (small F-stat and concentration parameter) can be associated with significant bias.

(d) Recall that in a structural model with one endogenous regressor that the bias of the 2SLS estimator will be less than 10 percent of the bias of the OLS estimator if $10 < \tilde{F}$

(e) Distribution of Two-Stage Least Squares Estimator.

The exact (finite) sample distribution of

$$\begin{pmatrix} \hat{\beta}_{1\cdot} \\ \hat{\gamma}_{1\cdot} \end{pmatrix}_{2SLS}$$

is not normal. In fact, the exact distributions of the 2SLS estimators were not derived until the 1970's. The distribution function involves multiple infinite series. The main findings from these studies will be summarized in another section. However, it is known that the distribution function of the 2SLS structural coefficient estimators approaches a normal distribution function as N grows indefinitely large with mean $(\beta_{1\cdot}, \gamma_{1\cdot})'$ and covariance matrix (Σ_{2SLS}) .

$$\left(\begin{array}{c} \beta'_{1\cdot} \\ \gamma'_{1\cdot} \end{array} \right)_{2SLS} \stackrel{a}{\sim} N \left[\left(\begin{array}{c} \beta'_{1\cdot} \\ \gamma'_{1\cdot} \end{array} \right); \Sigma_{2SLS} = \left[\frac{\omega_{11}^2}{N} \right] \left(\text{plim}_{N \rightarrow \infty} \left\{ \frac{1}{N} \begin{bmatrix} \hat{Y}'_1 \hat{Y}_1 & Y'_1 X_1 \\ X'_1 Y_1 & X'_1 X_1 \end{bmatrix} \right\}^{-1} \right) \right]$$

The covariance matrix of

$$\left(\begin{array}{c} \hat{\beta}'_{1\cdot} \\ \hat{\gamma}'_{1\cdot} \end{array} \right)_{2SLS} \quad \text{in the structural model}$$

$$y_1 = [Y_1, X_1] \begin{bmatrix} \beta'_{1\cdot} \\ \gamma'_{1\cdot} \end{bmatrix} + \varepsilon_1$$

$$= [\hat{Y}_1, X_1] \begin{bmatrix} \beta'_{1\cdot} \\ \hat{\gamma}'_{1\cdot} \end{bmatrix} + \varepsilon^*_{1\cdot} = Z \begin{bmatrix} \hat{\beta}'_{1\cdot} \\ \hat{\gamma}'_{1\cdot} \end{bmatrix} + \varepsilon'_{1\cdot}$$

is generally estimated by

$$\hat{\omega}_{11}(Z'Z)^{-1} = \hat{\omega}_{11} \begin{bmatrix} \hat{Y}'_1 \hat{Y}_1 & Y'_1 X_1 \\ X'_1 \hat{Y}_1 & X'_1 X_1 \end{bmatrix}^{-1}$$

where

$$\hat{\omega}_{11} = \left(\frac{1}{N - K + v} \right) \sum_{t=1}^n (\epsilon_t^2)$$

$$= \left(\frac{1}{N - K + v} \right) G_1(\hat{\beta}_1)$$

is used as an estimator of $\text{var}(\varepsilon_{11}) = \omega_{11}$. This is the formula used in most computer programs. Two other estimators of ω_{11} have been

considered:

$$\frac{G_2(\hat{\beta}_{1.})}{N - K + G_{\Delta} + 1} \quad \frac{G_1(\hat{\beta}_{1.}) - G_2(\hat{\beta}_{1.})}{v}$$

Dhrymes suggested the second estimator. The choice of the estimator of ω_{11} impacts the degrees of freedom used in performing t-tests. For example, in using Dhrymes' suggested estimator, one should use a t-table with v degrees of freedom.

(4) Limited Information Maximum Likelihood (LIML)

- (a) Development from a likelihood function
First we note that

$$y_{1.} = Y_1 \beta_{1.} + X_1 \gamma_{1.} + \varepsilon_{1.}$$

$$Y_{\Delta} = (y_{1.} Y_1) = X \Pi_{\Delta} + (v_{1.} V_1).$$

Consider the solution to the constrained optimization problem:

Maximize the likelihood function of $(y_{1.} Y_1)$ over the parameters $(\beta_{1.}, \gamma_{1.})$

Subject to the restrictions: $\text{rank } (\Pi_{\Delta 2}) = G_{\Delta} = -1$

The solution to this optimization problem yields the LIML estimates of $\beta_{1.}$ and $\gamma_{1.}$

- (b) Development of the LIML estimators as a least variance ratio (LVR) estimator.

$$\text{Define } W_{\Delta\Delta} = Y_{\Delta}'(I - X(X'X)^{-1}X')Y_{\Delta}$$

$$W^*_{\Delta\Delta} = Y_{\Delta}'(I - X_1(X_1'X_1)^{-1}X_1')Y_{\Delta}.$$

As noted previously, estimators of the variance $\omega_{11} = \text{var}(\varepsilon_{11})$ can be defined in terms of

the quadratic forms $G_1(\beta_{1.}) = (-1, \beta_{1.}) (W^*_{\Delta\Delta}) (-1, \beta_{1.})'$

$$G_2(\beta_1)$$

The LIML (LVR) estimator of $\beta_{1.}$, $\tilde{\beta}_{1.}$, can be obtained from

$$\underset{\beta_{1.}}{\text{minimize}} \left[\frac{G_1(\beta_{1.}) - G_2(\beta_{1.})}{G_2(\beta_{1.})} \right]$$

i.e., minimize

$$\frac{b'W_{\Delta\Delta}^* b}{b'W_{\Delta\Delta} b} \text{ s.t. } b'W_{\Delta\Delta} b = 1.$$

Solution:

The corresponding Lagrangian function is defined by

$$L(b, \lambda) = b'W_{\Delta\Delta}^* b + \lambda(1 - b'W_{\Delta\Delta} b)$$

$$\frac{\partial L}{\partial b} = 2(W_{\Delta\Delta}^* - \lambda W_{\Delta\Delta})b = 0.$$

In order to obtain a nontrivial solution for b we require the determinant

$$|W_{\Delta\Delta}^* - \lambda W_{\Delta\Delta}| = 0.$$

Premultiplying the first order condition by b' we note that

$$\hat{\lambda} = \frac{G_1(\tilde{\beta}_{1.})}{G_2(\tilde{\beta}_{1.})} - 1.$$

The LIML (LVR) estimator of $\beta_{1.}$ corresponds to the characteristic vector, normalized on $y_{1.}$, which is associated with the smallest characteristic root of $W_{\Delta\Delta}^$ in the metric $W_{\Delta\Delta}$ (where $|W_{\Delta\Delta}^* - \lambda W_{\Delta\Delta}| = 0$).

STATA Version 12 includes LIML capability using the command:

```
ivregress liml y1 (Y1=X1 X2) X1 or
ivregress liml y1 (Y1=X2) X1
```

Note:

$$(a) \tilde{F} = \left(\frac{N-K}{v} \right) \left(\frac{G_1(\tilde{\beta}_1) - G_2(\tilde{\beta}_1)}{G_2(\tilde{\beta}_1)} \right) \quad (5.11)$$

can be used to perform a statistical test of the necessary conditions for identification. F is approximately distributed $F(v, N-K)$. Note that this statistic is similar in form to the statistic associated with the Chow test.

If the command **estat overid** follows an **ivregress** command, then tests of overidentification are performed. If 2sls is used, Basmann's chi-square test is reported as is Woodridge's robust score test. If LIML is used, Anderson and Rubin's chi square test is reported as is Basmann's F test (above). If GMM is used, Hansen's J statistic and chi square test are given.

Statistically significant tests indicate that the exclusion restrictions are not valid, instruments are not valid.

(b) $\tilde{\beta}_1$ and $\tilde{\gamma}_1$ are consistent es β_1 and γ_1 but are bi

(c) $\text{Plim}_{N \rightarrow \infty} (\min \hat{\lambda}) = 1$

(d) $E(\hat{\beta}_{LML})$ is not defined for a fixed normalization (a particular endogenous y appears on the left hand side of the structural equation); hence, the associated pdf has thick tails, Mariano, R. and T. Sawa (1972, The exact finite sample distribution of the LIML estimator in the case of two included endogenous variables, JASA, 67, 159-163). Anderson (2010, The LIML estimator has finite moments!, Journal of Econometrics, 157, 359-361) demonstrates that LIML moments exist if the normalization is $b' \Phi b = 1$.

(e) The interquartile range of LIML can be considerably larger than that for 2SLS because of having thicker tails; however, the pdf of the LIML estimtor is often better "centered" than for 2SLS, particularly for large

values of $\left(\frac{\nu-1}{\mu^2} \right)$, weak instruments and large number of overidentifying

restrictions. There are tradeoffs between bias and variance (MSE).

Hansen, Heaton, and Yaron (1996, Finite sample properties of some alternative GMM estimators, *Journal of Business and Economic Statistics*, 14(3), 262-280) use a continuous updating (CUE) of a GMM-like

generalization of LIML to address this problem. Hausmann, Menzel, Lewis, and Newey (2007, A reduced bias GMM-like estimator with reduced estimator dispersion, MIT manuscript) modify the CUE to solve the no moments/large dispersion problem.

(f) Anderson, Kunitomo, and Matsushita (2010, “On the asymptotic optimality of the LIML estimator with possibly many instruments,” *Journal of Econometrics*, 157, 191-204) remind us that while 2SLS and LIML are asymptotically equivalent with large sample sizes, they are quite different with large K_2 (Many instruments) and argue that LIML may be an attractive option over the semiparametric methods of GMM and EL in situations with many instruments or many weak instruments.

(5) General k-Class Estimators.

The k-class estimators of $\tilde{\beta}_1$, and $\tilde{\gamma}_1$ are defined by

$$\begin{bmatrix} \beta'_{1.} \\ \gamma'_{1.} \end{bmatrix}_k = \begin{bmatrix} Y_1' Y_1 & -k \hat{V}_1' \hat{V}_1 & Y_1' X_1 \\ X_1' Y_1 & X_1' X_1 \end{bmatrix}^{-1} \begin{bmatrix} (Y_1' - k \hat{V}_1') y_{1.} \\ X_1' Y_1 \end{bmatrix}$$

where

$$\hat{V}_1 = Y_1 - X \hat{\Pi}_1 = Y_1 - X(X'X)^{-1}X'Y_1.$$

The k-class estimators can be obtained by

$$\min_{\beta_{1.}} [G_1(\beta_{1.}) - k G_2(\beta_{1.})]$$

with

$$(\hat{\gamma}_1)_k = \hat{\pi}'_{1A} \begin{bmatrix} -1 \\ \beta'_{1.} \end{bmatrix}, \text{ McDonald [1977, } \underline{\text{Econometrica}} \text{]}$$

It should be noted that $k = 0, 1$, respectively, yield OLS and 2SLS estimators. If k is selected to be λ from the previous section, then LIML estimates are obtained.

The k-class estimator of $\beta_{1.}$ and $\gamma_{1.}$ will be consistent if and only if $\text{plim}_{k=1}$. Note that it then follows that 2SLS and LIML estimates will be consistent; whereas, OLS estimates are inconsistent.

Stata will perform k-class estimators with the command
ivreg2 y1 (Y1=X1 X2 or X2) X1, kclass(#) fuller(#) liml

Exercises and Notes:

1. That if we select $Z = [Y_1 - k \hat{V}_1, X_1]$, then the associated instrumental variables estimator corresponding to Z is the k-class estimator.

$$\text{Hint: Solve } Z'y_{.1} = Z'[Y_1, X_1] \begin{bmatrix} \beta'_{1.} \\ \gamma'_{1.} \end{bmatrix}_k$$

for $(\beta_{1.}, \gamma_{1.})'$ and manipulate the derived expression to obtain the desired result.

2. Nagar demonstrates that if we select k to be $1 + \frac{v-1}{N}$, the k -class

estimator is "almost" unbiased. Demonstrate that this estimator is consistent. It can be shown that the expected value of k -class estimators (integer moments) are not defined if $k > 1$.

3. Zellner (Journal of Econometrics), Estimation of Functions of Population Means and Regression Coefficients Including Structural Coefficients (MELO), 8(1978), 127-158 (esp. p. 141).

Selecting k to be

$$k = 1 - \frac{K}{N-K-G_\Delta+1}$$

yields the minimum expected loss Bayesian estimator when based on a diffuse prior on (Π, Σ) . The random disturbances in the reduced form are assumed to be normally distributed. Note that this estimator is consistent ($\text{plim } k = 1$).

4. The k -class estimators ($0 \leq k \leq 1$) of $\beta_{1.}$ and in (1.1) can also be obtained using least squares algorithms by regressing

$$y_{.1} - a\hat{V}_1 \text{ on } (\hat{Y}_1 - a\hat{V}_1) \text{ and } X_1$$

where $a = 1 - \sqrt{1-k}$, $\hat{v}_1 = y_{.1} - \hat{y}_{.1}$ and

$$\hat{V}_1 = Y_1 - \hat{Y}_1$$

Thus least squares computer programs can be used to obtain k -class estimators, McDonald and Maynes (1980, Journal of Statistical Computation and Simulation).

(6) Sawa's combined estimator, Journal of Econometrics (1973):

$$\begin{pmatrix} \beta'_{1\cdot} \\ \gamma'_{1\cdot} \end{pmatrix}_{\text{comb}} = \left(1 + \frac{v-1}{N-K} \right) \begin{pmatrix} \beta'_{1\cdot} \\ \gamma'_{1\cdot} \end{pmatrix}_{\text{2SLS}} - \left(\frac{v-1}{N-K} \right) \begin{pmatrix} \beta'_{1\cdot} \\ \gamma'_{1\cdot} \end{pmatrix}_{\text{OLS}}$$

This estimator is "almost" unbiased and consistent. For equations having a single overidentifying restriction ($v = 1$), Sawa's combined estimator is merely the two stage least squares estimator.

(7) GMM and Instrumental variables have been discussed earlier and can be formulated to include the previous estimators as special cases. Depending on our schedule, we may spend more time on them. As noted earlier , there are many variations of GMM. The basic GMM form is to minimize a quadratic form

$$(\varepsilon' Z) \hat{Q}^{-1} (Z' \varepsilon) \quad \text{over the unknown } \hat{Q} \text{ parameters where } \hat{Q} \text{ denotes an estimate}$$

weighting matrix. The optimal weighting matrix is $\hat{Q} = \text{Var}(Z' \varepsilon)$. If \hat{Q} is ad

with each iteration, the estimators are referred to as continuous updating estimators (CUE) and reduce the bias of GMM. Hansen, McDonald, and Newey (2010) consider a variation of the IV (GMM) estimators previously described where the objective function is $(\rho(\varepsilon)' Z) \hat{Q}^{-1} (Z' \rho(\varepsilon))$

$\hat{Q} = \text{Var}(Z' \rho(\varepsilon))$ where $\rho(\varepsilon)$ is a vector of the derivatives of the log pdf evaluated at the vector of the estimated disturbances. If the errors are normal then this nonlinear general instrumental variables (NLIV) estimator includes regular GMM as a special case and has the potential to provide improved estimators for non normal errors distributions.

c. Simultaneous Equation Estimation Methods: FIML and 3SLS

- (1) Full Information Maximum Likelihood (FIML) Estimation.

$$\mathbf{BY}_t + \boldsymbol{\Gamma}\mathbf{X}_t + \boldsymbol{\varepsilon}_t = 0$$

$$\boldsymbol{\varepsilon}_t \sim N(\mathbf{0}, \boldsymbol{\Omega}).$$

The associated likelihood function is given by

$$L(\mathbf{Y}; \mathbf{B}, \boldsymbol{\Gamma}, \boldsymbol{\Omega}) = \frac{e^{-\frac{1}{2} \sum_t (\mathbf{BY}_t + \boldsymbol{\Gamma}\mathbf{X}_t)' \boldsymbol{\Omega}^{-1} (\mathbf{BY}_t + \boldsymbol{\Gamma}\mathbf{X}_t)}}{(2\pi)^{-\frac{N_G}{2}} |\boldsymbol{\Omega}|^{\frac{N}{2}} |\mathbf{B}|^N}$$

Solution: Maximize $L(\cdot)$, or $\ell(\cdot) = \ln L(\cdot)$ with respect to \mathbf{B} , $\boldsymbol{\Gamma}$, $\boldsymbol{\Omega}$ subject to any restrictions on the parameters, i.e., solve

$$\frac{\partial L}{\partial \mathbf{B}} = 0$$

$$\frac{\partial L}{\partial \boldsymbol{\Gamma}} = 0$$

$$\frac{\partial L}{\partial \boldsymbol{\Omega}} = 0$$

The first order conditions are a system of nonlinear equations in the unknown structural parameters. These estimators are asymptotically normal and efficient under quite general conditions.

The 2011 release of Stata allows FIML estimation of some structural models.

(2) Three-Stage Least Squares. References (3SLS)

Let the h^{th} structural equation in the system of structural equations

$$\mathbf{YB}' + \mathbf{XT}' + \boldsymbol{\varepsilon} = 0$$

be denoted

$$\mathbf{y}_{\cdot h} = \mathbf{Y}_1^{(h)} \boldsymbol{\beta}'_{\cdot h} + \mathbf{X}_1^{(h)} \boldsymbol{\gamma}'_{\cdot h} = \boldsymbol{\varepsilon}_{\cdot h} \quad (5.13)$$

where $\mathbf{Y}_1^{(h)}$ denotes the endogenous variables included on r.h.s. of h^{th} structural equation and $\mathbf{X}_1^{(h)}$ denotes the exogenous variables included in h^{th} structural equation

Most econometric packages can perform 3SLS estimation. The format for STATA 3SLS estimation is:

...

STATA Commands

`reg3 (depvar1 rhs_varlist1) (depvar2 rhs_varlist2) ... (depvarG rhs_varlistG), endog(list of endogenous variables)`

`reg3 (depvar1 rhs_varlist1) (depvar2 rhs_varlist2) ... (depvarG rhs_varlistG), endog(list of endogenous variables) ireg3 "iterates until estimates converge"`

2SLS adjusts for endogenous regressors, but does not take account of possible correlation between the error terms in the different equations. 3SLS takes account of both the endogeneity problem and possible correlation between the structural random disturbances. Hence, the 3SLS estimators are asymptotically efficient.

Development of 3SLS estimators: estimation equations and motivation.

Equation (5.13) can be rewritten as

$$\mathbf{y}_{\cdot h} = [\mathbf{Y}_1^{(h)} \mathbf{X}_1^{(h)}] \begin{bmatrix} \boldsymbol{\beta}'_{\cdot h} \\ \boldsymbol{\gamma}'_{\cdot h} \end{bmatrix} + \boldsymbol{\varepsilon}_{\cdot h} \quad (5.13)' \quad \text{or}$$

$$\mathbf{y}_{\cdot h} = \mathbf{Z}_{\cdot h} \boldsymbol{\delta}_{\cdot h} + \boldsymbol{\varepsilon}_{\cdot h} \quad h = 1, 2, \dots, G \quad \text{where } \boldsymbol{\varepsilon}_{\cdot h} \sim N[0, \omega_{\cdot h}] \quad (5.13)''$$

The G -equations given in (5.13)'' can be rewritten in matrix form as

(5.14)

$$\mathbf{y}^* = \begin{bmatrix} \mathbf{y}_{.1} \\ \mathbf{y}_{.2} \\ \vdots \\ \mathbf{y}_{.G} \end{bmatrix} = \begin{bmatrix} Z_1 & 0 & 0 & \dots & 0 \\ 0 & Z_2 & 0 & \dots & 0 \\ \vdots & & & & \\ 0 & 0 & 0 & \dots & Z_G \end{bmatrix} \begin{bmatrix} \boldsymbol{\delta}_1 \\ \boldsymbol{\delta}_2 \\ \vdots \\ \boldsymbol{\delta}_G \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_{.1} \\ \boldsymbol{\varepsilon}_{.2} \\ \vdots \\ \boldsymbol{\varepsilon}_{.G} \end{bmatrix}$$

$$= Z^* \boldsymbol{\delta}^* + \boldsymbol{\varepsilon}^* \quad \text{or} \quad (5.14)'$$

$$\boxed{\mathbf{y}^* = Z^* \boldsymbol{\delta}^* + \boldsymbol{\varepsilon}^*}$$

where

$$\boxed{\boldsymbol{\varepsilon}^* \sim N[0; \Omega \otimes I],}$$

$$\text{var}(\boldsymbol{\varepsilon}^*) = \begin{bmatrix} \text{var}(\boldsymbol{\varepsilon}_{.1}) & \text{cov}(\boldsymbol{\varepsilon}_{.1}, \boldsymbol{\varepsilon}_{.2}) & \dots & \text{cov}(\boldsymbol{\varepsilon}_{.1}, \boldsymbol{\varepsilon}_{.G}) \\ & \text{var}(\boldsymbol{\varepsilon}_{.2}) & \dots & \\ \vdots & \vdots & \vdots & \vdots \\ \text{cov}(\boldsymbol{\varepsilon}_{.G}, \boldsymbol{\varepsilon}_{.1}) & \dots & & \text{var}(\boldsymbol{\varepsilon}_{.G}) \end{bmatrix}$$

$$= \begin{bmatrix} \omega_{11}I & \omega_{12}I & \dots & \omega_{1G}I \\ \omega_{21}I & \omega_{22}I & \dots & \omega_{2G}I \\ \vdots & & & \\ \omega_{G1}I & \dots & & \omega_{G}I \end{bmatrix}$$

$$= \Omega \otimes I.$$

Note:

$$(a) \quad Z^* = \begin{bmatrix} Z_1 & 0 & 0 & \dots & 0 \\ 0 & Z_2 & 0 & \dots & 0 \\ \vdots & \vdots & & & \vdots \\ 0 & 0 & 0 & \dots & Z_G \end{bmatrix}$$

(b)

$$Z^* = \begin{bmatrix} Y_1^{(1)} & X_1^{(1)} & & 0 & 0 & \dots & 0 \\ 0 & & Y_1^{(2)} & X_1^{(2)} & 0 & \dots & 0 \\ \vdots & & & & & & \\ 0 & & \dots & 0 & \dots & Y_1^{(G)} & X_1^{(G)} \end{bmatrix} \quad \text{and}$$

(c)

$$\delta^* = \begin{bmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_G \end{bmatrix} = \begin{bmatrix} \beta'_{1\cdot} \\ \gamma'_{1\cdot} \\ \beta'_{2\cdot} \\ \gamma'_{2\cdot} \\ \vdots \\ \beta'_{G\cdot} \\ \gamma'_{G\cdot} \end{bmatrix}$$

δ^* is a column vector containing all structural coefficients in the entire model. The representation in (5.14) can be used to represent OLS, 2SLS, as well as 3SLS estimators.

