Econometric notes

Calzolari, Giorgio

Universita' di Firenze, Italy.

31 January 2012

ECONOMETRIC NOTES - Giorgio Calzolari - Università di Firenze - <calzolar@disia.unifi.it>
Undergraduate, Graduate, Doctoral Students - Laurea Triennale, Magistrale, Dottorato di Ricerca

Revised: May 2016

Greene, W. H. (2008): *Econometric Analysis* (6*th* edition). Prentice-Hall, Inc. Upper Saddle River, NJ. (Sec. 2, 3, 4, 5, 6.1, 6.2, 6.4, 7.1, 7.2, 8.1, 8.4, 8.5, 8.6, 19.7, App. A).

Johnston, J. (1984): *Econometric Methods* (3*rd* edition). New York: McGraw-Hill, Inc. Traduzione dalla lingua inglese a cura di M. Costa e P. Paruolo (1993): *Econometrica* (terza edizione). Milano: Franco Angeli. (Sec. 4, 5, 6; 8.1, 8.2, 8.4, 8.5.1, 8.5.2, 8.5.3, 8.5.4, 8.5.7).

Stock, J. H., and M. W. Watson (2015): *Introduction to Econometrics* (Updated 3*rd* edition). Boston: Person Education Limited. Edizione italiana a cura di F. Peracchi (2016): *Introduzione all' Econometria* (quarta edizione). Milano: Pearson Italia S.p.A. (Sec. 2, 3, 4, 5, 6, 7, 8, 12, 17, 18).

ELEMENTS OF LINEAR ALGEBRA

Vectors and matrices

Vectors with n components (or $n-$components vectors, or vectors in the Euclidean space of dimension n, or vectors with dimensions $n \times 1$, or simply with dimension n); notation using small letters and representation as *columns*.
Elements or components of vectors.
Null vector.
Representation with directed line segments (e.g. 2 or 3 dimensions).
Equality.
Sum of vectors (and graphical representation in 2 dimensions).
Opposite vector.
Difference of vectors.
Product of a vector with a scalar (and graphical representation in 2 dimensions).
Unit vectors (with n components).
Scalar product (or internal product) of two vectors.
Orthogonal vectors: the scalar product is zero (graphical example in 2 dimensions, based on similarity of triangles).
Linear combination of m vectors with n components: it is an $n-$ component vector.
Linear dependence or independence of m vectors with n components.
If m vectors are linearly dependent, someone of them can be represented as a linear combination of the others.
The m unit vectors with m components are linearly independent (example in 2 dimensions).
Two vectors are linearly dependent if and only if they have the same direction; 3 vectors if and only if they lay on the same plane.
If 2 vectors with 2 components are linearly independent, any other $2-$components vector is a linear combinations of them (graphical example); analogously in 3 dimensions, a fourth vector is always a linear combination of three linearly independent vectors; etc; in general, there cannot be more than m linearly independent $m-$components vectors; in particular, any $m-$components vector can be represented as a linear combination of the m unit vectors.
A *basis* of an $m-$dimensional space is a collection of m linearly independent $m-$components vectors; for instance, the m unit vectors.
Any $m-$components vector has a unique representation as a linear combination of m basis vectors; *ab absurdo*, suppose that there are two different linear combinations that produce the same vector; subtracting one from the other, there would be a linear combination of the basis vectors that produces a null vector.
Subsets of linearly independent vectors are linearly independent.
The vectors in a set that contains a subset of linearly dependent vectors are themselves linearly dependent.
If, in a set of n vectors (with the same dimensions), k vectors can be found linearly independent (but not more than k), and it is $k < n$, then all the other $n - k$ vectors in the set are linear combinations of these k vectors.
Matrices (with dimensions $m \times n$) and representation as *rectangles*.
Vectors can be considered matrices with a single column.
Row index and column index.
Columns can be called *column vectors*; rows can be called *row vectors*.
Notation for rows ($A_{i.}$) and columns ($A_{.j}$).
Null matrix.
Equality.
Multiplication by a scalar.
Linear combination of the columns of a matrix: it is a column vector.
Linear combination of the rows of a matrix: it is a row vector.
Sum, difference, linear combination of matrices with the same dimensions.
Matrix multiplication, or product *rows by columns* of two matrices conformable for multiplication; if the former (A) is an $m \times n$ matrix and the latter (B) has dimensions $n \times k$, the product AB is an $m \times k$ matrix; its $i, j - th$ element is the scalar product of the row vector $A_{i.}$ with the column vector $B_{.j}$.

The $i-th$ row of AB is the product of the $i-th$ row of A with the matrix B: $[AB]_{i.} = A_{i.}B$; the $j-th$ column of AB is the product of matrix A with the $j-th$ column of B: $[AB]_{.j} = AB_{.j}$.
Matrix multiplication is associative: $(AB)C = A(BC)$; it is distributive with respect to the sum: $D(E+F) = DE + DF$ (when the matrices are conformable for the operations above; example of proof with small dimensions).
Matrix multiplication of two matrices is not commutative (with examples: for different dimensions as well as equal dimensions); pre- and post- multiplication.
Square matrices ($n \times n$).
Identity matrix (I, or I_n); its n columns are the $n-$dimensional unit vectors; for any $m \times n$ matrix A, it is always $AI_n = A$; for any $n \times k$ matrix B, it is always $I_n B = B$.
Diagonal matrix.
Scalar matrix.
Transpose of a matrix and transpose of a vector (row vector).
Transpose of the sum of two matrices.
Transpose of the product of two matrices: $(AB)' = B'A'$ (example of proof with small dimensions).
Transpose of the product of 3 or more than 3 matrices.
Scalar (or internal) product of two vectors (a and b) as a particular case of matrix multiplication, using the transpose of the first vector ($a'b$).
External product of two vectors (ab') is a matrix.
Symmetric (square) matrix.
The product of a (rectangular) matrix with its transpose is always a square symmetric matrix: $A'A$ and AA' are both square symmetric matrices.
If b is a column vector, then Ab is a column vector, linear combination of the columns of the matrix A, the coefficients of the linear combination being the elements of the vector b; if c is a column vector, $c'D$ is a row vector, linear combination of the rows of matrix D.
In a matrix, the maximum number of linearly independent columns and of linearly independent rows are equal; to simplify the proof, given a 4×3 matrix A, with $r = 2$ maximum number of linearly independent rows, call $\tilde{A}$ one of the 2×3 sub-matrices with all rows linearly independent (for simplicity, let be the first two rows of A); the 3 columns of $\tilde{A}$ are $2-$dimensional vectors, thus they are linearly dependent; write explicitly the third column of $\tilde{A}$ as linear combination of the first two columns of $\tilde{A}$; write explicitly all the elements of the third and fourth rows of the matrix A as linear combinations of the first two rows (which are the two rows of $\tilde{A}$); making substitutions, it appears that the *whole* third column of A depends linearly on the first and second column of A, so that there cannot be 3 independent columns in A; independent columns in A will thus be 2 or 1; thus, the maximum number of linearly independent columns of A would be $c \leq r = 2$; repeating the whole procedure, but assuming that $c = 2$ is the maximum number of linearly independent columns of A, it will be $r \leq c$; thus the conclusion is $r = c$.
The maximum number of linearly independent rows or columns is called *rank* of the matrix (in the examples, use a rectangular matrix X with more rows than columns).
If X has dimensions $n \times k$, with $k \leq n$, the rank will be $\leq k$; if $r(X) = k$, X is called full rank matrix, or matrix with full rank.
A full rank square matrix (thus all columns are linearly independent and all rows are linearly independent) is called *non-singular*, otherwise it is called *singular*, and its columns (and rows as well) will be linearly dependent.
Definition of *inverse* of a square matrix: if A is an $n \times n$ (square) matrix, inverse of A is a matrix B (with the same dimensions) such that $AB = I$.
If an $n \times n$ (square) matrix A is non-singular, the inverse matrix exists and is unique; to prove it, remember that the columns of A form a basis for the $n-$dimensional vectors; as it must be $AB = I$, then for each $j-$th column it must be $AB_{.j} = e_j$ ($j-$th unit vector); thus, each e_j must be representable as a linear combination of the columns of A; as columns form a basis, this representation exists and is unique.
For the same matrix A just considered, there exists also a unique matrix C (with the same dimensions) such that $CA = I$; the proof is analogous to the proof above, remembering that, being linearly independent, also the n rows of the matrix form a basis for the $n-$dimensional row vectors; thus for each $i-$th row there is a unique linear combination of the rows of A that produces the $i-$th row of the identity matrix: $C_{i.}A = e_i'$ ($i-$th unit row vector).
The two matrices B and C, whose existence and uniqueness has just been proved, are equal; in fact, if $AB = I$ and $CA = I$; then CAB is equal to B and also equal to C, thus $B = C$ (no right or left inverse, just inverse).
A^{-1} is used to indicate the inverse of A.
Inverse of the transpose: $(A')^{-1} = (A^{-1})'$; thus also symbols like A'^{-1} or $A^{-1\prime}$ can be used.
Inverse of the product of two or more square matrices: $(ABC)^{-1} = C^{-1}B^{-1}A^{-1}$.
Inverse of a diagonal matrix.
Inverse of a 2×2 matrix.

Determinants

Permutations of n objects.
Factorial ($n!$).
Fundamental permutation and number of inversions.
Class of a permutation (even or odd).

Switching two elements, the class of the permutation changes (the proof is first for two consecutive elements, than for any pair of elements).
Product associated to a square matrix.
Definition of *determinant* of a square matrix as sum of the $n!$ products associated to the matrix.
Determinant of the transpose.
Switching two columns or two rows, the determinant changes sign.
If two rows or columns are equal, the determinant is zero.
Multiplying a row or a column by a scalar, the determinant is multiplied by the same scalar.
If a row (or a column) can be decomposed into the sum of two rows, the determinant is the sum of the determinants of two matrices.
If a row (or a column) is equal to the sum of two other rows (or columns) of the same matrix, the determinant is zero.
If a row (or column) is a linear combination of other rows (or columns) of the same matrix, the determinant is zero.
The determinant of the sum of two matrices *is not* the sum of the two determinants.
Multiplying the whole $n \times n$ matrix by a scalar, the determinant is multiplied by the $n-$th power of the scalar (for instance, the opposite matrix is obtained multiplying by -1, so the determinant remains unchanged if n is even, or it changes sign if n is odd).
Algebraic complements (or adjoints, or co-factors).
Expansion of the determinant using co-factors: it is equal to the scalar product of a row or column with the corresponding co-factors (only trace of the proof).
The scalar product of a row (or column) with the co-factors of another row (or column) is zero.
Adjoint matrix: it is the transpose of the matrix of co-factors.
Pre- or post-multiplying a matrix with its adjoint produces a diagnonal (scalar) matrix, whose diagonal elements are all equal to the determinant.
Singular matrix (determinant is zero).
Inverse of a non-singular matrix: it is obtained dividing the adjoint matrix by the determinant.
Determinant of the product (rows by columns) of two square matrices (only trace of the proof).
Determinant of the inverse.
Determinant of a diagonal matrix.

Equation systems

Solution of a linear system of n equations with n unknowns; Cramer's rule.
In a homogeneous system of n equations with n unknowns, if the coefficients matrix is non-singular, the unique solution is the null vector; other solutions are possible only if the matrix is singular.
In a non-singular matrix, rows (and columns) are linearly independent; if rows (and columns) are linearly dependent, the matrix is singular.

Partitioned matrices

Sottomatrici quadrate (di matrici rettangolari o quadrate) e minori.
Matrici partizionate.
Matrice rettangolare diagonale a blocchi.
Somma di matrici partizionate (uguali dimensioni delle matrici e uguali dimensioni dei blocchi corrispondenti): si sommano i blocchi corrispondenti.
Matrice quadrata A $(n \times n)$ partizionata in 4 blocchi, di cui quelli diagonali $A_{1,1}$ $(n_1 \times n_1)$ e $A_{2,2}$ $(n_2 \times n_2)$ quadrati (con $n_1+n_2=n$), mentre quelli non diagonali $A_{1,2}$ $(n_1 \times n_2)$ e $A_{2,1}$ $(n_2 \times n_1)$ non sono necessariamente quadrati; se una matrice B $(n \times n)$ viene partizionata in modo analogo, la matrice prodotto AB $(n \times n)$ può essere partizionata in modo analogo ad A e B; il blocco $1,1$ della matrice prodotto vale $(AB)_{1,1} = A_{1,1}B_{1,1} + A_{1,2}B_{2,1}$; il blocco $1,2$ vale $(AB)_{1,2} = A_{1,1}B_{2,1} + A_{1,2}B_{2,2}$, eccetera; si applicano cioè ai blocchi le stesse regole del prodotto righe per colonne.
La regola precedente vale anche per il prodotto di matrici rettangolari partizionate, purché le matrici siano di dimensioni compatibili, e i blocchi siano di dimensioni compatibili.
Se X è una matrice (rettangolare o quadrata) diagonale a blocchi, $X'X$ è una matrice quadrata diagonale a blocchi, con blocchi diagonali quadrati.
Inversa di una matrice quadrata A partizionata in quattro blocchi, di cui $A_{1,1}$ e $A_{2,2}$ quadrati; si indica con B la matrice inversa $B = A^{-1}$, e la si partiziona in maniera analoga; i quattro blocchi della matrice inversa valgono: $B_{1,1} = (A_{1,1} - A_{1,2}A_{2,2}^{-1}A_{2,1})^{-1}$; $B_{2,2} = (A_{2,2} - A_{2,1}A_{1,1}^{-1}A_{1,2})^{-1}$; $B_{1,2} = -A_{1,1}^{-1}A_{1,2}B_{2,2}$; $B_{2,1} = -A_{2,2}^{-1}A_{2,1}B_{1,1}$; sviluppando il prodotto delle due matrici partizionate, si verifica che $AB = I$.
Caso particolare: se la matrice è diagonale a blocchi (e i due blocchi diagonali sono quadrati), l'inversa è diagonale a blocchi; i blocchi diagonali dell'inversa sono gli inversi dei corrispondenti blocchi diagonali della matrice data; per la dimostrazione, basta osservare che $A_{1,2} = 0$ e $A_{2,1} = 0$; questa proprietà vale anche per matrici diagonali a blocchi con tre o più blocchi diagonali; per la dimostrazione, basta considerare la matrice come se fosse partizionata con due blocchi diagonali, ognuno dei quali eventualmente partizionato come una matrice diagonale a blocchi.

Eigenvalues and eigenvectors

Autovalori, o radici caratteristiche, o radici latenti di una matrice quadrata; autovettori, o vettori caratteristici, o vettori latenti corrispondenti; equazione caratteristica.

Autovalori e autovettori di una matrice quadrata reale possono essere reali o complessi (coniugati); ad autovalori reali corrispondono autovettori reali.
L'autovettore che corrisponde ad un certo autovalore non è definito in modo univoco; ad esempio, è certamente definito a meno di una costante moltiplicativa, nel qual caso è definita la direzione, ma non la lunghezza; nel caso di autovalori con molteplicità maggiore di uno non è definita in modo univoco nemmeno la direzione (ad esempio, per la matrice I, i cui autovalori sono tutti uguali a uno, qualsiasi vettore è autovettore, dunque l'autovettore non è definito univocamente nemmeno in direzione).
Una matrice quadrata simmetrica $n \times n$ ha n autovalori (distinti o multipli).
Una matrice quadrata simmetrica ha solo autovalori e autovettori reali.
In una matrice quadrata simmetrica, ad autovalori distinti corrispondono autovettori ortogonali.
In una matrice quadrata simmetrica, se un autovalore ha molteplicità k, esistono k autovettori ortogonali tra loro, corrispondenti a tale autovalore (senza dimostrazione).
Una matrice quadrata simmetrica $n \times n$, con autovalori non necessariamente distinti, ha n autovettori tra loro ortogonali; normalizzando ogni autovettore (lunghezza 1) si ottengono n autovettori ortonormali; questi n autovettori possono essere ordinati nelle colonne di una matrice quadrata Q, di ordine n, che gode della seguente proprietà: $Q'Q = I$, quindi $Q' = Q^{-1}$, quindi anche $QQ' = I$, quindi anche i vettori riga della matrice Q sono n vettori ortonormali.
Matrice ortogonale.
In una matrice ortogonale il determinante vale 1 o -1.
Se A è una matrice quadrata simmetrica, la matrice ortogonale degli autovettori diagonalizza A, cioè $Q'AQ = \Lambda$, dove Λ è la matrice diagonale degli n autovalori.
In una matrice quadrata simmetrica, il determinante è il prodotto degli autovalori.
Se A è una matrice quadrata simmetrica, gli autovalori di $A^2 = AA$ sono i quadrati degli autovalori di A, mentre gli autovettori di A^2 sono gli stessi di A.
Se A è una matrice quadrata simmetrica non singolare, gli autovalori di A^{-1} sono i reciproci degli autovalori di A, mentre gli autovettori di A^{-1} sono gli stessi di A.
Minore diverso da zero di ordine massimo.

Trace, idempotent matrices

$r(X'X) = r(XX') = r(X)$; if X has full rank $= k$ ($k < n$), also $X'X$ has full rank ($= k$), but not XX' (whose rank is k, but dimensions $n \times n$) (without proof).
$r(AB)$ is less than or equal to the smaller between $r(A)$ and $r(B)$ (without proof).
If B is a non-singular (thus, full rank) square matrix, then $r(AB) = r(A)$; in fact $r(AB) \leq r(A)$ and $r(A) = r[(AB)B^{-1}] \leq r(AB)$.
Il rango di una matrice quadrata simmetrica è uguale al numero degli autovalori diversi da zero.
Trace of a square matrix.
$Tr(AB) = Tr(BA)$ (if A e B have dimensions that allow both products).
La traccia di una matrice quadrata simmetrica è uguale alla somma degli autovalori.
Idempotent matrices.
Examples of idempotent matrices and their trace; matrix 0, I, $A = I - \iota\iota'/n$, the projection matrices $P_X = X(X'X)^{-1}X'$, $M_X = I - P_X = I - X(X'X)^{-1}X'$.
Use of the matrices A, P_X e M_X: if y is a vector, Ay is the vector containing the deviations of the elements of y from their arithmetical average ($Ay = y - \bar{y}$); $P_X y$ is the projection of the vector y on the plane (hyperplane) spanned by the columns of the matrix X (example with a 2−columns matrix X; first of all show what happens if y is one of the two columns of X, then show what happens if y is a generic vector of the plane and finally a generic vector y is decomposed into a component on the plane and a component orthogonal to the plane); $M_X y = y - P_X y$, that is the projection of the vector y on the straight line orthogonal to the plane (hyperplane) spanned by the columns of X.
In una matrice quadrata simmetrica idempotente gli autovalori valgono 0 o 1; il rango è quindi uguale alla traccia.

Quadratic forms

Quadratic form: if x is an n−dimensional vector and A is an $n \times n$ matrix, the *scalar* $x'Ax$ is called quadratic form; its value can be obtained from the (scalar) operation $\sum_i \sum_j a_{i,j} x_i x_j$.
Positive semidefinite square matrices.
Positive definite square matrices (a subset of the above).
A positive definite matrix is non-singular (columns are linearly independent).
$A'A$ and AA' are always symmetric positive semidefinite matrices, whatever the (square or rectangular) matrix A.
Inequality between matrices: given two square matrices, positive semidefinite (or definite) with the same dimensions, the former is said to be *greater than* the latter if the difference matrix is not null and positive semidefinite.
The inverse of a positive definite matrix is itself positive definite (the proof would be based on the properties of eigenvalues).
If a matrix is positive semidefinite, but not positive definite, it is singular (its columns are linearly dependent).
If P has dimensions $m \times n$, with $n \leq m$, and $r(P) = n$ (full rank), then $P'P$ (square $n \times n$ matrix) is positive definite; to prove it, given any non-null n−dimensional vector c, build the quadratic form $c'P'Pc = (Pc)'Pc$; it is a sum of squares, where Pc cannot be the null vector, being a linear combination of all the linearly independent columns of P; thus the result is always a strictly positive number; in addition, if A is an $n \times n$ symmetric positive definite matrix, $P'AP$ is also symmetric and positive definite.

In particular, if P is a non-singular square matrix (full rank), then both $P'P$ and PP' are positive definite.

First and second derivatives

Vettore delle derivate prime di una funzione scalare rispetto al vettore delle variabili (gradiente); casi particolari: derivare un prodotto scalare rispetto a uno dei due vettori $\partial(x'y)/\partial x = y$, $\partial(x'y)/\partial y = x$; derivare la forma quadratica $x'Ax$ rispetto al vettore x: $\partial(x'Ax)/\partial x = (A + A')x$.
Matrice delle derivate prime di un vettore di funzioni rispetto al vettore delle variabili (Jacobiano); caso particolare $\partial(Bx)/\partial x' = B$.
Matrice delle derivate seconde di una funzione scalare rispetto a un vettore di variabili (Hessiano): caso particolare: derivare due volte la forma quadratica $x'Ax$ rispetto al vettore x: $\partial^2(x'Ax)/\partial x \partial x' = A + A'$ (che è sempre una matrice simmetrica).
Massimi e minimi di funzioni di più variabili: gradiente zero e Hessiano definito negativo o positivo.
Massimi e minimi vincolati; vettore dei moltiplicatori di Lagrange.

ELEMENTS OF STATISTICAL ANALYSIS

Probability and discrete random variables.
Expectation (or expected value, or mean), variance, standard deviation of a discrete random variable.
Expectation and variance of a random variable *are not* random variables.
Transforming a random variable with a function produces a new random variable.
Expectation and variance of a function of random variable (same formula, but the original variable is replaced by the transformed variable).
A function of several random variables is itself a random variable.
Variance is always non-negative; it is zero if the random variable is *degenerate* (a constant).
Variance is equal to the expectation of the square minus the square of the expectation.
Expectation of the product of a constant with a random variable: $E(ax) = aE(x)$.
Expectation of the sum of two random variables.
Expectation of a linear combination of random variables with constant coefficients; it is equal to the linear combination of the expectations (expectation is a *linear* operator).
Variance of the product of a constant with a random variable: $Var(ax) = a^2 Var(x)$.
If k is a constant, $Var(x) = Var(x - k)$; in particular $Var(x) = Var[x - E(x)]$.
Continuous random variable.
Cumulated distribution function and probability density function.
Expectation, variance and standard deviation of a continuous random variable.
Bivariate and multivariate discrete random variables.
Bivariate and multivariate continuous random variables.
Probability density for bivariate and multivariate continuous random variables (also called joint probability density function).
Marginal probability density.
Conditional probability density.
The joint probability density for a bivariate random variable is the product of the marginal density of the former variable with the conditional probability density of the latter given the former.
Independent random variables: marginal and conditional probability densities are equal; the joint probability density is the product of the marginal probability densities.
Expectations and variances of the components of a multivariate random variable (discrete or continuous) are computed from the marginal probability densities.
Covariance of two random variables: $Cov(x, y)$.
Covariance *is not* a random variable.
$Cov(x, y) = Cov(y, x)$.
Covariance is equal to the expectation of the product minus the product of the two expectations $Cov(x, y) = E(xy) - E(x)E(y)$.
If a and b are constants, $Cov(ax, y) = Cov(x, ay) = aCov(x, y)$ and $Cov(ax, by) = abCov(x, y)$.
Expectation of the product of two random variables is equal to the product of the two expectations plus the covariance.
In a multivariate random variable, covariances are for *pairs* or *couples* of component elements.
Covariance may be positive, null or negative.
Correlation (or correlation coefficient) between two random variables: it is the covariance divided by the square root of the product of the two variances.
Correlation coefficient is a number between -1 and 1.
Two random variables are called *uncorrelated* when the covariance (and thus the correlation) is zero.
Independent random variables are always uncorrelated; not viceversa: uncorrelated random variables are not necessarily independent; for example sum of two dice and difference of the same dice are uncorrelated but not independent; (the multivariate normal variable is the most important counter-example; when two component elements are uncorralated, they are also independent).
Expectation of the product of two uncorrelated random variables is simply the product of the expectations (the covariance is zero).
Functions of independent random variables are themselves independent random variables (thus uncorrelated) functions of uncorrelated random variables are not necessarily uncorrelated.