The **OLS structural coefficient estimators** of δ in (5.14) can collectively be represented as

(5.15)

$$\boldsymbol{\delta}_{\text{OLS}} = \begin{bmatrix} \delta_{1(\text{OLS})} \\ \vdots \\ \delta_{G(\text{OLS})} \end{bmatrix} = (\mathbf{Z}^* \mathbf{Z}^*)^{-1} \mathbf{Z}^* \mathbf{y}^*$$

which can be shown to reduce to (5.2)

$$\begin{aligned} (\delta_h)_{\text{OLS}} &= \begin{bmatrix} \beta'_h \\ \gamma'_h \end{bmatrix}_{\text{OLS}} = (\mathbf{Z}'_h \mathbf{Z}_h)^{-1} \mathbf{X}'_h \mathbf{y}_h \\ &= \begin{bmatrix} \mathbf{Y}_1^{(h)} \mathbf{Y}_1^{(h)} & \mathbf{Y}_1^{(h)} \mathbf{X}_1^{(h)} \\ \mathbf{X}_1^{(h)} \mathbf{Y}_1^{(h)} & \mathbf{X}_1^{(h)} \mathbf{X}_1^{(h)} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{Y}_1^{(h)} \mathbf{y}_h \\ \mathbf{X}_1^{(h)} \mathbf{y}_h \end{bmatrix} \end{aligned} \quad (5.15)'$$

Recall that the OLS estimators are biased and inconsistent because the explanatory variables (right-hand side variables) are correlated with the error terms.

The **two-stage least squares estimators are consistent**. They can be thought of as being obtained by replacing the right-hand side endogenous variables by their LSNR predicted values and then applying least squares, i.e.,

(5.16)

$$\hat{\boldsymbol{\delta}}_{\text{2SLS}} = (\hat{\mathbf{Z}}^* \hat{\mathbf{Z}}^*)^{-1} \hat{\mathbf{Z}}^* \mathbf{y}^*$$

where

$$\hat{\mathbf{Z}}^* = \begin{bmatrix} \hat{\mathbf{Z}}_1 & 0 & 0 & \dots & 0 \\ 0 & \hat{\mathbf{Z}}_2 & 0 & \dots & 0 \\ \vdots & \vdots & & & \vdots \\ 0 & 0 & 0 & \dots & \hat{\mathbf{Z}}_G \end{bmatrix}$$

$$= \begin{bmatrix} \hat{Y}_1^{(1)} X_1^{(1)} & 0 & 0 & \dots & 0 \\ 0 & \hat{Y}_1^{(2)} X_1^{(2)} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \vdots \\ 0 & \dots & \dots & \dots & \hat{Y}_1^{(G)} X_1^{(G)} \end{bmatrix};$$

hence,

$$\begin{aligned} (\hat{\delta}_h)_{2SLS} &= \begin{bmatrix} \beta'_h \\ \gamma'_h \end{bmatrix}_{OLS} = (\hat{Z}'_h \hat{Z}_h)^{-1} \hat{Z}'_h y_h \\ &= \begin{bmatrix} \hat{Y}_1^{(h)} \hat{Y}_1^{(h)} & \hat{Y}_1^{(h)} X_1^{(h)} \\ X_1^{(h)} \hat{Y}_1^{(h)} & X_1^{(h)} X_1^{(h)} \end{bmatrix}^{-1} \begin{bmatrix} \hat{Y}_1^{(h)} y_h \\ X_1^{(h)} y_h \end{bmatrix} \end{aligned} \quad (5.16)'$$

which is in the same form as in (5.2).

The 2SLS estimator of δ^* can also be expressed as

$$\delta^*_{2SLS} = [Z^* X^* (X^* X^*)^{-1} X^* Z^*]^{-1} Z^* X^* (X^* X^*)^{-1} X^* y^*$$

since

$$\hat{Z}^* = X^* (X^* X^*)^{-1} X^* Z^*$$

$$\text{where } X^* = \begin{bmatrix} X & 0 & 0 & \dots & 0 \\ 0 & X & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & X \end{bmatrix} = I \otimes X.$$

The two-stage least squares estimators make corrections for the correlation between the right-hand side dependent variable and the structural error term, but do not take account of the covariances between the error terms in different equations.

Three-stage least squares makes adjustments for:

- (1) the correlation between the right-hand side endogenous variables and the error terms, and
- (2) correlations between the error terms in different structural equations. This is done by multiplying (5.14) by $\mathbf{X}^* = \mathbf{I} \otimes \mathbf{X}'$:

$$\boxed{\mathbf{X}^* \mathbf{y}^* = \mathbf{X}^* \mathbf{Z}^* \boldsymbol{\delta}^* + \mathbf{X}^* \boldsymbol{\epsilon}^*} \quad (5.17)$$

where $\mathbf{X}^* \boldsymbol{\epsilon}^* \sim N[0; \Omega \otimes \mathbf{X}' \mathbf{X}]$

since $\mathbf{X}^* (\Omega \otimes \mathbf{I}) \mathbf{X}^* = \Omega \otimes \mathbf{X}' \mathbf{X}$.

Note that the 2SLS estimator of $\boldsymbol{\delta}^*$ can be obtained by applying least squares to (5.17), three-stage least squares estimators can be obtained by applying a generalized least squares formula to (5.17) to yield

$$\begin{aligned} (\boldsymbol{\delta}^*)_{3SLS} &= [(\mathbf{X}^* \mathbf{Z}^*)' (\mathbf{X}^* (\Omega \otimes \mathbf{I}) \mathbf{X}^*)^{-1} \mathbf{X}^* \mathbf{Z}^*]^{-1} (\mathbf{X}^* \mathbf{Z}^*)' (\mathbf{X}^* (\Omega \otimes \mathbf{I}) \mathbf{X}^*)^{-1} \mathbf{X}^* \mathbf{y}^* \\ &= [\mathbf{Z}^* (\Omega^{-1} \otimes \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}') \mathbf{Z}^*]^{-1} \mathbf{Z}^* (\Omega^{-1} \otimes \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}) \mathbf{y}^* \end{aligned}$$

Hint: Recall properties of \otimes

Comments:

- (1) In practice Ω is not known and is estimated from 2SLS residuals.
- (2) If Ω is diagonal or each structural equation is exactly identified, then 3SLS = 2SLS.
- (3) $\hat{\boldsymbol{\delta}}_{3SLS}$ is consistent. The asymptotic distribution $\hat{\boldsymbol{\delta}}_{3SLS}$ is

$$N[\boldsymbol{\delta}; [\mathbf{Z}^* (\Omega^{-1} \otimes \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}') \mathbf{Z}^*]^{-1}]$$

Exercise: Demonstrate that 3SLS estimators of $\boldsymbol{\delta}$ can be obtained as instrumental variables estimators.

Hint: Select the instrumental variables, $Z = (\hat{\Omega}^{-1} \otimes X(XX)^{-1}X')Z^*$ equation (3.3) solve $Z'y^* = Z'Z^*\delta_{IV}$ for δ_{IV} .

The asymptotic distribution of

$$\sqrt{N} \left[\begin{bmatrix} \tilde{B}_{\text{vec}} \\ \tilde{\Gamma}_{\text{vec}} \\ \tilde{\Omega}_{\text{vec}} \end{bmatrix}_{3\text{SLS}} - \begin{bmatrix} B_{\text{vec}} \\ \Gamma_{\text{vec}} \\ \Omega_{\text{vec}} \end{bmatrix} \right] \quad \text{is}$$

$$N[0, \Sigma = -\text{plim } \frac{1}{N} \begin{bmatrix} \ell_{BB} & \ell_{BG} & \ell_{B\Omega} \\ \ell_{GB} & \ell_{GG} & \ell_{G\Omega} \\ \ell_{\Omega B} & \ell_{\Omega G} & \ell_{\Omega\Omega} \end{bmatrix}^{-1}]$$

where ℓ denotes the log likelihood function and subscripts denote derivatives. Thus, both 3SLS and FIML are asymptotically efficient.

Note:

1. David Hendry demonstrated that almost all simultaneous equation estimators (3SLS, 2SLS, instrumental variables, . . .) can be obtained as approximate solutions to the necessary conditions defining the FIML estimators.
Hendry, "The Structure of Simultaneous Equation Estimators," Journal of Econometrics 4 (1976), pp. 51-88.

6. Statistical Inference

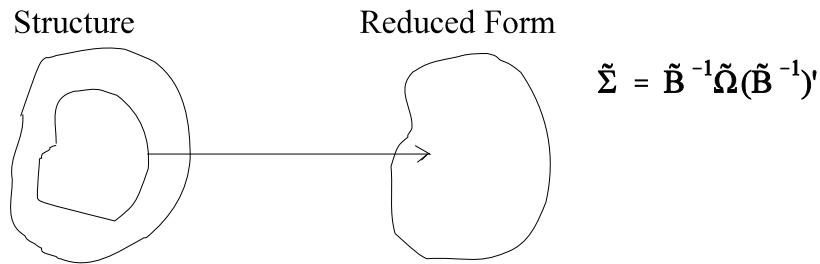
This section briefly summarizes procedures for testing hypotheses about (1) identification, (2) structural and reduced form coefficients, and (3) restrictions on structural coefficients.

a. Identification -- Logically Prior to Estimation

- (1) System tests of the validity of all overidentifying restrictions.
Hendry (1972, IER) Let

$$LR = \tilde{N} \ln(|\tilde{\Sigma}| / |\hat{\Sigma}|)$$

where



is the restricted estimator of the reduced form variance covariance matrix and $\hat{\Sigma}$ is the unrestricted estimator of the reduced form variance covariance matrix.

$$\hat{\Sigma} = [Y'(I - X(X'X)^{-1}X)Y]/(N - K)$$

The asymptotic distribution of LR is

$$\chi^2 \text{ (total number of overidentifying restrictions)}$$

Also see Hendry, The Econometric Analysis of Time Series, p. 338.

Questions:

1. Which estimation technique? Hendry suggests FIML or 3SLS.
2. How large sample size for the asymptotic distribution to be an accurate approximation?

(2) Single equation tests of over identifying restrictions

The null hypothesis is that the instruments (deleted exogenous variables) should not be included in the structural equation being estimated.

A common test statistic of this hypothesis is defined as follows:

$$\frac{(N - K)}{(v)} \frac{G_1(\tilde{\beta}_1) - G_2(\tilde{\beta}_1)}{G_2(\tilde{\beta}_1)} \stackrel{a}{\sim} F(v, N - K)$$

where $\tilde{\beta}_1$ denotes an estimator β_1 .

If the Stata command **estat overid** follows an **ivregress** command, then tests of overidentification are performed. If 2SLS is used, Basmann's chi-square test is reported as is Woodridge's robust score test. If LIML is used, Anderson and Rubin's chi square test is reported as is Basmann's F test (above). If GMM is used, Hansen's J statistic and chi square test are given. Statistically insignificant is consistent with the instruments not being included in the structural equation of interest; whereas, statistically significant tests indicate that the exclusion restrictions are not valid and the instruments are not valid.

**ivregress 2sls or liml y1 (Y1=X1 X2) X1
estat overid**

Some theoretical references on the exact distribution of the distribution of the identifiability test statistic. McDonald (1972, *Econometrica*, $G_{\Delta} = 2$, LIML), Basmann, Richardson ($G_{\Delta} = 2$, 2SLS) (1973, *Econometrica*), and Rhodes (1981, *Econometrica*, G_{Δ} arbitrary, LIML)

Simulation results suggest the following about the quality of approximation

$$F(v, N-K) \sim \chi^2(v) \text{ an}$$

OLS "F statistic" provides a poor approximation

2SLS and LIML identifiability test statistics are more nearly approximated by the F distribution than the OLS test statistic.

LIML identifiability test statistic seems to be closer to F than for 2SLS.

b. Reduced form coefficients

$$(1) \text{ LSNR: } \hat{\Pi}_{\text{Vec}} \sim N(\Pi_{\text{Vec}}; \Sigma \otimes (\mathbf{X}' \mathbf{X})^{-1})$$

- (a) t, F appropriate for LSNR
- (b) Chow, LR, and Wald tests are appropriate

$$(2) \text{ Derived RF: } \tilde{\Pi}_{\text{Vec}} \stackrel{A}{\sim} N(\Pi_{\text{Vec}}; \Sigma_{\tilde{\Pi}})$$

t and F statistics provide the *asymptotic* distributions for derived reduced from estimators

$$\frac{\tilde{\Pi}_{ij} - \Pi_{ij}}{S_{\tilde{\Pi}_{ij}}} \stackrel{a}{\sim} N(0, 1)$$

c. Structural coefficients

$$(1) \quad H_0: \beta_{ij} = \beta_{ij}^0$$

$$\frac{\hat{\beta}_{ij} - \beta_{ij}^0}{s_{\hat{\beta}_{ij}}} \sim_a N[0,1] \text{ or } t(\cdot)$$

$$(2) \quad H_0: \gamma_{ij} = \gamma_{ij}^0$$

$$\frac{\hat{\gamma}_{ij} - \gamma_{ij}^0}{s_{\hat{\gamma}_{ij}}} \sim_a N[0,1] \text{ or } t(\cdot)$$

Some observations based upon Monte Carlo simulations and analytic results suggest:

- (1) OLS “t-statistics” do not seem to have a distribution which is closely approximated by a t density.
- (2) The density functions of the 2SLS and LIML “t-statistics” seem to be much more closely approximated by a t density. (It all depends.)
 - “2SLS and LIML “t-statistics” seem reliable to use in testing significance of exogenous variables”
 - “For an endogenous variable, the distribution of the 2SLS t-statistic deviates far from Student’s t distribution if the non-centrality parameter is small (<13) and degree of over-identification (>7).”
 - “Student’s t approximation to the LIML t-statistic is the most accurate, and the two-sided test may be most appropriate in empirical studies (modest skewness).”
 - Morimune, K, t Test in a Structural Equation, Econometrica, 57 (1989), 1341-1360.
- (3) Caution: while the “t-ratios” for consistent structural coefficient estimators are asymptotically $N[0, 1]$, this does not imply that either the t or the $N(0,1)$ will give accurate results for sample sizes encountered in practice.
- (4) An alternative which may be practical in simple models is to transform the hypothesis to the reduced form and test using LSNR.
- (5) Still another approach to determining the distribution of the test statistic is to use the bootstrap.

c. “Testing a Subset of Coefficients in a Structural Equation”

Consider the structural model

$$y_1 = y_1\beta'_{10} + X_1\gamma'_{10} + \varepsilon_1$$

H_0 : Some of the coefficients in β_{10} and γ_{10} are zero.

- (1) Chow statistics based upon a comparison of SSE's are INVALID.
- (2) Wald tests can be fairly accurate—based on asymptotic results (sample size)

These tests can be implemented in Stata by obtaining consistent estimators of the structural coefficients and then using the test command, e.g., ivregress (2sls or liml) y1 (Y1=Z) X1, followed by test x3=0

- (3) Morimune and Tsukuda found that “likelihood ratio” tests based on 2SLS or LIML estimators were quite closely approximated with a chi square density.

$$\tilde{\lambda}_1 = \frac{G_1(\tilde{\beta}_{1.}) - G_2(\tilde{\beta}_{1.})}{G_2(\tilde{\beta}_{1.})}$$

$$\tilde{\lambda}_2 = \frac{G_1(\tilde{\beta}_{2.}) - G_2(\tilde{\beta}_{2.})}{G_2(\tilde{\beta}_{2.})}$$

where the subscript “1” corresponds to the case in which the variables with hypothesized zero coefficients have been deleted and “2” to the inclusion of the variables.

The test statistic is then given by

$$\frac{N - \# \text{ coeff. in unrestricted equation}}{\# \text{ restrictions}} \quad \frac{\tilde{\lambda}_1 - \tilde{\lambda}_2}{1 + \tilde{\lambda}_2}$$

This statistic is asymptotically distributed as

$F (\# \text{ restrictions}, N - \# \text{ Coefficients in unrestricted model})$

If there is a single restriction, as in the previous section, the corresponding LR test statistic is asymptotically as an $F(1, N - \# \text{ Coefficients in unrestricted model})$ or as a $\chi^2(N - \# \text{ Coefficients in unrestricted model})$ statistic.

Morimune, K. and Y. Tsukuda, "Testing a Subset of Coefficients in a Structural Equation," Econometrica, 52 (1984), 427-448.

(4) Testing Structural Parameters when using Instrumental Variables

Kleibergen, F, "Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression," Econometrica 70(2002), 1781-1803

Kleibergen proposes a test statistic, based on a quadratic form of the score of the concentrated log-likelihood, which can be used for performing joint tests on all of the structural parameters in instrumental variables regression. The test statistic is independent of "nuisance" parameters (pivotal statistic) and has an asymptotic chi distribution.

7. Simultaneous Equations Exercises

A. Identification

1. Using the supply and demand example (number 1) from section VI.3.a, express each of the six structural parameters in terms of reduced form parameters.
2. Attempt to replicate (1) for example 2 in section VI.3.a.
3. Attempt to replicate (1) for example 3 in section VI.3.a.
4. Verify that
 - (1) $v = K_2 - G_\Delta + 1 \geq 0$,
 - (2) $K_2 + G_{\Delta\Delta} + 1 \geq G$, and
 - (3) $K_2 \geq G_\Delta - 1$

are equivalent expressions for the order (necessary) conditions for identifying structural equations.

5. Obtain the reduced form of the following set of structural equations.

$$\begin{aligned} y_{1t} &= -2y_{2t} + 7x_{1t} + 4x_{2t} + x_{3t} - 8x_{4t} + u_{1t} \\ y_{2t} &= 2y_{1t} + y_{3t} - x_{1t} + 7x_{3t} - 9x_{5t} + u_{2t} \\ y_{3t} &= 2y_{1t} - 7x_{2t} + 7x_{3t} + 14x_{4t} + u_{3t} \end{aligned}$$

Investigate the identifiability of each equation, both (a) by using only the structural equations and (b) by using the reduced form equations. (L.S.E. 1966)

6. Consider the model defined by

$$C_t = \alpha_0 + \alpha_1 Y_t + u_t \quad (1)$$

$$I_t = \beta_0 + \beta_1 Y_t + \beta_2 I_{t-1} + v_t \quad (2)$$

$$Y_t = C_t + I_t + G_t \quad (3)$$

Discuss the identification of the above equations if G is the only exogenous variable (apart from the dummy variable for the constant term)

- a. as written
- b. as in (a) but over the sample period G is constant.

7. Discuss the restrictions, if any, which are implied for the reduced form by each of the following structural equations, assuming nothing is known about the covariance matrix of the disturbance terms.

$$y_{1t} = \gamma_{11}x_{1t} + \gamma_{12}x_{2t} + u_{1t}$$

$$y_{2t} = \beta_{21}y_{1t} + \gamma_{21}x_{1t} + \gamma_{22}x_{2t} + \gamma_{23}x_{3t} + u_{2t}$$

$$y_{3t} = \beta_{31}y_{1t} + \beta_{32}y_{2t} + u_{3t}$$

What additional restrictions, if any are implied by

a. $\gamma_{12} = 0$ (L.S.E. 1969)

B. A Little Theory

1. Verify that

$$\hat{\pi}_{\text{vec}} \sim N [\pi_{\text{vec}}, \sum \otimes (X' X)^{-1}]$$

simplifies to the regular regression result in the case of a model with a single reduced form equation.

2. What do you think the relationship between 2SLS and OLS estimators will be in the case of the reduced form equations having high R²'s. Hint: Compare (5.5) and (5.6) in VI.5.
3. Verify that

$$Y'_1 Y_1 - \hat{V}'_1 \hat{V}_1 = \hat{Y}'_1 \hat{Y}_1$$

in the discussion of 2SLS.

4. One estimator of $\omega_{11} = \text{var}(\varepsilon_{1t})$ is given by

$$\hat{\omega}_{11} = \left(\frac{1}{N - K + v} \right) G_1 (\hat{\beta}_{1.})$$

explain why the denominator $N - K + v$ is used. See the discussion of 2SLS.

5. In the discussion of LIML verify that

$$\hat{\lambda} = \frac{G_1(\hat{\beta}_1)}{G_2(\hat{\beta}_1)} - 1.$$

6. In the discussion of k-class estimators

verify $\text{Min}_{\beta} [G_1(\beta_1) - k G_2(\beta_1)]$ yields

- (1) OLS for $k = 0$
- (2) 2SLS for $k = 1$
- (3) LIML for $k = \lambda$ (problem 5 above)

7. Verify that $Z = [Y_1 - k \hat{V}_1, X_1]$ yields an instrumental variables equivalent to k- class estimators.
8. What value of k yields Zellner's MELO estimator for $N = 25, K = 5, K_2 = 1, G_\Delta = 2, v = 1$.
9. Verify that $\Omega = I$ implies that 3SLS = 2SLS.

C. Application: An exactly identified case

Consider the following Supply and Demand Model:

$$\text{Demand: } Q_t = a_{11} + B_{12} P_t + a_{12} Y_t + e_{t1}$$

$$\text{Supply: } Q_t = a_{21} + B_{22} P_t + a_{23} FC_t + e_{t2}$$

Where Q_t , P_t , Y_t and FC_t denote quantity, price, income and factor costs.

Observations on these variables are given by:

P_t	185	215	275	279	310	330	400	360	450	515
Q_t	320	360	460	460	480	540	600	570	680	780
Y_t	100	120	160	164	180	200	240	220	280	320
FC_t	10	12	14	15	20	16	24	20	28	30

1. Express the reduced form representation in terms of the structural coefficients.

2. Determine which of the structural coefficients can be expressed in terms of the reduced form coefficients and make this relationship explicit where possible.

3. Determine whether the supply and demand equations are identified. Use the necessary (order) conditions for identification in your analysis.

4. Estimate the reduced form equations for P and Q using the technique of Least Squares No Restrictions (LSNR), i.e., just use regular least squares {reg P D FC and reg Q D FC} .
 - a) Test for the presence of autocorrelation.

- b) Test for heteroskedasticity.
5. Estimate the supply and demand equations using OLS.
6. Estimate the supply and demand equations using 2SLS.
7. Comment on the properties of the estimators associated with (5) and (6).
8. Indicate how you could test the following hypotheses and discuss any related problems.
- a) $\beta_{12} = -2$
- b) $a_{12} = 0$
- c) $\Pi_{12} = 2.5$
- d) $\Pi_{22} = 0$
9. What implication does $\Pi_{22} = 0$ have with respect identification of any of the structural equations?
10. Indicate and perform a predictive test of your model when the last two observations are used to test the predictive ability of the model under consideration. Do these two observations lie in 95% confidence intervals?
11. Indicate how you would go about making price and quantity predictions for the next two periods after the sample data. Construct appropriate confidence intervals.

D. Application: an overidentified case

Consider the following model to explain variation in consumption and prices of food:

Demand: $Q_t = \beta_{12}P_t + \gamma_{11} + \gamma_{12}D_t + \varepsilon_t$

$$\text{Supply: } Q_t = \beta_{22}P_t + \gamma_{21} + \gamma_{23}F_t + \gamma_{24}A_t + \varepsilon_{t2}$$

where P = relative prices of food to consumer prices

Q=per capita food consumption

D=per capita disposable income

F=prices received by farmers last year/general consumer prices

A=time index in years

1. Check the necessary (order)conditions for each of the structural equations to be identified.
 2. Estimate the structural equations using the methods of OLS, 2SLS, LIML, and 3SLS.
 3. Estimate the corresponding reduced form equations using the methods of (1) LSNR and (2) 2SLS . Recall that the 2SLS estimates of the reduced form are obtained by estimating the structure using 2SLS and then deriving the corresponding reduced form estimates.
 4. Do the variables D, F, and A have statistically significant explanatory power in the reduced form equation for P? Use F and t-tests. What implications do these results have for the strength of the instruments?
 5. Perform and interpret a statistical test of the overidentifying restrictions for the demand equation
 6. Obtain predictions for Q corresponding to 2SLS and LSNR for $(D,F,A) = (140, 110, 25)$.

Q	P	D	F	A
98.485	100.323	87.4	98	1
99.187	104.264	97.6	99.1	2
102.163	103.435	96.7	99.1	3
101.504	104.506	98.2	98.1	4
104.24	98.001	99.8	110.8	5
103.243	99.456	100.5	108.2	6
103.993	101.066	103.2	105.6	7
99.9	104.763	107.87	109.8	8
100.35	96.446	96.6	108.7	9
102.82	91.228	88.9	100.6	10
95.435	93.085	75.1	81	11
92.424	98.801	76.9	68.6	12
94.5358	102.908	84.5	70.9	13
98.757	98.756	90.6	81.4	14
105.797	95.119	103.1	102.3	15
100.225	98.451	105.1	105	16
103.522	86.498	96.4	110.5	17
99.929	104.016	104.4	92.5	18
105.223	105.769	110.7	89.3	19
106.232	113.49	127.1	93	20

E. A simple dynamic model

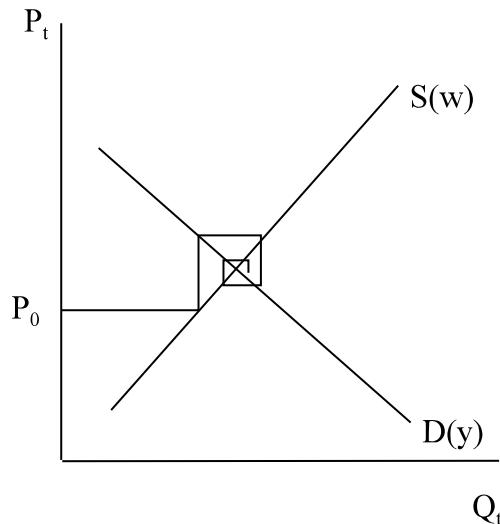
A very simple example of a dynamic model is the famous cobweb model defined by the **structural model**:

$$\text{Demand: } Q_t = \alpha + \beta P_t + \xi Y_t \quad (\text{E.1 a-b})$$

$$\text{Supply: } Q_t = \gamma + \delta P_{t-1} + \rho W_t$$

The variables appearing in the model can be classified as endogenous, and predetermined (exogenous and lagged endogenous) as follows:

Endogenous:	Q_t, P_t	}	Predetermined
Exogenous: Y_t, W_t			
Lagged Endog: P_{t-1}			



If the initial price (P_0) differs from the equilibrium price, then the time path of (Q_t, P_t) will resemble a “cobweb” -- converging to the equilibrium if

$$|\text{slope demand curve}| <$$

slope of the supply curve, or

$$|1/\beta| < 1/\delta \quad |\delta/\beta| < 1 \text{ or}$$

$|\delta/\beta| > 1$ diverging for

Note:

- Economic theory suggests that δ and ξ are positive and β and ρ are negative.
- For convenience, the random disturbances have been deleted.

1. Demonstrate that the supply and demand equations satisfy the necessary conditions for identification. Recall that the necessary condition is that the number of excluded predetermined (exogenous and predetermined) variables appearing in the model must be at least as large as the number of current endogenous regressors in each equation.

2. Demonstrate that the structural model can also be written as:

(E.2)

$$\begin{bmatrix} -1 & \beta \\ -1 & 0 \end{bmatrix} \begin{bmatrix} Q_t \\ P_T \end{bmatrix} + \begin{bmatrix} \alpha & 0 & 0 & \xi \\ \gamma & \delta & \rho & 0 \end{bmatrix} \begin{bmatrix} P_{t-1} \\ W_t \\ Y_t \end{bmatrix} = 0 ,$$

(E.3 a-c)

Structural equations

Generic form (Section VI.2)

$$\begin{bmatrix} -1 & \beta \\ -1 & 0 \end{bmatrix} \begin{bmatrix} Q_t \\ P_T \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & \delta \end{bmatrix} \begin{bmatrix} Q_{t-1} \\ P_{t-1} \end{bmatrix} + \begin{bmatrix} \alpha & 0 & \xi \\ \gamma & \rho & 0 \end{bmatrix} \begin{bmatrix} W_t \\ Y_t \end{bmatrix} = 0 \quad BY_t' + B_1Y_{t-1}' + \Gamma X_t' = 0$$

$$\left\{ \begin{bmatrix} -1 & \beta \\ -1 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & \delta \end{bmatrix} L \right\} \begin{bmatrix} Q_t \\ P_t \end{bmatrix} + \begin{bmatrix} \alpha & 0 & \xi \\ \gamma & \rho & 0 \end{bmatrix} \begin{bmatrix} W_t \\ Y_t \end{bmatrix} = 0 \quad (B + B_1L)Y_t' + \Gamma X_t' = 0$$

$$\begin{bmatrix} -1 & \beta \\ -1 & \delta L \end{bmatrix} \begin{bmatrix} Q_t \\ P_t \end{bmatrix} + \begin{bmatrix} \alpha & 0 & \xi \\ \gamma & \rho & 0 \end{bmatrix} \begin{bmatrix} W_t \\ Y_t \end{bmatrix} = 0 . \quad B(L)Y_t' + \Gamma X_t' = 0$$

3. Demonstrate that the reduced form representation can be obtained from (E.2)

(E.4)

$$\begin{bmatrix} Q_t \\ P_t \end{bmatrix} = -\begin{bmatrix} -1 & \beta \\ -1 & 0 \end{bmatrix}^{-1} \begin{bmatrix} \alpha & 0 & 0 & \xi \\ \gamma & \delta & \rho & 0 \end{bmatrix} \begin{bmatrix} 1 \\ P_{t-1} \\ W_t \\ Y_t \end{bmatrix}$$

$$= \begin{bmatrix} \gamma + \delta P_{t-1} + \rho W_t \\ \frac{\gamma - \alpha}{\beta} + \frac{\delta}{\beta} P_{t-1} + \frac{\rho}{\beta} W_t - \frac{\xi}{\beta} Y_t \end{bmatrix}$$

Note: The reduced form expresses each current dependent variable in terms of predetermined (exogenous and lagged endogenous) variables.

4. Demonstrate that the final form or transfer functions for P_t and Q_t can be written as are

Generic form

$$\begin{bmatrix} Q_t \\ P_t \end{bmatrix} = -\begin{bmatrix} -1 & \beta \\ -1 & \delta L \end{bmatrix}^{-1} \begin{bmatrix} \alpha & 0 & \xi \\ \gamma & \rho & 0 \end{bmatrix} \begin{bmatrix} 1 \\ W_t \\ Y_t \end{bmatrix}$$

$$Y'_t = -B^{-1}(L)\Gamma X'_t$$

$$= \begin{bmatrix} \frac{\gamma\beta + \alpha\delta}{\delta - \beta} + \rho \sum_{i=0} (\delta/\beta)^i W_{t-i} - \frac{\xi\delta}{\beta} \sum_{i=0} (\delta/\beta)^i Y_{t-i-1} \\ \frac{\alpha - \gamma}{\delta - \beta} - \frac{\rho}{\beta} \sum_{i=0} (\delta/\beta)^i W_{t-i} - \frac{\xi}{\beta} \sum_{i=0} (\delta/\beta)^i Y_{t-i} \end{bmatrix}$$

Hint: This result follows from the following relationship.

$$B^{-1}(L) = \begin{bmatrix} -1 & \beta \\ -1 & \delta L \end{bmatrix}^{-1} = \frac{1}{-\delta L + \beta} \begin{bmatrix} \delta L & -\beta \\ 1 & -1 \end{bmatrix}$$

$$= \frac{1}{\beta(1 - (\delta/\beta)L)} \begin{bmatrix} \delta L & -\beta \\ 1 & -1 \end{bmatrix}$$

$$= \frac{1}{\beta} \sum (\delta/\beta)^i L^i \begin{bmatrix} \delta L & -\beta \\ 1 & -1 \end{bmatrix}.$$

Note: In the final form, each current dependent variable is expressed in terms of current and lagged values of exogenous variables.

Further note that:

$$\frac{\partial Q_t}{\partial W_t} = \rho < 0 \quad \frac{\partial Q_t}{\partial W_{t-1}} = \frac{\rho\delta}{\beta} > 0,$$

$$\frac{\partial Q_t}{\partial Y_t} = 0 \quad \frac{\partial Q_t}{\partial Y_{t-1}} = \frac{-\xi\delta}{\beta} > 0, \dots$$

$$\frac{\partial P_t}{\partial W_t} = \frac{\rho}{\beta} > 0 \quad \frac{\partial P_t}{\partial W_{t-1}} = \frac{\rho\delta}{\beta^2} < 0, \dots$$

$$\frac{\partial P_t}{\partial Y_t} = \frac{-\xi}{\beta} > 0 \quad \frac{\partial P_t}{\partial Y_{t-1}} = \frac{-\xi\delta}{\beta^2} < 0, \dots .$$

5. Show that the short-run (impact) multipliers can be obtained from the reduced form coefficients as well as from the transfer function form. Hint: $-B^{-1}(L=0)\Gamma = -B^{-1}\Gamma$

6. Demonstrate that the long-run cumulative multipliers are given by

$$\begin{bmatrix} -1 & \beta \\ -1 & \delta(L=1) \end{bmatrix}^{-1} \begin{bmatrix} \alpha & 0 & \xi \\ \gamma & \rho & 0 \end{bmatrix} = \begin{bmatrix} \frac{\alpha\delta-\beta\gamma}{\delta-\beta} & \frac{-\beta\rho}{\delta-\beta} & \frac{\delta\xi}{\delta-\beta} \\ \frac{\alpha-\gamma}{\delta-\beta} & \frac{-\rho}{\delta-\beta} & \frac{\xi}{\delta-\beta} \end{bmatrix}$$

Hint: $-B^{-1}(L=1)\Gamma = -(B+B_1)^{-1}\Gamma$

7. What STATA commands would you use to obtain consistent estimators of

- a. the reduced form coefficients
- b. the structural coefficients in the cobweb model?

8. The cobweb formulations corresponds to expectations formed as follows

$E(P_t | P_{t-1}, \dots) = P_t^* = P_{t-1}$. Alternative models for forming expectations have been

including adaptive expectations, rational expectations, and the use of ARIMA models.

F. Consider the following model to describe returns to education for married working women:

$$\ln(wage) = \beta_1 + \beta_2 education + \varepsilon$$

This model can be estimated using data from Mroz (Econometrica 54 (1986, 765-799))

1. Estimate the model using OLS
2. Estimate the model using IV with Z=father's education
3. Estimate the model using IV with Z=(father's education and mother's education)
Use and interpret the **estat overid** command
4. Compare and interpret the different estimates of the impact of education on ln(wage).
5. Which estimate would you feel most comfortable with and why?
6. Comment on additional modifications you might include in the model.

The data is found in mroz.raw and is described as follows:

```
inlf    hours    kidslt6    kidsge6    age    educ    wage    repwage
hushrs    husage    huseduc    huswage    faminc    mtr    motheduc
fatheduc    unem    city    exper    nwifeinc    lwage    expersq
```

Obs: 753

1. inlf	=1 if in labor force, 1975
2. hours	hours worked, 1975
3. kidslt6	# kids < 6 years
4. kidsge6	# kids 6-18
5. age	woman's age in yrs
6. educ	years of schooling
7. wage	wife's estimated wage from earns., hours
8. repwage	reported wage at interview in 1976
9. hushrs	hours worked by husband, 1975
10. husage	husband's age
11. huseduc	husband's years of schooling
12. huswage	husband's hourly wage, 1975
13. faminc	family income, 1975
14. mtr	fed. marginal tax rate facing woman
15. motheduc	mother's years of schooling
16. fatheduc	father's years of schooling
17. unem	unem. rate in county of resid.
18. city	=1 if live in SMSA
19. exper	actual labor mkt exper
20. nwifeinc	(faminc - wage*hours)/1000
21. lwage	log(wage)
22. expersq	exper^2

G. An Application of Transfer Functions

Maloney and Ireland ("Fiscal Versus Monetary Policy: An Application of Transfer Functions," Journal of Econometrics, 1980, pp. 253-266) use transfer functions to study the relative importance of monetary and fiscal policy upon aggregate production. They examine the coefficients (dynamic multipliers) of the lagged exogenous variables in the transfer function representation. Maloney and Ireland do not begin their analysis by specifying a structural model, as done in Kmenta and Smith (RESTAT, 1973), but rather estimate a transfer function directly. Their model is given by

$$y_t = \frac{w_1(L)}{\delta_1(L)} m_t + \frac{w_2(L)}{\delta_2(L)} g_t + \zeta_t$$

where $w_i(L)$ and $\delta_i(L)$ are assumed to be polynomials in the lag operator (L) and y_t , g_t , and m_t respectively, denote the percentage change in real GDP, change in real government purchases of goods and services, and in the real monetary base. Using quarterly data from 1953 to 1975, they estimate the model to be

$$y_t = \frac{.243}{1 - 1.4847L + .70886L^2} m_t + \frac{.06648}{1 - 1.0311L + .83244L^2} g_t + .005 + \zeta_t.$$

The impact multipliers corresponding to m and g are .243 and .06648.

- a. What is the interpretation of the impact multipliers corresponding to m and g ?

- b. What is the interpretation of the long-run multipliers corresponding to m and g ?

- c. Evaluate the impact (short run) multipliers corresponding to m and g .

Helpful notes on the calculation of lagged weights for m and g in the Maloney-Ireland transfer function representation. Consider the calculation of the distributed lag coefficients for m .

$$\frac{w_1(L)}{\delta_1(L)} =$$

$$.24 + .36L + .37L^2 + .25L^3 + \dots$$

$$1 - .15L + .7L^2 | .24$$

The short run or impact multiplier for m is given by

$$\frac{w_1(L=0)}{\delta_1(L=0)} = .24$$

which is the leading coefficient of the expansion of

$$\frac{w_1(L)}{\delta_1(L)}$$

The long run multiplier is given by

$$\frac{w_1(L=1)}{\delta_1(L=1)} = \frac{.243}{1 - 1.4847 + .70886} = 1.08$$

which is the sum of the coefficients in the expansion of $\frac{w_1(L)}{\delta_1(L)}$

Summary Table

	Impact multiplier	Long run multiplier
m	.243	1.08
G	.066	.083

Polynomial distributed lags could have been used rather than a ratio of polynomials.

SOME NOTES ON UNIVARIATE TIME SERIES ANALYSIS*

O. Basic Problem

- The basic problem involves the use of past observations of a random variable to forecast future values of that variable.
- You might enjoy reading “Forecasting—looking back and forward: Paper to celebrate the 50th anniversary of the Econometrics Institute at the Erasmus University, Rotterdam” by Clive Granger (**Journal of Econometrics**, 138 (2007), 3-13) which reflects on the past and future of forecasting methods. Professor Granger received the Nobel Prize for his work in time series and causality.

A. A, B, C's of Time Series Models

1. Stationarity
2. Non-stationary: two important examples
 - a. Random walk with drift
 - b. Trend stationary model
 - c. Summary
 - d. Dickey-Fuller Tests
3. Random walks, trends, and spurious regressions
4. Cointegration

B. Basic ARIMA Models and methods

C. Characteristics of some ARIMA Models and Identification

1. An overview and simple models
 - a. AR(1)
 - b. MA(1)
 - c. Summary
2. AR(p)
3. MA (q)
4. ARMA (p,q)
5. ARIMA (p,d,q)

D. Diagnostic Analysis

E. Estimation

F. Forecasting

G. General comments

H.. Help: Computer Programs

* ***For a “short ARIMA course” read Sections A and B, look at the figures in Section C, review Section G (1-4), then work on the Homework(next two pages).***

Homework

- Select a time series with at least 60 observations: Search “time series data”
- Plot the data
- Explore ways to obtain a stationary time series
 - Plot differences
 - Consider taking the log of the data
- Perform a DF test
- Identify and estimate a time series model–using AC and PAC coefficients
- Conduct diagnostic tests of the estimated model (corrgram resid–use pac's, ac's, and Q-statistics corresponding to the estimated residuals)
- Write the equation of the estimated model
Stata reports the results of estimating :
$$(1 - \phi_1 L - \dots - \phi_p L^p)(\Delta^d Y_t - const) = (1 + \theta_1 L + \dots + \theta_q L^q)\varepsilon_t$$
- Determine forecasts for five future periods

Helpful STATA Homework Commands

Data in variable named “y” and time index “t”

tsset t Alerts to STATA to time series data. If a time index is not included in the data file, the command **gen t=_n** will generate the required variable.

scatter y t Plots y vs. t

scatter D.y t Plots the difference vs. t

dfuller y Helps determine whether the series is

dfuller D.y difference or trend stationary. The null hypothesis is difference stationary ($d=1$).

corrgram y, lags(#) Reports estimated ac and pac coefficients for Y and first differences.

corrgram D.y, lags(#) These can be used to identify the model (select p,d,q) based on the

pattern of the estimates at each and pac coefficients.

ac y, lags(#) **pac y, lags(#)**

arima y, arima(p,d,q) Estimates the identified arima model. The estimated model can be recovered from this output

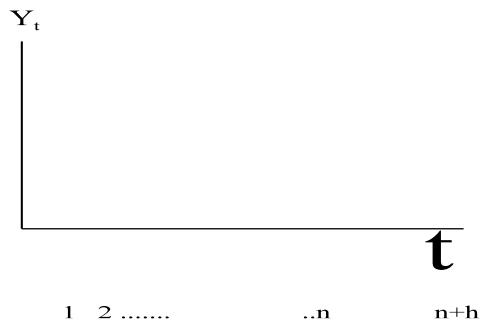
predict e, resid Stores the estimated residuals in variable e

corrgram e, lags(#)	Reports the estimated ac and pac coefficients corresponding to the estimated errors. Correct model specification would yield statistically insignificant ac and pac coefficients. The corresponding Q-statistics would be expected to be statistically insignificant.
tsappend, add(5)	Extrapolates the estimated model to obtain the next five predictions.
predict yhat	Stores predicted values in the variable <i>yhat</i>
list y yhat t	Prints the indicated (if available) values of <i>y</i> and indicated predicted values
list y yhat t if t>N*	

James B. McDonald
Brigham Young University

Univariate Time Series Analysis

0. Basic Problem: Consider the problem of, given n observations on a single variable Y (Y_1, Y_2, \dots, Y_n), obtain forecasts of Y at time period $n + h$, denoted by \hat{Y}_{n+h} or $Y_n(h)$. This might be pictorially represented as:



The data might correspond to GDP, consumer prices, foreign exchange rates, telephone calls, unemployment rates, product sales, the number of empty hospital beds in a hospital, commodity prices, or any other time series. The forecasting problem becomes that of attempting to obtain a prediction or forecast h periods in the future from known observations. Many techniques have been developed to extrapolate from past observations into the future. The techniques differ in structure and assumptions made, including the amount of past information used in forecasting. Some techniques assume that past (1) levels, (2) changes, or (3) percentage changes can be used to forecast the next period. Other techniques are based upon moving averages of past values and possibly allow for trends or seasonality in the underlying series. When forecasting, we would do well to remember the admonition found in 88th Section of the Doctrine and Covenants (79th verse) about studying things “which must shortly come to pass” and be very cautious about long-run forecasts.

Auto Regressive Integrated Moving Average (ARIMA) models are very general specifications which includes some of the previously mentioned forecasting approaches as special cases as well as including more general methods. These techniques will be studied in order of increasing "sophistication."

Before getting into the details of time series forecasting techniques, it is useful to provide a brief overview of potential applications of these techniques. They can be used (1) on a "stand alone" basis to predict future values of a dependent variable of interest, $Y_n(h)$, by extrapolating past trends ; (2) to predict a future value for an independent variable (X_t), $X_n(h)$ which can be substituted into an econometric model to yield

$$Y_n(h) = f(X_n(h))$$

to obtain forecasts of the dependent variable; and finally (3) these techniques can also be used to predict systematic components in the residuals (ε) , $Y_n(h) = f(X_n(\varepsilon_n(h)))$. Thus time series techniques can be used separately from the specification of an economic model or in conjunction with an estimated economic model.

A. A,B,C's of time series models

1. Stationarity

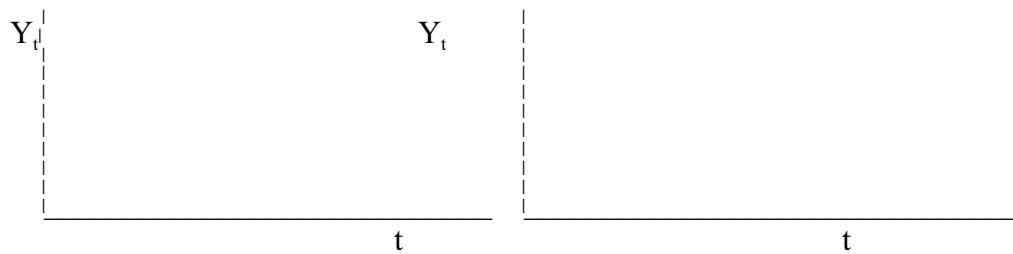
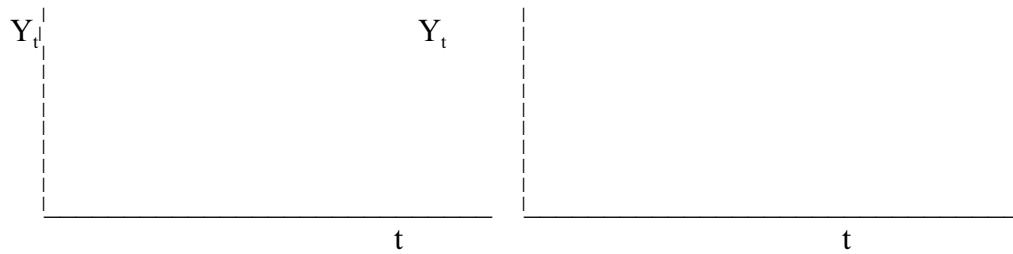
Consider a stochastic process Y_t which is defined for all integer values of t . Y_t is said to be **weakly or covariance stationary** if

- $E(Y_t) = E(Y_s) = \mu$ for all t and s (A.1a-c)
- $\text{Var}(Y_t) = \text{Var}(Y_s) = \sigma^2$ for all t and s
- $\text{Cov}(Y_t, Y_s) = \gamma_{t-s}$ depends only on $t-s$

for all t . A stronger definition of stationarity is that the joint distribution functions

$F(Y_{t+1}, \dots, Y_{t+n})$ and $F(Y_{s+1}, \dots, Y_{s+n})$ are identical for any value of t and s .

Consider the four following figures:

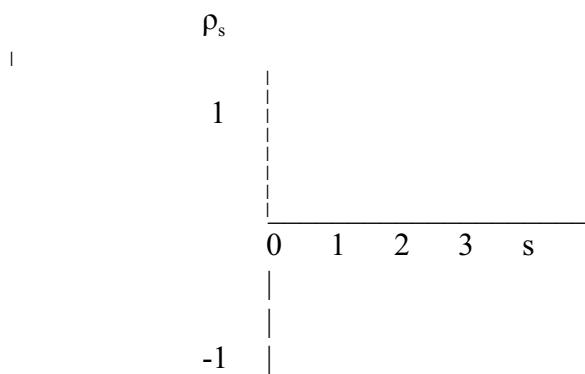


The first series would be classified as being stationary; whereas, the other three would not.