The n component elements $X_1, X_2, ..., X_n$ of an $n-$variate random variable, x, (or the n random variables X_1, X_2, ..., X_n) can be represented with an $n-$dimensional vector, called random vector, or vector of random variables.
Expactation of a random vector: $E(x)$.
Variance-covariance matrix of a random vector x is defined as $Var(x) = Cov(x) = E\{[(x - E(x)][x' - E(x')]\}$.
If a is a constant vector (non-random), then $a'x$ is a scalar random variable, linear combination of the elements of x; its expectation is therefore $E(a'x) = a'E(x)$, being the expectation of a linear combination.
The variance of the scalar random variable $a'x$ is $Var(a'x) = a'Var(x)a$.
The variance-covariance matrix of a random vector x is symmetric (because $Cov(x_i x_j) = Cov(x_j, x_i)$) and positive semidefinite infact, if a is a constant vector with the same dimension of x, then $a'Var(x)a$ is the variance of the scalar random variable $a'x$, therefore it is always non-negative; if it cannot happen that $a'x$ *degenerates* for some non-zero vector a (it cannot become a constant; its variance is therefore always strictly positive), then the variance-covariance matrix of x is positive definite.
If A is a constant (non-random) matrix (with conformable dimensions) $E(Ax) = AE(x)$ and $Var(Ax) = AVar(x)A'$.
The variance-covariance matrix of uncorrelated random variables is a diagonal matrix; if the variance is the same for each component element, then the matrix is a scalar matrix (constant elements on the diagonal).
Uniform distribution (discrete and continuous).
The sum of two uniform random variables *does not* have uniform ditribution.
The normal distribution (or Gaussian distribution).
The formula of the probability density function defines a family of probability distributions, indexed by two parameters, usually called μ and σ^2; computing expectation and variance of the random variable, they turn out to be exactly equal to those two parameters.
The probability of a value of the normal random variable to be between $\mu \pm \sigma$ is approximately 66%; to be between $\mu \pm 2\sigma$ is approximately 95%.
A normal random variable with expectazion zero and unit variance is called standard normal;
Any normal random variable is transformed into a standard normal subtracting the expected value and dividing by the standard deviation.
Tables for the standard normal distribution.
If a random vector (x) is such that any linear combination of its elements with constant coefficients $(a'x)$ is a random variable with normal distribution, the distribution of the random vector x is called *multivariate normal*.
Explicit formula for the probability density of a multivariate normal exists if and only if the variance-covariance matrix is non-singular; the formula involves the vector of expected values and the variance-covariance matrix; the usual notation is $x \sim N(\mu, \Sigma)$.
Random vectors obtained as linear combinations (with constant coefficients) of the elements of a multivariate normal vector are themselves multivariate normal vectors; for instance, if x is a vector $N(\mu, \Sigma)$, then Ax is a vector $N(A\mu, A\Sigma A')$.
The χ^2 distribution.
Summing the squares of n independent standard normal variables, the random variables obtained is called χ^2 with n degrees of freedom: χ^2_n.
The χ^2 family of probability distributions is indexed by one parameter (n, the number of degrees of freedom).
The expectation of a random variable χ^2 with n degrees of freedom is n; the variance is $2n$.
Tables for the χ^2 distributions, for varying degrees of freedom.
The Student's$-t$ distribution.
Given two independent random variables, the former with standard normal distribution, the latter distributed a χ^2 witn n degrees of freedom, divided by the constant n; the former divided by the square root of the latter produces a random variable called Studen's-t with n degrees of freedom: t_n.
The Student's$-t$ is a family of distributions indexed by one parameter (n, the number of degrees of freedom).
The probability density function is symmetric around zero.
Increasing n, the distribution becomes more and more close to the standard normal distribution (exactly equal when $n \to \infty$).
Tables for the Student's$-t$ distribution, t_n, for varying degrees of freedom.
Fisher's$-F$ distribution.
Given two independent random variables with χ^2 distribution, with n and k degrees of freedom, respectively, the ratio between the former (divided by n) and the latter (divided by k) is a random variable whose distribution is called Fisher's$-F$ with n, k degrees of freedom: $F_{n,k}$.
Fisher's$-F$ is a family of probability distribution indexed by two parameters (n and k, the numbers of degrees of freedom).
Tables for the Fisher's$-F$ distribution, $F_{n,k}$, for varying degrees of freedom.

MULTIPLE LINEAR REGRESSION MODEL (the simple linear regression as a particular case)

The assumptions of the multiple linear regression model

(1) A dependent or explained variable (also called regressand) is assumed to be a linear combination of some independent or explanatory variables (or regressors); the relationship is not exact, as it includes an additive error term (or unexplained disturbance); dependent variable and regressors are observable (no latent variables) and measured without errors (no measurement error); the coefficients of the linear combination are "fixed constants" (they are not random variables), but are unknown; the error terms are random variables and are not observable; the vector containing the n observations of the dependent variable (y_1, y_2, ..., y_i, ..., y_n) is called y ($n \times 1$); the n observations of the k explanatory variables are assembled

in a matrix X $(n \times k)$; if the model includes the intercept, matrix X has a column whose elements are all ones; the k coefficients of the linear combination are assembled in a vector β $(k \times 1)$; the vector u $(n \times 1)$ contains the error terms; the multiple linear regression model is represented as: $y = X\beta + u$; each element is $y_i = x_i'\beta + u_i$, being x_i' the $i-th$ row of X.
(2) All columns of X are linearly independent, thus $r(X) = k$; this implies, in particular, $n \geq k$; in other words, the number of observations (or sample size) cannot be smaller than the number of regressors (in practice, interest is confined to the case where strictly $n > k$).
(3) The expectation of all the error terms is zero: $E(u) = 0$.
(4) The error terms are uncorrelated (all covariances are zero) and homoskedastic (the variance of each error, called σ^2, is constant, but unknown): $Var(u_1) = Var(u_2) = \sigma^2$, etc; $Cov(u_1, u_2) = 0$, etc; $E(uu') = \sigma^2 I_n$ (scalar variance-covariance matrix).
(5) The contents of matrix X are known constants (*non-random* variables); since $E(u) = 0$, one gets $E(y) = X\beta$; the expected value (or "conditional" expected value) of each y_i is always a point on the regression line (or plane, or hyperplane).
(6) The vector of error terms u has a multivariate normal distribution; thus, combining assumption 6 with assumptions 3 and 4, u is distributed $N(0, \sigma^2 I_n)$.
The estimation method called "ordinary least squares" (OLS) provides an estimate of the unknown parameters of the model (coefficients β and variance σ^2); its algebraic properties are based on assumptions 1 and 2 (other assumptions being unnecessary); some statistical properties of the OLS estimation method are based on assumptions 1-5; finally, some other statistical properties need all assumptions (1-6).

OLS: algebraic properties (under assumptions 1-2)

Given a vector of coefficients β, the corresponding vector of "residuals" can be obtained as $u = y - X\beta$, thus each residual can be represented as a function of the variables y and X (observed) and coefficients (β, unknown); we look for the vector of coefficients (called $\hat{\beta}$) that minimize the sum of all squared residuals; the method is called OLS (ordinary least squares), and coefficients computed in this way are the OLS estimates of the regression coefficients (simply: OLS coefficients).
Under assumptions 1 and 2, OLS coefficients are available in closed form as $\hat{\beta} = (X'X)^{-1}X'y$; this expression is obtained equating to zero the first order derivatives of the sum of squared residuals with respect to the k coefficients β (first order conditions); it can then be verified that the $(k \times k)$ matrix of second order derivatives (Hessian) is positive definite (second order condition).
The vector that contains the computed values (or fitted values) of the dependent variable is $\hat{y} = X\hat{\beta}$.
The vector of OLS residuals is the difference between the vector of observed values and the vector of computed values of the dependent variable (computed with OLS coefficients): $\hat{u} = y - \hat{y} = y - X\hat{\beta} = y - X(X'X)^{-1}X'y = M_X y = M_X u$, where M_X is the idempotent symmetric matrix (or projection matrix) $M_X = I_n - X(X'X)^{-1}X'$, whose trace (=rank) is $n - k$.
If the number of observations (or sample size) is equal to the number of explanatory variables $n = k$ (instead of $n > k$), X would be a square matrix, thus $(X'X)^{-1} = X^{-1}(X')^{-1}$, thus $M_X = 0$, thus $\hat{u} = 0$; in other words, all the points of the sample would lie on the regression line (or plane, or hyperplane).
The vector of OLS residuals is orthogonal to each explanatory variable (or regressor): $X'\hat{u} = 0$; with different words, one can say that OLS residuals are uncorrelated in the sample with each regressor.
The vector of OLS residuals is orthogonal to the vector of computed value of the dependent variable: $\hat{y}'\hat{u} = 0$.
If the regression model includes the intercept, then the matrix of regressors includes a column whose values are all ones (a constant regressor); thus the sum of residuals is zero; if the model is without intercept, the sum of OLS residuals may be nonzero.
In particular, in a simple linear regression model with intercept $y = \beta_1 + \beta_2 z + u$, the point with coordinates ($\bar{z}$,$\bar{y}$, arithmetical averages) is on the regression line estimated by OLS; measuring variables y_i and z_i as deviations from their arithmetical averages is like shifting the origin of the Cartesian axes over the point ($\bar{z}$,$\bar{y}$); thus an OLS estimation of the model without intercept $y_i - \bar{y} = \beta_2(z_i - \bar{z}) + u$ would produce the same value $\hat{\beta}_2$ and the same residuals $\hat{u}$ as the OLS estimation of the original model (with intercept).
Coefficient of determination (R^2) for the model with intercept: it is a measure of the fit of the model (for the model without intercept the definition should be slightly modified; not done here).
Defining A as the symmetric idempotent matrix that produces deviations from the arithmetical average, $A = I_n - \iota\iota'/n$, the sum of squares of the deviations of the y_i from their arithmetical average is: $TSS = (Ay)'Ay = y'Ay$ (total sum of squares).
$ESS = \hat{y}'A\hat{y}$ = sum of squares of the deviations of the $\hat{y}_i$ from their arithmetical average (explained sum of squares).
$RSS = \hat{u}'\hat{u}$ = residual sum of squares (remembering that residuals have arithmetical average zero).
In the model with intercept, $TSS = ESS + RSS$; to prove it, from $y = \hat{y} + \hat{u}$, pre-multiplication by A gives $Ay = A\hat{y} + \hat{u}$, then transposition of this expression and multiplication by the expression itself gives $y'Ay = \hat{y}'A\hat{y} + \hat{u}'\hat{u}$ (the cross products are zero because $\hat{u}$ is orthogonal to $\hat{y}$, and $A\hat{u} = \hat{u}$ because the model has the intercept).
The coefficient of determination is defined as $R^2 = ESS/TSS = 1 - RSS/TSS$.
The sample correlation coefficient between y and $\hat{y}$ is $= \sqrt{R^2}$; the proof follows from observing that the sample variances of y and $\hat{y}$ are, respectively, TSS/n and ESS/n, and the sample covariance is $(Ay)'(A\hat{y})/n = (A\hat{y} + \hat{u})'(A\hat{y})/n = ESS/n$.
R^2 in the model with intercept is a number between 0 and 1.
$R^2 = 0$ means "no fit"; $R^2 = 1$ means "perfect fit"; as a rough indicator of goodness of fit; usually, the larger the R^2, the better the fit; a remarkable exception is when $k = 1$ and the only regressor is the constant (all values = 1), so that $\hat{\beta} = \bar{y}$; thus $TSS = RSS$, thus $R^2 = 0$, even if the fit is good.

Adding new explanatory variables to the same equation necessarily improves the R^2 (that cannot decrease); intuitively, if the additional regressors are "meaningful", the improvement will be large, but if they are meaningless the improvement will be small or even null; it is possible to define an "adjusted" R^2, that takes into account the reduction of degrees of freedom due to the introduction of new regressors: $1-[RSS/(n-k)]/[TSS/(n-1)]$; it might become smaller after the introduction of a new regressor without explanatory power.

OLS: some statistical properties (under assumptions 1-5; valid even without intercept)

The vector of estimated coefficients is a random vector (unlike the "true" coefficients vector β, which is a *non-random* vector).
The vector of coefficients estimation errors is $\hat{\beta}-\beta=(X'X)^{-1}X'u$.
Under assumptions 1-5 (6 is unnecessary), OLS estimator is linear and unbiased, as $E(\hat{\beta}-\beta)=(X'X)^{-1}X'E(u)$, being X *non-random*.
Under assumptions 1-5, the variance-covariance matrix of the OLS coefficients is $Var(\hat{\beta})=(X'X)^{-1}\sigma^2$; the proof follows from computing $E[(\hat{\beta}-\beta)(\hat{\beta}-\beta)']=E[(X'X)^{-1}X'uu'X(X'X)^{-1}]$ where *expectation* will be applied only to uu', being X *non-random*.
Gauss-Markov theorem: Under assumptions 1-5 (6 is unnecessary), OLS coefficients have the smallest variance-covariance matrix, among all linear unbiased estimators; thus, OLS estimator is the most efficient linear unbiased estimator.
Proof: any linear estimator of the coefficients vector would be $B'y$, where B is a matrix with the same dimensions of X and does not contain random variables; unbiasedness of the estimator is ensured if and only if $B'X=I_k$; defining $A'=B'-(X'X)^{-1}X'$, unbiasedness of the estimator is ensured if and only if $A'X=0$; the variance-covariance matrix of the coefficients obtained with the new estimation method is $B'B\sigma^2$, which is greater than the variance-covariance matrix of the OLS coefficients $(X'X)^{-1}\sigma^2$, the difference being $A'A\sigma^2$, positive semi-definite, having taken into account the unbiasedness condition $A'X=0$.
Corollary: With B satisfying the unbiasedness condition, defining the *non-random* $(n\times k)$ matrix $W=BX'X$, it follows that $B'y=(W'X)^{-1}W'y$; viceversa, if W is an arbitrary $(n\times k)$ matrix, not containing random variables, such that $W'X$ is non-singular, then the linear estimator $(W'X)^{-1}W'y$ is unbiased; thus, any linear unbiased estimator can be expressed as $\tilde{\beta}_W=(W'X)^{-1}W'y$; its variance-covariance matrix, being W *non-random*, is $(W'X)^{-1}W'W(X'W)^{-1}\sigma^2$; this matrix is always greater or equal to $(X'X)^{-1}\sigma^2$ (*Schwarz inequality*).
OLS estimator is BLUE (best linear unbiased estimator).
$RSS=\hat{u}'\hat{u}=u'M_Xu$ (even if the model has no intercept); its expected value is $E(RSS)=E(u'M_Xu)=E[tr(u'M_Xu)]=E[tr(M_Xuu')]=tr[E(M_Xuu')]=tr[M_XE(uu')]=tr(M_X\sigma^2)=tr(M_X)\sigma^2=(n-k)\sigma^2$.
Thus $E[RSS/(n-k)]=\sigma^2$; thus an unbiased estimator of the variance of the error terms is $\hat{\sigma}^2=RSS/(n-k)$.
Summarizing: $\hat{\sigma}^2=RSS/(n-k)=\hat{u}'\hat{u}/(n-k)=u'M_Xu/(n-k)$; its square root $(\hat{\sigma})$ is called "standard error" of the regression.
Since X does not contain random variables, $(X'X)^{-1}\hat{\sigma}^2$ is an unbiased estimator of the variance-covariance matrix of the OLS coefficients; the $j-th$ diagonal element $[(X'X)^{-1}]_{j,j}\hat{\sigma}^2$ is an unbiased estimator of the variance of the $j-th$ estimated coefficient $(\hat{\beta}_j)$, and its square root is the standard error of $\hat{\beta}_j$.
Forecast (or prediction) at time h (not belonging to the sample estimation period 1, 2, ..., n): given the vector of explanatory variables at time h, x_h, assumed known (conditional prediction), the best prediction at time h of the dependent variable y would be the expectation (conditional on x_h) of y at time h, that will be indicated as $\bar{y}_h=x_h'\beta$; if the model is correctly specified the "true" value of y at time h will be affected by a random error u_h and therefore will be $y_h=x_h'\beta+u_h$; being the "true" coefficients β unknown, and being $\hat{\beta}$ the available estimate, the actual prediction will be the estimated conditional expectation of y at time h, that is $\hat{y}_h=x_h'\hat{\beta}$; with a geometric notation, prediction would be the point on the estimated regression line (or plane, or hyperplane) corresponding to x_h.
Forecast error (or prediction error) at time h: it is the difference between prediction and the "true" value of y at time h, that is $\hat{y}_h-y_h$.
Variance of the forecast error (or simply variance of forecast): adding and subtracting the same quantity gives $\hat{y}_h-y_h=(\hat{y}_h-\bar{y}_h)+(\bar{y}_h-y_h)=x_h'(\hat{\beta}-\beta)-u_h=x_h'(X'X)^{-1}X'u-u_h$; it is the sum of two uncorrelated random variables (since the forecast period h does not belong to the sample estimation period, u_h is uncorrelated with the n "in sample" elements of the vector u, according to assumption 4), thus the variance is the sum of the two variances; the variance of the second component is σ^2 (constant for any x_h), while the variance of the first component is $x_h'(X'X)^{-1}x_h\sigma^2$, thus it depends on the values of the explanatory variables in the forecast period (x_h); in the simple linear regression model, with two variables $y=\beta_1+\beta_2z+u$, the variance has a minimum when z_h is equal to the arithmetical average of the elements of z in the sample, and becomes larger and larger as z_h moves far away from the average.

Distribution of linear and quadratic forms (built from multivariate normal vectors)

(1) If the random vector z (whose dimension is $n\times1$) has a multivariate standard normal distribution $N(0,I_n)$, then $z'z$ has a χ^2 distribution with n degrees of freedom (χ^2_n).
(2) If the vector z (whose dimension is $n\times1$) contains $k<n$ elements $=0$, while the other $n-k$ elements form a vector $N(0,I_{n-k})$, then $z'z$ has a χ^2 distribution with $n-k$ degrees of freedom (χ^2_{n-k}).
(3) If the random vector z (whose dimension is $n\times1$) has a multivariate standard normal distribution $N(0,I_n)$ and A is a matrix of constants, with dimensions $n\times n$, symmetric, idempotent with rank $n-k\leq n$, then the univariate random variable $z'Az$ has a χ^2 distribution with $n-k$ degrees of freedom; the proof is based on the decomposition $A=Q'\Lambda Q$, where Q is an orthogonal matrix $(Q'=Q^{-1})$ and Λ is the diagonal matrix containing the eigenvalues; among eigenvalues, there are

$n-k$ ones, while the others are zeroes; also Λ is idempotent ($\Lambda\Lambda = \Lambda$); the vector Qz has a multivariate normal distribution $N(0, I_n)$ (since $QQ' = I_n$); ΛQz is a random vector with n elements; $n-k$ elements have a $N(0, I_{n-k})$ distribution, while the other k elements are zeroes; finally, $z'Az = (\Lambda Qz)'\Lambda Qz$ and the results follows from applying (2).
(4) If the random vector x (whose dimension is $n \times 1$) has a multivariate normal distribution $N(0, \Sigma)$, where Σ is a $n \times n$ symmetric positive definite matrix, then the univariate random variable $x'\Sigma^{-1}x$ has a χ^2 distribution with n degrees of freedom; to prove it, first Σ must be decomposed as $\Sigma = P'P$, where P is a non-singular square matrix; it follows that $z = (P')^{-1}x$ has a zero mean multivariate normal distribution with variance-covariance matrix $(P')^{-1}\Sigma P^{-1} = (P')^{-1}P'PP^{-1} = I_n$; thus, z has a multivariate standard normal distribution, and the result follows from (1).
(5) If the random vector z (whose dimension is $n \times 1$) has a multivariate standard normal distribution $N(0, I_n)$, if A and B are two "constant" matrices with dimensions $m \times n$ and $k \times n$ respectively, and their product is $AB' = 0$ (null matrix), then the two random vectors Az ($m \times 1$) and Bz ($k \times 1$) are independent random vectors; to prove it, Az and Bz must be regarded as two sub-vectors of a multivariate normal random vector $[(m+k) \times 1]$, and the matrix that contains covariances between all the elements of Az and Bz is AB' (thus $= 0$); finally it is enough to remember that uncorrelated elements of a multivariate normal are independent.
(6) If A, B and z are as in (5), any transformation of the vector Az and any transformation of the vector Bz will produce independent random variables (or vectors).
(7) If A and B are as in (5), and the random vector x (with dimension $n \times 1$) is distributed $N(0, \sigma^2 I_n)$, the random vectors Ax and Bx will be independent multinormal random vectors, and any transformation of each of the two vectors will produce independent random variables (or vectors); the proof follows from (5) or (6) simply dividing each vector by the scalar constant σ, and remembering that $z = x/\sigma$ is $N(0, I_n)$.
(8) As a particular case of (7), if the random vector x (with dimension $n \times 1$) is distributed $N(0, \sigma^2 I_n)$, and A and B are both square symmetric idempotent matrices ($n \times n$) such that $AB = 0$, then the two quadratic forms $x'Ax/\sigma^2$ and $x'Bx/\sigma^2$ are independent scalar random variables; in addition, it follows from (3) that each of the two quadratic forms has a χ^2 distribution with degrees of freedom equal to the rank (therefore also equal to the trace) of the matrix A or B, respectively.

Statistical inference in the multiple linear regression model (under assumptions 1-6; also 6 is necessary)

Coefficients estimated by (Gaussian) maximum likelihood are equal to the OLS coefficients, and their variance-covariance matrix is the inverse of Fisher's information matrix (Cramér-Rao bound); remember that Gauss-Markov theorem did not use the assumption of normality, and proved efficiency among "linear unbiased" estimators; here, under the additional assumption of normality (6), OLS is efficient with respect to "any unbiased" estimator (proof, see sect. 17.6).
The vector of coefficient estimation errors $\hat{\beta} - \beta = (X'X)^{-1}X'u$ is a linear combination of u (multivariate normal); thus it has a multivariate normal distribution $N[0, (X'X)^{-1}\sigma^2]$.
The $j-th$ estimated coefficient ($\hat{\beta}_j$) has a normal distribution with mean β_j and variance $[(X'X)^{-1}]_{j,j}\sigma^2$.
The vector u/σ has a multivariate normal distribution $N(0, I)$.
$RSS/\sigma^2 = \hat{u}'\hat{u}/\sigma^2 = (u'/\sigma)M_x(u/\sigma)$, where M_x is symmetric, idempotent and its rank is $n-k$, is a random variable with distribution χ^2 with $(n-k)$ degrees of freedom.
Since $\hat{\sigma}^2 = RSS/(n-k)$, then the ratio $\hat{\sigma}^2/\sigma^2$ is a random variable χ^2_{n-k} divided by the number of degrees of freedom $n-k$.
The two random vectors $\hat{\beta} - \beta$ and $\hat{u}$ are independent, since $\hat{\beta} - \beta = (X'X)^{-1}X'u$, $\hat{u} = M_x u$, and the product of the two matrices $(X'X)^{-1}X'M_x = 0$.
Any transformation of $\hat{\beta} - \beta$ and of $\hat{u}$ will produce independent random variables; in particular, $\hat{\beta} - \beta$ is independent of $\hat{\sigma}^2$.
(1) $\hat{\beta}_j - \beta_j$, divided by the square root of its variance $[(X'X)^{-1}]_{j,j}\sigma^2$, is a standard normal random variable.
(2) If R is a constant row vector, the scalar variable $(R\hat{\beta} - R\beta)$ divided by the square root of its variance $[R(X'X)^{-1}R']\sigma^2$, is a standard normal random variable.
(3) The quadratic form $(\hat{\beta} - \beta)'(X'X/\sigma^2)(\hat{\beta} - \beta)$ is a random variable χ^2_k.
(4) If R is a constant matrix with dimensions $(q \times k)$ and rank q, the quadratic form $(R\hat{\beta} - R\beta)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - R\beta)/\sigma^2$ is a random variable χ^2_q; the proof follows observing that $R\hat{\beta} - R\beta = R(\hat{\beta} - \beta)$ is a $q \times 1$ random vector with multivariate normal distribution, zero mean and variance-covariance matrix $[R(X'X)^{-1}R']\sigma^2$.
Case (1) is a particular case of (2), obtained when R is a row vector of all zeroes, but the $j-th$ element $= 1$.
Case (3) is a particular case of (4), obtained when R is the identity matrix $k \times k$.
If r is a constant vector with dimension $q \times 1$, then $R\beta = r$ is a system of q linear restrictions (or linear constraints) on the k coefficients; in particular, if $q = 1$ (that is matrix R is a row vector and r is a scalar constant), $R\beta = r$ represents "one" linear restriction on coefficients.
Suppose that σ^2 is known, then a test of "one" coefficient or a test of "one" linear restriction on coefficients (cases 1 and 2) could be done using the standard normal distribution.
Suppose that σ^2 is known, then a test of q linear restrictions on coefficients (also called multiple restriction, case 4) could be done using the χ^2_q distribution; in particular a test of all coefficients (case 3) would use the χ^2_k distribution.
Usually σ^2 is unknown, and the formulas of cases 1, 2, 3 and 4 can be applied replacing σ^2 with its unbiased estimate $\hat{\sigma}^2$; as a consequence, the test statistics that had a standard normal distribution (cases 1 and 2) are now distributed as a Student's$-t$ with $n-k$ degrees of freedom; the test statistics that had χ^2 distributions with k or q degrees of freedom (cases 3 and 4) are now distributed as by Fisher's$-F$ with $k, n-k$ or $q, n-k$ degrees of freedom, after the expressions of the test statistics are divided by k or q, respectively.
The proof follows observing that, in all cases, σ^2 is always at the denominator (under square root in cases 1 and 2); replacing σ^2 with $\hat{\sigma}^2$ is equivalent to multiplying the denominator by the ratio $\hat{\sigma}^2/\sigma^2$, that is a random variable $\chi^2_{n-k}/(n-k)$ (under

square root in cases 1 and 2) independent of the numerator; thus, the standard normal will produce a Student's$-t$ with $n-k$ degrees of freedom (cases 1 and 2); in case 3, the random variable χ^2_k will be divided by an independent random variable $\chi^2_{n-k}/(n-k)$, thus a further division by k will produce a Fisher's$-F$ with $k, n-k$ degrees of freedom; in case 4, the random variable χ^2_q will be divided by an independent random variable $\chi^2_{n-k}/(n-k)$, thus a further division by q will produce a Fisher's$-F$ with $q, n-k$ degrees of freedom.
(1bis) $(\hat{\beta}_j - \beta_j)/\sqrt{[(X'X)^{-1}]_{j,j}\hat{\sigma}^2}$, is a random variable with Student's$-t$ distribution (t_{n-k}).
(2bis) If R is a row vector of constants, the scalar $(R\hat{\beta} - R\beta)/\sqrt{R(X'X)^{-1}R'\hat{\sigma}^2}$ is a random variable with Student's$-t$ distribution (t_{n-k}).
(3bis) The quadratic form $(\hat{\beta} - \beta)'[X'X/(k\hat{\sigma}^2)](\hat{\beta} - \beta)$ is a random variable with Fisher's$-F$ distribution ($F_{k,n-k}$).
(4bis) If R is a matrix of constants, with dimensions $(q \times k)$ and rank q, the quadratic form $(R\hat{\beta} - R\beta)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - R\beta)/(q\hat{\sigma}^2)$ is a random variable with Fisher's$-F$ distribution ($F_{q,n-k}$).
Examples of tests that use the Student's$-t$ distribution.
The null hypothesis concerns the exact value of the $j-th$ coefficient, while the alternative is that such a coefficient has a different value; this is usually written as $H_0: \ \beta_j = r$; $H_1: \ \beta_j \neq r$, where r is a given constant; under the null hypothesis the ratio between $(\hat{\beta}_j - r)$ and the standard error of $\hat{\beta}_j$ will be a random variable with Student's$-t$ distribution (t_{n-k}); as a "default" option, all software packages test the null hypothesis $\beta_j = 0$, thus they simply compute the ratio between $\hat{\beta}_j$ and its standard error; under the null hypothesis such a ratio is a random variable with Student's$-t$ distribution (t_{n-k}); if this ratio (in absolute value) is greater than the critical value (at 5%, for instance), the null hypothesis is rejected in favour of the alternative hypothesis ($\beta_j \neq 0$, thus concluding that the $j-th$ regressor is significant).
The null hypothesis concerns the "equality" of two coefficients, that is $H_0: \ \beta_1 = \beta_2$; $H_1: \ \beta_1 \neq \beta_2$; the null hypothesis is a linear restriction that can be represented as $R\beta = r$, where $r = 0$ (scalar) and R is a row vector whose first two elements are 1 and -1, while all the others are zeroes; then, under the null hypothesis, the ratio between the scalar random variable $(R\hat{\beta} - r)$ and the square root of $[R(X'X)^{-1}R']\hat{\sigma}^2$ is a t_{n-k}; if this ratio (in absolute value) is greater than the critical value (at 5%, for instance), the null hypothesis is rejected in favour of the alternative hypothesis (thus concluding that the two coefficients are different).
The null hypothesis concerns the "sum" of two coefficients: H_0: $\beta_1 + \beta_2 = 1$; H_1: $\beta_1 + \beta_2 \neq 1$; for instance, the exponents of the two production factors in a Cobb-Douglas log-linear function become the coefficients of a linear regression model after variables have been transformed into logarithms, and the constant returns to scale hypothesis has to be tested; the null hypothesis is a linear restriction representable as $R\beta = r$, where $r = 1$ (scalar) and R is a row vector whose first two elements are 1, while all the others are zeroes; then the procedure is the same as in the previous case.
Examples of tests that use the Fisher's$-F$ distribution.
If the matrix R has dimensions $1 \times k$ (row vector) and its elements are all zeroes with the only exception of the $j-th$ element, which is 1, then the test statistic is distributed as a $F(1, n-k)$ and it is exactly equal to the square of the test statistic discussed above (when testing the hypothesis $\beta_j = 0$), which was distributed as a Student's$-t$ (t_{n-k}); the two tests always give the same result, since the critical value (for instance at 5%) of the $F(1, n-k)$ is exactly the square of the t_{n-k} critical value.
If r is a vector of k zeroes, and R is the identity matrix with dimensions $k \times k$, then the system of linear restrictions $R\beta = r$ means $\beta_1 = 0$, $\beta_2 = 0$, ..., $\beta_k = 0$, and the $F(k, n-k)$ test statistic is obtained from $\hat{\beta}'X'X\hat{\beta}/(k\hat{\sigma}^2)$; this could be considered a significance test of the whole regression; in fact, the null hypothesis would be accepted if no regressor is significant; in practice, this is done only for linear regressions without intercept.
Significance test for a subset of regressors; the usual procedure is applied using a suitable matrix R $(q \times k)$ with elements zeroes or ones, and a $(q \times 1)$ vector $r = 0$; as a "default" option for regression models with intercept, software packages test the null hypothesis that "all coefficients but the intercept" are zeroes, and this is the usual significance test of the whole regression.

Restricted least squares estimation

The method aims at producing coefficient values that minimize the sum of squared residuals satisfying, at the same time, q≤k linear restrictions $R\beta = r$; λ indicates a $q \times 1$ vector of *Lagrange multipliers* and is used to build the *Lagrangean function*: $f = (y - X\beta)'(y - X\beta) - 2\lambda'(R\beta - r)$ (minus sign and the factor 2 are introduced to simplify computation); estimates of β and λ are the solution of the system of first order conditions, obtained differentiating f with respect to β and λ; differentiating with respect to β gives $\partial f/\partial\beta = -2X'y + 2X'X\beta - 2R'\lambda$; differentiating with respect to λ gives $\partial f/\partial\lambda = -2(R\beta - r)$; the first order conditions are obtained equating to zero the two vectors of derivatives: (1) $X'X\hat{\alpha} - X'y - R'\hat{\lambda} = 0$; (2) $R\hat{\alpha} - r = 0$ (the system (1) and (2) is a system of $k + q$ linear equations with $k + q$ unknowns; to avoid confusion with the OLS coefficients of the unrestricted model, $\hat{\alpha}$ will be used to indicate the solution for coefficients, while $\hat{\lambda}$ will be the solution for the multipliers); pre-multiplying (1) by $R(X'X)^{-1}$ gives $R\hat{\alpha} - R(X'X)^{-1}X'y - R(X'X)^{-1}R'\hat{\lambda} = 0$, where substitution of $R\hat{\alpha}$ with r gives $r - R\hat{\beta} - R(X'X)^{-1}R'\hat{\lambda} = 0$, that produces the solution for the vector of Lagrange multipliers $\hat{\lambda} = [R(X'X)^{-1}R']^{-1}(r - R\hat{\beta})$; this expression of $\hat{\lambda}$ can be substituted into (1) giving $X'X\hat{\alpha} - X'y - R'[R(X'X)^{-1}R']^{-1}(r - R\hat{\beta}) = 0$, that provides the solution $\hat{\alpha} = \hat{\beta} - W(R\hat{\beta} - r)$, having indicated $W = (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}$.
After estimating coefficients that satisfy the system of linear restrictions, $\hat{\alpha}$, the corresponding residuals are $\hat{e} = y - X\hat{\alpha} = y - X\hat{\beta} - X(\hat{\alpha} - \hat{\beta}) = \hat{u} - X(\hat{\alpha} - \hat{\beta})$ where $\hat{u}$ is the vector of OLS residuals (unrestricted); pre-multiplication by the transpose gives $\hat{e}'\hat{e} = \hat{u}'\hat{u} + (\hat{\alpha} - \hat{\beta})'X'X(\hat{\alpha} - \hat{\beta})$ (the cross products vanish, because $X'\hat{u} = 0$); substituting the value of $\hat{\alpha} - \hat{\beta}$ computed above gives $\hat{e}'\hat{e} - \hat{u}'\hat{u} = (R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)$.

The above expression is the difference between the sums of squared residuals in the least squares estimations with restrictions and without restrictions, respectively; $\hat{e}'\hat{e} = RRSS$ is the restricted residual sum of squares, $\hat{u}'\hat{u} = URSS$ is the unrestricted residual sum of squares, (obviously the former is always greater or equal to the latter); the explicit formula just obtained for such a difference is equal to the numerator of the Fisher's$-F$ statistic when testing the system of q linear restrictions $R\beta = r$ (4bis).

Thus, an alternative expression of the Fisher's$-F$ test statistic can be used: $[(RRSS - URSS)/q]/[URSS/(n-k)]$, where $RRSS$ is the restricted residual sum of squares (that is, after restrictions have been imposed to the model), $URSS$ is the unrestricted residual sum of squares (the model estimated by OLS, without imposing any restriction), q is the number of restrictions; the denominator, as above, is the OLS unbiased estimate of the variance of the unrestricted model: $\hat{\sigma}^2 = URSS/(n-k)$; obviously, it is always $RRSS \geq URSS$, so that the value cannot be negative.

The above alternative expression of the Fisher's$-F$ test statistic can always be applied when testing a set of linear restrictions $R\beta = r$; instead of estimating by OLS the unrestricted model, and then applying the formula (4bis), the alternative procedure needs two OLS estimations, one of the original model (unrestricted) and one of the model after restrictions have been imposed (restricted model); $URSS$ and $RRSS$ are computed from the two set of residuals.

When restrictions produce a model whose OLS estimation is simple (in other words, when restricted least squares can be performed easily) the alternative procedure can be easier than the application of the formulas (3bis and 4bis); it does not lead to a simplification when restricted least squares is of difficult application.

As a "default" option for regression models with intercept (β_1), software packages test the null hypothesis that "all coefficients but the intercept" are zeroes, and this is the usual significance test of the whole regression, that is $H_0 : \beta_2 = 0$, $\beta_3 = 0$, ..., $\beta_k = 0$; thus the number of restrictions is $q = k - 1$, and the alternative expression of the Fisher's$-F$ is quite simple; the restricted model has a unique regressor (the constant), thus its coefficient is the arithmetical average of the dependent variable; thus $RRSS$ is the sum of squared deviations of the observed elements of y from their arithmetical average (called TSS, when dealing with the definition of R^2); therefore in this particular case the computation of the $F_{k-1,n-k}$ test statistic is quite similar to the computation of the R^2 (in particular, the R^2 adjusted for the degrees of freedom); a value of the test statistic greater than the critical value (for instance at 5%) of the $F_{k-1,n-k}$ distribution leads to rejection of the null hypothesis, thus accepting some sort of significance of the whole regression.

Other cases where it is simple to estimate the restricted least squares coefficients (thus the alternative form of the Fisher's$-F$ test statistic is of simple computation): when the null hypothesis is $\beta_1 = \beta_2$, or $\beta_1 + \beta_2 = 0$, or $\beta_1 + \beta_2 = 1$, or the hypothesis concerns a structure of *distributed lags* where weights decrese linearly.

Expected values of coefficients estimated by restricted least squares: since $\hat{\alpha} = \hat{\beta} - W(R\hat{\beta} - r)$, it is $E(\hat{\alpha}) = E(\hat{\beta}) - W(RE(\hat{\beta}) - r) = \beta - W(R\beta - r)$; thus, if the restrictions are "valid" so that $R\beta - r = 0$, the consequence is $E(\hat{\alpha}) = \beta$, thus estimation is unbiased; on the contrary, if restrictions are not valid, $R\beta - r \neq 0$ and the restricted least squares estimates are usually biased, the bias being $E(\hat{\alpha}) - \beta = -W(R\beta - r)$.

The variance-covariance matrix of coefficients estimated by restricted least squares is $Var(\hat{\alpha}) = Var[\hat{\beta} - W(R\hat{\beta} - r)] = Var(\hat{\beta} - WR\hat{\beta}) = Var[(I - WR)\hat{\beta}] = (I - WR)(X'X)^{-1}(I - WR)'\sigma^2 = (X'X)^{-1}\sigma^2 - WR(X'X)^{-1}\sigma^2 - (X'X)^{-1}R'W'\sigma^2 + WR(X'X)^{-1}R'W'\sigma^2$ (when W is replaced by its full expression, two terms will cancel out) $= (X'X)^{-1}\sigma^2 - (X'X)^{-1}R'$ $[R(X'X)^{-1}R']^{-1}$ $R(X'X)^{-1}\sigma^2$; thus it is equal to the variance-covariance matrix of OLS coefficients "minus" a symmetric positive semidefinite matrix; thus restricted least squares coefficients always have a variance-covariance matrix smaller than (at most equal to) the unrestricted OLS coefficients, no matter if restrictions are valid (thus the restricted estimate is unbiased) or not valid (thus the restricted estimate is usually biased); it must be noticed that the vector r does not appear in the expression of this matrix.

Specification error due to omission of relevant explanatory variables: let $X = [X_1, X_2]$, but instead of the correctly specified model $y = X_1\beta_1 + X_2\beta_2 + u$, with $k = k_1 + k_2$ regressors, the model that is estimated is $y = X_1\alpha_1 + e$, with k_1 regressors, omitting k_2 relevant regressors (X_2); it is like estimating the original model with restricted least squares, after imposing the restrictions $\beta_2 = 0$, which are not valid; thus there will usually be a bias in the estimated coefficients; explicit computation gives $E(\hat{\alpha}_1) = E[(X_1'X_1)^{-1}X_1'y] = E[(X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2 + u)] = \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2$; thus, bias depends, among other things, on the omitted variables and coefficients ($X_2\beta_2$); however, in a particular case the restricted estimate might be unbiased for the included coefficients: when the omitted regressors are orthogonal to the included regressors, that is when $X_1'X_2 = 0$; also the estimate of the variance is biased (overestimated); in fact $\hat{\sigma}^2 = (\hat{e}'\hat{e})/(n - k_1) = (y'M_{X_1}y)/(n - k_1) = (X_1\beta_1 + X_2\beta_2 + u)'M_{X_1}(X_1\beta_1 + X_2\beta_2 + u)/(n - k_1) = (X_2\beta_2 + u)'M_{X_1}(X_2\beta_2 + u)/(n - k_1)$ (being $M_{X_1}X_1 = 0$), and its expectation is $E(\hat{\sigma}^2) = \sigma^2 + \beta_2'X_2'M_{X_1}X_2\beta_2/(n - k_1)$; this sistematic overestimation occurs also when the omitted regressors are orthogonal to the included regressors: in fact, if $X_1'X_2 = 0$, the expected value is $\sigma^2 + \beta_2'X_2'X_2\beta_2/(n - k_1)$.

Specification error due to the inclusion of irrelevant explanatory variables (regressors that do not help explanation of the dependent variable): the correctly specified model is $y = X_1\beta_1 + e$, but OLS estimation is applied to a model that includes additional regressors, X_2, which are not relevant; this is like saying that the "unrestricted" model $y = X_1\beta_1 + X_2\beta_2 + u$ is correctly specified, with "true" values of the β_2 coefficients = 0; thus OLS estimation of the unrestricted model is unbiased (in particular with $E(\hat{\beta}_2) = 0$); the original model, $y = X_1\beta_1 + e$, can be viewed as obtained from the unrestricted model imposing the set of "valid" restrictions $\beta_2 = 0$, so OLS estimation is also unbiased and, having imposed restrictions, it has a variance-covariance matrix smaller than the unrestricted OLS coefficients; thus, including some irrelevant explanatory variables on the right hand side of the equation does not produce any bias, but reduces efficiency.

Test of structural change (also called Chow test): in a linear regression model, where the sample size is n, a change in the structure may have occurred; it is possible that the coefficients in the first sub-sample (n_1 observations) are different from coefficients in the second sub-sample (n_2 observations, where $n_1 + n_2 = n$); the null hypothesis to be tested is that coefficients

remained constant over the whole sample; vector y (n elements) is divided into the two sub-vectors y_1 e y_2, corresponding to the two sub-samples of the dependent variable; analogously, the matrix of regressors is divided into the two sub-matrices X_1 ($n_1 \times k$) and X_2 ($n_2 \times k$) and two vectors of coefficients are considered, β_1 and β_2, both $k \times 1$; the coefficients vector β ($2k \times 1$) contains the elements of β_1 followed by the elements of β_2; matrix X ($n \times 2k$) is a block-diagonal matrix whose diagonal blocks are X_1 and X_2, while off-diagonal blocks contain all zeroes; OLS estimation is applied to the model $y = X\beta + u$; the coefficients in the vector $\hat{\beta}$ ($2k \times 1$) are exactly the same that would be obtained from two separate OLS estimations of the model $y_1 = X_1\beta_1 + u_1$ using the first n_1 observations, and the model $y_2 = X_2\beta_2 + u_2$ on the last n_2 observations (the proof is straightforward, remembering that $X'X$ is a block-diagonal square matrix, thus it can be inverted simply inverting the two diagonal blocks that are $X_1'X_1$ and $X_2'X_2$); the sum of the n squared residuals is $URSS$; the null hypothesis is that $\beta_1 = \beta_2$ (no structural change); this test is based on the Fisher's$-F$ distribution and is usually applied using the alternative form of the test; so it is now necessary to estimate the model after imposing the k linear restrictions (or constraints) $\beta_1 = \beta_2$; the restricted model has a matrix of regressors X ($n \times k$) where X_1 and X_2 are two consecutive blocks (rather than diagonal blocks), and has a vector of coefficients β containing k elements; this restricted model $y = X\beta + u$ is estimated by OLS on the whole sample period (n observations) computing $RRSS$ as the sum of n squared residuals; it must be noticed that the unrestricted model has $2k$ regressors, and that the restricted model is obtained imposing k restrictions; thus the Fisher's$-F$ test statistic $(k, n-2k)$ is obtained as $[(RRSS - URSS)/k]/[URSS/(n-2k)]$; a value of the test statistic greater than the critical value (for instance at 5%) of the $F_{k,n-2k}$ distribution leads to rejection of the null hypothesis (that coefficients did not change in the two sub-samples), thus evidencing a structural change.

The test of structural change can be applied to a subset of coefficients; this can be done in the unrestricted model by "duplicating" only the coefficients that might change in the sample, and splitting the corresponding regressors, while the other regressors and coefficients remain unchanged; the degrees of freedom of the Fisher's$-F$ depend on the number of coefficients that are tested; for instance, when testing only one coefficient, the degrees of freedom of the Fisher's$-F$ will be $1, n-k-1$ (1 restriction, k regressors in the original model, $k+1$ regressors in the unrestricted model, after the split of one column into two columns).

The test of structural change can be applied also when two or more changes may have occurred in the sample; for instance it is possible that all coefficients have changed their values when passing from the first sub-sample (n_1 observations) to the second sub-sample (n_2 observations) and again to the third sub-sample (n_3 observations, where $n_1 + n_2 + n_3 = n$); the unrestricted model will have a block-diagonal matrix of regressors with 3 diagonal blocks (its dimensions will be $n \times 3k$), and a $(3k \times 1)$ vector of coefficients.

Remark. Estimation of the unrestricted model would be impossible if the block-diagonal matrix $X'X$ is singular; this happens if one of the sub-periods has a number of observations $< k$ (insufficient observations); for instance, if $n_2 < k$, $X_2'X_2$ is singular, so $X'X$ cannot be inverted, being $X_2'X_2$ its second diagonal block; a solution to this problem is the Chow predictive test, that estimates the restricted model on the longer sub-sample, uses it to predict the shorter sub-period, and finally considers the distribution of the prediction error.

Multiple linear regression model where the variance-covariance matrix is not scalar and it is "known"

(4-bis) If assumption (4) is not valid, the variance-covariance matrix of the error terms is represented as $E(uu') = \sigma^2\Omega$, where Ω ($n \times n$) is symmetric and positive definite; OLS estimation is unbiased, since $E(\hat{\beta} - \beta) = E[(X'X)^{-1}X'u] = 0$; the variance-covariance matrix of the OLS coefficients is $(X'X)^{-1}X'\Omega X(X'X)^{-1}\sigma^2$; if assumption (4) is not valid, Gauss-Markov may be not applicable; thus OLS may be inefficient.

Aitken's theorem: under assumptions 1, 2, 3, 4-bis and 5 (6 is unnecessary), if Ω is known (that is, the variance-covariance matrix is almost completely known, the only unknown being a scalar multiplicative constant called σ^2), coefficient estimated by generalized least squares (GLS) $\dot{\beta} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$ have the smallest variance-covariance matrix among all linear unbiased estimators; in other words, GLS is efficient with respect to any other linear unbiased estimator.