It will be useful to define the autocorrelation coefficients corresponding to Y_t in terms of the autocovariances (γ_s)

$$\rho_s = \frac{\gamma_{t-s}}{\gamma_0} = \text{correlation}(Y_t, Y_s) \text{ w } \gamma_0 e \text{ } Var(Y_t)$$

and the plot of ρ_s against s is referred to as the correlogram.



An example of a stationary series is the autoregressive model, AR(1), where

$$Y_t = \phi_1 Y_{t-1} + \varepsilon_t \quad (\text{A.3 a-b})$$

where

$$\varepsilon_t \sim N[0, \sigma^2].$$

From this specification, it can be shown that

$$E(Y_t) = 0$$

$$\text{Var}(Y_t) = \frac{\sigma^2}{1 - \phi_1^2} = \gamma_0$$

$$\text{Cov}(y_t, y_{t-s}) = (\phi_1^s) \frac{\sigma^2}{1 - \phi_1^2} = \phi_1^s \gamma_0 = \gamma_s$$

with autocorrelation coefficients given by

$$\rho_s = \frac{\gamma_s}{\gamma_0} = \phi_1^s \quad \text{which decay exponentially for } |\phi_1| < 1.$$

While the AR(1) model is stationary if $|\phi_1| < 1$, many economic series are not stationary.

Remember a time series is not stationary if either the mean, variance, or covariance change with time. Thus a series which increases over time or is characterized by heteroskedasticity is not stationary. Since many forecasting techniques are based upon the stationarity

assumption, it is comforting to note that some non-stationary series can be transformed into stationary series. We now consider two such series.

2. Non-stationary series: two important examples

a. Consider the **random walk**, with drift:

$$Y_t = \mu + Y_{t-1} + \varepsilon_t \quad (\text{A.4})$$

where $\varepsilon_t \sim N[0, \sigma^2]$. By recursive substitution this model can be rewritten as

$$Y_t = \mu t + \sum_{i=0}^t (\varepsilon_{t-i}) \quad (\text{A.5})$$

Note that Y_t is not stationary because the mean of Y_t (μt) increases (linearly) with t and the variance of Y_t , about the linear trend, increases with time. **It is important to note that the first difference of Y_t , denote ΔY_t , of a random walk**

$$\Delta Y_t = Y_t - Y_{t-1} = \mu + \varepsilon_t \quad (\text{A.6})$$

is stationary. In working with time series it can be useful to use lag operators (L) or backshift operators (B) defined as follows:

$$L Y_t = Y_{t-1} \quad B Y_t = Y_{t-1} \quad \Delta = (1 - L) Y_t, \text{ thi} (1 - B) Y_t$$

Y_t is said to be integrated of order one, $I(1)$, when the first difference of the series is stationary. A series is said to be integrated of order d , $I(d)$, if $\Delta^d Y_t$ is stationary.

Note that if $d=1$, then $\Delta^d Y_t = Y_t - Y_{t-1}$ and if $d=0$ then $\Delta^0 Y_t = Y_t$.

b. Now consider the **trend stationary** model defined by

$$Y_t = \mu + \beta t + \varepsilon_t \quad (\text{A.7})$$

where $\varepsilon_t \sim N[0, \sigma^2]$. Like the random walk with drift, the trend stationary model has a mean ($\mu + \beta t$) which increases linearly with t ; however, the variance of Y_t , about its trend, is a constant σ^2 which is in contrast to the random walk with drift whose variance increases with t . The first difference of a trend stationary model

$$Z_t = Y_t - Y_{t-1} = (\mu + \beta t + \varepsilon_t) - (\mu + \beta(t-1) + \varepsilon_{t-1})$$

has a constant mean (β) and variance $2\sigma^2$, but the errors involve a moving average

$(\varepsilon_t - \varepsilon_{t-1})$ of error terms. The correlation between Z_t and Z_{t-1} is $-1/2$ and correlation

between Z_t and Z_{t-s} for $s > 1$ is zero.

c. The optimal estimation procedures for these two series are different. The best way to estimate the random walk with drift is to take first differences and then estimate the unknown parameters associated with the differenced series. The best way to work with the trend stationary series is to estimate a polynomial in “ t ” and then analyze the residuals. Thus, the two non stationary series are treated in different ways. The difference between these two important series can be summarized as in the following table.

Series type/ Approach	Differences: $\Delta^d Y_t$	Detrend [regress y on a polynomial time trend]
Random walk with drift <ul style="list-style-type: none"> • $Y_t = \mu + Y_{t-1} + \varepsilon_t$ • $Y_t = \mu t + \sum_{i=0}^t (\varepsilon_{t-i})$ • Behavior: Increasing mean and variance • The impact of innovations (ε_j) persist. 	<ul style="list-style-type: none"> • Optimal 	<ul style="list-style-type: none"> • Not optimal

Trend stationary	<ul style="list-style-type: none"> $Y_t = \mu + \beta t + \varepsilon_t$ Behavior: increasing mean with constant variance The impact of innovations (ε_t) pass. 	<ul style="list-style-type: none"> Not optimal OLS optimal Standard statistical tests are problematical
------------------	------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------------------------------------------

If a series appears to increase exponentially, the previous approaches could be applied to the logarithms of the data.

The implications of the two different models can be important in some applications. A practical problem is that it may be difficult to differentiate between trend and difference stationary models and, hence, to know the most efficient estimation technique to use. One approach to discriminating between the two models is to consider the following “nesting” regression model:

$$Y_t - Y_{t-1} = \mu(1-\gamma) + \beta\gamma + \beta(1-\gamma)t + (\gamma-1)Y_{t-1} + \varepsilon_t \quad (\text{A.9})$$

What does this nested regression model simplify to with r

(1) $\gamma = 0?$ T_____ S_____ (fill in the blanks)

(2) $\gamma = 1?$ R_____ W____ or DS (fill in the blanks)

There are a series of tests known as **Dickey-Fuller tests** which explore the null hypothesis, $H_0: \gamma = 1$, which are based on the regression model:

$$Y_t - Y_{t-1} = \alpha_0 + [\alpha_1] = (\gamma-1)Y_{t-1} + [\beta(1-\gamma)]t \quad (\text{A.10})$$

The hypothesis $\gamma = 1$ implies that the coefficients of the variables t and \mathbf{Y}_{t-1} equal 0. Standard t-tests are not appropriate, but appropriate tables have been constructed using Monte Carlo methods. If the coefficient of \mathbf{Y}_{t-1} , α_1 , is

negative, the evidence favors trend stationarity.

In some applications there is concern that the ε_t will be characterized by autocorrelation. In these cases, the **augmented Dickey_Fuller test** is based on estimating the equation

$$\mathbf{Y}_t - \mathbf{Y}_{t-1} = \alpha_0 + [\alpha_1 = (\gamma-1)]\mathbf{Y}_{t-1} + [\alpha_2 = \beta(1-\gamma) \sum_{j=1}^{T-t} \phi_j \Delta Y_{t-j}] + \varepsilon_t$$

where the term involving the summation has been added to pick up the impact of autocorrelation. The test for differencing is performed by testing the null hypothesis

$$H_0: \alpha_1 = (\gamma-1) = 0 \text{ (with or without a time trend in (A.10))}$$

This and other related tests can be performed using STATA (version 9 and 10) with the commands:

dfuller y

dfuller y, noconst suppresses the constant term in the regression

dfuller y, trend includes a trend term in the regression

dfuller y, lags(#) includes # lagged differences

Notes:

- (1) Tests of $\alpha_1 = 0$, null hypothesis of a unit root, are one tailed tests. The hypothesis of difference stationarity is rejected if the estimated test statistic is less than the

reported critical value. For example, for series without time trends, Asymptotic critical values for a unit root t - test [with no time trend] are given by

Significance level	1%	2.5%	5%	10%
Critical value	-3.43	-3.12	-2.86	-2.57

If $\frac{\hat{\alpha}_1}{s_{\hat{\alpha}_1}} < -3.43$, then we would reject the hypothesis of a unit root. In the case of unit

roots, one approach is to work with first differences of the data. There are other approaches.

(2).Peter Phillips and others have explored the use of fractional differences where Δ^d where $0 < d < 1$ and provides an intermediate ground between working with levels ($d=0$) and working with differences ($d=1$).

3. Random Walks, trends, and Spurious Regressions

Have you ever wondered what happens if you regress one series on an unrelated series when both series grow over time? This was the question explored by Granger and Newbold in a classic paper published in the Journal of Econometrics in 1974. Their finding may not be surprising. Often, in these cases, standard statistical tests suggest statistical significance when, in fact, there is no relationship. They concluded that in such cases much larger t-statistics would be needed than suggested by traditional t-tables. Granger and Newbold find a critical value of 11.2 more appropriate than 1.96. Their study was based on Monte Carlo simulations. Peter Phillips worked with a more general model and used analytical

methods, rather than simulation studies, and recommends the use of a critical value given by $(n^5)(t_{critical_value})$.

The bottom line of all of this, is that we may want to consider fitting relationships to appropriately differenced or detrended data. However, there are alternatives.

4. Cointegration

An interesting problem arises if Y and X are integrated of different orders. This makes it impossible for the error term, $\varepsilon_t = Y_t - X_t \beta$, to be stationary. Greene gives a very abbreviated treatment of this important issue.

B. Basic ARIMA Models/methods

1. Basic models and special cases

Autoregressive-integrated moving average (ARIMA) models are an important general class of stochastic models which have been widely adopted as models for time series. These models include several of the models in the previous section as special cases and are extremely versatile in terms of their statistical properties. An ARIMA model with parameters (p,d,q) is defined as follows:

$$Y_t^* - \phi_1 Y_{t-1}^* - \dots - \phi_p Y_{t-p}^* = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \dots - \theta_q \varepsilon_{t-q}^1, \quad (B.1)$$

¹Some authors use θ_i rather than θ_i as coefficients on the right hand side. These notes were written using $-\theta_i$. Stata uses θ_i 's and reports corresponding estimates. .

$$\left(1 - \sum_{i=1}^p \phi_i L^i\right) Y_t^* = \left(1 - \sum_{i=1}^q \theta_i L^i\right) \varepsilon_t$$

, or

$$\varphi(L)Y_t^* = \theta(L)\varepsilon_t$$

where

- $Y_t^* = (\Delta^d - (\mu - \dots)) = ((I - \mu^{-1}Y_{t-1} - \dots))$, e.g., for $d=1$ $Y_t^* = (\mu \cdot Y_{t-1} - \dots)$
- $\varepsilon_t \sim N(0, \sigma^2)$
- $\varphi(L) = 1 - \varphi_1 L - \dots - \varphi_p L^p$
- $\theta(L) = 1 - \theta_1 L - \dots - \theta_q L^q$

Note that the portion of the expression on the left hand side of the equal sign in equation (B.1) is the **Autoregressive (AR)** portion with p lags and that on the right hand side is the **Moving average component (MA)** with q lags. The d refers to the number of times the series is differenced. This model is denoted ARIMA(p,d,q) and includes many useful models as special cases.

Some special cases include

$$(1) \quad \text{ARIMA } (1,0,0) = \text{AR}(1)$$

$$Y_t - \varphi_1 Y_{t-1} = \varepsilon_t$$

This is a common form used to model autocorrelation in regression models.

$$(2) \quad \text{ARIMA } (p,0,0) = \text{AR}(p)$$

$$Y_t - \varphi_1 Y_{t-1} - \dots - \varphi_p Y_{t-p} = \varepsilon_t$$

$$(3) \quad \text{ARIMA } (0,0,1) = \text{MA}(1)$$

$$Y_t = \varepsilon_t - \theta_1 \varepsilon_{t-1}$$

$$(4) \quad \text{ARIMA } (0,0,q) = \text{MA}(q)$$

$$Y_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \dots - \theta_q \varepsilon_{t-q}$$

(5) ARIMA (p, 0, q) = ARMA (p,q)

(6) ARIMA(0, 0, 0) White noise
 $Y_t = \varepsilon_t$

(7) ARIMA(0,1,0) Random Walk
 $Y_t - Y_{t-1} = \varepsilon_t$

Other special cases include exponential smoothing ARIMA (0,1,1) and the Holt-Winters nonseasonal model corresponding to an ARIMA (0,2,2).

2. Stationarity and invertibility

An ARIMA(p,d,q) models is said to be **invertible** if it can be expressed as an

$ARI(d, \infty) \text{ moe } \varepsilon_t = \frac{\phi(L)}{\theta(L)} \Delta^d (Y_t - \mu)$. Similarly, an ARIMA(p,d,q) models is

stationary if it can be expressed as an IMA(d, ∞) moe $\Delta^d (Y_t - \mu) = \frac{\theta(L)}{\phi(L)} \varepsilon_t$

IMA representation is referred to as the **Wold decomposition**. Necessary and sufficient conditions for invertibility/stationarity require that each of the zeros of the respective polynomials, $\theta(z)$ and $\phi(z)$, have modulus greater than one. Certain coefficient restrictions insure invertibility/stability.

3. Basic steps in application of ARIMA models

Traditional applications of ARIMA models as forecasting tools involves the following four steps: (1) identification, (2) estimation, (3) diagnostics, and (4) forecasting.

(1) IDENTIFICATION--determine values for p, d, and q. In other words the appropriate number of autoregressive lags (p) and moving average lags (q) as well as the determination degree of differencing needs to be determined.

"d" is selected to be the number of times that the series must be differenced to obtain a stationary series.

Probably the most common method of "identifying" or selecting p and q is by analyzing the behavior of the estimated autocorrelation coefficients (ρ_i 's) and the partial autocorrelation coefficients ϕ_{ii} (related to the ϕ_i and will be defined shortly).

Different ARIMA models will be associated with different behavioral patterns for the true autocorrelation and partial autocorrelation coefficients. These patterns depend upon the values for p and q as well as the corresponding numerical values of φ_i , and θ_i in the model. See the charts in section C.

A summary and preview: the patterns of the autocorrelation and partial autocorrelation coefficients for an AR(p) and MA(q) "appear" as follows

	pac patterns	ac patterns
AR(p) :	p “spikes”, then cut off	decline
MA(q):	decline	q “spikes”, then cut off
ARMA(p,q)	decline	decline

Thus, an inspection of the autocorrelation coefficients and partial autocorrelation coefficients can help identify an ARIMA model (p, d, and q) much like fingerprints or DNA can be used to “identify” a suspect. Values for p and q will be selected so that patterns of the corresponding autocorrelation and partial autocorrelation coefficients will be mimic patterns of the observed behavior of the estimated autocorrelation and partial autocorrelation coefficients.

The behavior of the autocorrelation coefficients and partial correlation coefficients corresponding to different values of p and q can be derived mathematically or using a computer program like (on 588 Blackboard)

THEORYTS

THEORYTS generates **true** population auto and partial autocorrelation coefficients corresponding to arbitrary values for p, q, and user provided values for φ_i , θ_i and can be used to provide a useful **tutorial** in the identification process. Sample outputs are included in these notes.

Many programs are available which will **estimate** autocorrelation and partial autocorrelation coefficients corresponding to a series of data.

The following **STATA** command will report the indicated number (#lags) of estimated auto and partial autocorrelation coefficients:

corrgram y, lags(#lags)

Now what are the partial autocorrelation coefficients?

The partial autocorrelation coefficients are denoted by φ_{ii} and are useful in determining the order of the autoregressive process.

ϕ_{ii} is equal to the coefficient ϕ_i if the model has an i th order autoregressive component, AR(i). For example,

$$\begin{aligned}\phi_{11} &= \phi_1 \text{ in an AR(1)} \\ \phi_{22} &= \phi_2 \text{ in an AR(2)} \\ \phi_{33} &= \phi_3 \text{ in an AR(3)} \\ &\vdots \\ \phi_{pp} &= \phi_p \text{ in an AR}(p)\end{aligned}$$

Unfortunately, in applications we do not know the true values of ϕ_{ii} which must be estimated from the data. The partial autocorrelation coefficients can be estimated using (1) OLS to estimate AR(1), AR(2), AR(3), ..., AR(#lags) models or (2) Yule Walker equations which will be discussed later.

The estimated partial autocorrelation coefficients provide a tool in determining the value of p , the number of autoregressive lags in the model. For example, if the model is an AR(2), we would expect to find $\hat{\phi}_{11}$ and $\hat{\phi}_{22}$ to be statistically significant, with $\hat{\phi}_{33}$ being statistically insignificant. More on this later.

Alternative and complimentary approaches to the identification process involve the use of the spectral density function or the specification of an objective function which is optimized over p and q and ϕ and θ . The Akaike information criterion $AIC = -2 \ln(\text{likelihood}) + 2(p+q)$ is an example of this procedure. AIC is then minimized over p and q as well as the coefficients in the ARIMA specification.

After the model has been identified (values for p,d,q selected), the second step involves estimating the specified model.

- (2) Estimation--Given values for p and q , nonlinear estimation techniques can be employed to estimate $\sigma^2 = \text{Var}(\epsilon_t)$, ϕ_1, \dots, ϕ_p , $\theta_1, \dots, \theta_q$. Conditional maximum likelihood or nonlinear least square estimators can be obtained. Some of the associated details are discussed in section E.

The STATA estimation command for estimating an ARIMA(p,d,q) for the variable Y is given by

arima y, arima(p,d, q)

The estimation routine is based upon an equivalent AR(∞) model representations of the errors ϵ_t :

$$\begin{aligned}\epsilon_t &= \Delta^d (Y_t - \mu) - \phi_1 \Delta^d (Y_{t-1} - \mu) - \dots - \phi_p \Delta^d (Y_{t-p} - \mu) \\ &\quad - \theta_q \epsilon_{t-q} \quad \text{or}\end{aligned}$$

$$= \theta^{-1}(L)\varphi(L) \Delta^d (Y_t - \mu)$$

In either representation ε_t depends upon the φ_i 's and θ_i 's (autoregressive and moving average parameters). The associated sum of squared errors is given by

$$\text{SSE}(\varphi, \theta) = \sum \varepsilon_t^2$$

$$= \sum_t \{\Delta^d (Y_t - \mu) - \phi_1 \Delta^d (Y_{t-1} - \mu) - \dots - \phi_p \Delta^d (Y_{t-p} - \mu) + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q}\}^2$$

$$= \sum \{\theta^{-1}(L) \varphi(L) \Delta^d (Y_t - \mu)\}^2.$$

This formulation gives MLE for normally distributed error terms. Alternative formulations based on different probability density functions can be employed.

- (3) Diagnostics--Given that the hypothesized model has been estimated, tests are performed to check the validity of the conjectured model.

- Given estimated values for the φ 's and θ 's , we can obtain estimated residuals from:

$$\varepsilon_t = \theta^{-1}(L)\varphi(L) \Delta^d (Y_t - \mu)$$

- The estimated residuals (ε_t) can then be tested for the existence of underlying patterns. This can be done by checking for patterns in the autocorrelation and partial autocorrelation coefficients associated with the estimated errors. In a correctly specified model, the estimated residuals should be white noise without statistically significant autocorrelation coefficients or partial autocorrelation coefficients. A "Q statistic" provides the basis for a statistical test of the hypothesis that the autocorrelation coefficients of the residuals are zero. Using the Box-Pierce Q-statistic

$$Q = T \sum_{k=1}^p r_k^2 \sim \chi^2(p)$$

or the Box-Ljung Q statistic

$$Q = T(T+2) \sum_{k=1}^p \frac{r_k^2}{T-k} \sim \chi^2(p)$$

- This analysis could also be performed by using the ARIMA command on the residuals and investigating the associated patterns of the auto and partial autocorrelation coefficients.
- Extra AR or MA parameters may be included and then test for the statistical significance of the extra terms using a "t-type" statistic or likelihood ratio test.

(4) Forecasting

The forecasts can be thought of as being generated from the equivalent $MA(\infty)$ representation or Wold decomposition, if it exists, corresponding to the identified and estimated ARIMA model. This form is given by:

$$\begin{aligned}\Delta^d(Y_t - \mu) &= \frac{\Theta(L)}{\Phi(L)} \varepsilon_t \\ &= \Gamma(L) \varepsilon_t = \varepsilon_t + \Gamma_1 \varepsilon_{t-1} + \Gamma_2 \varepsilon_{t-2} + \dots\end{aligned}$$

To illustrate how the estimation can be performed consider the STATA forecasting commands illustrated below:

Consider the Australian unemployment data from <http://www-personal.bus eco.monash.edu.au/~hyndman/TSDL/>. Let the unemployment rate be denoted by "y" and create a variable t which is the observation number. There are 211 observations in the time series.

```
generate t = _n           //this makes t = to the obs number
tsset t                  // sets t as the time variable
arima y, arima(1,0,1)    //this runs arima on the 211 observations
tsappend, add(9)         //this adds 9 obs to the data
predict yhat             //this predicts y for all 220 observations
list y yhat t            //this lists the 220 values for y, yhat and t
```

Alternatively, to estimate the model on a subset of the data (the first 200 observations) and to obtain forecasts on a holdout sample (the eleven observations not used in estimation), we could use the commands

```
arima y, arima(1,0,1), if t<201
predict yhat
list y yhat t
```

The previous material has attempted to give a brief overview of alternative models, their characteristics and identification, estimation, model diagnostics, and

forecasting procedures. Each of these issues will be considered in additional detail in subsequent sections. We turn to a more thorough analysis of a number of simple ARIMA(p,d,q) models—including an investigation of the behavior of the associated autocorrelation coefficients considered.

C. Characteristics of some ARIMA Models and Identification -- more detail

1. An overview and introduction—more details and some illustrations.

In this section we consider the behavior of the autocorrelation (ac) coefficients and partial autocorrelation (pac) coefficients for AR(1), MA(1), AR(p), MA(q), and ARMA(p,0,q) models along with a brief discussion of the Yule-Walker equations which can be used to provide estimates of the pac coefficents. In each case we will review the basic mathematics illustrate the underlying patterns of corresponding autocorrelation and partial autocorrelation coefficients. The reader may only want to review the AR(1) and MA(1) discussion to get an idea of the underlying mechanics.