The proof follows from a straightforward application of Gauss-Markov theorem to an appropriate transformation of the model; first of all the GLS estimator is unbiased because $E(\dot{\beta} - \beta) = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}u = 0$, and its variance-covariance matrix is $(X'\Omega^{-1}X)^{-1}\sigma^2$; to prove efficiency, first of all matrix Ω must be decomposed as $\Omega = P'P$, where P is a non-singular square matrix ($n \times n$); the model is then transformed pre-multiplying by P'^{-1}, which gives $P'^{-1}y = P'^{-1}X\beta + P'^{-1}u$; the transformed variables are now called $q = P'^{-1}y$, $Q = P'^{-1}X$ and the transformed error terms are called $\varepsilon = P'^{-1}u$; the transformed variable are thus related through the linear regression model $q = Q\beta + \varepsilon$, where coefficients are the same as in the original model, and the error terms are such that $E(\varepsilon) = 0$, $E(\varepsilon\varepsilon') = P'^{-1}E(uu')P^{-1} = P'^{-1}\Omega P^{-1}\sigma^2 = \sigma^2 I_n$; thus the transformed model satisfies all the conditions underlying Gauss-Markov theorem, thus OLS is efficient when it is applied to the transformed model (instead of the original model), which gives: $\dot{\beta} = (Q'Q)^{-1}Q'q = (X'P^{-1}P'^{-1}X)^{-1}X'P^{-1}P'^{-1}y = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$, which is the GLS estimation of the original model.

Substituting $y = X\beta + u$ into $\dot{\beta} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$, it follows that $\dot{\beta} - \beta = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}u$, so that the variance $Var(\dot{\beta}) = (X'\Omega^{-1}X)^{-1}\sigma^2$.

If the model is estimated by OLS, being $\hat{\beta} - \beta = (X'X)^{-1}X'u$, it follows that $Var(\hat{\beta}) = (X'X)^{-1}(X'\Omega X)(X'X)^{-1}\sigma^2$. This variance-covariance matrix is greater or equal to the GLS matrix, the proof being obtained computing the product $[(X'X)^{-1}X'P' - (X'\Omega^{-1}X)^{-1}X'P^{-1}]\ [PX(X'X)^{-1} - P'^{-1}X(X'\Omega^{-1}X)^{-1}] = (X'X)^{-1}(X'\Omega X)(X'X)^{-1} - (X'\Omega^{-1}X)^{-1}$ which is positive semi definite, being the product of a matrix with its transpose.

Remark. The result does not change if, in the final GLS formula, the whole expression of the variance-covariance matrix $\sigma^2\Omega$ is used instead of Ω (the scalar constant σ^2 would cancel out).

A list of other topics

Dummy variables.
Regression specification error test (Reset).
Heteroskedastic errors: OLS is unbiased but not efficient; heteroskedasticity of "known" form and weighted least squares; Breusch and Pagan test; heteroskedasticity of "unknown" form and "sandwich" estimator of the variance-covariance matrix; White's test.
Autocorrelated errors: OLS is unbiased but not efficient; first order autoregressive errors AR(1); Cochrane-Orcutt estimation method; Breusch and Godfrey test (LM test or Fisher's$-F$ test); optimal prediction when errors are AR(1).
Multicollinearity; perfect and near multicollinearity.

1 OBVIOUS AND BANAL MATTERS - Ma con voi non si sa mai.....

1.1 Products of matrices and vectors

$$\underset{(k\times n)}{X'} = \begin{bmatrix} x_{1,1} & x_{2,1} & ... & x_{i,1} & ... & x_{n,1} \\ x_{1,2} & x_{2,2} & ... & x_{i,2} & ... & x_{n,2} \\ ... & ... & ... & ... & ... & ... \\ x_{1,k} & x_{2,k} & ... & x_{i,k} & ... & x_{n,k} \end{bmatrix} \qquad \underset{(n\times k)}{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & ... & x_{1,k} \\ x_{2,1} & x_{2,2} & ... & x_{2,k} \\ ... & ... & ... & ... \\ x_{i,1} & x_{i,2} & ... & x_{i,k} \\ ... & ... & ... & ... \\ x_{n,1} & x_{n,2} & ... & x_{n,k} \end{bmatrix} \qquad \underset{(n\times 1)}{u} = \begin{bmatrix} u_1 \\ u_2 \\ ... \\ u_i \\ ... \\ u_n \end{bmatrix}$$

$$\underset{(k\times 1)}{x_i} = \begin{bmatrix} x_{i,1} \\ x_{i,2} \\ ... \\ x_{i,k} \end{bmatrix} = \text{ column } i \text{ of } X' \qquad \underset{(1\times k)}{x_i'} = [x_{i,1} \quad x_{i,2} \quad ... \quad x_{i,k}] = \text{ row } i \text{ of } X \qquad u_i = \text{is a scalar}$$

There are n matrices $x_i x_i'$ of dimensions $(k \times k)$.

$X'X = \sum_{i=1}^{n} x_i x_i'$ is the matrix $(k \times k)$ *sum* of these n matrices.

$\frac{1}{n} X'X = \frac{1}{n} \sum_{i=1}^{n} x_i x_i'$ is the matrix $(k \times k)$ *arithmetical average* of these n matrices.

There are n vectors $x_i u_i$ of dimensions $(k \times 1)$.

$X'u = \sum_{i=1}^{n} x_i u_i$ is the vector $(k \times 1)$ *sum* of these n vectors.

$\frac{1}{n} X'u = \frac{1}{n} \sum_{i=1}^{n} x_i u_i$ is the vector $(k \times 1)$ *arithmetical average* of these n vectors.

$$\underset{(k\times n)}{W'} = \begin{bmatrix} w_{1,1} & w_{2,1} & ... & w_{i,1} & ... & w_{n,1} \\ w_{1,2} & w_{2,2} & ... & w_{i,2} & ... & w_{n,2} \\ ... & ... & ... & ... & ... & ... \\ w_{1,k} & w_{2,k} & ... & w_{i,k} & ... & w_{n,k} \end{bmatrix} \qquad \underset{(n\times k)}{W} = \begin{bmatrix} w_{1,1} & w_{1,2} & ... & w_{1,k} \\ w_{2,1} & w_{2,2} & ... & w_{2,k} \\ ... & ... & ... & ... \\ w_{i,1} & w_{i,2} & ... & w_{i,k} \\ ... & ... & ... & ... \\ w_{n,1} & w_{n,2} & ... & w_{n,k} \end{bmatrix}$$

$$\underset{(k\times 1)}{w_i} = \begin{bmatrix} w_{i,1} \\ w_{i,2} \\ ... \\ w_{i,k} \end{bmatrix} = \text{ column } i \text{ of } W' \qquad \underset{(1\times k)}{w_i'} = [w_{i,1} \quad w_{i,2} \quad ... \quad w_{i,k}] = \text{ row } i \text{ of } W$$

There are n matrices $w_i x_i'$ of dimensions $(k \times k)$.

$W'X = \sum_{i=1}^{n} w_i x_i'$ is the matrix $(k \times k)$ *sum* of these n matrices.

$\frac{1}{n} W'X = \frac{1}{n} \sum_{i=1}^{n} w_i x_i'$ is the matrix $(k \times k)$ *arithmetical average* of these n matrices.

There are n vectors $w_i u_i$ of dimensions $(k \times 1)$.

$W'u = \sum_{i=1}^{n} w_i u_i$ is the vector $(k \times 1)$ *sum* of these n vectors.

$\frac{1}{n} W'u = \frac{1}{n} \sum_{i=1}^{n} w_i u_i$ is the vector $(k \times 1)$ *arithmetical average* of these n vectors.

1.2 Quadratic forms and rectangular forms

If u is a $(n \times 1)$ vector and A is a $(n \times n)$ matrix, the quadratic form has the following scalar expression:
$u'Au = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{i,j} u_i u_j$
If v is a $(k \times 1)$ vector and B is a $(n \times k)$ matrix, the rectangular form has the following scalar expression:
$u'Bv = \sum_{i=1}^{n} \sum_{j=1}^{k} b_{i,j} u_i v_j$

1.3 A special product of three matrices

If X is a $(n \times k)$ matrix as above, and Σ is a $(n \times n)$ diagonal matrix

$$\underset{(n\times n)}{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & ... & 0 & ... & 0 \\ 0 & \sigma_2^2 & ... & 0 & ... & 0 \\ ... & ... & ... & ... & ... & ... \\ 0 & 0 & ... & \sigma_i^2 & ... & 0 \\ ... & ... & ... & ... & ... & ... \\ 0 & 0 & ... & 0 & ... & \sigma_n^2 \end{bmatrix}$$

then $X'\Sigma X = \sum_{i=1}^{n} x_i x_i' \sigma_i^2$

1.4 Schwarz inequality

Between scalars, with scalar notation: $\left(\sum_{i=1}^{n} a_i b_i\right)^2 \leq \left(\sum_{i=1}^{n} a_i^2\right) \left(\sum_{i=1}^{n} b_i^2\right)$

Between scalars, with vector notation (if $a'b \neq 0$): $(a'b)^2 \leq (a'a)(b'b) \Rightarrow (a'b)^{-1} a'a (b'a)^{-1} \geq (b'b)^{-1}$

It is a particular case of the following inequality between positive semidefinite matrices (with dimensions of X and W as above, and provided $n \geq k$ and $W'X$ is non-singular): $(W'X)^{-1}W'W(X'W)^{-1} \geq (X'X)^{-1}$

Proof: $(W'X)^{-1}W'W(X'W)^{-1} - (X'X)^{-1} = \left[(W'X)^{-1}W' - (X'X)^{-1}X'\right]\left[(W'X)^{-1}W' - (X'X)^{-1}X'\right]'$, being the product of a matrix with its transpose, is a positive semidefinite matrix.

2 AN EXAMPLE OF SIMULTANEOUS EQUATIONS: KLEIN-I MODEL

The *structural form* of the model is

$$\left[\begin{array}{ll} C_t = \alpha_1 + \alpha_2\ P_t + \alpha_3\ P_{t-1} + \alpha_4\ W_t + u_{1,t} & \text{Consumption} \\ I_t = \alpha_5 + \alpha_6\ P_t + \alpha_7\ P_{t-1} + \alpha_8\ K_{t-1} + u_{2,t} & \text{Investment (net)} \\ W_t^p = \alpha_9 + \alpha_{10}\ X_t + \alpha_{11}\ X_{t-1} + \alpha_{12}\ A_t + u_{3,t} & \text{Private wages} \\ X_t = C_t + I_t + G_t & \text{Equilibrium demand} \\ P_t = X_t - T_t - W_t^p & \text{Profits} \\ K_t = K_{t-1} + I_t & \text{Capital stock} \\ W_t = W_t^p + W_t^g. & \text{Total wages} \end{array}\right. \tag{2.1}$$

The model is usually presented as a system of 6 equations with 6 endogenous variables, as it was originally proposed in Klein (1950). The last (seventh) equation and endogenous variable, introduced here in addition to the original 6, avoids the need of equality restrictions in the first equation (otherwise the coefficient α_4 would multiply the sum of two variables $W_t^p + W_t^g$). Also, the third and fourth equations are usually presented in a slightly different way; the representation adopted here (2.1), taken from Greene (2008, 15.2), is perfectly equivalent to the original, but can be treated more easily.

This is an excellent example of a small, linear and manageable macroeconomic dynamic model, widely used in the literature as a test ground for estimation methods. The original Klein's data set contains data of the U.S. economy from 1920 to 1941 (interwar years, including the *depression* years; all variables are *at constant prices*). Due to the lag-1 variables, the estimation period (or sample period) is 1921-1941.

Endogenous variables are the 7 variables appearing on the left hand side of each structural equation, labeled on the right.

The exogenous variables are 5:

1 = Constant

W_t^g = Government wages

T_t = Business taxes

A_t = Linear time trend, measured as annual deviations from 1931, positive or negative; it is used as a *proxy* for increased bargaining power of labour (or union strength) during the sample period

G_t = Government nonwage expenditure.

The model also includes 3 lagged endogenous variables:

X_{t-1}

P_{t-1}

K_{t-1}.

The model contains 3 behavioural stochastic equations (the first 3 equations) and 4 identities, the first of which is an equilibrium condition, while the last three equations are accounting (or definitional) identities.

No variable appears in this model with an order lag greater than one; also there are no lagged exogenous variables. The original data set and detailed numerical results are in Appendix (14).

3 SIMULTANEOUS EQUATIONS: STRUCTURAL FORM AND REDUCED FORM

There are problems for which it is necessary to distinguish between exogenous and lagged endogenous variables (for example, dynamic solution, *multi steps ahead* forecast, delay or cumulated multipliers). In these cases we use an explicit *dynamic* notation.

There are cases where such a distinction is unnecessary, for example when studying identification, estimation methods and solution of the model *one step ahead*, and the notation can be slightly simplified, becoming essentially a *static* notation.

3.1 Dynamic notation

Structural form and *reduced form* of a system of dynamic simultaneous equations can be represented as

$$\left[\begin{array}{l} \text{Structural form} \\ \\ By_t + Cz_t + Dy_{t-1} = u_t \\ u_t:\ i.i.d. \\ E(u_t) = 0 \\ Var(u_t) = \Sigma \end{array}\right. \qquad \left[\begin{array}{l} \text{Reduced form} \\ \\ y_t = \Pi_1 z_t + \Pi_0 y_{t-1} + v_t \\ v_t:\ i.i.d. \\ E(v_t) = 0 \\ Var(v_t) = \Psi \end{array}\right. \tag{3.2}$$

The $(G \times 1)$ vector of endogenous variables at time t is called y_t (7×1 in the example); with the same dimensions, y_{t-1} is the vector of lagged endogenous variables; z_t is the $(K \times 1)$ vector of exogenous variables at time t (5×1 in the example); u_t and v_t are the $(G \times 1)$ vectors of error terms at time t, Σ and Ψ their variance-covariance matrices (assumed constant $\forall t$). With reference to the model used as example it is

$$\underset{\substack{(G\times 1)\\(7\times 1)}}{y_t} = \begin{bmatrix} C_t \\ I_t \\ W_t^p \\ X_t \\ P_t \\ K_t \\ W_t \end{bmatrix} \quad \underset{(7\times 1)}{y_{t-1}} = \begin{bmatrix} C_{t-1} \\ I_{t-1} \\ W_{t-1}^p \\ X_{t-1} \\ P_{t-1} \\ K_{t-1} \\ W_{t-1} \end{bmatrix} \quad \underset{(5\times 1)}{z_t} = \begin{bmatrix} 1 \\ W_t^g \\ T_t \\ A_t \\ G_t \end{bmatrix} \quad \underset{\substack{(G\times 1)\\(7\times 1)}}{u_t} = \begin{bmatrix} u_{1,t} \\ u_{2,t} \\ u_{3,t} \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \underset{\substack{(G\times 1)\\(7\times 1)}}{v_t} = \begin{bmatrix} v_{1,t} \\ v_{2,t} \\ v_{3,t} \\ v_{4,t} \\ v_{5,t} \\ v_{6,t} \\ v_{7,t} \end{bmatrix} \tag{3.3}$$

B is the $(G \times G)$ matrix of structural form coefficients of the endogenous variables (7×7 in the example); C is the $(G \times K)$ matrix of structural form coefficients of the exogenous variables (7×5 in the example); D is the $(G \times G)$ matrix of structural form coefficients of the lagged endogenous variables (7×7 in the example). Although being of dimensions (7×7), the matrix C has 4 columns of zeroes, corresponding to the 4 endogenous variables that *do not appear* in the model with lag-1.

$$\underset{\substack{(G\times G)\\(7\times 7)}}{B} = \begin{bmatrix} 1 & 0 & 0 & 0 & -\alpha_2 & 0 & -\alpha_4 \\ 0 & 1 & 0 & 0 & -\alpha_6 & 0 & 0 \\ 0 & 0 & 1 & -\alpha_{10} & 0 & 0 & 0 \\ -1 & -1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\underset{(7\times 5)}{C} = \begin{bmatrix} -\alpha_1 & 0 & 0 & 0 & 0 \\ -\alpha_5 & 0 & 0 & 0 & 0 \\ -\alpha_9 & 0 & 0 & -\alpha_{12} & 0 \\ 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 \end{bmatrix} \qquad \underset{(7\times 7)}{D} = \begin{bmatrix} 0 & 0 & 0 & 0 & -\alpha_3 & 0 & 0 \\ 0 & 0 & 0 & 0 & -\alpha_7 & -\alpha_8 & 0 \\ 0 & 0 & 0 & -\alpha_{11} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{3.4}$$

Remark. Higher order lags as well as lagged exogenous could be easily accomodated for by *augmenting* the vector of endogenous variables and including appropriate definitional identities. For example, should we want to include the lagged exogenous T_{t-1} into some equations, the simplest technique would be to introduce an $8-th$ endogenous variable (called for instance H_t), complete the structural model with the $8-th$ equation $H_t = T_t$, and replace everywhere in the model T_{t-1} with the lagged endogenous H_{t-1}.

The vector of error terms in the structural form equations, u_t, has some elements identically zero (4 in the example), corresponding to the identities. So, the $(G \times G)$ matrix of variances and covariances of the structural form errors Σ (7×7 in the example) has a 3×3 nonzero block (Σ_3, assumed symmetric and positive definite), while all the other elements are zero.

$$\underset{\substack{(G\times G)\\(7\times 7)}}{\Sigma} = \begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} & \sigma_{1,3} & 0 & 0 & 0 & 0 \\ \sigma_{2,1} & \sigma_{2,2} & \sigma_{2,3} & 0 & 0 & 0 & 0 \\ \sigma_{3,1} & \sigma_{3,2} & \sigma_{3,3} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} \Sigma_3 & 0 \\ 0 & 0 \end{bmatrix} \tag{3.5}$$

Π_1 is the $(G \times K)$ matrix of reduced form coefficients of exogenous variables (7×5 in the example). Π_0 is the $(G \times G)$ matrix of reduced form coefficients of lagged endogenous variables (7×7 in the example). Ψ is the $(G \times G)$ variance-covariance matrix of the reduced form error terms v_t (7×7 in the example; of course, its rank cannot be greater than 3).

$$\underset{(7\times 5)}{\Pi_1 = -B^{-1}C} = \begin{bmatrix} \pi_{1_{1,1}} & \pi_{1_{1,2}} & \pi_{1_{1,3}} & \pi_{1_{1,4}} & \pi_{1_{1,5}} \\ \pi_{1_{2,1}} & \pi_{1_{2,2}} & \pi_{1_{2,3}} & \pi_{1_{2,4}} & \pi_{1_{2,5}} \\ \pi_{1_{3,1}} & \pi_{1_{3,2}} & \pi_{1_{3,3}} & \pi_{1_{3,4}} & \pi_{1_{3,5}} \\ \pi_{1_{4,1}} & \pi_{1_{4,2}} & \pi_{1_{4,3}} & \pi_{1_{4,4}} & \pi_{1_{4,5}} \\ \pi_{1_{5,1}} & \pi_{1_{5,2}} & \pi_{1_{5,3}} & \pi_{1_{5,4}} & \pi_{1_{5,5}} \\ \pi_{1_{6,1}} & \pi_{1_{6,2}} & \pi_{1_{6,3}} & \pi_{1_{6,4}} & \pi_{1_{6,5}} \\ \pi_{1_{7,1}} & \pi_{1_{7,2}} & \pi_{1_{7,3}} & \pi_{1_{7,4}} & \pi_{1_{7,5}} \end{bmatrix} \qquad \underset{(7\times 7)}{\Pi_0 = -B^{-1}D} = \begin{bmatrix} 0 & 0 & 0 & \pi_{0_{1,4}} & \pi_{0_{1,5}} & \pi_{0_{1,6}} & 0 \\ 0 & 0 & 0 & \pi_{0_{2,4}} & \pi_{0_{2,5}} & \pi_{0_{2,6}} & 0 \\ 0 & 0 & 0 & \pi_{0_{3,4}} & \pi_{0_{3,5}} & \pi_{0_{3,6}} & 0 \\ 0 & 0 & 0 & \pi_{0_{4,4}} & \pi_{0_{4,5}} & \pi_{0_{4,6}} & 0 \\ 0 & 0 & 0 & \pi_{0_{5,4}} & \pi_{0_{5,5}} & \pi_{0_{5,6}} & 0 \\ 0 & 0 & 0 & \pi_{0_{6,4}} & \pi_{0_{6,5}} & \pi_{0_{6,6}} & 0 \\ 0 & 0 & 0 & \pi_{0_{7,4}} & \pi_{0_{7,5}} & \pi_{0_{7,6}} & 0 \end{bmatrix} \tag{3.6}$$

$$\underset{\substack{(G\times G)\\(7\times 7)}}{\Psi = B^{-1}\Sigma B'^{-1}} \tag{3.7}$$

Matrices B, C and D are usually *sparse* matrices. Zeroes and ones represent *a-priori* restrictions on the structural form. In particular, considering only the behavioural stochastic equations (the first three equations in the example) zeroes represent *exclusion* restrictions, ones represent *normalization* restrictions. For instance, in the first equation the coefficient of the endogenous variable *Consumption* (C_t) is 1, and not a generic $b_{1,1}$ (normalization); the coefficient of I_t, is 0, and not a generic $b_{1,2}$ (exclusion); the coefficient of T_t, is 0, and not a generic $c_{1,3}$ (exclusion); etc.
Matrix Π_1 is usually a *full* matrix; some of its elements are zeroes only exceptionally.
Matrix Π_0 has 4 columns of zeroes (like matrix D); the other 3 columns have usually no zeroes.

3.2 Static notation

When it is unnecessary to distinguish between exogenous and lagged endogenous variables, a simplified notation can be adopted. We still use the same vectors y_t, u_t, v_t and the matrix B exactly as in the dynamic notation. but z_t becomes a (8×1) vector containing the 5 exogenous variables at time t and the 3 lagged endogenous variables that *really* appear in the model. The matrix of structural form coefficients of the exogenous and lagged endogenous variables has therefore dimensions (7×8) and will be called Γ; in the example, the first 5 columns of Γ will be the columns of the matrix C adopted with the dynamic notation,, while the last 3 columns will be the *nonzero* columns of D. When using the static notation, K indicates the total number of exogenous and lagged endogenous variables (8 in the example).

$$\begin{array}{l} z_t = \\ (K \times 1) \\ (8 \times 1) \end{array} \begin{bmatrix} 1 \\ W_t^g \\ T_t \\ A_t \\ G_t \\ X_{t-1} \\ P_{t-1} \\ K_{t-1} \end{bmatrix} \qquad \begin{bmatrix} By_t + \Gamma z_t = u_t & \text{Structural form} \\ \\ y_t = \Pi z_t + v_t & \text{Reduced form} \end{bmatrix} \tag{3.8}$$

$$\begin{array}{l} B = \\ (G \times G) \\ (7 \times 7) \end{array} \begin{bmatrix} 1 & 0 & 0 & 0 & -\alpha_2 & 0 & -\alpha_4 \\ 0 & 1 & 0 & 0 & -\alpha_6 & 0 & 0 \\ 0 & 0 & 1 & -\alpha_{10} & 0 & 0 & 0 \\ -1 & -1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 1 \end{bmatrix} \quad \begin{array}{l} \Gamma = \\ (G \times K) \\ (7 \times 8) \end{array} \begin{bmatrix} -\alpha_1 & 0 & 0 & 0 & 0 & 0 & -\alpha_3 & 0 \\ -\alpha_5 & 0 & 0 & 0 & 0 & 0 & -\alpha_7 & -\alpha_8 \\ -\alpha_9 & 0 & 0 & -\alpha_{12} & 0 & -\alpha_{11} & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{3.9}$$

$$\begin{array}{l} \Pi = -B^{-1}\Gamma = \\ (G \times K) \\ (7 \times 8) \end{array} \begin{bmatrix} \pi_{1,1} & \pi_{1,2} & \pi_{1,3} & \pi_{1,4} & \pi_{1,5} & \pi_{1,6} & \pi_{1,7} & \pi_{1,8} \\ \pi_{2,1} & \pi_{2,2} & \pi_{2,3} & \pi_{2,4} & \pi_{2,5} & \pi_{2,6} & \pi_{2,7} & \pi_{2,8} \\ \pi_{3,1} & \pi_{3,2} & \pi_{3,3} & \pi_{3,4} & \pi_{3,5} & \pi_{3,6} & \pi_{3,7} & \pi_{3,8} \\ \pi_{4,1} & \pi_{4,2} & \pi_{4,3} & \pi_{4,4} & \pi_{4,5} & \pi_{4,6} & \pi_{4,7} & \pi_{4,8} \\ \pi_{5,1} & \pi_{5,2} & \pi_{5,3} & \pi_{5,4} & \pi_{5,5} & \pi_{5,6} & \pi_{5,7} & \pi_{5,8} \\ \pi_{6,1} & \pi_{6,2} & \pi_{6,3} & \pi_{6,4} & \pi_{6,5} & \pi_{6,6} & \pi_{6,7} & \pi_{6,8} \\ \pi_{7,1} & \pi_{7,2} & \pi_{7,3} & \pi_{7,4} & \pi_{7,5} & \pi_{7,6} & \pi_{7,7} & \pi_{7,8} \end{bmatrix} \tag{3.10}$$

4 MULTIPLIERS, FORECASTS, GOODNESS OF FIT MEASURES

Recursive substitution into the reduced form system (3.2) gives

$$\begin{aligned} y_t &= \Pi_1 z_t + \Pi_0 y_{t-1} + v_t \\ &= \Pi_1 z_t + \Pi_0 \Pi_1 z_{t-1} + \Pi_0^2 y_{t-2} + v_t + \Pi_0 v_{t-1} \\ &= \Pi_1 z_t + \Pi_0 \Pi_1 z_{t-1} + \Pi_0^2 \Pi_1 z_{t-2} + \Pi_0^3 y_{t-3} + v_t + \Pi_0 v_{t-1} + \Pi_0^2 v_{t-2} \\ &= etc. \end{aligned} \tag{4.11}$$

The long run dynamic behaviour of the system depends on the powers of Π_0. In the example, Π_0 is a (7×7) nonsymmetric matrix, with 4 columns of zeroes; it has therefore 3 nontrivial eigenvalues (one of which is necessarily real, the other two can be real or conjugate complex) that ensure stability if all are less than one in modulus.

$$\begin{array}{lll} \Pi_1 & = \partial E[y_t | z_t, z_{t-1}, ...]/\partial z_t & \text{matrix of } \textit{impact multipliers}; \\ \Pi_0\Pi_1 & = \partial E[y_t | z_t, z_{t-1}, ...]/\partial z_{t-1} & \text{matrix of } \textit{lag-1 delay multipliers}; \\ \Pi_0^2\Pi_1 & = \partial E[y_t | z_t, z_{t-1}, ...]/\partial z_{t-2} & \text{matrix of } \textit{lag-2 delay multipliers}; \\ & & \text{etc.} \end{array} \tag{4.12}$$

In the example, all these matrices have dimensions (7×5). Their sum, up to a given lag, is the matrix of *cumulated* or *sustained multipliers*.

Multipliers are fundamental tools for economic policy simulations.
The reduced form equations (3.8 or 3.2), coefficients (Π, Π_0 or Π_1) and variance-covariance matrix (Ψ) are called *restricted* if they derive (i.e. are computed) from the structural form (inverting matrix B, etc.). Otherwise they are called *unrestricted* (when we consider directly each reduced form equation as a the linear regression of an endogenous variable against all the exogenous and lagged endogenous variables).
All the expressions *restricted reduced form*, *reduced form derived from the structural form*, *simultaneous solution of the structural form equations* have exactly the same meaning. When talking of *static* or *one step ahead* solution of the structural form equations, reference is always done to (3.8). When talking of *dynamic* solution of the structural form equations, reference is always done to (3.2).
Forecasts, simulations and economic policy experiments are usually conducted using the restricted reduced form, deriving it from the structural form after a convenient estimate of the unknown coefficients has been computed (α_1, ..., α_{12} in the example).
There is much more *economic theory* in the structural form than in the unrestricted reduced form.
Forecasts *one step ahead* are usually produced using the static notation (3.8), setting the random error terms v_t to zero (expected value). When data are available till time n (last sample observation) and forecast is performed for time $n+1$, some elements of the vector z_{n+1} are available from the sample (the lagged endogenous variables, since they are related to time n). But the other elements of z_{n+1} (the exogenous variables) must be supplied from outside (usually, financial plans of the government, forecasts produced by central banks, etc.).
To produce forecasts *multi steps ahead* it is necessary to resort to the dynamic notation (3.2), still setting the random error terms v_t to zero at any time. For example, if data are available till time n (last sample observation) and forecast is performed for time $n+1$ and $n+2$, we first forecast at $n+1$ as above. Then, to forecast at $n+2$, the values of the lagged endogenous variables in z_{n+2} are taken from the forecast at $n+1$, while the exogenous variables at $n+2$ must be supplied from outside. Etc.
In sample forecasts (historical tracking) and *goodness of fit* measures over the sample period can be both static or dynamic. In the static case, the structural form is always solved one step ahead, taking values of the lagged endogenous variables from the observed sample. In the dynamic case, the dynamic notation is used (3.2) and the values of the lagged endogenous variables, in each period, are taken from the solution of the previous period. Initial values of the endogenous variables are always taken from the sample, so there is no difference between static and dynamic solution at the beginning of the sample period.
The following are the most common univariate measures of goodness of fit. Each formula is related to *one* endogenous variable; O_t is the *observed* value of the variable at time t; C_t is the value of the variable at time t *computed* with the model (static or dynamic solution); o_t is the *observed growth rate* of the variable at time t (the annual percentage change, in the example); c_t is the growth rate of the variable at time t *computed* with the model (static or dynamic solution).

$$
\begin{array}{lll}
RMSE & = \sqrt{\frac{\sum_{t=1}^{n}(O_t-C_t)^2}{n}} & \text{Root Mean Squared Error} \\
RMSE(dim) & = \sqrt{\frac{\sum_{t=1}^{n}(O_t-C_t)^2}{\sum_{t=1}^{n}O_t^2}} & \text{Dimensionless Root Mean Squared Error} \\
MAPE & = \frac{1}{n}\sum_{t=1}^{n}\frac{|O_t-C_t|}{O_t}\times 100 & \text{Mean Absolute Percentage Error} \\
Theil's\ U_1 & = \sqrt{\frac{\sum_{t=1}^{n}(o_t-c_t)^2}{\sum_{t=1}^{n}o_t^2}} & \text{Theil Inequality Coefficient (1966, eq.4.5)} \\
Theil's\ U_2 & = \sqrt{\frac{\sum_{t=1}^{n}(o_t-c_t)^2}{\sum_{t=1}^{n}(o_t-\bar{o})^2}} & \text{Theil Inequality Coefficient (1966, eq.4.6)}
\end{array}
\tag{4.13}
$$

$MAPE$ and Theil's inequality coefficients are not computed for variables that change sign over the sample period (such as I_t in the example).

5 IDENTIFICATION BY MEANS OF A-PRIORI RESTRICTIONS

We consider in detail the case of *a-priori* restrictions on coefficients, in particular *exclusion* restrictions. Covariance restrictions are briefly discussed at the end.

Definition. Two *different* structural form systems

$$
\left[\begin{array}{l} By_t+\Gamma z_t=u_t \\ Var(u_t)=\Sigma \end{array}\right] \qquad and \qquad \left[\begin{array}{l} B^*y_t+\Gamma^* z_t=u_t^* \\ Var(u_t^*)=\Sigma^* \end{array}\right] \tag{5.14}
$$

are called *observationally equivalent* if they have the *same reduced form*.
More precisely, let the corresponding reduced form systems be

$$\left[\begin{array}{l} y_t = \Pi z_t + v_t \\ v_t = B^{-1}u_t \\ \Pi = -B^{-1}\Gamma; \\ \Psi = Var(v_t) = B^{-1}\Sigma B'^{-1} \end{array}\right. \quad and \quad \left[\begin{array}{l} y_t = \Pi^* z_t + v_t^* \\ v_t^* = B^{*-1}u_t^* \\ \Pi^* = -B^{*-1}\Gamma^*; \\ \Psi^* = Var(v_t^*) = B^{*-1}\Sigma^* B^{*\prime-1} \end{array}\right. \tag{5.15}$$

then, the two structural form systems (5.14) are called *observationally equivalent till the second moments* if $\Pi^* = \Pi$ and $\Psi^* = \Psi$. If this happens, it will be impossible to discriminate between the two different structural forms on the basis of the observed data, since data (the values of y_t) are produced by the reduced form.

Definition. A parameter (or an equation) of the structural form is *identified* if it (or the equation's parameters) can be deduced from knowledge of the reduced form parameters Π and Ψ.

Theorem. Two structural forms (5.14) are observationally equivalent if and only if there exists a non-singular square matrix F (same dimensions as B) such that $B^* = FB$, $\Gamma^* = F\Gamma$, and $\Sigma^* = F\Sigma F'$.

The proof is straightforward. In fact, if $B^* = FB$, $\Gamma^* = F\Gamma$, and $\Sigma^* = F\Sigma F'$, then $\Pi^* = -B^{*-1}\Gamma^* = -(FB)^{-1}(F\Gamma)$ $= -B^{-1}F^{-1}F\Gamma = -B^{-1}\Gamma = \Pi$; moreover $\Psi^* = B^{*-1}\Sigma^* B^{*\prime-1} = (FB)^{-1}F\Sigma F'(FB)'^{-1} = B^{-1}F^{-1}F\Sigma F'F'^{-1}B'^{-1} = B^{-1}\Sigma B'^{-1} = \Psi$.
Viceversa, if $\Pi^* = \Pi$, then $B^{*-1}\Gamma^* = B^{-1}\Gamma$ and pre-multiplication of both sides by B^* gives $\Gamma^* = B^*B^{-1}\Gamma$; thus, $B^* = FB$ and $\Gamma^* = F\Gamma$, having defined $F = B^*B^{-1}$. Also, $\Psi^* = \Psi$ implies that $B^{*-1}\Sigma^* B^{*\prime-1} = B^{-1}\Sigma B'^{-1}$, where pre-multiplication of both sides by B^* and post-multiplication by $B^{*\prime}$ gives $\Sigma^* = B^*B^{-1}\Sigma B'^{-1}B^{*\prime} = F\Sigma F'$, having defined F as above.

The theorem implies that, given a structural form $By_t + \Gamma z_t = u_t$, there will be an infinity of other structural forms, different from it, but observationally equivalent to it: any non-singular square matrix F, arbitrarily chosen, will in fact produce the matrices B^*, Γ^* and Σ^* of an observationally equivalent structural form. Notice that each row of B^* and Γ^* would be a linear combination ot the rows of B and Γ.

5.1 Restrictions and admissible transformations

It may happen that pre-multiplication by F produces matrices B^* and Γ^* that do not satisfy the *a-priori* restrictions of B, Γ and Σ. For instance, it may happen that $\beta^*_{1,2}$ is nonzero, while $\beta_{1,2} = 0$ in the original model; this means that the variable I_t was *excluded*, by the economic theory of the model's builder, from the first equation of $By_t + \Gamma z_t = u_t$, but is *included* in the first equation of $B^*y_t + \Gamma^* z_t = u_t^*$. Exclusion restrictions are a particular case of *homogeneous* restrictions. They are the only type of homogeneous restrictions in the model used as example.

Definition. A linear transformation produced by a non-singular matrix F is *admissible* if the transformed structural form coefficients $[B^*; \Gamma^*]$ satisfy all the *a-priori* restrictions on $[B\ ;\ \Gamma]$ and the transformed variance-covariance matrix Σ^* satisfies all the *a-priori* restrictions on Σ.

To simplify the problem, we only consider restrictions on coefficients, without considering possible restrictions on the variance-covariance matrix (which are rather unusual); for further simplification, we first consider only homogeneous restrictions (and later normalization restrictions). To fix ideas, we focus on the first equation of the structural form model (consumption, in the example).
The $(G\times G)$ *unit* (or *identity*) matrix obviously produces an admissible transformation. Any $(G\times G)$ *scalar* matrix (the identity matrix multiplied by a nonzero scalar) also produces an admissible transformation (exclusion restrictions are preserved, as well as homogeneous restrictions in general).
If an admissible matrix F exists, and it is different from a scalar matrix, this implies that an alternative structural form $B^*y_t + \Gamma^* z_t = u_t^*$ exists, which is observationally equivalent to the original model $By_t + \Gamma z_t = u_t$, satisfies all the *a-priori* restrictions on the original model, but has one or more coefficients different from the original model: thus, some coefficients (or equations) are not identified. If this happens for some coefficients of the first equation, then the first equation is not identified. If this *cannot* happen in the first equation (even if it may happen in other equations), then the first equation is identified.
Notice that, when pre-multiplying by F the matrices B and Γ to produce B^* and Γ^*, the first row (structural coefficients of the consumption equation, in the example) is

$$[B^*\ ;\ \Gamma^*]_{1,\bullet} = [F(B\ ;\ \Gamma)]_{1,\bullet} = F_{1,\bullet}[B\ ;\ \Gamma] \tag{5.16}$$

If such a $F_{1,\bullet}$ produces a first row of structural coefficients satisfying all the restrictions on the first equation of the original model, the first equation is not identified. If only the first row of the unit matrix or the first row of a scalar matrix can do it, and no other $F_{1,\bullet}$, then the first equation is identified.
A representation of the exclusion restrictions in the first equation can be obtained introducing the matrix Φ_1. In the example such a matrix is

$$\underset{\substack{[(G+K)\times R_1]\\(15\times 10)}}{\Phi_1 =} \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{5.17}$$

The number of columns, R_1, is the number of exclusion restrictions from the first structural form equation of the model. The whole system includes $G + K = 15$ variables, and the first structural form equation includes 5 variables (a dependent variable, and 4 variables on the right hand side); therefore the number of exclusions is $R_1 = 15 - 5 = 10$. Each column corresponds to an excluded variable, and contains a unique nonzero element (= 1), indicating which variable is excluded.
For example, considering the list of all the variables of the model (beginning with the endogenous variables, followed by the exogenous and lagged endogenous variables) the first variable of the list is C_t, which is included in the first equation, and therefore corresponds to no column. The following variable in the list is I_t and it is excluded from the equation. Since it is the first variable excluded from the equation, and it is the second variable in the list of all variables, then in the first column there is a value = 1 in the second row. The following variable in the list is W_t^p, and it is also excluded from the equation. Since it is the second variable excluded from the equation, and it is the third in the list of all variables, then in the second column there is a value = 1 in the third row. And so on.
Matrix Φ_1 is such that the coefficients of the first structural form equation satisfy the system of R_1 (= 10 in the example) linear homogeneous equations

$$[B\ ;\ \Gamma]_{1,\bullet}\ \Phi_1 = 0 \tag{5.18}$$

5.2 Rank condition - Order condition

An admissible transformation matrix F has its first row ($F_{1,\bullet}$) that must produce transformed coefficients satisfying the analogous system of equations

$$[B^*;\Gamma^*]_{1,\bullet}\ \Phi_1 = 0 \tag{5.19}$$

that is, substituting (5.16)

$$F_{1,\bullet}\ [B\ ;\ \Gamma]\ \Phi_1 = 0 \tag{5.20}$$

This can be viewed as a system of R_1 (= 10) homogeneous equations with G (= 7) unknowns (the elements of $F_{1,\bullet}$). The matrix of coefficients is $[B\ ;\ \Gamma]\Phi_1$, and its dimensions are $(G \times R_1)$ (7×10 in the example). If the rank of this matrix of coefficients is G (= 7), then the system (5.20) has the unique solution $F_{1,\bullet} = 0$. This is obviously impossible, because also $F_{1,\bullet} = I_{1,\bullet}$ (the first row of the unit matrix) is for sure a solution of the system.
If the rank of this matrix of coefficients is $G - 1$ (= 6 in the example), then the system (5.20) would have ∞^1 solutions for $F_{1,\bullet}$, and all these solutions would be proportional to $I_{1,\bullet}$, the first row of the unit matrix (in other words, the first row of an arbitrary scalar matrix would be a solution, and there would be no other solutions). These ∞^1 solutions are reduced to a single solution after imposing the *normalization* restriction $\beta_{1,1} = 1$ (the coefficient of C_t in the first equation of the example is not a generic $\beta_{1,1}$, but is fixed to 1). Thus, the first equation is identified.
If the rank of this matrix of coefficients is $G - 2$ or less (5 or less than 5 in the example), then the system (5.20) would have ∞^2 or more solutions for $F_{1,\bullet}$, and not all would be proportional to $I_{1,\bullet}$, the first row of the unit matrix. These solutions would not be reduced to a single solution after imposing the *normalization* restriction $\beta_{1,1} = 1$. Thus, there would be other observationally equivalent structural forms whose first equation satisfies all the *a-priori* restrictions on the original first equation, but coefficients would be different. Thus the first equation would not be identified.
The above discussion can be summarized in the following condition, which is necessary and sufficient.

Theorem (*rank condition*). The $i - th$ structural form equation is identified if and only if $rank[(B\ ;\ \Gamma)\Phi_i] = G - 1$.

Considering that B is non-singular, so that necessarily $rank[B\ ;\ \Gamma] = G$ (= 7 in the example), and that the rank of a product of matrices is smaller than or equal to the smallest of the ranks, it is necessary that $rank[\Phi_1] \geq G - 1$, otherwise the rank condition cannot hold. This is a *necessary but not sufficient* condition, known as the *order condition*.
For the case of exclusion restrictions, the order condition can be stated in very simple terms. We first consider that the rank of Φ_1 equals the number of columns of Φ_1, or number of exclusion restrictions R_1 (= 10 in the example). Thus the order

condition becomes $R_1 \geq G-1$. We then consider that the whole structural form system includes $G+K$ variables, but R_1 of these variables are excluded from the first equation, so the number of included variables in the first equation is $G-R_1+K$. Considering now that one of these variables is on the left hand side of the structural equation (the dependent variable, C_t in the example), there are $G-1-R_1+K$ variables on the right hand side of the first equation (the regressors, or explanatory variables in the structural equation of consumption). When $R_1 \geq G-1$ (as stated above), then the number of regressors on the right hand side of the strucural equation $G-1-R_1+K$ will be $\leq K$. Summarizing

Theorem (*order condition*). A necessary (but not sufficient) condition for a structural form equation to be identified is that the number of regressors on the right hand side of the equation must not exceed K, the total number of exogenous and lagged endogenous variables of the system.

Stated as above, the order condition is quite intuitive. We cannot have in a single structural equation more regressors than *independent inputs* in the whole system, which are the exogenous and lagged endogenous variables of the system.
The same condition (order) is presented by some textbooks in the following (equivalent) way: the number of exogenous (and lagged endogenous) variables excluded from a structural equation must be at least as large as the number of endogenous variables included, less one.

Definitions. The first structural form equation is called *just-identified* (or *exactly identified*) if it is identified (i.e. the rank condition is satisfied), and $rank[\Phi_1] = G-1$.
The first structural form equation is called *over-identified* if it is identified (i.e. the rank condition is satisfied), and $rank[\Phi_1] > G-1$.
If $rank[\Phi_1] < G-1$, then the rank condition cannot be satisfied and the equation is not identified; it can also be called *under-identified*.

5.3 Remarks

Matrix Φ_1 would be more complex if restrictions other than exclusion were introduced. For instance, the first structural equation of the Klein-I model (the private consumption equation) is more usually presented as

$$C_t = \alpha_1 + \alpha_2\ P_t + \alpha_3\ P_{t-1} + \alpha_4\ (W_t^p + W_t^g) + u_{1,t}$$

This would avoid the need of the seventh equation, and the endogenous variables would be 6 rather than 7. But at the same time it would make more complex the structure of the matrices. In fact, the *same* coefficient α_4 would multiply the sum of two variables. This would be an additional restriction, and should be properly considered either in the matrices B and Γ (α_4 should appear in both matrices) or in matrix Φ_1 (where a column should contain a $+1$ and a -1 in the proper rows).
Other types of restrictions can be found in econometric models. For example, the *constant return to scale* hypothesis in a Cobb-Douglas production function implies that two (or more) structural coefficients sum to one. This is a *nonhomogeneous* restriction on coefficients and would require some changes in the rank condition (5.20 would be replaced by a nonhomogenous equation system and the normalization restriction should be directly introduced into the system, rather than at the end as above).

5.4 Demand - supply model

If we consider the following model in structural form

$$\left[\begin{array}{ll} Q_i^d = \alpha_1 + \alpha_2\ P_i + u_{1,i} & \text{Demand} \\ Q_i^s = \alpha_3 + \alpha_4\ P_i + u_{2,i} & \text{Supply} \\ Q_i^d = Q_i^s & \text{Equilibrium} \end{array}\right. \tag{5.21}$$

neither demand nor supply equations are identified, since they both fail the order condition (necessary). The model, in fact, has 3 endogenous variables (Q_i^d = demand, Q_i^s = supply, P_i = equilibrium price) and only one exogenous variable (the constant); so $K=1$, while each of the first two equations would have two explanatory variables (regressors).
Intuitively the lack of identification has a quite simple explanation. We expect that demand is a decreasing function of price, thus $\alpha_2 < 0$, while supply is expected to be an increasing function of price, thus $\alpha_4 > 0$. In the two-dimensional plane (Q, P) the two straight lines would cross in a unique point (the equilibrium value of Q and P). This point would remain fix for any i (the simultaneous solution of the equation system). If the model is correctly specified, observations at various times would be points scattered around this unique solution point (maybe very close to it, if the random error terms $u_{1,i}$ and $u_{2,i}$ are small). Such a scatter diagram makes it impossible to distinguish between a demand (decreasing) function and a supply (increasing) function of price.
We could introduce an additional explanatory variable (exogenous) into the supply equation, for instance L_i = cost of labour and/or raw materials: $Q_i^s = \alpha_3 + \alpha_4\ P_i + \alpha_5\ L_i + u_{2,i}$. The order condition would now be satisfied by the *first* structural equation (demand); there would be in fact $K=2$ exogenous variables in the system, the constant and L_i, and there would be two regressors in the equation. Some simple algebra could show that also the rank condition is satisfied by the first equation if $\alpha_5 \neq 0$. The second equation (supply) would be still under-identified.
There is again an intuitive explanation of all this. We could still represent the demand and supply functions in the two dimensional plane. Values of the additional exogenous variable L_i, changing with i, would *shift* the supply line in the plane. There would be, therefore, several intersection points between demand and supply (equilibrium values of Q and P for

different values of L_i). All these points would be *on the demand* line (that does not shift). If the model is correctly specified, observations at various i-s would be points scattered around the solution points, therefore they would be scattered along the demand line, making it *visible*. There would be, however, no chance to identify the supply function.