2. Autoregressive model of order 1 [ARIMA (1,0,0) or AR(1)]

$$Y_t - \phi_1 Y_{t-1} = \varepsilon_t \quad (C.1)$$

where $\varepsilon_t \sim N[0, \sigma^2]$. Through recursive substitution, (C.1) can be rewritten as

$$Y_t = \frac{\varepsilon_t}{1 - \phi_1 L} = \sum_{i=0}^{\infty} \phi_1^i L^i \varepsilon_t = \sum_{i=0}^{\infty} \phi_1^i \varepsilon_{t-i}$$

From this specification, it can be shown that

$$E(y_t) = 0 \quad (C.2)$$

$$\text{Var}(Y_t) = \frac{\sigma^2}{1 - \phi_1^2} = \gamma_0$$

$$\text{Cov}(y_t, y_{t-s}) = (\phi_1^s) \frac{\sigma^2}{1 - \phi_1^2} = \phi_1^s \gamma_0 = \gamma_s$$

details

Details associated with these results:

- $E(Y_t) = E(\varepsilon_t + \phi_1 \varepsilon_{t-1} + \phi_1^2 \varepsilon_{t-2} + \dots) = 0$
- $\text{Var}(Y_t) = \text{Var}(\sum \phi_1^i \varepsilon_{t-i})$
 $= \sum \phi_1^{2i} \text{Var}(\varepsilon_{t-i})$

$$= \sigma^2 \sum \varphi_1^{2i} = \sigma^2 / (1 - \varphi_1^2)$$

- $\gamma_s = E(Y_t Y_{t-s}) = E(\varepsilon_t + \varphi_1 \varepsilon_{t-1} + \dots + \varphi_1^s \varepsilon_{t-s} + \dots)$
- $(\varepsilon_{t-s} + \varphi_1 \varepsilon_{t-s-1} + \dots)$
- $= \varphi_1^s E(\varepsilon_{t-s} + \varphi_1 \varepsilon_{t-s-1} + \dots)^2$
- $= \varphi_1^s \text{Var}(y_{t-s})$
- $$\frac{\varphi_1^s \sigma^2}{1 - \varphi_1^2}$$

$$\rho_s = \gamma_s / \gamma_0 = \varphi_1^s$$

The autocorrelation coefficients

$$\rho_s = \frac{\gamma_s}{\gamma_0} = \varphi_1^s \quad \text{decay exponentially.}$$

Note $\varphi_1 = \rho_1$

The partial autocorrelation coefficients denoted (φ_{ii}) can be shown to be $\varphi_{11} = \rho_1$ and $\varphi_{ii} = 0$ $i > 1$.

Some sample output from the THEORYTS program illustrates these patterns. The reader might experiment with some other values of φ_i . The first example corresponds to a positive value of φ_i and the second illustrates the impact of a negative value of φ_i

Example 1: $\phi_i = .8$

DOS mode: THEORYTS

ENTER THE NUMBER OF AUTOREGRESSIVE PARAMETERS: 1

ENTER THE NUMBER OF MOVING AVERAGE PARAMETERS: 0

ENTER THE NUMBER OF AUTOCORRELATION COEFFICIENTS: 15

ENTER THE VALUE OF PHI(1): .8

****AUTO CORRELATION COEFFICIENTS****

LAGS	VALUES
15	I* 0.035
14	I* 0.044
13	I* 0.055
12	I* 0.069
11	I * 0.086
10	I * 0.107
9	I * 0.134
8	I * 0.168
7	I * 0.210
6	I * 0.262
5	I * 0.328
4	I * 0.410
3	I * 0.514
2	I * 0.640
1	I 0.800

-1 0 +1

ENTER THE NUMBER OF PARTIALS: 15

* * * PARTIAL AUTOCORRELATION COEFFICIENTS * * *

LAGS	VALUES
15	*
14	*
13	*
12	*
11	*
10	*
9	*
8	*
7	*
6	*
5	*
4	*
3	*
2	*
1	*

-1 0 +1

Example 2: $\phi_i = -.8$

ENTER 1 FOR NEW PROCESS. 2 FOR THE SAME
PROCESS: 1

ENTER THE NUMBER OF AUTOREGRESSIVE PARAMETERS:

ENTER THE NUMBER OF MOVING AVERAGE PARAMETERS: 0

ENTER THE NUMBER OF AUTOCORRELATION COEFFICIENTS: 15

ENTER THE VALUE OF PHI(1): -.8

* * * * AUTO CORRELATION COEFFICIENTS * * * *

LAGS

		VALUE
15	*	-0.035
14	I*	0.044
13	*	-0.055
12	I*	0.069
11	*	-0.086
10	I *	0.107
9	*	-0.134
8	I *	0.168
7	*	-0.210
6	I *	0.262
5	*	-0.320
4	I *	0.410
3	*	-0.512
2	I *	0.640
1	*	-0.800

-1

0

1

ENTER THE NUMBERS OF PARTIALS 15

* * * PARTIAL AUTOCORRELATION COEFFICIENTS * * * *

LAGS

		VALUES
15	*	-0.000
14	*	-0.000
13	*	-0.000
12	*	-0.000
11	*	-0.000
10	*	-0.000
9	*	-0.000
8	*	-0.000
7	*	-0.000
6	*	-0.000
5	*	-0.000
4	*	-0.000
3	*	-0.000
2	*	-0.000
1	*	-0.800

-1

0

+1

Notes

- (1) In each case of the previous two cases, the autocorrelation coefficients decline geometrically as φ_1^s . The partial autocorrelation coefficients are all equal to zero except for φ_{11} (the first) which is equal to φ_1 .
- (2) It will also be instructive to note that an AR(1) with $|\varphi_1| < 1$ can be written as an infinite moving average MA(∞). This is the reason that the autocorrelation coefficients decline geometrically.
- (3) If the process is AR(p), then the first p partial autocorrelation coefficients may be nonzero and all others zero. The corresponding autocorrelation coefficients decline to zero.

3. Moving average of order 1, ARIMA (0,0,1) or MA(1)

The MA(1) model is defined by

$$Y_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} = (1 - \theta_1 L) \varepsilon_t = \theta(L) \varepsilon_t$$

The MA(1) model is invertible (can be written as an AR(∞)) if $|\theta_1| < 1$. In particular,

$$\varepsilon_t = \theta^{-1}(L) Y_t = (1 - \theta_1 L)^{-1} Y_t$$

$$\sum_{i=0}^{\infty} \theta^i L^i Y_t$$

$$\sum_{i=0}^{\infty} \theta^i Y_{t-i}$$

or

$$Y_t + \theta_1 Y_{t-1} + \theta_1^2 Y_{t-2} + \dots = \varepsilon_t$$

which is an AR(∞). Looking at the form of the coefficients of the lagged Y_t 's, suggests that the partial autocorrelation coefficients will decline geometrically. This is in fact what happens for a MA(1).

In order to determine the behavior of the autocorrelation coefficients, we derive expressions for ρ_s . From the form for a moving average model we obtain the following

- $E(Y_t) = E(\varepsilon_t - \theta_1 \varepsilon_{t-1}) = E(\varepsilon_t) - \theta_1 E(\varepsilon_{t-1})$
 $= 0$
- $Var(Y_t) = Var(\varepsilon_t - \theta_1 \varepsilon_{t-1})$
 $= Var(\varepsilon_t) + \theta_1^2 Var(\varepsilon_{t-1})$
 $= \sigma^2 + \theta_1^2 \sigma^2$
 $= \sigma^2 (1 + \theta_1^2)$

- $$\begin{aligned} \text{Cov}(Y_t Y_{t-1}) &= E(Y_t Y_{t-1}) \\ &= E(\varepsilon_t - \theta_1 \varepsilon_{t-1})(\varepsilon_{t-1} - \theta_1 \varepsilon_{t-2}) \\ &= -\theta_1 E(\varepsilon_{t-1})^2 \\ &= -\theta_1 \sigma^2 \end{aligned}$$
- $$\begin{aligned} \text{Cov}(Y_t Y_{t-2}) &= E(\varepsilon_t - \theta_1 \varepsilon_{t-1})(\varepsilon_{t-1} - \theta_1 \varepsilon_{t-3}) \\ &= 0 \end{aligned}$$
- $$\begin{aligned} \text{Cov}(Y_t Y_{t-s}) &= E(\varepsilon_t - \theta_1 \varepsilon_{t-1})(\varepsilon_{t-s} - \theta_1 \varepsilon_{t-s-1}) \\ &= 0 \end{aligned}$$

Therefore

$$\rho_s = \begin{cases} -\frac{\theta}{1 + \theta_1^2} & s = 1 \\ 0 & s = 2, 3, \dots \end{cases}$$

These results demonstrates that a moving average process of order 1 (MA(1)) only has a "memory" of one period. Similarly a moving average process of order q (MA(q)) only has a "memory" of q periods. In other words for a MA(q) model $\rho_s = 0$ for $s > q$. The impact of Y_t on the Y 's will completely die out after q periods.

The following computer printout illustrates typical behavior of the autocorrelation and partial autocorrelation coefficients corresponding to a MA(1) with $\theta_1 = .9$. Note that there is only one nonzero autocorrelation coefficient and the partial autocorrelation coefficients decline geometrically. Computational details will be reviewed following the graphs.

Consider the following example of an MA(1) with $\theta_1 = .9$. The autocorrelations coefficients can be shown to be:

$$\rho_i = 0 \quad \text{for } i > 2$$

$$\rho_1 = \frac{-\theta_1}{1 + \theta_1^2}$$

$$= \frac{-0.9}{1 + 0.81} = -0.497$$

The following figures illustrate the corresponding ac and pac coefficients.

Example: $\theta_1 = .9$

ENTER THE NUMBER OF AR PARAMETERS: 0

ENTER THE NUMBER OF MA PARAMETERS: 1

ENTER THE NUMBER OF AC COEFFICIENTS: 15

ENTER THE VALUE OF THETA (1): .9

* * * * AUTO CORRELATION COEFFICIENTS * * * *

LAGS		VALUES
15	*	0.000
14	*	0.000
13	*	0.000
12	*	0.000
11	*	0.000
10	*	0.000
9	*	0.000
8	*	0.000
7	*	0.000
6	*	0.000
5	*	0.000
4	*	0.000
3	*	0.000
2	*	0.000
1	*	-0.497

-1

0

+1

ENTER THE NUMBER OF PAC COEFFICIENTS (15)

* * * PARTIAL AUTOCORRELATION COEFFICIENTS * * * *

LAGS		VALUES
15	* I	-0.041
14	* I	-0.045
13	* I	-0.051
12	* I	-0.057
11	* I	-0.065
10	* I	-0.073
9	* I	-0.084
8	* I	-0.096
7	* I	-0.112
6	* I	-0.131
5	* I	-0.156
4	*	-0.191
3	*	-0.243
2	*	-0.329
1	*	-0.497

-1

0

+1

Important note:

Sampling variation and the possible presence of autoregressive and moving average components makes the identification process more difficult.

We now turn to an analysis of higher order AR and mixed processes.

4. Autoregressive model of order p, ARIMA (p,0,0) or AR(p)

$$Y_t - \varphi_1 Y_{t-1} - \dots - \varphi_p Y_{t-p} = \varepsilon_t \quad (C.3)$$

$$\varphi(B)y_t = \varepsilon_t \quad (C.4)$$

- (a) An AR(p) will be stationary and can be written as a MA(∞) if the roots of $\varphi(z)$ are greater than one in absolute value.

Details: Factoring the polynomical

$$\begin{aligned} \varphi(z) &= 1 - \varphi_1 z - \varphi_2 z^2 - \dots - \varphi_p z^p \\ &= (1 - \tilde{\varphi}_1 z)(1 - \tilde{\varphi}_2 z) \dots (1 - \tilde{\varphi}_p z) \end{aligned}$$

The roots of $\varphi(z)$ are then given by

$$(1 - \tilde{\varphi}_i z) = 0, z = 1/\tilde{\varphi}_i.$$

Now consider

$$\begin{aligned} y_t &= \{1/\varphi(L)\}\varepsilon_t \\ &= \left(\frac{1}{1 - \tilde{\varphi}_1 L} \right) \dots \left(\frac{1}{1 - \tilde{\varphi}_p L} \right) \varepsilon_t \\ &= \left(\sum_{i=0}^{\infty} \tilde{\varphi}_1^i L^i \right) \dots \left(\sum_{j=0}^{\infty} \tilde{\varphi}_p^j L^j \right) \varepsilon_t \\ &= \{1 + \tilde{\varphi}_1 \beta + \tilde{\varphi}_1^2 \beta^2 + \dots\} \dots \{1 + \tilde{\varphi}_p L + \tilde{\varphi}_p^2 L^2 \dots\} \varepsilon_t \\ &= \{1 + (\tilde{\varphi}_1 + \tilde{\varphi}_2 + \dots + \tilde{\varphi}_p)L \end{aligned}$$

$$+ (\tilde{\phi}_1^2 + \tilde{\phi}_2^2 + \dots + \tilde{\phi}_p^2 + \tilde{\phi}_1\tilde{\phi}_2 + \dots + \tilde{\phi}_1\tilde{\phi}_p$$

$$+ \dots + \tilde{\phi}_{p-1}\tilde{\phi}_p) L^2 \} \varepsilon_t + \dots$$

which is valid if the $|\tilde{\phi}| < 1$, i.e., the roots of $\varphi(z)$ are greater than one in absolute value.

(b) $E(Y_t) = 0$ if the series (C.3) is stationary.

(c) Yule Walker equations (an alternative approach to evaluating the pac's)

The relationship between the autocorrelation coefficients and φ_i 's in (C.3) is given by

$$\begin{pmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_p \end{pmatrix} = \begin{pmatrix} 1 & \rho_1 & \dots & \rho_{p-1} \\ \rho_1 & 1 & \dots & \rho_{p-2} \\ \vdots & \vdots & & \\ \rho_{p-1} & \rho_{p-2} & \dots & 1 \end{pmatrix} \begin{pmatrix} \varphi_1 \\ \varphi_2 \\ \vdots \\ \varphi_p \end{pmatrix}$$

or

$$\rho_k = \sum_{j=1}^p \varphi_j \rho_{k-j} \quad \text{for } k > 0$$

(C.5) is referred to as the system of Yule Walker equations. The square matrix on the right hand side of (C.5) is a Toeplitz matrix in the ρ_i 's. The φ_i 's can then be expressed in terms of the ρ_i 's and the ρ_i 's can be expressed in terms of the φ_i 's. For example, for

$$\underline{p=1}: \quad \rho_1 = \varphi_1$$

$$\underline{p=2}: \quad \begin{pmatrix} \rho_1 \\ \rho_2 \end{pmatrix} = \begin{pmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{pmatrix} \begin{pmatrix} \varphi_1 \\ \varphi_2 \end{pmatrix}$$

$$\begin{aligned}\varphi_1 &= \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2}, & \rho_1 &= \frac{\varphi_1}{1 - \rho_2} \\ \varphi_2 &= \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2}, & \rho_2 &= \frac{\varphi_2(1 - \varphi_2) + \varphi_1^2}{1 - \rho_2}\end{aligned}$$

Derivation of the Yule Walker Equations: Multiply (C.3) by y_{t-k} and take the expected value

$$\begin{aligned}E(y_t y_{t-k}) - \varphi_1 E(y_{t-1} y_{t-k}) - \dots - \varphi_p E(y_{t-p} y_{t-k}) \\ = E(\varepsilon_t y_{t-k})\end{aligned}$$

or

$$\gamma_k - \varphi_1 \gamma_{k-1} - \dots - \varphi_p \gamma_{k-p} = 0$$

Dividing by $\gamma_0 = \text{Var}(y_t)$ yields

$$\rho_k - \varphi_1 \rho_{k-1} - \dots - \varphi_p \rho_{k-p} = 0.$$

(d) Partial Autocorrelation Coefficients

If different values of p are selected and the "last coefficient" φ_p obtained for each p , using the Yule Walker equations, these coefficients are referred to as partial autocorrelation coefficients and are useful in determining the order of the autoregressive process. This is analogous to deciding on how many terms to include in a multiple regression. For an autoregressive process of order p , the first p partial autocorrelations will be nonzero and higher order partial autocorrelation coefficients will equal zero. The partial autocorrelation coefficients are denoted by φ_{ii} .

- As an example for $p = 1$, the Yule Walker equations are

$$\rho_1 = \varphi_1 = \varphi_{11}$$

- For $p = 2$, the Yule Walker equations are

$$\begin{pmatrix} \rho_1 \\ \rho_2 \end{pmatrix} = \begin{pmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{pmatrix} \begin{pmatrix} \varphi_1 \\ \varphi_2 \end{pmatrix};$$

therefore, using Cramer's Rule to solve for φ_2 yields

$$\varphi_{22} = \frac{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & \rho_2 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{vmatrix}} = \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2}$$

- For $p = 3$, the Yule Walker equations are

$$\begin{pmatrix} \rho_1 \\ \rho_2 \\ \rho_3 \end{pmatrix} = \begin{pmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_2 \\ \rho_2 & \rho_1 & 1 \end{pmatrix} \begin{pmatrix} \varphi_1 \\ \varphi_2 \\ \varphi_3 \end{pmatrix}.$$

The corresponding third partial autocorrelation coefficient is

$$\varphi_{33} = \frac{\begin{vmatrix} 1 & \rho_1 & \rho_1 \\ \rho_1 & 1 & \rho_2 \\ \rho_2 & \rho_1 & \rho_3 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1 \\ \rho_2 & \rho_1 & 1 \end{vmatrix}}$$

- More generally

$$\varphi_{ii} = \begin{vmatrix} 1 & \rho_1 & \dots & \rho_1 \\ \rho_1 & 1 & \dots & \rho_2 \\ \vdots & \vdots & & \vdots \\ \rho_{i-1} & \rho_{i-2} & & \rho_i \\ \hline 1 & \rho_1 & \dots & \rho_{i-1} \\ \rho_1 & 1 & \dots & \rho_{i-2} \\ \vdots & \vdots & & \vdots \\ \rho_{i-1} & \rho_{i-2} & \dots & 1 \end{vmatrix}$$

Note:

If the actual value of p is 1, then

$$\varphi_{11} = \rho_1$$

$$\varphi_{22} = \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2} = \frac{\rho_1^2 - \rho_1^2}{1 - \rho_1^2} = 0$$

since $\rho_i = \varphi_i^i$ for a first order autoregressive process. Therefore,

$$\varphi_{ii} = 0 \quad i \geq 2.$$

The determination of an appropriate estimate for p becomes a statistical question. The ρ_k can be estimated by

$$\sum_{t=1}^{n-k} (y_t - \bar{y})(y_{t+k} - \bar{y}) / \sum (y_t - \bar{y})^2$$

The Yule-Walker equations can then be used to estimate the corresponding partial autocorrelation coefficients. The associated asymptotic standard errors were shown to be

$$\hat{s}_{\varphi_{kk}} = \frac{1}{\sqrt{T}} \quad k \geq +1$$

by Quenouille. If one assumes that n (number of observations used in fitting) is large enough for $\hat{\varphi}_{kk}$ to be approximately normally distributed, we have a procedure which can be used in determining a reasonable value of p .

Consider the following AR(2) example $(\varphi_1, \varphi_2) = (.2, .7)$ using the Yule Walker
 $\rho_k = \varphi_1 \rho_{k-1} + \varphi_2 \rho_{k-2}$

$$\rho_0 = 1$$

$$\rho_1 = \frac{\varphi_1}{1 - \varphi_2} = \frac{.02}{1 - .7} = \frac{2}{3}$$

$$= .667$$

$$\rho_2 = \frac{\varphi_2(1 - \varphi_2) + \varphi_1^2}{1 - \varphi}$$

$$= \frac{(.7)(.3) + .04}{1 - .7}$$

$$= \frac{.25}{.30} = .833$$

$$\varphi_{11} = \rho_1 = .667$$

$$\varphi_{22} = \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2} = \frac{\left(\frac{5}{6}\right) - \left(\frac{2}{3}\right)^2}{1 - \left(\frac{2}{3}\right)^2} = \frac{21}{30} = .7$$

$$\varphi_{33} = 0$$

Example: $(\phi_1, \phi_2) = (.2, .7)$

ENTER THE NUMBER OF AUTOREGRESSIVE PARAMETERS: 2

ENTER THE NUMBER OF MOVING AVERAGE PARAMETERS: 0

ENTER THE NUMBER OF AUTOCORRELATION COEFFICIENTS: 15

ENTER THE VALUE OF PHI(1): .2

ENTER THE VALUE OF PHI(2): .7

* * * * AUTO CORRELATION COEFFICIENTS * * * *

LAGS		VALUES
15	I	*
14	I	*
13	I	*
12	I	*
11	I	*
10	I	*
9	I	*
8	I	*
7	I	*
6	I	*
5	I	*
4	I	*
3	I	*
2	I	*
1	I	*

-1 0 +1

ENTER THE NUMBER OF PARTIALS 15

* * * PARTIAL AUTOCORRELATION COEFFICIENTS * * * *

LAGS		VALUES
15		-0.000
14		0.000
13		0.000
12	*	0.000
11	*	0.000
10	*	-0.000
9	*	0.000
8	*	-0.000
7	*	-0.000
6	*	0.000
5	*	-0.000
4	*	0.000
3	*	-0.000
2	I	*
1	I	*

-1 0 +1

5. q^{th} order moving average model, ARIMA (0,0,q) or MA(q)

$$y_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \dots - \theta_q \varepsilon_{t-q} \\ = \theta(L) \varepsilon_t \quad (C.8)$$

In order for the process defined by (C.8) to be invertible, the roots of $\theta(L)$ must have modulus greater than one,

$$\theta(z) = \prod_{i=1}^q (1 - \theta_i z) = 0$$

$$\text{Hint: } \theta^{-1}(L) = \prod_{j=1}^q (1 - \tilde{\theta}_j L)^{-1} = \prod_{j=1}^q \sum_{i=0}^{\infty} (\tilde{\theta}_j^i L^i)$$

all j .

The roots of $\theta(z)$ are equal to $1/\tilde{\theta}_j$.

The autocovariances and autocorrelations can be evaluated by considering

$$\gamma_k = E(y_t y_{t-k}) \\ \gamma_k = E[\varepsilon_t - \theta_1 \varepsilon_{t-1} - \dots - \theta_q \varepsilon_{t-q})(\varepsilon_{t-k} - \theta_1 \varepsilon_{t-k-1} - \dots - \theta_q \varepsilon_{t-k-q})] \quad (C.9)$$

$$\gamma_0 = (1 + \theta_1^2 + \dots + \theta_q^2)\sigma^2.$$

$$\gamma_k = (-\theta_k + \theta_1 \theta_{k+1} + \theta_2 \theta_{k+2} + \dots + \theta_q \theta_{q-k})\sigma^2 \quad (C.10)$$

$$\text{for } k = 1, 2, \dots, q$$

$$= 0 \quad \text{for } k > q.$$

$$\rho_k = \frac{-\theta_k + \theta_{k+1}\theta_1 + \dots + \theta_q\theta_{q-k}}{1 + \theta_1^2 + \dots + \theta_q^2}$$

k=1,2,...,q (C.11)

$$= 0 \quad \text{for } k > q$$

From (C.10) we see that the autocorrelation function of a MA(q) has a "cut-off" at lag q. We might say that a MA(q) has a memory of Length q.

Bartlet's approximation for the standard error of estimators of ρ_k is useful in determining an estimate of q. For any process for which the autocorrelation's ρ_k are zero for $k > q$, Bartlet's approximation is given by

$$s_{\hat{\rho}_k}^2 = \left(\frac{1}{T} \right) \left\{ 1 + 2 \sum_{i=1}^q \rho_i^2 \right\} \text{ for } k > q$$

The reader should be reminded that the appropriateness of the convention of using the limiting normal density with (C.7) or (C.12) for purposes of assessing statistical significance is questionable for small samples.

Consider the following example:

The auto and partial autocorrelation coefficients corresponding to

$$\text{MA}(2): \quad y_t = \varepsilon_t - .5\varepsilon_{t-1} - .3\varepsilon_{t-2}$$

are given in the next figure.

$$\begin{aligned} \rho_1 &= \frac{-\theta_1 + \theta_1 \theta_2}{1 + \theta_1^2 + \theta_2^2} &= -.26 \quad \text{and} \\ \rho_2 &= \frac{-\theta_1}{1 + \theta_1^2 + \theta_2^2} &= -.224 \end{aligned}$$

Example of an MA(2) with $(\theta_1 = .5, \theta_2 = .3)$

ENTER THE NUMBER OF AUTOREGRESSIVE PARAMETERS: 0

ENTER THE NUMBER OF MOVING AVERAGE PARAMETERS: 2

ENTER THE NUMBER OF AUTOCORRELATION COEFFICIENTS: 15

ENTER THE VALUE OF THETA(1): .5

ENTER THE VALUE OF THETA(2): .3

* * * AUTO CORRELATION COEFFICIENTS * * *

LAGS		VALUES
15	*	0.000
14	*	0.000
13	*	0.000
12	*	0.000
11	*	0.000
10	*	0.000
9	*	0.000
8	*	0.000
7	*	0.000
6	*	0.000
5	*	0.000
4	*	0.000
3	*	0.000
2	*	I -0.224
1	*	I -0.261

-1

0

+1

ENTER THE NUMBER OF PARTIALS: 15

* * * PARTIAL AUTOCORRELATION COEFFICIENTS * * *

LAGS		VALUES
15	*I	-0.023
14	* I	-0.027
13	* I	-0.032
12	* I	-0.037
11	* I	-0.044
10	* I	-0.052
9	* I	-0.062
8	* I	-0.074
7	*	I -0.088
6	*	I -0.107
5	*	I -0.127
4	*	I -0.165
3	*	I -0.189
2	*	I -0.313
1	*	I -0.261

-1

0

+1

6. Autoregressive Moving Average Processes, ARIMA (p,0,q), ARMA (p,q)

$$Y_t - \phi_1 Y_{t-1} - \dots - \phi_p Y_{t-p} = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \dots - \theta_q \varepsilon_{t-q}$$

$$\phi(L)Y_t = \phi(L)\varepsilon_t$$
(C.13)

The process defined by (C.13) will be stationary if the roots of $\phi(L)$ have modulus greater than one and will be invertable if the roots of $\theta(L)$ have modulus greater than one.

Given that the conditions of stationarity and invertibility are satisfied, we note that an ARMA (p,q) process can be expressed as

$$AR(\infty): \theta^{-1}(L)\phi(L)Y_t = \varepsilon_t$$
(C.14)

$$MA(\infty): Y_t = \phi^{-1}(L)\theta(L)\varepsilon_t$$
(C.15)

Multiplying (C.13) by y_{t-k} and taking expected values we see that

$$\rho_k = \varphi_1 \rho_{k-1} + \dots + \varphi_p \rho_{k-p} + \rho_k(Y, u) - \theta_1 \rho_{k-1}(Y, u)$$

$$- \theta_q \rho_{k-q}(Y, u)$$
(C.16)

where $\rho_i(Y, u) = E(Y_{t-i}, u_t) = 0 \quad i > 0$.

(C.16) simplifies to

$$\rho_k = \varphi_1 \rho_{k-1} + \varphi_2 \rho_{k-2} + \dots + \varphi_p \rho_{k-p} \text{ for } k \geq q+1$$

The first q autocorrelation coefficients will depend upon the moving average parameters θ_i as well as the autoregressive parameters φ_j . Therefore, the autocorrelation coefficients will exhibit an irregular pattern at lags 1 through g , then tail off according to (C.16).

The process (C.13) is equivalent to an $AR(\infty)$; hence, the partial autocorrelation coefficients tail off eventually in the same manner as with a pure autoregressive process.

These results will assist in the determination of p, q . The following tables are taken from Nelson (p. 89) and Box and Jenkins (176,7) and provide useful summary information to assist in the determination of p and q . It should be noted that the autocorrelation coefficients and partial autocorrelations provide the basis for this determination. Recall that the asymptotic standard errors of $\hat{\rho}_k$ and $\hat{\varphi}_{kk}$ are given by

$$s_{\hat{\rho}_k} = \left(\frac{1}{\sqrt{T}} \right) \left\{ 1 + 2(\rho_1^2 + \dots + \rho_q^2) \right\}^{1/2}$$

$$s_{\hat{\varphi}_{kk}} = \frac{1}{\sqrt{T}} \quad \text{for } k > p.$$

Table 1. Characteristic Behavior of Autocorrelations and Partial Autocorrelations for Three Classes of Processes.

Class of processes	Autocorrelations	Partial autocorrelations
Moving average(q)	Spikes at lags 1 through q, then cut off	Tail off
Autoregressive (p)	Tail off according to $\rho_j = \phi_1\rho_{j-1} + \dots + \phi_p\rho_{j-p}$	Spikes at lags 1 through p, then cut off
Mixed auto-regressive-moving average	Irregular pattern at lags 1 through q, then tail off according to $\rho_j = \phi_1\rho_{j-1} + \dots + \phi_p\rho_{j-p}$	

Table 2. Behavior of the autocorrelation functions for the dth difference of an ARIMA process of order (p,d,q). (Table A and Charts B, C, and D are included at the end of this volume to facilitate the calculation of approximate estimates of the parameters for first-order moving average, second-order autoregressive, second-order moving average, and for the mixed (ARMA) (1,1) process.)

Order	(1,d,0)	(0,d,1)
Behavior of ρ_k	decays exponentially	only ρ_1 nonzero
Behavior of ϕ_{kk}	only ϕ_{11} nonzero	decay exponentially
Preliminary estimates	$\phi_1 = \rho_1$	$\rho_1 = \frac{\theta_1}{1 + \theta^2}$
Admissible region	$-1 < \phi_1 < 1$	$-1 < \theta_1 < 1$
Order	(2,d,0)	(0,d,2)
Behavior of ρ_k	mixture of exponentials or damped sine wave	only ρ_1 and ρ_2 nonzero
Behavior of ϕ_{kk}	only ϕ_{11} and ϕ_{22} are nonzero	dominated by mixture of exponentials or damped sine wave
Preliminary estimates	$\phi_1 = \frac{\rho_1(1 - \rho_2)}{1 - \rho_1^2}$ $\phi_2 = \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2}$	$\rho_1 = \frac{-\theta_1(1 - \theta_2)}{1 + \theta_1^2 \theta_2^2}$ $\rho_2 = \frac{\theta_2}{1 + \theta_1^2 \theta_2^2}$
Admissible region	$-1 < \phi_2 < 1$	$-1 < \theta_2 < 1$
	$\phi_2 + \phi_1 < 1$	$\theta_2 + \theta_1 < 1$
	$\phi_2 - \phi_1 < 1$	$\theta_2 - \theta_1 < 1$

Order	(1,d,1)
Behavior of ρ_k	decays exponentially from first lag
Behavior of ϕ_{kk}	dominated by exponential decay from first lag
Preliminary estimates from	$\rho_1 = \frac{(1-\theta_1\phi_1)(\phi_1-\theta_1)}{1 + \theta_1^2 - 2\phi_1\theta_1}$
Admissible region	$-1 < \phi_1 < 1 \quad -1 < \theta_1 < 1$

- (5) ARIMA (p,d,q) The foregoing discussion has assumed that the underlying series is stationary. If this is not the case, for any of several reasons, the previous approaches are not strictly appropriate. For example if a time series didn't have a fixed mean, then the series wouldn't be stationary. It will frequently be the case that $Y_t - Y_{t-1} = (I-L)Y_t$, or $(I-L)^2Y_t$, or $(I-L)^dY_t$, will be stationary. If such a value of d can be determined then the previous techniques can be applied to the "differenced" series, $(I-L)^dY_t$. For example if Y_t is basically distributed with constant variance about a linear (quadratic) trend, then $(I-L)Y_t$, $(I-L)^2Y_t$, will be stationary. If Y_t shows evidence of trend stationarity, then Y could be regressed on a polynomial in "t" and the previously described techniques can be applied to the resulting residuals. Sometimes nonlinear transformation on Y_t , such as $\ln Y_t$, may facilitate the search for a stationary process.

D. Diagnostic Analysis

There are several approaches which can be utilized to determine the "validity" of the estimated ARIMA models. Three of the most common involve : (1) considering more general models, (2) an analysis of estimated residuals, and (3) the Q-statistic.

1. Generalized Model. Assume that ARIMA(p,d,q) has been "identified." The researcher might estimate an ARIMA(p',d,q') where p' and/or q' are larger than p, q and then check the statistical significance of the additional coefficients. This approach has at least two limitations. The validity of the statistical inference (test statistics) is questionable for small samples and an ARIMA (p,d,q) process is uniquely determined by the autocorrelation structure up to a multiple of polynomials in L. t "type" statistics or likelihood ratio tests may be used.

2. Analysis of Estimated Residuals

The estimated residuals can be obtained from the stimated ARIMA model as follows:

$$\hat{\epsilon}_t = \theta^{-1}(L) \hat{\phi}(L)(I-L)^d y_t$$

One might consider an analysis of the behavior of autocorrelation coefficients

$$\hat{\rho}_k(\hat{\varepsilon}_t) = \sum_t \hat{\varepsilon}_t \hat{\varepsilon}_{t-k} / \sum_t \hat{\varepsilon}_t^2$$

and partial autocorrelation coefficients corresponding to the estimated residuals.

It should be mentioned that the distributional characteristics of $\hat{\varepsilon}_t$ are not necessarily exactly the same as for ε_t . If the model has been correctly specified, the estimated residuals and the associated auto and partial autocorrelation coefficients should correspond to white noise. However, if the estimated residuals appear to have an AR or MA component, then the model specification should be respecified. For example, if the $\hat{\varepsilon}_t$'s appear to be an AR(1), then one more AR component should be included in the specification for the Y_t series.

3. Q-Statistics.

Box and Pierce and Box and Ljung, respectively, define the following Q-statistics

$$Q = n \sum_{k=1}^M \hat{\rho}_k^2(\hat{\varepsilon}_t)$$

$$Q = T(T+2) \sum_{k=1}^M \frac{\hat{\rho}_k^2}{T-k}$$

which can be used to test the hypothesis that the autocorrelations coefficients are zero. The two Q-statistics are asymptotically distributed as $\chi^2(M-p-q)$ under the hypotheses of ε_t being independently and identically distributed as $N(0, \sigma^2)$. This follows because $\hat{\rho}_k \sim N[0, 1/T^5]$ if the model is correctly specified. The hypothesis is rejected by large values of Q.

E. Estimation

1. Background. Once values for p and q have been determined, the coefficients in the ARIMA (p,d,q) need to be estimated in order to use the model.

$$(1-\phi_1 L - \dots - \phi_p L^p) Y_t = (1-\theta_1 L - \dots - \theta_q L^q) \varepsilon_t \quad (\text{E.1})$$

or

$$\phi(L) Y_t = \theta(L) \varepsilon_t$$

Note that ε_t can be explicitly expressed as

$$\begin{aligned} \varepsilon_t &= Y_t - \phi_1 Y_{t-1} - \dots - \phi_p Y_{t-p} + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q} \\ &= \varphi(B) Y_t + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q} \end{aligned} \quad (\text{E.2})$$

and also

$$= \theta^{-1}(L)\varphi(L)Y_t \quad (E.3)$$

Note that in both representations, ε_t depends upon the φ_i 's and θ_i 's (autoregressive and moving average parameters). The associated sum of squared errors is given by

$$\begin{aligned} SSE(\varphi, \theta) &= \sum \varepsilon_t^2 \\ &= \sum_t (Y_t - \varphi_1 Y_{t-1} - \varphi_2 Y_{t-2} - \dots - \varphi_p Y_{t-p} + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q})^2 \\ &= \sum \{\theta^{-1}(L) \varphi(L) Y_t\}^2 \end{aligned} \quad (E.4)$$

Under the assumption that the ε_t 's are independently and identically distributed as $N(0, \sigma^2)$ the log likelihood function is

$$\begin{aligned} l(\theta, \varphi, \sigma^2) &= \ln \frac{e^{-\sum \varepsilon_t^2 / 2\sigma^2}}{(2\pi)^{N/2} (\sigma^2)^{N/2}} \\ &= \frac{-SSE(\varphi, \theta)}{2\sigma^2} - \frac{N}{2} \ln(2\pi) - \frac{N}{2} \ln(\sigma^2) \end{aligned}$$

Hence, minimizing $SSE(\varphi, \theta)$ with respect to φ and θ is a necessary part of obtaining maximum likelihood estimators and the normality assumption.

2. Numerical optimization.

A close inspection of the first expression for SSE in (E.4) reveals that ε_t depends on the previous random disturbances and observations on the variable Y , e.g.

$$\varepsilon_1 = Y_1 - \varphi_1 Y_0 - \varphi_2 Y_{-1} - \dots - \varphi_p Y_{1-p} + \theta_1 \varepsilon_0 + \dots + \theta_q \varepsilon_{1-q}$$

Several approaches to these problems have been proposed. This is frequently referred to as the question as to how to initialize the series. One approach is to replace the unobservable values of Y_t and ε_t by their expected values--in this case 0; hence

$$\varepsilon_1 = Y_1$$

$$\varepsilon_2 = Y_2 - \varphi_1 Y_1 + \varphi_1 \varepsilon_1$$

$$\varepsilon_3 = Y_3 - \varphi_1 Y_2 - \varphi_2 Y_1 + \varphi_1 \varepsilon_2 + \varphi \varepsilon_1$$

The associated sum of squared error terms is

$$\sum_{t=1}^N \varepsilon_t^2$$

Another approach is to start the sum at $t = p + 1$

$$SSE = \sum_{t=p+1}^N (y_t - \phi_1 y_{t-1} - \dots - \phi_p y_{t-p} + \theta_1 \varepsilon_{t-1} \dots + \theta_q \varepsilon_{t-q})^2$$

where $\varepsilon_p, \varepsilon_{p-1}, \dots, \varepsilon_{p+1-q}$ are set equal to zero.

The minimization of (E.4) requires a nonlinear optimization routine if there is a moving average component to the series, i.e., some of the θ_i 's are nonzero. *If there is not a moving average component, (all θ_i 's = 0), then a regular linear regression package can be used in the estimation process. A number of studies have found that autoregressive models perform very well. Recall that an ARIMA(p,d,q) model can be expressed as an AR(∞).

Nonlinear optimization routines require that initial estimates of the parameters be provided. The Yule Walker equations are frequently used for this purpose.

$$\begin{pmatrix} \phi_1 \\ \vdots \\ \phi_p \end{pmatrix} = \begin{pmatrix} 1 & \rho_1 & \dots & \rho_{p-1} \\ \rho_1 & 1 & & \cdot \\ \cdot & \cdot & \ddots & \cdot \\ \rho_{p-1} & \dots & \dots & 1 \end{pmatrix} \begin{pmatrix} \rho_1 \\ \vdots \\ \rho_p \end{pmatrix}$$

where ρ_i is estimated by

$$\hat{\rho}_i = \frac{\sum_t (y_t - \bar{y})(x_t - \bar{x})/n}{\sum_t (y_t - \bar{y})^2/n} = \frac{\text{cov}(y_t, y_{t-i})}{\text{var}(y_t)}$$

3. Nonnormal data

MLE estimates of the ARIMA model can be obtained by maximizing the log-likelihood function

$$\ell(\) = \sum_{t=1}^n \ell n f(\Theta^{-1}(L) \Phi^{-1}(L) \Delta^d Y_t; \psi)$$

over the ARIMA and distributional parameters.

F. Forecasts

1. Forecasts

Let $Y_n(h)$ denote a forecast of Y_{n+h} made at time $t = n$. Assume that Y_t is an ARIMA(p,d,q) stochastic process, i.e.,

$$\varphi(L)Y_t = \theta(L)\varepsilon_t \quad (F.1)$$

Y_t can be expressed in terms of ε_t as

$$\begin{aligned} Y_t &= \frac{\theta(L)}{\varphi(L)} \varepsilon_t \\ &= \Gamma(L) \varepsilon_t \\ &= 1\varepsilon_t + \Gamma_1\varepsilon_{t-1} + \Gamma_2\varepsilon_{t-2} + \dots \end{aligned} \quad (F.2)$$

From (F.2) we can express Y_{n+h} as

$$Y_{n+h} = \varepsilon_{n+h} + \Gamma_1\varepsilon_{n+h-1} + \dots + \Gamma_{h-1}\varepsilon_{n+1} + \Gamma_h\varepsilon_n + \Gamma_{h+1}\varepsilon_{n-1} + \dots \quad (F.3)$$

Unknown at $t=n$	Can be estimated at $t=n$
------------------	------------------------------

It can be shown that the optimal (minimum mean squared error) forecast of Y_{N+h} at time N is given by

$$\begin{aligned} Y_n(h) &= \Gamma_h\varepsilon_n + \Gamma_{h+1}\varepsilon_{n-1} + \Gamma_{h+2}\varepsilon_{n-2} + \dots \\ &= \sum_{j=0}^{\infty} \Gamma_{h+j} \varepsilon_{n-j} \end{aligned} \quad (F.4)$$

A recurrence relationship can be developed from (F.3) which facilitates evaluation of forecasts.

$$\begin{aligned} Y_n(h) &= \sum_{i=1}^p \varphi_i Y_{n-(h-i)} + \sum_{i=0}^{\infty} \theta_{i+h} \varepsilon_{n-i} \\ &= \varphi_1 Y_n(h-1) + \dots + \varphi_p Y_n(h-p) \\ &\quad + \theta_h \varepsilon_n + \theta_{h+1} \varepsilon_{n-1} + \dots + \theta_{q+h} \varepsilon_{n-q} \end{aligned}$$

As an example consider forecasts corresponding to an **AR(1) model**.

$$Y_n(h) = \varphi_1 Y_n(h-1)$$

$$Y_n(1) = \varphi_1 Y_n$$

$$Y_n(h) = \varphi_1^h Y_n$$

2. Forecast errors and confidence intervals.

The forecast error is defined by

$$e_N(h) = Y_{n+h} - Y_n(h) \quad (F.6)$$

$$= \sum_{i=0}^{\infty} \Gamma_i \varepsilon_{n+h-i} - \sum_{j=0}^{\infty} \Gamma_{h+j} \varepsilon_{n-j}$$

$$= \varepsilon_{n+h} + \Gamma_1 \varepsilon_{n+h-1} + \dots + \Gamma_{h-1} \varepsilon_{n+1}$$

(see equations F.3 and F.4), where Y_{n+h} and $Y_n(h)$, denote the actual forecast at $t+n$, respectively.

The variance of the forecast error is given by

$$\text{Var}(e_N(h)) = \text{Var}(\varepsilon_{n+h} + \dots + \Gamma_{h-1} \varepsilon_{n+1}) \quad (F.7)$$

$$= \text{Var}(\varepsilon_{n+h}) + \dots + \Gamma_{h-1}^2 \text{Var}(\varepsilon_{n+h})$$

$$= \sigma_\varepsilon^2 (1 + \Gamma_1^2 + \dots + \Gamma_{h-1}^2)$$

$$\text{Var}(e_n(h)) = \sigma_\varepsilon^2 (1 + \Gamma_1^2 + \dots + \Gamma_{h-1}^2) \quad (F.7)$$

Note: (1)

That the variance of the forecast error increases as the lead time increases

(2) σ_ε^2 can be estimated by

$$\hat{\sigma}_\varepsilon^2 = \frac{\text{SSE}}{N-p-q}$$

(3) "Asymptotic" confidence intervals are given by

$$Y_n(h) \pm Z\alpha \sqrt{\hat{\sigma}_\varepsilon^2 (\Gamma_0^2 + \dots + \Gamma_{h-1}^2)}$$

The forecasts will eventually converge to the "trend."

(4) The expression for the variance of the forecast error in (F.7) doesn't take account of parameter uncertainty.

G. General Comments

1. It should be mentioned that

$$\phi(L)(I-L)^d y_t = \theta(L) \varepsilon_t$$

is uniquely determined by a given autocorrelation structure (given d) up to a multiple of a polynomial in L . For example,

$$Y_t - .5Y_{t-1} = \varepsilon_t \text{ and}$$

$$Y_t - .75Y_{t-1} + .125Y_{t-2} = (1 - .5L)(1 - .25L)Y_t = \varepsilon_t - .25\varepsilon_{t-1}$$

Are observationally equivalent. This suggests working with the simplest possible model, the Principle of Parsimony.

2. If the time series exhibits seasonal behavior [large ρ_{12} (monthly) or ρ_4 (quarterly)] then the previously developed model can be modified to incorporate this behavior into the modeling process. Seasonal components can be incorporated into an ARIMA model in several different ways.

- $\Phi(L)(1-L)^d(Y_t - \mu) = \Theta(L)\varepsilon_t \quad \Delta = I - L \quad \text{basic model}$
- $\Phi(L)(1-L)^d(1-L^s)(Y_t - \mu) = \Theta(L)\varepsilon_t$
- $\Phi_s(L)\Phi(L)(1-L)^d(Y_t - \mu) = \Theta_s(L)\Theta(L)\varepsilon_t$
Seasonal differencing with seasonal AR and MA components

STATA can accommodate these variations of the basic ARIMA model.

3. Just a reminder: the simple exponential smoothing is an ARIMA(0,1,1) model and the Holt-Winters nonseasonal predictor is an ARIMA (0,2,2) model.

4. A number of texts suggest that time series models based on ARIMA formulations should only be expected to be "successful" if at least 40 observations are available. For smaller samples, use other techniques—such as exponential smoothing or Holt Winters. . .

5. A complementary method of analysis is that of spectral analysis. The spectral density of a stationary series is given by

$$g(f) = 2 \left\{ 1 + 2 \sum_{k=1}^{\infty} \rho_k \cos 2\pi fk \right\}$$

$$0 \leq f \leq 1/2$$

In practice this is estimated using "lag windows" by

$$g(f) = 2 \left\{ 1 + 2 \sum_{k=1}^{n-1} \lambda_k \rho_k \cos 2\pi f k \right\}$$

The spectral density function provides information about the cyclical behavior of a series and shows how the variance of a stochastic process is distributed between a continuous range of frequencies.