5.5 Some remarks on variance-covariance restrictions

Restrictions on variances and covariances help identification (but they are quite unusual). For example, if the first equation is not identified by means of exclusion restrictions, it could be identified by imposing restrictions on the first row (and first column) of the Σ matrix. In particular, the variance could be known, or some covariances could be zero.
We still consider linear restrictions, but not necessarily homogeneous, so that the right hand side of equations (5.18) and (5.20) will be a constant vector r_1, not necessarily equal to zero. If we impose S_1 restrictions on the first row of Σ (Θ_1 is the matrix of restrictions)

$$\Sigma_{1,\bullet}\ \Theta_1 = s_1 \tag{5.22}$$

an admissible transformation matrix F must have the first row satisfying the system of $R_1 + S_1$ equations

$$\left[\begin{array}{l} F_{1,\bullet}\ [B\ ;\ \Gamma]\ \Phi_1 = r_1 \\ F_{1,\bullet}\ \Sigma\ F'\ \Theta_1 = s_1 \end{array}\right. \tag{5.23}$$

The system is *nonlinear* in the unknown $F_{1,\bullet}$. We can nevertheless discuss the problem considering what could be its solution if F was treated as *fixed*, and the unknown $F_{1,\bullet}$ was considered as *not belonging* to F. In such a way we derive a condition which is necessary, but not sufficient.
The first equation of the model is identified if, and only if, for any admissible F, the unique solution of the system (5.23) is $F_{1,\bullet} = I_{1,\bullet}$ (unit row vector). The matrix $F = I$ is surely admissible, so we fix $F = I$; then, the solution $F_{1,\bullet} = I_{1,\bullet}$ must be unique, and this holds when $rank[(B\ ;\ \Gamma)\ \Phi_1\ ;\ \ \Sigma\ \Theta_1] = G$. This is called *generalized rank condition*. It is a necessary condition. It ensures that there cannot be a solution different from $F_{1,\bullet} = I_{1,\bullet}$, but having treated $F_{1,\bullet}$ and F separately, the condition does not ensure the existence of a solution; thus, the condition is not sufficient.
To ensure a $rank = G$, the matrix, which has G rows, must have at least G columns. The number of columns is $R_1 + S_1$. Therefore it must be $R_1 + S_1 \geq G$, which is the *generalized order condition* (also necessary, but not sufficient). When R_1 is too small, and thus identification is not ensured by means of restrictions on coefficients, $R_1 + S_1$ could be large enough, and the first equation of the model might be identified.

6 ASYMPTOTIC PROPERTIES OF ORDINARY LEAST SQUARES (OLS)

Let's consider the linear regression model

$$y = X\beta + u \tag{6.24}$$

where

$$\underset{(n\times 1)}{y = \begin{bmatrix} y_1 \\ y_2 \\ ... \\ y_t \\ ... \\ y_n \end{bmatrix}} \quad \underset{(n\times k)}{X = \begin{bmatrix} x_{1,1} & x_{1,2} & ... & x_{1,k} \\ x_{2,1} & x_{2,2} & ... & x_{2,k} \\ ... & ... & ... & ... \\ x_{t,1} & x_{t,2} & ... & x_{t,k} \\ ... & ... & ... & ... \\ x_{n,1} & x_{n,2} & ... & x_{n,k} \end{bmatrix}} \quad \underset{(k\times 1)}{\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ ... \\ \beta_k \end{bmatrix}} \quad \underset{(n\times 1)}{u = \begin{bmatrix} u_1 \\ u_2 \\ ... \\ u_t \\ ... \\ u_n \end{bmatrix}} \quad \left[\begin{array}{l} u_t :\ i.i.d. \\ \\ E[u] = 0 \\ \\ Var[u] = E[uu'] = \sigma^2 I_n \end{array}\right. \tag{6.25}$$

OLS estimator of coefficients is

$$\hat{\beta} = (X'X)^{-1}X'y \tag{6.26}$$

Substituting $y = X\beta + u$ into the above expression, we get the *estimation error*

$$\hat{\beta} - \beta = (X'X)^{-1}X'u = \left(\frac{X'X}{n}\right)^{-1}\frac{X'u}{n} \tag{6.27}$$

and the *estimation error rescaled by* $\sqrt{n}$

$$\sqrt{n}(\hat{\beta} - \beta) = \left(\frac{X'X}{n}\right)^{-1}\frac{X'u}{\sqrt{n}} \tag{6.28}$$

We assume that the $k \times k$ matrix $X'X/n$ is non-singular for any n (classical hypothesis), and converges to a non-singular and finite limit as $n \to \infty$. If no random variables are contained in the matrix X, then *convergence* for $X'X/n$ is in *mathematical* sense (*lim*), otherwise we are dealing with convergence in *probability* (*plim*) to a constant matrix. *Trends* are therefore excluded (otherwise the limit would not be finite).
We *roughly* consider four different cases.

6.1 First case

The matrix of explanatory variables, X, does not contain random variables. In this case, from (6.27) we get

$$E[\hat{\beta} - \beta] = (X'X)^{-1}E[X'u] = (X'X)^{-1}X'E(u) = 0$$

and

$$plim[\hat{\beta} - \beta] = lim\left(\frac{X'X}{n}\right)^{-1} plim\frac{X'u}{n} = lim\left(\frac{X'X}{n}\right)^{-1} plim\frac{1}{n}\sum_{t=1}^{n} x_t u_t = 0$$

The last equality follows directly from the weak law of large numbers (WLLN) observing that the *plim* is the probability limit of the average of n vectors $(k \times 1)$, each of which has zero expected value: $E(x_t u_t) = x_t E(u_t) = 0$.
So in this case the OLS estimator is *unbiased* and *consistent*.

6.2 Second case

The matrix of explanatory variables, X, contains some random variables, but these random variables are *independent* from the error terms u (*strictly exogenous*).
Also in this case the OLS estimator is *unbiased* and *consistent*. The only difference, with respect to the previous case, is that the limit of $(X'X/n)$ must be a probability limit, rather than a limit in mathematical sense.

6.3 Third case

Contemporaneous explanatory variables and error terms are independent, but some explanatory variables at time t (elements of the vector x_t) may be not independent of u_s for some $s \neq t$. This is, for example, the case of a model where the *lagged* dependent variable y_{t-1} is one of the explanatory variables, or, more generally, when x_t contains some *lagged endogenous* variables.
In this case the OLS estimator is *biased*, but *consistent*.
Bias follows from $E[\hat{\beta} - \beta] = E[(X'X)^{-1}X'u] \neq (X'X)^{-1}X'E(u)$, because the *whole* vector u is not independent of the *whole* matrix X, and so the result is generally $\neq 0$.
Consistency follows still from (6.27) observing that

$$plim\frac{X'u}{n} = plim\frac{1}{n}\sum_{t=1}^{n} x_t u_t = lim\frac{1}{n}\sum_{t=1}^{n} E[x_t u_t] \tag{6.29}$$

which is $= 0$, because each of the n vectors $(k \times 1)$ in the sum has zero expected value: $E(x_t u_t) = E(x_t)E(u_t) = 0$, being the *contemporaneous* x_t and u_t independent.
Of course, the vectors in the sum are not independent of each other; for instance, x_t may contain a lagged endogenous variable, that is a function of u_{t-1}, so that $x_t u_t$ and $x_{t-1}u_{t-1}$ are not independent vectors. So it is necessary to resort to a *suitable* form of the weak law of large numbers (WLLN) for *non-independent* sequences.

6.4 Fourth case

Contemporaneous explanatory variables and error terms are not independent. In other words, some of the explanatory variables at time t (elements of the vector x_t) are not independent of u_t. This is, for example, the case of a model where a *current endogenous* variable is one of the explanatory variables of the equation (for example, a structural form equation of a simultaneous equation model).
In this case the OLS estimator is *biased* (as in the previous case) and *inconsistent*.
Inconsistency follows observing that (6.29) usually produces a result $\neq 0$, because the *contemporaneous* x_t and u_t are not independent, so that each of the n vectors $(k \times 1)$ has *nonzero* expected value: $E[x_t u_t] \neq 0$,
Therefore, in this last case it is necessary to resort to estimation methods different from OLS.

7 INSTRUMENTAL VARIABLES (I.V.)

Let W be a $n \times k$ matrix (same dimensions as X), such that the two $k \times k$ matrices $W'X/n$ and $W'W/n$ are both non-singular for any n, and both converge to finite, non-singular, *constant* limits as $n \to \infty$.
Temporarily we assume that the matrix W *does not contain random variables*. This assumption will help in simplifying the first proofs of the next section; then it will be relaxed, and random variables (with some limitations) will be admitted into W. Thus, *convergence* for $W'W/n$ is in *mathematical* sense (lim), while for $W'X/n$ is in *probability* ($plim$), if X contains random variables.
Define the *instrumental variable estimator* (that makes use of W as a matrix of instruments) as

$$\tilde{\beta}_W = (W'X)^{-1}W'y \tag{7.30}$$

Substituting into the above expression $y = X\beta + u$ we get the estimation error

$$\tilde{\beta}_W - \beta = (W'X)^{-1}W'u = \left(\frac{W'X}{n}\right)^{-1}\frac{W'u}{n} \tag{7.31}$$

and the estimation error rescaled by $\sqrt{n}$

$$\sqrt{n}(\tilde{\beta}_W - \beta) = \left(\frac{W'X}{n}\right)^{-1}\frac{W'u}{\sqrt{n}} \tag{7.32}$$

8 ASYMPTOTIC PROPERTIES OF INSTRUMENTAL VARIABLE ESTIMATOR

We have the following preliminary results.

$$plim\frac{W'u}{n} = 0 \tag{8.33}$$

This follows from the weak law of large numbers (WLLN) observing that $W'u/n = \sum_{t=1}^{n} w_t u_t/n$ is the $(k \times 1)$ vector *arithmetical average* of the n vectors $w_t u_t$, each of which has zero expected value: $E(w_t u_t) = w_t E(u_t) = 0$.

$$\frac{W'u}{\sqrt{n}} \quad \underset{n \to \infty}{\overrightarrow{distr}} \quad N\left(0, \sigma^2 lim\frac{W'W}{n}\right) \tag{8.34}$$

This can be easily proved considering that the $(k \times 1)$ vector $W'u/\sqrt{n}$ has zero expected value and variance-covariance matrix $E(W'uu'W)/n = W'E(uu')W/n = W'\sigma^2 I_n W/n = \sigma^2 W'W/n$. This expression of the variance-covariance matrix is valid for any n, therefore also in the limit. The normal distribution is obtained by a straightforward application of the central limit theorem.

8.1 Consistency and asymptotic normality of the Instrumental Variable estimator

If we consider the estimation error (7.31), then

$$plim(\tilde{\beta}_W - \beta) = plim\left[\left(\frac{W'X}{n}\right)^{-1}\frac{W'u}{n}\right] = plim\left(\frac{W'X}{n}\right)^{-1} plim\frac{W'u}{n} = 0 \tag{8.35}$$

as it follows from (8.33). If we consider the estimation error rescaled by $\sqrt{n}$ (7.32), then

$$\sqrt{n}(\tilde{\beta}_W - \beta) = \left(\frac{W'X}{n}\right)^{-1}\frac{W'u}{\sqrt{n}} \quad \underset{n \to \infty}{\overrightarrow{distr}} \quad N\left[0, \sigma^2\left(plim\frac{W'X}{n}\right)^{-1} lim\frac{W'W}{n}\left(plim\frac{X'W}{n}\right)^{-1}\right] \tag{8.36}$$

This follows from considering separately the limits of the two elements of the product: $(W'X/n)^{-1}$, whose limit is the inverse of the *constant* $plim W'X/n$, and $W'u/\sqrt{n}$ whose limit is the multivariate normal distribution, with zero mean, given in (8.34).

8.2 Efficient instrumental variables: expectations of regressors

Since w_t does not contain random variables, it is

$$E(w_t x_t') = w_t E(x_t') \tag{8.37}$$

Then

$$plim\frac{W'X}{n} = lim\frac{W'E(X)}{n} \tag{8.38}$$

which follows from a straightforward application of some *suitable version* of the weak law of large numbers (WLLN), observing that $W'X/n = \sum_{t=1}^{n} w_t x_t'/n$ it the $(k \times k)$ matrix *arithmetical average* of the n matrices $w_t x_t'$, each of which has expected value given by (8.37). Finally, $W'E(X)/n = \sum_{t=1}^{n} w_t E(x_t')/n$ is the $(k \times k)$ matrix *arithmetical average* of the n matrices containing the expected values $w_t E(x_t')$. By assumption, the limit exists and is a finite non-singular matrix. Moreover, it is clear from the right hand side of (8.38) that it does not contain random variables (so it can be treated as a *constant*). Applying (8.38), the asymptotic variance-covariance matrix in (8.36) can be written as

$$\sigma^2\left(plim\frac{W'X}{n}\right)^{-1} lim\frac{W'W}{n}\left(plim\frac{X'W}{n}\right)^{-1} = \sigma^2\left(lim\frac{W'E(X)}{n}\right)^{-1} lim\frac{W'W}{n}\left(lim\frac{E(X')W}{n}\right)^{-1} \tag{8.39}$$

If we choose $W = E(X)$, the above asymptotic variance-covariance matrix becomes

$$\sigma^2 \left(lim \frac{E(X')E(X)}{n} \right)^{-1} \tag{8.40}$$

For any other choice of W, the asymptotic variance-covariance matrix (8.39) cannot be smaller than (8.40)

$$\sigma^2 \left(lim \frac{W'E(X)}{n} \right)^{-1} lim \frac{W'W}{n} \left(lim \frac{E(X')W}{n} \right)^{-1} \geq \sigma^2 \left(lim \frac{E(X')E(X)}{n} \right)^{-1} \tag{8.41}$$

according to Schwarz inequality.
Thus $W = E(X)$ can be called the matrix of *efficient instrumental variables*.

8.3 Efficient instrumental variables: conditional expectations of regressors

Sections 8, 8.1 and 8.2 proved consistency, asymptotic normality and efficieny, confining to "*non-random* variables only" the choice of the instrumental variables (elements of the matrix W). Quite similar results hold, still under the assumptions (6.25), if we "enlarge" the choice of the instrumental variables. We admit also random variables among the elements of W, provided that, at time t, all the elements of w_t are *independent* from the random error terms u_t, u_{t+1}, u_{t+2}... Notice that the "independence" requirement is stronger than strictly necessary, and is here assumed to simplify the proofs. It is, however, important to notice that it would *not be enough* to assume that u_t and w_t are not correlated. The same consideration holds for the *strong* assumption on the u_t (i.i.d., eq. 6.25, rather than simply not autocorrelated).
Exogenous variables can be random variables, but they satisfy the requirement, and so they can be used as elements of W.
At time t, *lagged* endogenous variables (lagged one or more periods) also satisfy the requirement, so they can be used as elements of w_t. On the contrary, the value of *current* endogenous variables (or future endogenous variables) cannot be used as elements of w_t.
Since all the variables in the simplified world summarized by the model are included in the vectors y_t and z_t, for varying t, the vector of instrumental variables at time t, w_t, can include any element of z_t, z_{t-1}, z_{t-2}, etc., but no element of y_t, y_{t+1}, etc. In principle, it might also contain any exogenous element of z_{t+1}, z_{t+2}, etc., but no lagged endogenous element of z_{t+1}, z_{t+2}, etc. The set of variables that can be used as elements of w_t will be indicated as $\Im_t$. It contains, as a subset, all the *non-random* variables that were considered as the only possible elements of W in the previous sections.
With some simple changes, the main results of sections 8, 8.1 and 8.2 can now be proved under the new, less restrictive conditions on the instrumental variables choice. The differences will be $E(x_t|\Im_t)$ replacing $E(x_t)$ in all the formulas, and *plim* replacing *lim* when the sequences contain random variables.
Analogously to (8.33) we have

$$plim \frac{W'u}{n} = 0 \tag{8.42}$$

because $W'u/n = \sum_{t=1}^{n} w_t u_t/n$ is the $(k \times 1)$ vector *arithmetical average* of the n vectors $w_t u_t$, each of which has zero expected value: $E(w_t u_t) = E(w_t)E(u_t) = 0$.
Analogously to (8.34) we have

$$\frac{W'u}{\sqrt{n}} \quad \underset{n \to \infty}{distr} \quad N\left(0, \sigma^2 plim \frac{W'W}{n}\right) \tag{8.43}$$

This follows from some *suitable version* of the central limit theorem (CLT, for non-independent sequences), considering that the $(k \times 1)$ vector $W'u/\sqrt{n} = \sum_{t=1}^{n} w_t u_t/\sqrt{n}$, where each term has zero expected value. Computing its variance-covariance matrix, we get $E(W'uu'W)/n = E\left[\left(\sum_{t=1}^{n} w_t u_t\right)\left(\sum_{t=1}^{n} w_t' u_t\right)\right]/n = E\left[\sum_{t=1}^{n} w_t w_t' u_t^2\right]/n + E\left[\sum_{r \neq s} w_r w_s' u_r u_s\right]/n$ (notice that each element in the the second sum is zero being always *one of the* u_r or u_s independent of all the other terms of the product) $= \sum_{t=1}^{n} E[u_t^2 w_t w_t']/n$ (notice also that the independence of u_t from w_t implies independence of u_t^2 as well; it would not happen if they were simply not correlated) $= \sum_{t=1}^{n} [E(u_t^2)E(w_t w_t')]/n = \sigma^2 \sum_{t=1}^{n} E[w_t w_t']/n$, whose limit is $\sigma^2 plim W'W/n$ (having applied some *suitable* WLLN for non-independent sequences).
Analogously to (8.36) we have

$$\sqrt{n}(\tilde{\beta}_W - \beta) = \left(\frac{W'X}{n}\right)^{-1} \frac{W'u}{\sqrt{n}} \quad \underset{n \to \infty}{distr} \quad N\left[0, \sigma^2 \left(plim \frac{W'X}{n}\right)^{-1} plim \frac{W'W}{n} \left(plim \frac{X'W}{n}\right)^{-1}\right] \tag{8.44}$$

that follows from considering separately the limits of the two elements of the product and applying the previous results.
Analogously to (8.37) we have

$$E(w_t x_t'|\Im_t) = w_t E(x_t'|\Im_t) \tag{8.45}$$

because w_t is $\sigma(\Im_t) - measurable$; roughly speaking, when $\Im_t$ is known, also w_t is known, thus it can be *moved outside* conditional expectation. However it must be noticed that, unlike (8.37), here w_t and $E(x_t|\Im_t)$ are *random variables*.
A new simbol must be introduced to indicate the matrix whose $t - th$ row is $E(x_t'|\Im_t)$

$$E_{\Im}(X) = \underset{(n \times k)}{\begin{bmatrix} E(x_{1,1}|\Im_1) & E(x_{1,2}|\Im_1) & ... & E(x_{1,k}|\Im_1) \\ E(x_{2,1}|\Im_2) & E(x_{2,2}|\Im_2) & ... & E(x_{2,k}|\Im_2) \\ ... & ... & ... & ... \\ E(x_{t,1}|\Im_t) & E(x_{t,2}|\Im_t) & ... & E(x_{t,k}|\Im_t) \\ ... & ... & ... & ... \\ E(x_{n,1}|\Im_n) & E(x_{n,2}|\Im_n) & ... & E(x_{n,k}|\Im_n) \end{bmatrix}} = \begin{bmatrix} E(x'_1|\Im_1) \\ E(x'_2|\Im_2) \\ ... \\ E(x'_t|\Im_t) \\ ... \\ E(x'_n|\Im_n) \end{bmatrix} \tag{8.46}$$

Notice that in each row the expectation is conditional on a *different*, time varying information set.
Analogously to (8.38) we have

$$plim\frac{W'X}{n} = plim\frac{W'E_{\Im}(X)}{n} \tag{8.47}$$

This can be proved observing that $plim W'X/n = plim \sum_{t=1}^{n} w_t x'_t/n$ (applying some *suitable* WLLN) $= lim \sum_{t=1}^{n} E(w_t x'_t)/n$ (thus it is not random; we assume that the limit exists, and is a finite non-singular matrix. Applying now iterated expectations) $= lim \sum_{t=1}^{n} E[E(w_t x'_t|\Im_t)]/n$ (applying 8.45) $= lim \sum_{t=1}^{n} E[w_t E(x'_t|\Im_t)]/n$ (WLLN) $= plim \sum_{t=1}^{n} w_t E(x'_t|\Im_t)/n$ $= plim W'E_{\Im}(X)/n$.
Applying (8.47), the asymptotic variance-covariance matrix in (8.44) can be written as

$$\sigma^2 \left(plim\frac{W'X}{n}\right)^{-1} plim\frac{W'W}{n} \left(plim\frac{X'W}{n}\right)^{-1} = \sigma^2 \left(plim\frac{W'E_{\Im}(X)}{n}\right)^{-1} plim\frac{W'W}{n} \left(plim\frac{E_{\Im}(X')W}{n}\right)^{-1} \tag{8.48}$$

Choosing $W = E_{\Im}(X)$, that is, at time t, $w_t = E(x_t|\Im_t)$, the above asymptotic variance-covariance matrix becomes

$$\sigma^2 \left(plim\frac{E_{\Im}(X')E_{\Im}(X)}{n}\right)^{-1} \tag{8.49}$$

which is the *smallest possible*, being for any other choice of W

$$\sigma^2 \left(plim\frac{W'E_{\Im}(X)}{n}\right)^{-1} plim\frac{W'W}{n} \left(plim\frac{E_{\Im}(X')W}{n}\right)^{-1} \geq \sigma^2 \left(plim\frac{E_{\Im}(X')E_{\Im}(X)}{n}\right)^{-1} \tag{8.50}$$

according to Schwarz inequality; this is analogous to (8.41).
Thus $W = E_{\Im}(X)$ can be called the matrix of *efficient instrumental variables*.
Notice that, being this the most efficient choice in the new class of instrumental variables, that include the *previous* instrumental variables (*non-random*) as a subset, it must be *more* efficient than (or at least as efficient as) the previous choice. This follows also considering directly that

$$\left(plim\frac{E_{\Im}(X')E_{\Im}(X)}{n}\right)^{-1} \leq \left(lim\frac{E(X')E(X)}{n}\right)^{-1} \tag{8.51}$$

because $plim E_{\Im}(X')E_{\Im}(X)/n = plim \sum_{t=1}^{n} [E(x_t|\Im_t)E(x'_t|\Im_t)]/n = lim \sum_{t=1}^{n} E[E(x_t|\Im_t)E(x'_t|\Im_t)]/n$ (each term of the sum is the "expectation of a square", that is always $\geq$ the "square of the expectation") $\geq lim \sum_{t=1}^{n} E[E(x_t|\Im_t)]E[E(x'_t|\Im_t)]/n$ $= lim \sum_{t=1}^{n} E(x_t)E(x'_t)/n = lim E(X')E(X)/n$. The variance-covariance matrices are obtained inverting the expressions, so that the inequality would be inverted, as in (8.51).
To conclude, we observe that if a regressor at time t (an element of x_t) is exogenous or lagged endogenous (thus it is an element of z_t), it coincides with its conditional expectation, given $\Im_t$, because all elements of z_t belong to $\Im_t$. Thus it remains unchanged in the vector of efficient instrumental variables w_t. If a regressor at time t is a *current* endogenous, its conditional expectation, given $\Im_t$, follows immediately from the reduced form: $y_t = \Pi z_t + v_t$, thus $E[y_t|\Im_t] = \Pi z_t$.
In all cases we obtain as efficient instrumental variables the same values that would be obtained by treating exogenous variables and lagged endogenous variables *as if they were non-random*. In such a case, in fact, we could simply say that $z_t = E[z_t]$, with a notational simplification over $E[z_t|\Im_t]$; also, we can say that $\Pi z_t = E[y_t]$, with a notational simplification over $E[y_t|\Im_t]$. This suggest to adopt a *trick* (8.4) to simplify notations.

8.4 A simplification trick

What has been proved above is that, if we treat exogenous and lagged endogenous variables *as if they were non-random variables*, the main results remain valid, with a considerable simplification of notations.