Spectral Density Functions for

$$\text{ARMA}(p,q) \quad \varphi(L)y_t = \theta(L)\varepsilon_t$$

$$y_t = \varphi^{-1}(L)\theta(L)\varepsilon_t$$

The spectral density function is given by

$$f(u) = \sigma^2 \frac{\theta(e^{iu})\theta(e^{-iu})}{\phi(e^{iu})\phi(e^{-iu})}$$

$$-\pi < u < \pi$$

DeMoivre's Theorem.

$$e^{i\theta} = \cos\theta + i \sin\theta$$

$$(\cos\theta + i \sin\theta)^m = e^{i\theta m} = \cos m\theta + i \sin m\theta$$

<u>Model</u>	Spectral Density Function
ARMA(0,0)	σ_{ε}^2
AR(1)	$\sigma_{\varepsilon}^2 / (1 - 2\varphi_1 \cos \omega + \sigma_1^2)$
AR(2)	$\sigma^2 / [1 + \varphi_1^2 + \sigma_2^2 - 2\varphi_1(1-\varphi_2)\cos\omega - 2\varphi_2 \cos 2\omega]$
MA(1)	$\sigma^2(1 + \theta_1^2 - 2\theta_1 \cos \omega)$
MA(2)	$\sigma^2(1 + \theta_1^2 + \theta_2^2 - 2\theta_1(1-\theta_2)\cos \omega - 2\theta_2 \cos 2\omega)$

$$\text{ARMA}(1,1) \quad \frac{\sigma^2(1 + \theta_1^2 - 2\theta_1 \cos\omega)}{(1 + \varphi_1^2 - 2\varphi_1 \cos\omega)}$$

$$\text{ARMA}(1,2) \quad \frac{\sigma^2(1 + \theta_1^2 + \theta_2^2 - 2\theta_1(1-\theta_2)\cos\omega - 2\theta_2(\cos 2\omega))}{(1 + \varphi_1^2 - 2\varphi_1 \cos\omega)}$$

$$\text{ARMA}(2,1) \quad \frac{\sigma^2(1 + \theta_1^2 - 2\theta_1 \cos\omega)}{1 + \varphi_1^2 + \varphi_2^2 - 2\varphi_1(1-\varphi_2)\cos\omega}$$

$$\text{ARMA}(2,2) \quad \frac{\sigma^2(1 + \theta_1^2 + \theta_2^2 - 2\theta_1(1-\theta_2)\cos\omega - 2\theta_2\cos 2\omega)}{1 + \varphi_1^2 + \varphi_2^2 - 2\varphi_1(1-\varphi_2)\cos\omega - 2\varphi_2\cos\omega}$$

APPENDIX: COMPUTER PROGRAMS (TIME SERIES ANALYSIS)

I. THEORYTS—look at ac and pac

ENTER THE NUMBER OF AUTOREGRESSIVE PARAMETERS: 2

ENTER THE NUMBER OF MOVING AVERAGE PARAMETERS: 0

ENTER THE NUMBER OF AUTOCORRELATION COEFS:25

ENTER THE NUMBER OF PARTIAL AUTO CORRELATION COEFFICIENTS: 25

ENTER THE VALUE OF PHI(1): .2

ENTER THE VALUE OF PHI(2): .7

II. RUN SHR:DATATSF—generates ARIMA series

HOW MANY SERIES DO YOU WANT? 1

HOW MANY OBSERVATIONS ARE TO BE IN EACH SERIES? 300

WHAT ARE THE AR PARAMETERS? (UP TO TWO) .9

WHAT ARE THE MA PARAMETERS? (UP TO TWO)

DO YOU WANT THE SIBYL RUNNER HEADING? N

WHAT IS THE SEASONAL AR PARAMETER? 0

WHAT IS THE SEASONAL MA PARAMETER?

HOW MANY TIMES SHOULD THE SERIES BE SUMMED? 0

HOW MANY TIMES SEASONALLY SUMMED? 0

WHAT IS THE SPAN OF SEASONALITY? 12

WHAT IS THE MEAN OF THE SERIES? 0

WHAT TYPE OF ERROR TERMS DO YOU WANT?

(N=NORMAL, L=LOG-NORMAL, P-PARETO, E=EXPONENTIAL) N

WHAT VALUE OF SIGMA DO YOU WANT? 1. **(remember the decimal)**

STOP

III. RUN SHR:BYUTSF-estimates ARIMA models with different pdf's

* * * * BYU TIME SERIES FORECASTING PACKAGE * * * *

WOULD YOU LIKE TO TEST LEVELS OF DIFFERENCING? N (see next section,
used for identification)

WHAT FORECASTING TECHNIQUE WOULD YOU LIKE TO RUN?
FOR HELP TYPE HELP

THE FOLLOWING FORECASTING TECHNIQUES ARE AVAILABLE

BOXJN	=	BOX-JENKINS MODEL
EXPS	=	SINGLE PARAMETER EXPONENTIAL SMOOTHING
WINTR	=	WINTER'S 3-PARAMETER MODEL
MARM	=	TRANSFER FUNCTIONS (MARMA MODELS)

WHICH OF THESE WOULD YOU LIKE TO RUN? BOXJN

WHAT IS THE DATA FILE NAME? SER1.DAT

HOW MANY OBSERVATIONS ARE TO BE USED? 300

IS THIS A SIBYL-RUNNER FILE? N

WOULD YOU LIKE TO TEST LEVELS OF DIFFERENCING? N

WHAT IS THE DATA FILE NAME? L.DAT

HOW MANY OBSERVATIONS ARE TO BE USED? 50

IS THIS A SIBYL-RUNNER FILE? N

DO YOU WANT A GRAPH OF THE DATA? N

HOW MANY TIMES MUST THE SERIES BE DIFFERENCED? 1

HOW MANY TIMES MUST THE SERIES BE SEASONALLY DIFFERENCED?

WHAT IS THE SPAN OF SEASONALITY?

DO YOU WANT TO SEE GRAPH OF THE NEW DATA? N

WOULD YOU LIKE TO SEE THE SPECTRUM? N

HOW MANY AUTOCORRELATION COEFFICIENTS ARE TO BE SEEN? 10

HOW MANY PARTIAL AUTO'S ARE TO BE SEEN? 10

WOULD YOU LIKE TO REPEAT THIS PROCEDURE? N

VIII. Dynamic models revisited: VAR

- 1. Introduction**
- 2. Multivariate autoregression moving average models**
- 3. Dynamic structural models-as a special case**
- 4. Vector autoregression representations (VAR)**
- 5. Vector moving average representations (transfer functions)**
- 6. Impulse response functions**
- 7. Identification**
- 8. Statistical inference**
- 9. A review and applications**
- 10. Exercises**

James B. McDonald
Brigham Young University

VIII. Dynamic Econometric Models Revisited: VAR's

1. Introduction

This section will consider multivariate autoregressive moving average models for the vector $Z_t = \begin{pmatrix} Y_t \\ X_t \end{pmatrix}$. This formulation includes the dynamic structural models discussed in section (VI.2), as well as showing their relationship to vector autoregression (VAR) formulations, and lays the foundation for exogeneity tests.

2. Multivariate Autoregressive Moving Average Representation

Let $Z_t = \begin{pmatrix} Y_t \\ X_t \end{pmatrix}$ denote a multivariate autoregressive moving average (MARMA)

process which can be written as

$$F(L) Z_t = G(L) \varepsilon_t$$

(2.1)

or using partitioned matrices as

$$\begin{pmatrix} F_{11}(L) & F_{12}(L) \\ F_{21}(L) & F_{22}(L) \end{pmatrix} \begin{pmatrix} Y_t \\ X_t \end{pmatrix} = \begin{pmatrix} G_{11}(L) & G_{12}(L) \\ G_{21}(L) & G_{22}(L) \end{pmatrix} \begin{pmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{pmatrix} \quad (2.1)$$

where Y_t and X_t denote $G \times 1$ and $K \times 1$ vectors of variables, respectively.

3. Dynamic Structural Equation Representation as a Special Case

Consider the special case of equation (2.1) corresponding to

$$F_{11}(L) = B(L), F_{12}(L) = \Gamma(L)$$

$$F_{21}(L) = 0, F_{22}(L) = \Phi(L)$$

$$G_{11}(L) = I, G_{12}(L) = 0$$

$$G_{21}(L) = 0, G_{22}(L) = \theta(L)$$

which yields

$$\begin{aligned} B(L)Y_t + \Gamma(L)X_t &= \epsilon_{t1} \\ \phi(L)X_t &= \theta(L)\epsilon_{t2} \end{aligned}$$

(3.1) is a generalization of the model considered in (VI.2) in which Y_t is endogenous and X_t is exogenous. There are G structural equations with G endogenous variables.

Zellner's condition for dynamic structural econometric models, $F_{21}(L) = 0$ with $G_{21}(L) = 0$ and $G_{12}(L) = 0$, provides the basis for exogeneity tests to be considered later.

Dynamic structural econometric models are sometimes criticized because of

“exogeneity” assumptions and the ad hoc nature of some of the “exclusion restrictions”(deleting some variables from the structural equations) used to identify the model. These concerns have lead to the consideration of vector autoregressive (VAR) models. VAR models provide a method for testing exogeneity and forecasting, but are also associated with estimation and identification problems when used for economic analysis. Two tests for exogeneity developed by Granger (*Investigating Causal Relations by Econometric Models and Cross -Spectral Methods*, *Econometrica*, 1969) and Sims (*Money Income and Causality*, *AER*,1972) are briefly summarized then alternative representations of dynamic econometric models are discussed.

Test 1. Regress Y on lagged values of Y and lagged values of X and then test for the collective explanatory power of the lagged X's. If the lagged X's are not statistically significant, then X is said to fail to "Granger-cause" Y. "Granger-causality" implies the rejection of the collective explanatory power of the X's.

Test 2. Regress Y on current levels of X, past levels of X, and future levels of X. The hypothesis of the coefficients of future X's being equal to zero is consistent with X being "exogenous" to Y for failing to "Granger-cause" Y. Thus if "causality" runs one way, from X to Y, we would expect the coefficients of the "future" X's to all be zero.

Simple F tests can be used for the Granger and Sims tests. Granger-Sims causality tests based on Test 1 are built into the STATA package (version 10) and can be performed by typing the command "**vargranger**" on a line following a VAR estimation command (**var** or **svar**).

This command will be discussed again later.

In previous discussions of univariate time series models (section VII), alternative moving average (MA) and autoregressive (AR) representations were often helpful in model analysis. This will also be true with dynamic econometric models, vector autoregression models, and in exploring the implications of notions of "causality" and "exogeneity."

4. Autoregressive Representation: VAR's.

The multivariate autoregressive representation can be derived from (3.1) and is given by

$$\mathbf{B}(L)Y_t + \Gamma(L)X_t = \varepsilon_{t1} \quad (4.1)$$

$$\mathbf{B}_0 Y_t + \mathbf{B}_1 Y_{t-1} + \dots + \mathbf{B}_p Y_{t-p} + \Gamma(L)X_t = \varepsilon_{t1} \quad (4.2)$$

$$\mathbf{B}_0 Y_t + \mathbf{B}_1 Y_{t-1} + \dots + \mathbf{B}_p Y_{t-p} = \lambda + \varepsilon_{t1} \quad (4.3)$$

where $\lambda = -\Gamma(L)X_t$. Often the “exogenous” variables are “suppressed” in the model

and included in the λ term. More formally, the exogenous variables might be explicitly included in the model as in (4.2). Enders (1995, p.295) refers to this form as the **structural vector autoregressive form** (structural VAR form or SVAR). This representation can include current and lagged values of Y_t and X_t in each equation; hence, **OLS estimation of the VAR form would yield biased and inconsistent estimators.**

The structural error terms, ε_{it} 's, are viewed as being pure innovations or structural shocks which are uncorrelated, both across time and across equations; hence, **Var (ε_{t1}) = Ω is assumed to be a diagonal matrix.** Each structural VAR equation is associated with one innovation.

The **reduced form vector autoregressive representation** (VAR) can be obtained by premultiplying equation (4.2) by \mathbf{B}_0^{-1} and then expressing Y in terms of predetermined variables to yield

$$Y_t = -B_0^{-1}B_1 Y_{t-1} - \dots - B_0^{-1}B_p Y_{t-p} - B_0^{-1}\Gamma(L)X_t + B_0^{-1}\varepsilon_{t1}$$

$$Y_t = \Pi_1 Y_{t-1} + \dots + \Pi_p Y_{t-p} + \mu + \eta_t \quad (4.5)$$

$$Y_t = \Pi(L)Y_t + \mu + \eta_t \quad (4.6)$$

where $\Pi_i = -B_0^{-1}B_i$, $\mu = B_0^{-1}\lambda = -B_0^{-1}(L)\Gamma(L)X_t$, $\eta_t = B_0^{-1}\varepsilon_{t1}$

and $\Pi(L) = \Pi_1 L + \Pi_2 L^2 + \dots + \Pi_p L^p$

μ term. It is important to note that η_t represents the reduced form random disturbances. STATA

and EVIEWS are two fairly “friendly” econometric program which estimate VAR models and allows for the inclusion of X’s into the model.

(4.6) is referred to as the **VAR in standard form** by Enders (1995, p. 295).

Hamilton (1994, p. 327) notes that (4.6) can be viewed as a **reduced form** of a general dynamic structural model because the right hand side variables are all predetermined.

Each equation in the standard or reduced form VAR representation includes one current endogenous variable and possibly lagged values of that and other predetermined variables. (4.6) can be used for forecasting. **The reduced form VAR’s can be**

estimated using OLS. Even though there may be correlation between error terms, SURE will not yield more efficient estimators because each equation will normally include the same regressors. The STATA command for estimating the system of VAR equations is as follows:

```
var y1 y2 ... yg, lags(1 p) (X1 X2 ...Xk)
```

where corresponding Granger Tests can be performed by using the command

vargranger

The Akaike Information Criterion (AIC) is often used to determine the lag length.

5. Vector Moving Average Representation (transfer functions)

In order to study the dynamics of a VAR, it is useful to consider the vector moving average representation (VMA). Solving equation (4.1) or (4.6) for Y_t yields

$$Y_t = -B^{-1}(L)\Gamma(L)X_t + B^{-1}(L)\varepsilon_{t1} \quad (5.1)$$

$$= (I - \Pi(L))^{-1}\mu + (I - \Pi(L))^{-1}\eta_t$$

$$= \Phi(L)X_t + \Psi(L)\eta_t$$

$$Y_t = \sum_{i=0}^{\infty} \Phi_i X_{t-i} + \sum_{i=0}^{\infty} \Psi_i \eta_{t-i} \quad (5.2)$$

or

$$Y_t = \gamma + u_t \quad (5.3)$$

where

$$\gamma = \Phi(L)X_t = (I - \Pi(L))^{-1}\mu$$

- $= (I - \Pi(L))B_0^{-1}\Gamma(L)X_t = -B^{-1}(L)\Gamma(L)X_t$

- $u_t = (I - \Pi(L))\eta_t = \sum_{i=1}^{\infty} \Psi_i \eta_{t-i}$

- $= \sum_{i=0}^{\infty} \Psi_i B_0^{-1} \varepsilon_{t-i}$

- γ Can be interpreted as the equilibrium for Y if the X 's remain unchanged

- The random disturbances in the transfer function representation can be expressed in terms of the reduced form or the structural random disturbances

The **impact, interim, and long run multipliers**, respectively, can be obtained from (5.2) as

$$\frac{dY_t}{dX_t} = \Phi_0 = -B_0^{-1}\Gamma(L=0), \quad \frac{dY_t}{dX_{t-i}} = \Phi_i, \text{ and}$$

$$\sum_{i=0}^{\infty} \Phi_i = \Phi(L=1) = -B^{-1}(L=1)\Gamma(L=1)$$

(5.2) is a vector moving average representation expressed in terms of the structural form VAR innovations ε_t , $\sin \eta_t = B_0^{-1} \varepsilon_t$. Equation (5.2) is the **transfer function** discussed in Section VI.2 of the notes.

6. Impulse Response Functions

The impulse response function describes the impact of an innovation or shock on future values of variables in the model. The moving average representation of the VAR in standard form given by (5.2) yields

$$\left(\frac{dY_t}{d\eta_{t-i}} \right) = \Psi_i \quad (6.1)$$

Since the Ψ_s 's are independent of B_0 , only the reduced form VAR representation is necessary to estimate (6.1). However, since the structural shocks (ε_t) are related to the

reduced form errors (η_t 's) by the equation, $\eta_t = B_0^{-1} \varepsilon_t$ and it will be necessary to estimate B_0 in order to “unravel” the impact of the structural innovations or shocks on the time path of Y_t to explore “causal” links between the variables. For cases in which B_0 is

known or can be consistently estimated, we can use the result

$$\begin{aligned} \left(\frac{dY_t}{d\varepsilon_{t-i}} \right) &= \left(\frac{dY_t}{d\eta_{t-i}} \right) \left(\frac{d\eta_{t-i}}{d\varepsilon_{t-i}} \right) \\ &= \Psi_i B_0^{-1} \end{aligned} \quad (6.2)$$

Economic analysis and estimation of the impulse response function is conditional on solving the identification problem, i.e. being able to estimate the matrix B_0 from observed data. The matrix of impulse response coefficients in equation (6.2) provides information about how “structural shocks” associated with different structural equations will impact the endogenous variables over time.

7. Identification

Recall that the identification problem in econometric models deals with the question as to whether the structural parameters can be determined from the reduced form parameters:

- $B_0\Pi + \Gamma$ (7.1 a-b)
- $\Omega = Var(\varepsilon_{tl}) = B_0 Var(\eta_t) B_0' = B_0 \Sigma B_0'$

In structural dynamic econometric models, the common practice is to impose sufficient restrictions (variable exclusions or to use at least as many instruments as there are endogenous regressors) on B and Γ to enable us to solve $B\Pi + \Gamma = 0$ for B and Γ in

terms of the π_{ij} 's. In structural vector autoregressive formulations the identification restrictions may take different forms. However, the problem is the same: can the structural parameters be uniquely determined from the reduced form. Equation (7.1 b) is frequently the focal point of identification of VAR models. If the structure is identified and B_0 can be determined, then the structural impulse response functions can be evaluated using equation (6.2). If economic theory has provided information about ψ_s and B_0 , then the impulse response function reflects the implications of economic theory about important "causal" relationships. The impulse response functions are often plotted as a function of "s" to visualize the inter-temporal impact of structural shocks.

Evaluating typically does not involve imposing (exclusion) restrictions on the structural coefficient matrices to identify the structural model. models, typ, then the impulse response function can be identified as well as the impact multipliers from the transfer function representation. If in a given application the structural model is not identified, the criticism of VARS not having any economic content would be valid. However, if one is only interested in obtaining forecasts, rather than economic analysis, identification may not be important.

Let's consider the identification problem associated with equations (7.1 a-b) in VAR formulations in more detail: Can the $[G + G^2 + GK]$ structural VAR parameters

- (1) Ω G unknown diagonal elements, the variance of each structural error in each structural equation. The structural shocks are assumed to be independent of other structural shocks.
- (2) B_0 G^2 unknown parameters
- (3) Γ GK unknown parameters

be recovered from the $[G(G+1)/2] + GK$ reduced form VAR parameters of the variance

$$(1) \Sigma = \text{Var}(\eta_t) \quad \frac{G(G+1)}{2} \quad \text{unknown parameters}$$

$$(2) \Pi \quad \text{GK unknown parameters}$$

A necessary condition for identification is that there are at least

$$(G^2 + G + GK) - \left(\frac{G(G+1)}{2} + GK \right) = \frac{G(G+1)}{2}$$

restrictions imposed on B_0 and on Ω .

Furthermore, if the matrix B_0 is normalized on the G diagonal elements (each structural equation has one dependent variable with a coefficient of "1"), then a necessary identification condition is that there are at least

$$\frac{G(G+1)}{2} \quad \frac{G(G-1)}{2}$$

additional “structural” restrictions on Ω or B_0 , for example if the VAR involves three endogenous variables, then $(3(3-1)/2=3)$ additional restrictions would need to be imposed on Ω or B_0 .

One approach to the identification problem is what is termed “structural” decomposition which involves determining a matrix A (B_0) which solves equation (7.1b):

$$\Omega = A \Sigma A'$$
 (7.2)

Recall that the structural covariance matrix (Ω) of the structural shocks is assumed to diagonal

$$\begin{pmatrix} \omega_{\varepsilon_1}^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \omega_{\varepsilon_G}^2 \end{pmatrix} = A \begin{pmatrix} \sigma_{\eta_1}^2 & \sigma_{\eta_1 \eta_2} & \dots & \sigma_{\eta_1 \eta_G} \\ \sigma_{\eta_2 \eta_1} & \sigma_{\eta_2}^2 & \dots & \sigma_{\eta_2 \eta_G} \\ \vdots & \vdots & \dots & \vdots \\ \sigma_{\eta_G \eta_1} & \sigma_{\eta_G \eta_2} & \dots & \sigma_{\eta_G}^2 \end{pmatrix} A'$$

The matrix A can then be used to construct a set of orthogonal “structural” shocks

$$\varepsilon = A \eta$$

such that equation (7.2) , $\text{var}(\varepsilon) = A \text{var}(\eta) A'$, is satisfied. It is important to remember that (1) the resultant innovations or shocks only have economic meaning if $A=B_0$ and (2) the “decomposition” or “factorization” involved in (7.2) is not unique.

The Cholesky decomposition is **one** approach to obtaining an orthogonal decomposition (unraveling). The matrix A is selected to be of the form

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ a_{21} & 1 & \dots & 0 \\ a_{31} & a_{32} & 1 & \\ \vdots & & & \vdots \\ a_{G1} & a_{G2} & \dots & 1 \end{pmatrix}$$

Note that there are $\frac{G(G-1)}{2}$ zero restrictions imposed in the A matrix. Mathematically,

there is a degree of arbitrariness in where the block of zero's is placed in the matrix A .

The argument for the “ordering” of the variables is the block diagonal matrix A is often based on “timing” of economic interactions, for example monetary policy can respond more rapidly to economic conditions than can the implementation of fiscal policies.

Other decompositions have been considered by Sims (1986) in his discussion of a six-variable VAR macro model, by Beveridge and Nelson (1981), and by Blanchard and Quah (1989) among others. Once again, the reader is reminded that the interpretation of “structural” impulse response functions shocks depend on the selection of $A \underline{\equiv} (\varepsilon = A\eta)$, and only have economic interpretations if $A = B_0$.

Markku Lanne and Helmut Lutkepohl (2010, Structural Vector Autoregressions with nonnormal residuals, JBES, 159-168) demonstrate that distributional assumptions can sometimes be used to identify structural shocks.

8. Estimation and Testing Hypotheses

a. Short and long-run VAR models

The STATA command **varbasic** performs a Cholesky decomposition and reports the corresponding “structural” impulse response functions.

varbasic depvar_list , lags(p) exog(var_list)

The STATA command **svar** allows for more flexibility in solving the identification problem and in estimation. These commands are organized according to estimation of short-run and long-run formulations. The format for estimating the short-run formulation is as follows:

svar depvarlist, aconstraints lags(p) exog(var_list),

svar depvarlist, bconstraints lags(p) exog(var_list), or

or **svar depvarlist, aconstraints bconstraints lags(p) exog(var_list)**

where the acon or bcon commands can be written in different forms (aeq() or acns ()), but allow the imposition of identifying restrictions and correspond to GxG matrices. If either *acon* or *bcon*. is deleted, it is assumed to correspond to an identity matrix. Each of the acon or bcon options be written in different ways and facilitates structural identification and estimation. The purpose of unknown diagonal elements in the b_constraints is to scale the structural innovations to have unit variance. The a_constraints take into account identifying restrictions on the B_0 matrix and also satisfy (7.1b),

$$\Omega = \text{Var}(\varepsilon_{t1}) = B_0 \text{Var}(\eta_t) B_0' = B_0 \Sigma B_0'$$

If the B_0 matrix in a bivariate VAR model with one lag is assumed to be of the form

$$B_0 = \begin{pmatrix} -1 & 0 \\ \beta_{21} & -1 \end{pmatrix},$$

then the **svar** command could be written as follows

```
mat A=(1,0\.,1)
mat B=(.,0\0,.)
svar y1 y2, aeq(A) lags(1) exog(x's) or
svar y1 y2, aeq(A) beq(B) lags(1) exog(x's)
```

where the “.” in the matrix command indicates that the corresponding coefficient in the

B_0 matrix needs to be estimated and the other parameters are set equal to the indicated

values. The **aeq** option is a matrix alternative to the **aconstraints** option.

To discuss the estimation of long-run VAR models, it will be helpful to consider VMA or transfer function representation (equation (5.1)) corresponding to long-run adjustments

$$\begin{aligned} Y_t &= (I - \Pi(L = 1))^{-1} \mu + (I - \Pi(L = 1))^{-1} \eta_t \\ &= -B^{-1}(L = 1)\Gamma(L = 1)X_t + (I - \Pi(L = 1))^{-1} B_0^{-1}\varepsilon_t \end{aligned}$$

The coefficient matrix of the structural disturbances, shocks, or innovations corresponds to the C matrix in STATA,

$$C = (I - \Pi(L = 1))^{-1} B_0^{-1}$$

The matrix C can be interpreted as the matrix of long-run impact of structural shocks on the endogenous variables. For example, in a bivariate VAR model of the money supply (m) and output (gdp) it might be expected that an unexpected shock to the money supply

would not have a long-run impact on output and similarly, an unexpected shock to output would not have a long-run impact on the money supply. The corresponding C matrix would have zeros off-diagonal elements. The corresponding STATA commands could be

mat C=(.,0\0,.)

svar gdp m, lreq(C) or alternatively as

svar d.ln_m d.ln_gdp, lreq(C) to take account of non-stationarity of m and gdp

b. Hypothesis testing

Hypotheses about reduced form (standard form) VAR coefficients,

$$Y_t = \mu + \Pi_1 Y_{t-1} + \dots + \Pi_p Y_{t-p} + \eta_t,$$

can be tested using the likelihood ratio statistic

$$LR = N (\ln |\Sigma_R| - \ln |\Sigma_{UR}|) \sim \chi^2 (d.f.)$$

where Σ_R and Σ_{UR} denote the estimated variance-covariance matrix corresponding to the restricted and unrestricted estimates and df = degrees of freedom. Sim's suggests using

$$(N-c) (\ln |\Sigma_R| - \ln |\Sigma_{UR}|)$$

for small samples where c denotes the number of estimated parameters in the unrestricted model.

9. A Review and an Application

- (a) A dynamic structural econometric model may be expressed as a special case

$$\begin{aligned} B(L)Y_t + \Gamma(L)X_t &= \varepsilon_{t1} \\ \varphi(L)X_t &= \theta(L)\varepsilon_{t2} \end{aligned}$$

A structural vector autoregressive model can be written in the form

$$B_0 Y_t + B_1 Y_{t-1} + \dots + B_p Y_{t-p} + \Gamma(L) X_t = \varepsilon_{t1} \quad (9.2)$$

Note that the structural VAR generalizes the dynamic structural econometric model.

$$\begin{aligned} Y_t &= \mu + B_0^{-1}\Gamma(L)X_t + \varepsilon_{t1} \\ &= \mu + [B_0^{-1}\Gamma(L)]X_t + \varepsilon_{t1} \end{aligned} \quad (9.3)$$

where $\mu = -B_0^{-1}\Gamma(L)X_t$, with impulse response functions given by

$$\left(\frac{dY_t}{d\varepsilon_{t-i}} \right) = \Psi_i B_0^{-1} \quad (9.4)$$

where $\Psi_i = \frac{d^i}{dt^i}(I - \Pi_1 t - \dots - \Pi_s t^s)^{-1}$ with $t=0$.

(b) An application: cobweb model

We now consider an application of these approaches to analyzing the Cobweb Model defined in the homework in section (VI)

$$Q_t = \beta P_t + \xi Y_t + \delta_d + \varepsilon_{td}$$

$$Q_t = \delta P_{t-1} + \alpha w_t + \delta_s + \varepsilon_{ts}.$$

This dynamic structural model can also be written in the following manner

$$\begin{pmatrix} -1 & \beta \\ -1 & 0 \end{pmatrix} \begin{pmatrix} Q_t \\ P_t \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & \delta \end{pmatrix} \begin{pmatrix} Q_{t-1} \\ P_{t-1} \end{pmatrix} + \begin{pmatrix} \alpha & 0 & \xi \\ \gamma & \rho & 0 \end{pmatrix} \begin{pmatrix} 1 \\ W_t \\ Y_t \end{pmatrix} + \begin{pmatrix} \varepsilon_{td} \\ \varepsilon_{ts} \end{pmatrix} = 0$$

Multiplying the dynamic structural equation by the inverse of the leading coefficient matrix yields the reduced form:

$$\begin{pmatrix} Q_t \\ P_t \end{pmatrix} = \begin{pmatrix} 0 & \delta \\ 0 & \frac{\delta}{\beta} \end{pmatrix} \begin{pmatrix} Q_{t-1} \\ P_{t-1} \end{pmatrix} + \begin{pmatrix} \gamma & \rho & 0 \\ \frac{\gamma-\alpha}{\beta} & \frac{\rho}{\beta} & \frac{\xi}{\beta} \end{pmatrix} \begin{pmatrix} 1 \\ W_t \\ Y_t \end{pmatrix} + \begin{pmatrix} \varepsilon_{td} \\ \frac{\varepsilon_{ts} - \varepsilon_{td}}{\beta} \end{pmatrix}$$

The reduced form includes 7 non-zero coefficients which are functions of 6 structural parameters.

Some practitioners may object to the imposition of “economic” hypotheses (exclusions) on coefficients of variables in the model. Vector Auto Regression (VAR) models attempt to circumvent these assumptions.

The structural VAR Model corresponding to the structural cobweb model could be written as

$$\begin{pmatrix} -1 & \beta_{12}^0 \\ -1 & \beta_{22}^0 \end{pmatrix} \begin{pmatrix} Q_t \\ P_t \end{pmatrix} + \begin{pmatrix} \beta_{11}^1 & \beta_{12}^1 \\ \beta_{21}^1 & \beta_{22}^1 \end{pmatrix} \begin{pmatrix} Q_{t-1} \\ P_{t-1} \end{pmatrix} + \begin{pmatrix} \gamma_{11} & \gamma_{12} & \gamma_{13} \\ \gamma_{21} & \gamma_{22} & \gamma_{23} \end{pmatrix} \begin{pmatrix} 1 \\ W_t \\ Y_t \end{pmatrix} + \begin{pmatrix} \varepsilon_{td} \\ \varepsilon_{ts} \end{pmatrix} = 0$$

Note that the structural VAR model includes the cobweb model as a special case. The structural VAR includes 12 free parameters compared to 6 parameters for the cobweb model.

The corresponding reduced form VAR for the cobweb model can be written in the form

$$\begin{pmatrix} Q_t \\ P_t \end{pmatrix} = \begin{pmatrix} \pi_{11}^1 & \pi_{12}^1 \\ \pi_{21}^1 & \pi_{22}^1 \end{pmatrix} \begin{pmatrix} Q_{t-1} \\ P_{t-1} \end{pmatrix} + \begin{pmatrix} \pi_{11} & \pi_{12} & \pi_{13} \\ \pi_{21} & \pi_{21} & \pi_{23} \end{pmatrix} \begin{pmatrix} 1 \\ W_t \\ Y_t \end{pmatrix} + \begin{pmatrix} \eta_{t1} \\ \eta_{t2} \end{pmatrix}$$

Note that this form is a generalization of the reduced form of the cobweb model. The reduced form VAR involves 10 parameters compared to 7 parameters for the reduced form corresponding to the structure or 6 parameters associated with the reduced form estimates derived from the estimated structure. Thus the reduced form VAR can be looked at as over-fitting the underlying model. The hypotheses associated with the original specification (exclusions and exogeneity) can be collectively tested using the likelihood ratio test. Let $\hat{\Sigma}_{10}$, $\hat{\Sigma}_7$, $\hat{\Sigma}_6$, denote the variance covariance matrices associated with the unrestricted and two restricted VAR's.

$$(T-10) \quad (\ln|\hat{\Sigma}_7| - \ln|\hat{\Sigma}_{10}|)$$

and

$$(T-10) \quad (\ln|\hat{\Sigma}_6| - \ln|\hat{\Sigma}_{10}|)$$

can be used to test the hypotheses of interest. These test statistics are asymptotically distributed as $\chi^2(3)$ and $\chi^2(4)$, respectively.

If the researcher is only interested in forecasting, the reduced form (structure or

VAR based reduced form) can be used. If the researcher is also interested in economic analysis of the structure, transfer functions or impulse functions will be of interest.

The form for the impulse response functions is particularly simple in the case of a first-order VAR

$$\left(\frac{d \begin{pmatrix} Q_{t+i} \\ P_{t+i} \end{pmatrix}}{d \varepsilon_t} \right) = \Pi_1^t B_0^{-1}$$

Estimates of B_0 can be obtained by solving

$$\begin{pmatrix} \omega_1^2 & 0 \\ 0 & \omega_2^2 \end{pmatrix} = \begin{pmatrix} -1 & \beta_{12}^0 \\ -1 & \beta_{22}^0 \end{pmatrix} \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix} \begin{pmatrix} -1 & -1 \\ \beta_{12}^0 & \beta_{22}^0 \end{pmatrix}$$

for the structural ω 's and β 's in terms of the reduced form σ 's. There are four unknown structural parameters and three reduced form parameters; hence, there needs to be one restriction ($G(G-1)/2 = 1$) restriction imposed. Let's assume that $\beta_{22} = 0$ as in the original specification, which was motivated by the timing of the economic decisions.

The STATA commands to estimate the reduced form and structural VAR's are as follows:

var q p, lags(1) exog(w f)

mat A =(1,\1,0)

svar q p, aeq(A) exog(w f)

Other STATA commands which might be of interest are

varirf calculates and analyzing impulse response functions

varstable	checks stability conditions of var or svar estimates
varsoc	obtain lag-order statistics for a set of VAR's
varwle	obtain Wald lag exclusion statistics after var or svar
vargranger	performs pairwise Granger causality tests after var or svar
varlmar	obtain LM statistics for residual autocorrelation after var or svar
varnorm	tests for normally distributed disturbances after var or svar
varforecast compute	computes dynamic forecasts of dependent variables after var/svar

EXERCISES

1. Cover, Enders, and Hueng (“Using the Aggregate Demand-Aggregate Supply Model to Identify Structural Demand-side and Supply side Shocks (using a bivariate VAR),” Journal of Money Credit and Banking, 38(2006), 777-790) consider a bivariate VAR

Reduced form VAR: $Z_t = Z_0 + A(L)Z_t + \eta_t$ $Z_t = \begin{pmatrix} y_t \\ p_t \end{pmatrix}$ $Var(\eta_t) = \Sigma$

Structural VAR: $B_0 Z_t - B_0 A(L) Z_t = B_0 Z_0 + B_0 \eta_t$ ε_t where
 $= B_0 Z_0 + \varepsilon_t$

shocks with $Var(\varepsilon_t) = \Omega$. The 2x2 matrix of long-run response multipliers for

Z_t corresponding to the structural shocks can be expressed as (see p.15)

$$C = \left(B_0 (I - A(L=1)) \right)^{-1} = (I - A(L=1))^{-1} B_0^{-1} = (I - A(L=1))^{-1} D \quad ***$$

Identification and estimation of the four parameters in the D matrix (hence in the B_0 =

D^{-1} matrix) requires at least four restrictions, which were chosen to be satisfied by

- $\Omega = I_2$
- Demand shock have no permanent impact on output, i.e. the element in the first row and second column of *** is zero. This is known as the Blanchard-Quah long-run neutrality restriction. Show that this is equivalent to $d_{12}(1 - a_{22}(L=1)) + d_{22}a_{12}(L=1) = 0$

2. Consider the Kmenta-Smith(RESTAT, (August 1973, 299-307)) structural form defined by

$$\begin{aligned} C_t &= \gamma_{10} + \beta_{12}Y_t + \beta_{15}L_t + \beta_{111}C_{t-1} + \varepsilon_{t1} \\ I^d_t &= \gamma_{20} + \beta_{25}Y_t + \gamma_{25}(S_{t-1} - S_{t-2}) \\ &\quad + \gamma_{212}t + \gamma_{27}I_{t-1}^d + \varepsilon_{t2} \\ I^r_t &= \gamma_{30} + \beta_{33}r + \gamma_{35}(S_{t-1} - S_{t-2}) \\ &\quad + \gamma_{312}t + \gamma_{38}I_{t-1}^r + \varepsilon_{t3} \\ I^i_t &= \gamma_{50} + \beta_{43}r + \gamma_{45}(S_{t-1} - S_{t-2}) \\ &\quad + \gamma_{412}t + \gamma_{49}I_{t-1}^i + \varepsilon_{t4} \\ Y_t &= \gamma_{52}Y_t + \gamma_{53}M_t + \gamma_{54}M_{t-1} + \varepsilon_{t5} \\ Y_t &= C_t + I_t^d + I_t^i + G_t \\ S_t &= Y_t - I_t^i \\ L_t &= M_t + R_t \end{aligned}$$

where

Y = gross national product (\$ bill.)

C = consumption expenditures (\$ bill.)

I^d = producer's outlays on durable plant and equipment (\$ bill.)

I^r = residential construction (\$ bill.)

Iⁱ = investment in inventories (\$ bill.)

G = government purchases of goods and services plus net foreign investment (\$ bill.)

S = final sales of goods and services (\$ bill.)

t = time in quarters (first quarter of 1954 = 0)

r = yield on all corporate bonds (%)

M = money supply, i.e., demand deposits plus currency outside banks (\$ bill.)
 R = time deposits in commercial banks (\$ bill.)
 L = money supply plus time deposits in commercial banks (\$ bill.)

- a. Write the form of the reduced form representations.
- b. Investigate the form of the associated transfer functions.
- c. Compare the form of the transfer functions associated with those used by Maloney and Ireland.
- d. Discuss how you could compare restricted and unrestricted estimated transfer functions, also see exercise 2 in Section V)

- a. Write out a structural and reduced form VAR representation of the Kmenta-Smith model.

- b. Discuss how, given data, you could use the VAR representation to test the restrictions imposed in the Kmenta-Smith formulation.

- c. What is the relationship between the coefficients in the transfer functions and the impulse response functions.

3. Samuelson proposed a model known as the multiplier-accelerator model which has the potential to generate a business cycle. This model is defined by

$$y_t = c_t + I_t + G_t$$

$$\begin{aligned} c_t &= \gamma y_{t-1} & 0 < \gamma < 1 \\ I_t &= \alpha (c_t - c_{t-1}) & 0 < \alpha \end{aligned}$$

- a. Write out the forms for the
 - 1. structural and
 - 2. reduced form
 vector autoregressive models.

- b. How can the exogeneity assumptions be tested?

- c. Describe how, given data, you could use the VAR representation to test the validity of the economic hypothesis imposed by the Multiplier-Accelerator model.

4. Given the relationships

$$\begin{aligned} (E.1) \quad F(L) Z_t &= G(L) \varepsilon_t & \text{MARMA} \\ (E.3) \quad Z_t &= C(L) \varepsilon_t & \text{Multivariate MA} \\ (E.4) \quad A(L) Z_t &= \varepsilon_t & \text{Multivariate AR} \\ C(L) &= F^{-1}(L) G(L) \\ A(L) &= C^{-1}(L) = G^{-1}(L) F(L) \end{aligned}$$

Investigate the relationship between the conditions:

$$\begin{array}{ll} \text{Zellner:} & F_{21}(L) \equiv 0, G_{21}(L) \equiv 0, G_{12}(L) \equiv 0 \\ \text{Granger (MA):} & C_{21}(L) \equiv 0 \\ \text{AR:} & A_{21}(L) \equiv 0 \end{array}$$

Hint: Recall that the inverse of a partitioned matrix is given by

$$D^{-1} = \begin{pmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{pmatrix}^{-1} = \begin{pmatrix} D^{11} & D^{12} \\ D^{21} & D^{22} \end{pmatrix}$$

$$D^{11} = (D_{11} - D_{12}D_{22}^{-1}D_{21})^{-1}$$

$$D^{12} = -D^{11}D_{12}D_{22}^{-1}$$

$$D^{21} = -D_{22}^{-1}D_{21}D^{11}$$

$$D^{22} = (D_{22} - D_{21}D_{11}^{-1}D_{12})^{-1}$$

Appendix Exogeneity tests: more details

1. Introduction

An important objective of research in economics is to determine relationships between variables and explain which variables are exogenous and which are endogenous. Specification of causal relationships is usually viewed as being the domain of the theoretician and not determined statistically. A number of statistical tests have been proposed which claim to test for causality or exogeneity. There has been considerable discussion as to whether the proposed tests do indeed check what is alleged. One of the key issues in the discussion centers around what is meant by causality. The paper by Zellner (1979, Carnegie Rochester Conference series on Public Policy) summarizes the major positions in this debate.

The purpose of this section is to outline the suggested tests and associated definitions. Before outlining Granger's test (1969, Econometrica), some notation needs to be discussed.

Let A_t be a stationary stochastic process, let \bar{A}_t represent the set of past values $\{A_{t-j}, j=1, 2, \dots, \infty\}$, and $\bar{\bar{A}}_t$ represent the set of past and present values $\{A_{t-j}, j=0, 1, \dots, \infty\}$. Further, let $\bar{A}_t(k)$ represent the set $\{A_{t-j}, j=k+1, \dots, \infty\}$.

Denote the optimum, unbiased, least squares predictor of A_t that uses the set of values of B_t by $P_t(A|B)$. Thus, for instance, $P_t(X|\bar{X}_{-t})$ will be the optimum predictor of X_t using only past X_{t-1} . The prediction error will be denoted by $\varepsilon_t(A|B) = A_t - P_t(A|B)$. Let $\sigma^2(A|B)$ be the variance of $\varepsilon_t(A|B)$. Let U_t denote all the information in the universe accumulated as of time t , and let $U_t - Y$ denote all this information other than the series Y_t .

We then have the following definition.

Definition: Causality

$$\text{If } \sigma^2(Y|\bar{U}) < \sigma^2(Y|\bar{U} - \bar{X})$$

, then X is said to cause Y , denoted by $X \rightarrow Y$; t

is “causing” Y if we are better able to predict Y_t using all available information, including past values of X and Y , rather than if all information other than X had been used.

Definition: Feedback

$$\text{If } \sigma^2(X|\bar{U}) < \sigma^2(X|\bar{U} - \bar{Y})$$

and

$$\sigma^2(Y|\bar{U}) < \sigma^2(Y|\bar{U} - \bar{X})$$

, then

we say that feedback is occurring, which is denoted $Y \leftrightarrow X$, that is, feedback is said to occur when X is “causing” Y and Y is “causing” X .

Definition: Instantaneous Causality

$$\text{If } \sigma^2(Y|\bar{U}, \bar{X}) < \sigma^2(Y|\bar{U})$$

, we say that instantaneous causality between X and

occurring. In other words, the current value of Y_t is better “predicted” if the present value of X_t is included in the “prediction” than if it is not.

There are a number of impediments to operational interpretations of the previous definitions. The feasibility of optimal minimum variance forecasts may be questionable, and the notion of using all information in the universe is untractable. The relationship between these definitions and the philosophical notion of causality is an important question.

We now consider a representation of a dynamic econometric model which facilitates a discussion of tests proposed by Granger (1969, Econometrica) and Sims (1972, AER). The statistical tests will then be defined and an application considered.

2. Operational Definition of Tests

Zellner's definition of X_t being exogenous is that

$$F_{21}(L) = 0 \text{ with } G_{12}(L) = 0 \text{ and } G_{21}(L) = 0$$

This implies $C_{21}(L) = 0$ is the moving average representation

$$\begin{pmatrix} Y_t \\ X_t \end{pmatrix} = \begin{pmatrix} C_{11}(L) & C_{12}(L) \\ 0 & C_{22}(L) \end{pmatrix} \begin{pmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{pmatrix}$$

which implies

$$Y_t = C_{11}(L)\epsilon_{1t} + C_{12}(L)C_{22}^{-1}(L)X_t$$

Thus Y_t depends on current and lagged X_t 's. Sim's proposes regressing

Y_t on current, lagged, and future X_t 's and performing a joint hypothesis test on the coefficients of future X 's being zero.

Failure to reject this hypothesis is consistent with X_t being exogenous to Y_t and not the other way around. Rejecting the null hypothesis raises questions about Y_t being exogenous. Two problems associated with implementing this test are observed: (1) how many lags should be used and (2) the random disturbance is $C_{11}(L)\epsilon_{1t}$ which will likely exhibit autocorrelation and raise questions about the validity of the usual F-statistic in performing such tests. Reference: Hamilton (1995, p. 304).

A second method of testing for causality can be motivated from the multivariate autoregressive representation of the model

$$\begin{pmatrix} A_{11}(L) & A_{12}(L) \\ A_{21}(L) & A_{22}(L) \end{pmatrix} \begin{pmatrix} Y_t \\ X_t \end{pmatrix} = \begin{pmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{pmatrix}$$

$A_{21}(L) = 0$ if and only if $C_{21}(L) = 0$ (Granger Causality). Zellner's definition

implies $C_{21}(L) = 0$ which is equivalent to $A_{21}(L) = 0$. (A good exercise). Hence,

$$A_{11}(L)Y_t + A_{12}(L)X_t = \epsilon_{1t}$$

$$A_{22}(L)X_t = \epsilon_{2t}$$

$$A_{11}(L) = A_0 + A_1 L + \dots$$

Premultiplying by A_0^{-1} and solving for Y_t yields

$$Y_t = A_0^{-1} A_1 Y_{t-1} + A_0^{-1} A_2 Y_{t-2} + \dots + A_0^{-1} A_{12}(L)X_t + A_0^{-1} \epsilon_{1t}$$

This form suggests a second test*

Regress Y on lagged Y and current and lagged X's. If the coefficients on the X's are significantly different from zero, then the X's "cause" Y. If not, then X does not "cause" Y. A Chow test is asymptotically distributed as an F. Hamilton (1995, p. 305)

*This test is probably most commonly applied by regressing Y on lagged Y and lagged X's.

3. An Application of Causality Tests

A widely used model in monetary analysis is given by

$$Y_t = \sum_{j=0}^{\infty} h_j m_{t-j} + \eta_t \quad \text{where}$$

y_t = log of nominal income

m_t = log of money

$E(\eta_t m_{t-j}) = 0, j = 0, 1, 2, \dots$

The h_j 's have been interpreted as dynamic multipliers (Anderson and Jordan, 1968, Federal Reserve Bank of St. Louis Review).

The two major criticisms of this model and its interpretation are that this equation is not the final form or transfer function because many other variables have an impact, and secondly, m_t need not be exogenous because the monetary authority considers the behavior of lagged y_i 's in the determination of m_t .

Sims (1972, AER) considered the model

$$Y_t = \sum_{j=-\infty}^{\infty} \delta_j m_{t-j} + \epsilon_t$$

and tested the joint hypothesis that the coefficients of future m_t 's were equal to zero,

$$H_0: \delta_{-1} = \delta_{-2} = \dots = 0$$

Sim's conclusion was that he could not reject the hypothesis with a high degree of

confidence. Sim's paper and statistical results have generated considerable discussion in the literature. Some of the limitations were discussed in the previous notes.