Let's consider an equation where a *current endogenous* variable is among the regressors (an endogenous variable at time t is one of the elements of the vector x_t). For example, in the first structural equation of the Klein-I model (the private consumption equation)

$$C_t = \alpha_1 + \alpha_2\ P_t + \alpha_3\ P_{t-1} + \alpha_4\ W_t + u_{1,t}$$

the vector of regressors (explanatory variables) at time t is $x_t = [1, P_t, P_{t-1}, W_t]'$. Current profits ($P_t$) is the second regressor of C_t in the structural form. In the *reduced form* system $y_t = \Pi z_t + v_t$, the equation of P_t is the $5th$, being P_t the $5th$ element of the 7×1 vector y_t. Therefore $P_t = [y_t]_5 = \Pi_{5,\bullet} z_t + v_{5,t}$, being $\Pi_{5,\bullet}$ the $5th$ row of the 7×8 matrix Π. Since $\Im_t$ contains all the elements of z_t, it is, $E[P_t|\Im_t] = \Pi_{5,\bullet} z_t$. Exactly the same value would be obtained using the *simplification trick*:

$E[P_t] = \Pi_{5,\bullet} z_t$, because z_t contains only exogenous and lagged endogenous variables (thus *non-random*), and the error term $v_{5,t}$ has zero mean. Thus the vector of *efficient instrumental variables* at time t, $w_t = E(x_t)$, should contain, as a second element, $E[P_t] = \Pi_{5,\bullet} z_t$.
We do analogously for the fourth element of the vector w_t, that should be filled by $E[W_t] = \Pi_{7,\bullet} z_t$, being *total wages and salaries* the $7th$ endogenous variable of the model.
The first (1) and the third (P_{t-1}) element of the vector w_t are equal to the corresponding elements of x_t, because they are *non-random* (simplification trick).
Notice finally that being $[P_t] = \Pi_{5,\bullet} z_t + v_{5,t}$, the scalar $E[P_t] = \Pi_{5,\bullet} z_t$ can be viewed as a linear combination of the elements of z_t, but also as the observed value of the endogenous variable P_t *purged* of its reduced form error $E[P_t] = P_t - v_{5,t}$.
In all the formulas that follow, $E(X)$ implicitly means $E_{\Im}(X)$, and *expectation* implicitly means *conditional expectation.*

8.5 Instrumental variables for Klein-I model

The model has 3 stochastic behavioural equations. We call X_1 the $(n \times k_1)$ matrix of the explanatory variables in the structural form equation of consumption. X_2 $(n \times k_2)$ and X_3 $(n \times k_3)$ are the matrices of explanatory variables in the structural form equations of investment and private wages, respectively. For this particular model the three matrices have the same dimensions (21×4). The $t-th$ row of these matrices are as follows

$$\underset{(1\times k_1)}{x'_{1t} =} \begin{bmatrix} 1 & P_t & P_{t-1} & W_t \end{bmatrix} \qquad \underset{(1\times k_2)}{x'_{2t} =} \begin{bmatrix} 1 & P_t & P_{t-1} & K_{t-1} \end{bmatrix} \qquad \underset{(1\times k_3)}{x'_{3t} =} \begin{bmatrix} 1 & X_t & X_{t-1} & A_t \end{bmatrix}$$

The matrices W_1, W_2 and W_3 have the same dimensions as the corresponding matrices X_1, X_2 and X_3. Their $t-th$ rows are as follows

$$\underset{(1\times k_1)}{w'_{1t} =} \begin{bmatrix} 1 & \Pi_{5,\bullet} z_t & P_{t-1} & \Pi_{7,\bullet} z_t \end{bmatrix} \qquad \underset{(1\times k_2)}{w'_{2t} =} \begin{bmatrix} 1 & \Pi_{5,\bullet} z_t & P_{t-1} & K_{t-1} \end{bmatrix} \qquad \underset{(1\times k_3)}{w'_{3t} =} \begin{bmatrix} 1 & \Pi_{4,\bullet} z_t & X_{t-1} & A_t \end{bmatrix}$$

8.6 Feasible instrumental variable estimator

Unfortunately, the method discussed above is asymptotically efficient just in principle; in practice the method is *not feasible.* To make the method *feasible*, we shall replace the $(n \times k)$ matrix $E(X)$ with a matrix that contains *good* estimates of the expected values of the elements of X. So, in practice, we shall use as a matrix of instrumental variables

$$W = \widehat{E(X)} \tag{8.52}$$

More or less all the estimation methods proposed in the literature use instrumental variables of this type (8.52). The differences from one another are due to different ways of computing the *estimated expected values* $\widehat{E(X)}$.
Concerning the consumption equation, being Π (and therefore $\Pi_{5,\bullet}$) unknown, to make the estimation method *feasible* in practice we first estimate Π (or at least $\Pi_{5,\bullet}$), obtaining $\widehat{\Pi}$, and then plug into w_t, as its second element, the scalar $\widehat{E[P_t]} = \widehat{\Pi}_{5,\bullet} z_t$.
If a *consistent* estimator of Π is used to build the matrix of instrumental variables, then the resulting *feasible* instrumental variable estimator has the same asymptotic variance-covariance matrix as the *not feasible* efficient estimator (the one that would use the *true* matrix Π).
To prove it, we can consider how the estimation error (eq. 8.36) changes if we use $W = E(X)$ (the *not feasible* estimator that uses the *true* Π) or if we use $W = \widehat{E(X)}$ (the *feasible* estimator that uses a consistent estimator $\widehat{\Pi}$). Let's first consider the $(k \times k)$ matrix $W'X/n$ of equation (8.36). It has exactly the same *plim* whether we use $W = E(X)$, or we use $W = \widehat{E(X)}$

$$plim\left(\frac{E(X')X}{n}\right) = plim\left(\frac{\widehat{E(X')}X}{n}\right)$$

The above equality can be easily proved element by element. For example, still with reference to the consumption equation of the Klein-I model, the element (1,2) of such a matrix is $\sum_{t=1}^{n} E(P_t)/n = \sum_{t=1}^{n} \Pi_{5,\bullet} z_t/n = \Pi_{5,\bullet} (\sum_{t=1}^{n} z_t/n)$ in the *not feasible* case, while in the *feasible* case it is $\sum_{t=1}^{n} \widehat{E(P_t)}/n = \sum_{t=1}^{n} \widehat{\Pi}_{5,\bullet} z_t/n = \widehat{\Pi}_{5,\bullet} (\sum_{t=1}^{n} z_t/n)$. The two expressions have obviously the same limit if $plim\widehat{\Pi}_{5,\bullet} = \Pi_{5,\bullet}$.
Analogously, the equality can be proved for all the other elements of the $(k \times k)$ matrix (4×4, in the example).
Considering now the $(k \times 1)$ vector $W'u/\sqrt{n}$ in equation (8.36), again it is straightforward to verify that each element converges to the same distribution whether we use $W = E(X)$, or we use $W = \widehat{E(X)}$.

We conclude, therefore, that also the *feasible* estimator is asymptotically efficient.

9 LIMITED INFORMATION ESTIMATION METHODS (or Single Equation Estimation Methods)

Most of the different *traditional* estimation methods of the literature are based on equation (7.30), with different ways of computing the feasible $W = \widehat{E(X)}$ (more precisely, $W = \widehat{E_{\Im}(X)}$). Its computation always uses a *previously computed* estimator ($\widehat{\Pi}$) of the matrix of reduced form coefficients, such that $plim\widehat{\Pi} = \Pi$ (consistent estimator of Π). All estimation

methods are performed in several *stages* (or steps, two or more than two): the final stage is always equation (7.30), while the previous stages aim at providing a consistent estimator of Π.
Limited information methods do not exploit information contained in the correlation between error terms of different equations.

9.1 2SLS - Two Stage Least Squares: *Basmann (1957), Theil (1958)*

We first select all the *current* endogenous variables appearing somewhere on the right hand side of the structural form equations. Then we regress, with OLS, each of these current endogenous variables against *all the exogenous and lagged endogenous variables* of the system (*first stage*). The *fitted* values of these variables are used in the matrices of instrumental variables, where exogenous and lagged endogenous variables are left at their observed value. Then we apply the instrumental variables formula (7.30) to each structural form equation (*second stage*).
The first stage is an OLS estimation of each reduced form equations, *unrestricted*. Each OLS provides a consistent estimate of a row of Π, since the variables on the right hand side of each equation are only exogenous and lagged endogenous variables. The *fitted* values of the dependent variables can therefore be used in the matrices of instrumental variables, to *replace the current endogenous regressors* of the structural form equations.
Having built the matrices of instrumental variables *in this particular way*, the results remain *algebraically equal* if, instead of the I.V. formula, in the second stage we again apply the OLS formula. For instance, in the first equation, $W_1'X_1 = W_1'W_1$, thus $(W_1'X_1)^{-1}W_1'y_1 = (W_1'W_1)^{-1}W_1'y_1$. For this reason the method is called *two stage least squares*.
2SLS is perhaps the most *popular* among limited information methods. It cannot be applied to large scale systems. In fact, when the number of exogenous and lagged endogenous variables in the system is too large ($> n$), the first stage OLS estimation is not feasible.

9.2 LIVE - Limited information Instrumental Variables Efficient: *Brundy and Jorgenson (1971), Dhrymes (1971)*

In the first stage of this method some *arbitrary* matrices of instrumental variables are used, and equation (7.30) is applied to each structural form equation. In the example, we use three matrices W_1, W_2 and W_3 that only need to satisfy the quite general requirements for the matrices of instrumental variables given in section 7.
This first stage provides, for each structural form equation, coefficient estimates which are consistent, but not asymptotically efficient. Estimated coefficients are then *plugged into* the matrices of structural form coefficients, producing a consistent (but inefficient) estimate of B and Γ.
Inverting the estimated B and multiplying by the estimated Γ (with minus sign) provides a consistent estimate of the matrix of reduced form coefficients Π. This estimate of Π is now used to build, for each equation, the matrix of the estimated expected values of the regressors, to be used as new matrices of instrumental variables (as in section 8.5 for the example model).
Then the second stage applies equation (7.30) to each structural form equation, producing coefficient estimates which are consistent and asymptotically efficient.
Unlike 2SLS, this method estimates Π from the *restricted* reduced form. The *estimation formula* is only applied to the structural form equations, (usually with a *small* number of regressors), thus the method can be applied also to large scale models. It is, however, less *robust* than 2SLS. A specification error in a structural form equation may have consequences in the estimation of the other equations as well, even if correctly specified. This does not happen for 2SLS, where a specification error in one equation has consequences only for such equation.
Notice finally that the estimated expected values of the endogenous regressors, to be used in the i.v. matrices of the second stage, are the values of the endogenous variables computed from the simultaneous solution of the structural form model, using the terminology of section 4. Solution is, of course, static (or one-step-ahead), since lagged endogenous are considered fixed (section 8.4).
The instrumental variables used in the first stage can be completely arbitrary, as already observed. A simple technique is customarily (even if not necessarily) adopted to build them. A preliminary estimation is done, using OLS on the structural form equations. Estimates would therefore be inconsistent, but *presumably better* than if we *invent* them from scratch. From these estimates, filling the matrices B and Γ we compute an estimate of Π (still inconsistent, of course), and use it to fill the matrices of instrumental variables to be used in the first stage. Then, first and second stage are as above.

9.3 IIV - Iterative Instrumental Variables: *Dutta and Lyttkens (1974), Lyttkens (1974)*

The final stage of LIVE can be applied iteratively, till convergence is achieved. At the end of each iteration, estimated coefficients are *plugged into* the matrices B and Γ; a new estimate of Π is then computed; new matrices of instrumental variables are then computed and used in the next iteration.
Each new iteration (or stage) may change the numerical values of the estimates, but not their asymptotic distribution: efficiency has been already achieved at the second stage.

9.4 k-class Estimator: *Theil (1958), Nagar (1959)*

It is convenient here to interpret the instrumental variables as at the end of section 8.4, that is the observed value of each regressor *purged* of its reduced form error.

With reference to the first structural form equation of the example model, we may replace P_t, in the matrix of instrumental variables, with $P_t - k\ \hat{v}_{5,t}$, where k is a scalar random variable, function of the data. Analogously, we replace W_t with $W_t - k\ \hat{v}_{7,t}$. If $\hat{v}_{5,t}$ and $\hat{v}_{7,t}$ are residuals of OLS applied to the unrestricted reduced form (as in the first stage of 2SLS), the instrumental variable estimator is called *k-class* estimator. It is straightforward to prove that the estimator is consistent if $plim\ (k-1) = 0$, and also asymptotically efficient if $plim\ \sqrt{n}\ (k-1) = 0$. Roughly speaking, k must converge to 1 *fast enough*.
2SLS is the particular case when $k = 1$; as well known, it is consistent and asymptotically efficient. OLS is the particular case when $k = 0$, and it is inconsistent.

9.5 GIVE - Generalized Instrumental Variable Estimator: *Sargan (1958)*

9.6 LIML - Limited Information Maximum Likelihood: *Anderson and Rubin (1949, 1950)*

10 SEEMINGLY UNRELATED REGRESSION EQUATIONS (SURE)

A system of G linear regression models, *without* endogenous regressors,

$$\begin{bmatrix} y_1 & = X_1\beta_1 + u_1 \\ & ... \\ y_i & = X_i\beta_i + u_i \\ & ... \\ y_G & = X_G\beta_G + u_G \end{bmatrix} \quad \text{where} \quad \underset{(n\times 1)}{y_i} = \begin{bmatrix} y_{i,1} \\ y_{i,2} \\ ... \\ y_{i,t} \\ ... \\ y_{i,n} \end{bmatrix} \quad \underset{(n\times k_i)}{X_i} = \begin{bmatrix} x_{i,1,1} & x_{i,1,2} & ... & x_{i,1,k_i} \\ x_{i,2,1} & x_{i,2,2} & ... & x_{i,2,k_i} \\ ... & ... & ... & ... \\ x_{i,t,1} & x_{i,t,2} & ... & x_{i,t,k_i} \\ ... & ... & ... & ... \\ x_{i,n,1} & x_{i,n,2} & ... & x_{i,n,k_i} \end{bmatrix} \quad \underset{(k_i\times 1)}{\beta_i} = \begin{bmatrix} \beta_{i,1} \\ \beta_{i,2} \\ ... \\ \beta_{i,k_i} \end{bmatrix}$$

$$\underset{(n\times 1)}{u_i} = \begin{bmatrix} u_{i,1} \\ u_{i,2} \\ ... \\ u_{i,t} \\ ... \\ u_{i,n} \end{bmatrix} \qquad \begin{bmatrix} E[u_{i,t}] = 0 \ \ \forall\ i,t \\ Var[u_{i,t}] = \sigma_i^2 = \sigma_{i,i} \ \ \forall\ t \\ Cov[u_{i,t}, u_{j,t}] = \sigma_{i,j} \ \ \forall\ t \\ Cov[u_{i,t_1}, u_{j,t_2}] = 0 \ \ \forall\ i,j,t_1 \neq t_2 \end{bmatrix} \tag{10.53}$$

can be represented as a *single* linear regression model, $y = X\beta + u$ with Gn observations, defining the vectors and matrices

$$y = X\beta + u \qquad \underset{(Gn\times 1)}{y} = \begin{bmatrix} y_1 \\ y_2 \\ ... \\ y_i \\ ... \\ y_G \end{bmatrix} \quad \underset{[Gn\times(k_1+...+k_G)]}{X} = \begin{bmatrix} X_1 & 0 & ... & 0 \\ 0 & X_2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & ... & X_i & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & X_G \end{bmatrix} \quad \underset{[(k_1+...+k_G)\times 1]}{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ ... \\ \beta_i \\ ... \\ \beta_G \end{bmatrix} \quad \underset{(Gn\times 1)}{u} = \begin{bmatrix} u_1 \\ u_2 \\ ... \\ u_i \\ ... \\ u_G \end{bmatrix} \tag{10.54}$$

where the vector of error terms, with Gn elements, has expected value zero and variance-covariance matrix

$$\underset{(Gn\times Gn)}{Var(u)} = \underset{(G\times G)}{\Sigma} \otimes \underset{(n\times n)}{I_n} = \begin{bmatrix} \sigma_{1,1} & 0 & .. & 0 & \sigma_{1,2} & 0 & .. & 0 & .. & \sigma_{1,G} & 0 & .. & 0 \\ 0 & \sigma_{1,1} & .. & 0 & 0 & \sigma_{1,2} & .. & 0 & .. & 0 & \sigma_{1,G} & .. & 0 \\ ... & ... & .. & ... & ... & ... & .. & ... & .. & ... & ... & .. & ... \\ 0 & 0 & .. & \sigma_{1,1} & 0 & 0 & .. & \sigma_{1,2} & .. & 0 & 0 & .. & \sigma_{1,G} \\ \sigma_{2,1} & 0 & .. & 0 & \sigma_{2,2} & 0 & .. & 0 & .. & \sigma_{2,G} & 0 & .. & 0 \\ 0 & \sigma_{2,1} & .. & 0 & 0 & \sigma_{2,2} & .. & 0 & .. & 0 & \sigma_{2,G} & .. & 0 \\ ... & ... & .. & ... & ... & ... & .. & ... & .. & ... & ... & .. & ... \\ 0 & 0 & .. & \sigma_{2,1} & 0 & 0 & .. & \sigma_{2,2} & .. & 0 & 0 & .. & \sigma_{2,G} \\ ... & ... & .. & ... & ... & ... & .. & ... & .. & ... & ... & .. & ... \\ ... & ... & .. & ... & ... & ... & .. & ... & .. & ... & ... & .. & ... \\ \sigma_{G,1} & 0 & .. & 0 & \sigma_{G,2} & 0 & .. & 0 & .. & \sigma_{G,G} & 0 & .. & 0 \\ 0 & \sigma_{G,1} & .. & 0 & 0 & \sigma_{G,2} & .. & 0 & .. & 0 & \sigma_{G,G} & .. & 0 \\ ... & ... & .. & ... & ... & ... & .. & ... & .. & ... & ... & .. & ... \\ 0 & 0 & .. & \sigma_{G,1} & 0 & 0 & .. & \sigma_{G,2} & .. & 0 & 0 & .. & \sigma_{G,G} \end{bmatrix} \tag{10.55}$$

whose inverse is $\Sigma^{-1} \otimes I_n$.
There is no explicit relationship among equations, since there are no current endogenous variables on the right hand side of the equations (no *simultaneity*). There is, however, a relationship due to the correlations among contemporaneous error terms (or cross-equations correlations).

10.1 An example of SURE model: *Zellner (1962)*

The model is a system of 2 equations, each with 3 explanatory variables (regressors). Dependent variables are annual gross investments of two corporations, during the period 1935-1954.

$$\left[\begin{array}{lll} I_t^{GE} & = \beta_{1,1} + \beta_{1,2}\ F_{t-1}^{GE} + \beta_{1,3}\ C_{t-1}^{GE} + u_{1,t} & \text{General Electric} \\ I_t^{W} & = \beta_{2,1} + \beta_{2,2}\ F_{t-1}^{W} + \beta_{2,3}\ C_{t-1}^{W} + u_{2,t} & \text{Westinghouse} \end{array}\right. \tag{10.56}$$

F_{t-1} is the *market value of the firm*, defined as the total value of the outstanding stock at end-of-year market quotations. C_{t-1} is the existing capital stock.

10.2 GLS and Feasible GLS estimation of SURE models

If X does not contain random variables and Σ is *known*, the GLS estimator

$$\dot{\beta}_{GLS} = [X'(\Sigma^{-1} \otimes I_n)X]^{-1}X'(\Sigma^{-1} \otimes I_n)y \tag{10.57}$$

is BLUE (Aitken's theorem), as well as consistent and asymptotically efficient. The variance-covariance matrix of the GLS estimator is

$$Var(\dot{\beta}_{GLS}) = E[(\dot{\beta}_{GLS} - \beta)(\dot{\beta}_{GLS} - \beta)'] = [X'(\Sigma^{-1} \otimes I_n)X]^{-1} \tag{10.58}$$

If Σ is not known, a *feasible* GLS estimator can be obtained from the same equation, having previoulsy computed a consistent estimate $\widehat{\Sigma}$. $\widehat{\Sigma}$ is usually computed from residuals of a preliminary OLS estimation; it is consistent, being OLS consistent for a model without endogenous regressors. It is common practice to compute $\widehat{\Sigma}$ without *degrees of freedom correction*, that is dividing by n the sums of squared residuals (variances) or the sums of cross products of contemporaneous residuals (covariances).

$$\dot{\beta}_{FGLS} = [X'(\widehat{\Sigma}^{-1} \otimes I_n)X]^{-1}X'(\widehat{\Sigma}^{-1} \otimes I_n)y \tag{10.59}$$

If $plim\widehat{\Sigma} = \Sigma$, both estimation errors (rescaled by $\sqrt{n}$) have the same asymptotic distribution (multivariate normal)

$$\left.\begin{array}{ll} \sqrt{n}\left[\dot{\beta}_{GLS} - \beta\right] & = \left[\frac{X'(\Sigma^{-1}\otimes I_n)X}{n}\right]^{-1} \quad \frac{X'(\Sigma^{-1}\otimes I_n)u}{\sqrt{n}} \\ \sqrt{n}\left[\dot{\beta}_{FGLS} - \beta\right] & = \left[\frac{X'(\widehat{\Sigma}^{-1}\otimes I_n)X}{n}\right]^{-1} \quad \frac{X'(\widehat{\Sigma}^{-1}\otimes I_n)u}{\sqrt{n}} \end{array}\right\} \xrightarrow[n\to\infty]{distr} N\left\{0, \left[\frac{X'(\Sigma^{-1} \otimes I_n)X}{n}\right]^{-1}\right\} \tag{10.60}$$

Each estimation error, in fact, is the product of two terms: the first term of the product has the same limit, in the two cases; the second term of the product has, in the two cases, the same asymptotic normal distribution.

10.3 Remarks and special cases

1. Kronecker product is a convenient algebraic operator that permits a closed form representation of the variance-covariance matrix. Its use, however, is *not recommended* in the computational practice. Software algorithms should avoid its use, because of its computational inefficiency.

   Indicating with $\hat{\sigma}^{i,j}$ the generic element of $\widehat{\Sigma}^{-1}$, it is easier and faster to compute the matrix $[X'(\widehat{\Sigma}^{-1} \otimes I_n)X]$ block by block, the $i,j-th$ block being $\hat{\sigma}^{i,j}X_i'X_j$ (of dimensions $k_i \times k_j$).

   The vector $X'(\widehat{\Sigma}^{-1} \otimes I_n)y$ would be analogously partitioned, the $i-th$ sub-vector being $X_i' \sum_{j=1}^{G} \hat{\sigma}^{i,j}y_j$.

2. GLS (or Feasible GLS) obviously gives the same results as OLS (algebraically and numerically) when Σ (or $\widehat{\Sigma}$) is diagonal (all cross equation covariances are zero).

3. Even if Σ (or $\widehat{\Sigma}$) is not diagonal (cross equation covariances are not zero), GLS (or Feasible GLS) gives the same results as OLS (algebraically and numerically) if the explanatory variables (regressors) are the same in each equation. In such a case, if we call Z the $(n \times k)$ matrix of explanatory variables common to all equations, then the block-diagonal matrix X could be represented as $X = I_G \otimes Z$ (with dimensions $Gn \times Gk$), and some straightforward algebra would give

$$\begin{aligned} \dot{\beta}_{FGLS} &= [X'(\widehat{\Sigma}^{-1} \otimes I_n)X]^{-1}X'(\widehat{\Sigma}^{-1} \otimes I_n)y = [(I_G \otimes Z)'(\widehat{\Sigma}^{-1} \otimes I_n)(I_G \otimes Z)]^{-1}(I_G \otimes Z)'(\widehat{\Sigma}^{-1} \otimes I_n)y \\ &= \{I_G \otimes [(Z'Z)^{-1}Z']\}y = \left[\begin{array}{l} \widehat{\beta}_{OLS_1} \\ ... \\ \widehat{\beta}_{OLS_i} \\ ... \\ \widehat{\beta}_{OLS_G} \end{array}\right] \end{aligned} \tag{10.61}$$

4. As a final remark, it can be shown that, with a simple transformation, current endogenous regressors appear explicitly, while they seem to be absent from (10.53), thus explaining why the equations are unrelated only *seemingly* and not really. If the contemporaneous error terms u_t, in a 2 equations model, have a bivariate normal distribution

$$\begin{bmatrix} y_{1,t} &= x'_{1,t}\beta_1 + u_{1,t} \\ y_{2,t} &= x'_{2,t}\beta_2 + u_{2,t} \end{bmatrix} \quad \text{where} \quad \begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix} \sim N[0,\Sigma] \quad \text{therefore} \quad \begin{bmatrix} u_{1,t} \sim N[0,\sigma_{1,1}] \\ u_{2,t}|u_{1,t} \sim N\left[\frac{\sigma_{1,2}}{\sigma_{1,1}}\, u_{1,t},\ \sigma_{2,2} - \frac{\sigma^2_{1,2}}{\sigma_{1,1}}\right] \end{bmatrix}$$

we can write

$$\begin{bmatrix} u_{1,t} &= \sqrt{\sigma_{1,1}}\, e_{1,t} \\ u_{2,t} &= \frac{\sigma_{1,2}}{\sigma_{1,1}}\, u_{1,t} + \sqrt{\sigma_{2,2} - \frac{\sigma^2_{1,2}}{\sigma_{1,1}}}\, e_{2,t} \end{bmatrix} \quad \text{where} \quad \begin{bmatrix} e_{1,t} \\ e_{2,t} \end{bmatrix} \sim N[0, I_2]$$

Replacing $u_{1,t} = y_{1,t} - x'_{1,t}\beta_1$ into the expression of $u_{2,t}$, then the two equations become

$$\begin{bmatrix} y_{1,t} &= x'_{1,t}\beta_1 + \sqrt{\sigma_{1,1}}\, e_{1,t} \\ y_{2,t} &= x'_{2,t}\beta_2 + \frac{\sigma_{1,2}}{\sigma_{1,1}}\, (y_{1,t} - x'_{1,t}\beta_1) + \sqrt{\sigma_{2,2} - \frac{\sigma^2_{1,2}}{\sigma_{1,1}}}\, e_{2,t} \end{bmatrix}$$

where an endogenous regressor explicitly appears in the second equation. Notice that, after transformation, the error terms are no more correlated; a system of this type is called "recursive".

10.4 Iterative Feasible GLS and Maximum Likelihood

Feasible GLS (10.59) can be applied *iteratively*, each time re-computing an estimate of Σ from residuals of the last iteration. Let $\dot\beta_{FGLS(m)}$ be the coefficient estimates at the end of iteration m, $\dot u_{FGLS(m)}$ the corresponding residuals and $\dot\Sigma_{FGLS(m)}$ the variance-covariance matrix computed from residuals. It is therefore $y = X\dot\beta_{FGLS(m)} + \dot u_{FGLS(m)}$, that can be introduced into equation (10.59) to replace y, obtaining

$$\begin{aligned} \dot\beta_{FGLS(m+1)} &= [X'(\dot\Sigma^{-1}_{FGLS(m)} \otimes I_n)X]^{-1}X'(\dot\Sigma^{-1}_{FGLS(m)} \otimes I_n)(X\dot\beta_{FGLS(m)} + \dot u_{FGLS(m)}) \\ &= \dot\beta_{FGLS(m)} + [X'(\dot\Sigma^{-1}_{FGLS(m)} \otimes I_n)X]^{-1}X'(\dot\Sigma^{-1}_{FGLS(m)} \otimes I_n)\dot u_{FGLS(m)} \end{aligned} \tag{10.62}$$

Convergence is achieved when $\dot\beta_{FGLS(m+1)} = \dot\beta_{FGLS(m)}$, therefore when $X'(\dot\Sigma^{-1}_{FGLS(m)} \otimes I_n)\dot u_{FGLS(m)} = 0$. This expression is the *gradient of the concentrated log-likelihood*, under the additional assumption that the error terms have a multivariate normal distribution. Thus, iterative feasible GLS converges to maximum likelihood (ML). Proof is in Appendix (12).
Notice that (10.60) holds for each iteration, therefore the asymptotic efficiency is the same at each iteration, as well as when convergence is achieved.

11 FULL INFORMATION ESTIMATION METHODS (or System Estimation Methods)

Matrix Σ is completely *ignored* by limited information methods, but it may contain useful information, that may improve the estimator's efficiency. Full information methods take into account also this information.

11.1 Remark

It must be noticed that estimation concerns only the behavioural *stochastic* equations of the model (the first three equations, in the example). What is called Σ in this section is therefore the variance-covariance matrix of the error terms, at time t, of the stochastic equations only, excluding the identities. In the example, it is the 3×3 positive definite matrix previously called Σ_3 (equation 3.5), and not the full 7×7 matrix that, being singular, could not be inverted. The whole system of 7 equations must be considered when computing expected values of endogenous regressors (or reduced form coefficients, or simultaneous solution of the system). For calculations involving residuals, like estimation of the Σ matrix, only the 3 stochastic equations must be considered.

11.2 Efficient instrumental variables in the full information context

We first decompose the positive definite variance-covariance matrix as the product of a non-singular square matrix P with its transpose: $\Sigma \otimes I_n = P'P$, so that $\Sigma^{-1} \otimes I_n = P^{-1}P'^{-1}$. Equations, coefficients, variables and error terms are represented as in section 10, but in some or all G equations the matrices of regressors, X_1, X_2, ..., X_G, may contain current endogenous variables. We build, therefore, the corresponding matrices of instrumental variables W_1, W_2, ..., W_G, containing the same variables, but with current endogenous variables replaced by their expected values (conditional expectations, to be more precise).
The matrices of instrumental variables W_1, W_2, ..., W_G are used as blocks of the matrix W, while the matrices of explanatory variables (regressors) X_1, X_2, ..., X_G are used as blocks of the matrix X

$$\underset{[Gn\times(k_1+...+k_G)]}{W=}\begin{bmatrix} W_1 & 0 & ... & 0 \\ 0 & W_2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & W_G \end{bmatrix}=\begin{bmatrix} E(X_1) & 0 & ... & 0 \\ 0 & E(X_2) & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & E(X_G) \end{bmatrix}=E\begin{bmatrix} X_1 & 0 & ... & 0 \\ 0 & X_2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & X_G \end{bmatrix}=E(X) \quad (11.63)$$

Analogously to section 10, we represent the whole system as a single equation $y = X\beta + u$ and pre-multiply each term by P'^{-1}, obtaining $P'^{-1}y = P'^{-1}X\beta + P'^{-1}u$. Defining $q = P'^{-1}y$, $Q = P'^{-1}X$ and $\varepsilon = P'^{-1}u$, the equation becomes $q = Q\beta + \varepsilon$, where variables and errors terms have been transformed, but coefficients are still the same as in the model of interest. Of course there will be correlation between explanatory variables Q and error terms ε.
Some simple algebra shows that the transformed error terms have zero mean and variance-covariance matrix I_{Gn}, therefore homoskedastic and not correlated. Thus, instrumental variable estimator of the transformed equation would be consistent and asymptotically efficient, if instrumental variables are the expected values of regressors

$$H = E(Q) = E(P'^{-1}X) = P'^{-1}E(X) = P'^{-1}W \quad (11.64)$$

and applying the instrumental variable formula we get

$$\breve{\beta} = [H'Q]^{-1}H'q = [W'P^{-1}P'^{-1}X]^{-1}W'P^{-1}P'^{-1}y = [W'(\Sigma^{-1}\otimes I_n)X]^{-1}W'(\Sigma^{-1}\otimes I_n)y \quad (11.65)$$

Analogously to equation (10.60), considering that $W = E(X)$, the estimation error (rescaled by $\sqrt{n}$) is

$$\sqrt{n}\left[\breve{\beta}-\beta\right]=\left[\frac{W'(\Sigma^{-1}\otimes I_n)X}{n}\right]^{-1}\frac{W'(\Sigma^{-1}\otimes I_n)u}{\sqrt{n}} \quad \underset{n\to\infty}{\overset{distr}{\longrightarrow}} \quad N\left\{0,\left[\frac{W'(\Sigma^{-1}\otimes I_n)W}{n}\right]^{-1}\right\} \quad (11.66)$$

To make the estimator *feasible* we replace Σ with an estimate, and fill matrix W with estimates of the expected values of regressors. If the estimate of Σ is computed from residuals of a preliminary consistent (even if inefficient) estimation, and estimates of expected values of regressors are computed using a preliminary consistent (even if inefficient) estimate of Π, then the feasible estimator would have the same asymptotic distribution as the *theoretical* estimator (11.66), and would be therefore asymptotically efficient.

11.3 3SLS - Three Stage Least Squares: *Zellner and Theil (1962)*

The structural form equations of the model are first estimated by two stage least squares (2SLS), obtaining a consistent estimate of all the structural form coefficients ($\tilde{\alpha}_1$, ..., $\tilde{\alpha}_{12}$ in the example), and the corresponding residuals. Sums of squares and sums of cross products of structural residuals are used to produce $\tilde{\Sigma}$, a consistent estimate of Σ, while second stage coefficients are no more used.
Third stage is simply the application of the feasible full information estimator

$$\breve{\beta} = [W'(\tilde{\Sigma}^{-1}\otimes I_n)X]^{-1}W'(\tilde{\Sigma}^{-1}\otimes I_n)y \quad (11.67)$$

where the blocks of matrix W are the same computed at the end of the first stage, and already used in the second stage.

11.4 Iterative Three Stage Least Squares

The final formula of 3SLS (11.67) can be applied *iteratively*, each time re-computing an estimate of Σ from residuals of the last iteration. Matrix W is not updated during the iterations, but remains fixed. Iterations continue till convergence is achieved.

11.5 FIVE - Full information Instrumental Variables Efficient: *Brundy and Jorgenson (1971), Dhrymes (1971)*

The first stage is the same as in the corresponding limited information method (LIVE, section 9.2). First of all it produces, for each equation, the new matrix of instrumental variables (the blocks of W). Also residuals of the structural form equations are used to produce a consistent estimate of Σ.
The next (final) stage is simply the application of the feasible full information estimator (analogous to 11.67).

11.6 FIML - Full Information Maximum Likelihood: *Koopmans, Rubin and Leipnik (1950), Chernoff and Divinsky (1953)*

11.7 FIML from iterative instrumental variables: *Durbin (1963, 1988), Hausman (1974, 1975)*

The last stage of FIVE can be applied *iteratively*, each time re-computing estimates of B and Γ, from which a new estimate of Π and new matrices of instrumental variables are derived, and re-estimating Σ from structural form residuals. Let $\breve{\beta}_{(m)}$ be the coefficient estimates at the end of iteration m, $W_{(m)}$ the matrix of instrumental variables built by means of such coefficients, $\breve{u}_{(m)}$ the corresponding residuals and $\breve{\Sigma}_{(m)}$ the variance-covariance matrix computed from residuals. It is therefore $y = X\breve{\beta}_{(m)} + \breve{u}_{(m)}$, that can be introduced into equation (11.67) to replace y, obtaining

$$\begin{aligned}\breve{\beta}_{(m+1)} &= \left[W'_{(m)}\left(\breve{\Sigma}^{-1}_{(m)}\otimes I_n\right)X\right]^{-1}W'_{(m)}\left(\breve{\Sigma}^{-1}_{(m)}\otimes I_n\right)\left(X\breve{\beta}_{(m)}+\breve{u}_{(m)}\right)\\ &= \breve{\beta}_{(m)}+\left[W'_{(m)}\left(\breve{\Sigma}^{-1}_{(m)}\otimes I_n\right)X\right]^{-1}W'_{(m)}\left(\breve{\Sigma}^{-1}_{(m)}\otimes I_n\right)\breve{u}_{(m)}\end{aligned} \tag{11.68}$$

Convergence is achieved when $\breve{\beta}_{(m+1)}=\breve{\beta}_{(m)}$, therefore when $W'_{(m)}\left(\breve{\Sigma}^{-1}_{(m)}\otimes I_n\right)\breve{u}_{(m)}=0$. This expression is the *gradient of the concentrated log-likelihood*, under the additional assumption that the error terms have a multivariate normal distribution. Proof is in Appendix (13) . Thus, iterative FIVE converges to full information maximum likelihood (FIML).
Notice that (10.60) holds for each iteration, therefore the asymptotic efficiency is the same at each iteration, as well as when convergence is achieved.
This method for computing FIML estimates was proposed by Durbin in 1963 (published in 1988), discussed in Hausman (1974, 1975), Hendry (1976), Calzolari and Sampoli (1993), and extended to nonlinear models in Amemiya (1977).

12 APPENDIX. Complements of linear algebra: some useful derivatives

Lemma: If A is a non-singular square matrix $(n\times n)$, then

$$\frac{\partial\ ln||A||}{\partial A}=A^{-1'} \tag{12.69}$$

where $||A||$ is the absolute value of the determinant and $A^{-1'}$ is the transpose of the inverse matrix.
To prove it, it is enough to observe that the Laplace expansion of the determinant for a generic row (i) is

$$|A|=a_{i,1}A_{i,1}+...+a_{i,j}A_{i,j}+...+a_{i,n}A_{i,n}$$

where none of the cofactors $A_{i,1}$, $A_{i,2}$, etc. *contains* $a_{i,j}$. Thus $\partial|A|/\partial a_{i,j}=A_{i,j}$. Application of the chain rule gives

$$\frac{\partial\ ln||A||}{\partial a_{i,j}}=\frac{\partial\ ln||A||}{\partial|A|}\ \frac{\partial|A|}{\partial a_{i,j}}=\frac{1}{|A|}A_{i,j}$$

which is the $j,i-th$ element of A^{-1}.

13 APPENDIX

14 APPENDIX. Data set and numerical results for Klein-I model

	C	I	Wp	X	P	K	W	1	Wg	T	A	G
1920	39.8	2.7	28.8	44.9	12.7	182.8	31.0	1.0	2.2	3.40	-11.	2.40
1921	41.9	-.20	25.5	45.6	12.4	182.6	28.2	1.0	2.7	7.70	-10.	3.90
1922	45.0	1.9	29.3	50.1	16.9	184.5	32.2	1.0	2.9	3.90	-9.0	3.20
1923	49.2	5.2	34.1	57.2	18.4	189.7	37.0	1.0	2.9	4.70	-8.0	2.80
1924	50.6	3.0	33.9	57.1	19.4	192.7	37.0	1.0	3.1	3.80	-7.0	3.50
1925	52.6	5.1	35.4	61.0	20.1	197.8	38.6	1.0	3.2	5.50	-6.0	3.30
1926	55.1	5.6	37.4	64.0	19.6	203.4	40.7	1.0	3.3	7.00	-5.0	3.30
1927	56.2	4.2	37.9	64.4	19.8	207.6	41.5	1.0	3.6	6.70	-4.0	4.00
1928	57.3	3.0	39.2	64.5	21.1	210.6	42.9	1.0	3.7	4.20	-3.0	4.20
1929	57.8	5.1	41.3	67.0	21.7	215.7	45.3	1.0	4.0	4.00	-2.0	4.10
1930	55.0	1.0	37.9	61.2	15.6	216.7	42.1	1.0	4.2	7.70	-1.0	5.20
1931	50.9	-3.4	34.5	53.4	11.4	213.3	39.3	1.0	4.8	7.50	.00	5.90
1932	45.6	-6.2	29.0	44.3	7.00	207.1	34.3	1.0	5.3	8.30	1.0	4.90
1933	46.5	-5.1	28.5	45.1	11.2	202.0	34.1	1.0	5.6	5.40	2.0	3.70
1934	48.7	-3.0	30.6	49.7	12.3	199.0	36.6	1.0	6.0	6.80	3.0	4.00
1935	51.3	-1.3	33.2	54.4	14.0	197.7	39.3	1.0	6.1	7.20	4.0	4.40
1936	57.7	2.1	36.8	62.7	17.6	199.8	44.2	1.0	7.4	8.30	5.0	2.90
1937	58.7	2.0	41.0	65.0	17.3	201.8	47.7	1.0	6.7	6.70	6.0	4.30
1938	57.5	-1.9	38.2	60.9	15.3	199.9	45.9	1.0	7.7	7.40	7.0	5.30
1939	61.6	1.3	41.6	69.5	19.0	201.2	49.4	1.0	7.8	8.90	8.0	6.60
1940	65.0	3.3	45.0	75.7	21.1	204.5	53.0	1.0	8.0	9.60	9.0	7.40
1941	69.7	4.9	53.3	88.4	23.5	209.4	61.8	1.0	8.5	11.6	10.	13.8

OLS estimation of stochastic structural equations. Estimation (sample) period: 1921 - 1941. Annual data

Consumption equation				Investment equation				Private wages equation			
'C' = 1., 'P', 'P-1', 'W'				'I' = 1., 'P' ,'P-1', 'K-1'				'Wp' = 1., 'X', 'X-1', 'A'			
Coeff.	Std.err	t-Stud.	Average variab.	Coeff.	Std.err	t-Stud.	Average variab.	Coeff.	Std.err	t-Stud.	Average variab.
16.2366	1.30270	12.4638	1.00000	10.1258	5.46555	1.85266	1.00000	1.49704	1.27004	1.17875	1.00000
.192934	.091210	2.11527	16.8905	.479636	.097114	4.93886	16.8905	.439477	.032408	13.5609	60.0571
.089885	.090648	.991588	16.3762	.333039	.100859	3.30202	16.3762	.146090	.037423	3.90371	57.9857
.796219	.039944	19.9334	41.4810	-.111795	.026728	-4.18275	200.495	.130245	.031910	4.08160	.0
Coeff. of determin. (R**2) .981008				Coeff. of determin. (R**2) .931348				Coeff. of determin. (R**2) .987414			
Standard error of equation 1.02554				Standard error of equation 1.00945				Standard error of equation .767149			
Durbin-Watson statistic 1.36747				Durbin-Watson statistic 1.81018				Durbin-Watson statistic 1.95843			

```
C ********************************************************************************************
C * Fortran code (part) for solution with Gauss-Seidel method                                *
C * (with simple changes, it could be used also for Jacobi method)                           *
C                                                                                            *
C                   List of endogenous variables                                             *
C  Y(1,t) = C       Consumption.                                                             *
C  Y(2,t) = I       Investment.                                                              *
C  Y(3,t) = Wp      Private wages.                                                           *
C  Y(4,t) = X       Equilibrium demand (=National product + T)                               *
C  Y(5,t) = P       Profits.                                                                 *
C  Y(6,t) = K       End-of-year Capital stock.                                               *
C  Y(7,t) = W       Total wages (=Wp+Wg)                                                     *
C                                                                                            *
C                   List of exogenous variables                                              *
C  Z(1,t) = 1.      Constant                                                                 *
C  Z(2,t) = Wg      Government wages.                                                        *
C  Z(3,t) = T       Business taxes.                                                          *
C  Z(4,t) = A       Proxy for bargaining power of labour = time trend (1931=0)               *
C  Z(5,t) = G       Government nonwage spending.                                             *
C                                                                                            *
C                   List of equations                                                        *
C     Consumption.                                                                           *
      Y(1,t) = A(1)*Z(1,t) + A(2)*Y(5,t) + A(3)*Y(5,t-1) + A(4)*Y(7,t) + U(1,t)
C                                                                                            *
C     Investment                                                                             *
      Y(2,t) = A(5)*Z(1,t) + A(6)*Y(5,t) + A(7)*Y(5,t-1) + A(8)*Y(6,t-1) + U(2,t)
C                                                                                            *
C     Private wages                                                                          *
      Y(3,t) = A(9)*Z(1,t) + A(10)*Y(4,t) + A(11)*Y(4,t-1) + A(12)*Z(4,t) + U(3,t)
C                                                                                            *
C     Equilibrium demand                                                                     *
      Y(4,t) = Y(1,t) + Y(2,t) + Z(5,t)
C                                                                                            *
C     Profits                                                                                *
      Y(5,t) = Y(4,t) - Z(3,t) - Y(3,t)
C                                                                                            *
C     Capital stock                                                                          *
      Y(6,t) = Y(6,t-1) + Y(2,t)
C                                                                                            *
C     Total wages                                                                            *
      Y(7,t) = Y(3,t) + Z(2,t)
C ********************************************************************************************
```

```
 7 Endogenous variables - 3 Stochastic equations - 5 Exogenous variables
 12 Estimated coefficients - 1920 1941 Time range - 1921 1941 Sample period

            ****************************************************************
            * One-Step-Ahead (Static) Solution - OLS structural coefficients *
            ****************************************************************
            Year 1921                                  Year 1922

Variable  Observed  Computed  % Error        Variable  Observed  Computed  % Error
Y(1)=C     41.9000   43.9284  4.84101        Y(1)=C     45.0000   48.1869  7.08189
Y(2)=I    -.200000  -.211785  5.89235        Y(2)=I     1.90000   3.33087  75.3092
Y(3)=Wp    25.5000   27.6804  8.55070        Y(3)=Wp    29.3000   31.0337  5.91713
Y(4)=X     45.6000   47.6166  4.42236        Y(4)=X     50.1000   54.7177  9.21702
Y(5)=P     12.4000   12.2362 -1.32121        Y(5)=P     16.9000   19.7840  17.0651
Y(6)=K     182.600   182.588 -.645383e-2     Y(6)=K     184.500   185.931  .775542
Y(7)=W     28.2000   30.3804  7.73202        Y(7)=W     32.2000   33.9337  5.38422

 ......................other years ......................

Output for Variable Y(1)=C for Years 1921 - 1941    Output for Variable Y(5)=P for Years 1921 - 1941

Year Observed Comput.  % Error  Observed Comput.    Year Observed Comput.  % Error  Observed Comput.
     Value    Value             %Change  %Change         Value    Value             %Change  %Change
1921 41.9000  43.9284  4.84101                      1921 12.4000  12.2362 -1.32121
1922 45.0000  48.1869  7.08189  7.39857  9.69411    1922 16.9000  19.7840  17.0651  36.2903  61.6846
1923 49.2000  50.3380  2.31309  9.33333  4.46427    1923 18.4000  19.9412  8.37594  8.87574  .794414
1924 50.6000  54.2978  7.30784  2.84553  7.86627    1924 19.4000  23.0849  18.9945  5.43478  15.7651
1925 52.6000  52.2601 -.646148  3.95257 -3.75271    1925 20.1000  18.8844 -6.04758  3.60825 -18.1958
1926 55.1000  50.6623 -8.05385  4.75285 -3.05739    1926 19.6000  14.3922 -26.5704 -2.48756 -23.7880
1927 56.2000  51.8835 -7.68067  1.99637  2.41034    1927 19.8000  14.8901 -24.7972  1.02041  3.45980
1928 57.3000  55.2600 -3.56019  1.95730  6.50794    1928 21.1000  20.4843 -2.91787  6.56566  37.5697
1929 57.8000  56.5899 -2.09352  .872600  2.40669    1929 21.7000  21.5775 -.564732  2.84360  5.33639
1930 55.0000  53.8983 -2.00304 -4.84429 -4.75636    1930 15.6000  14.3352 -8.10762 -28.1106 -33.5639
1931 50.9000  50.9713  .140127 -7.45455 -5.43060    1931 11.4000  12.2391  7.36033 -26.9231 -14.6223
1932 45.6000  45.7654  .362793 -10.4126 -10.2134    1932 7.00000  6.98673 -.189582 -38.5965 -42.9146
1933 46.5000  44.8969 -3.44754  1.97368 -1.89781    1933 11.2000  10.4154 -7.00577  60.0000  49.0734
1934 48.7000  48.9169  .445435  4.73118  8.95392    1934 12.3000  12.9839  5.55996  9.82143  24.6609
1935 51.3000  51.3647  .126210  5.33881  5.00403    1935 14.0000  14.0607  .433760  13.8211  8.29376
1936 57.7000  52.4316 -9.13068  12.4756  2.07701    1936 17.6000  11.6524 -33.7931  25.7143 -17.1280
1937 58.7000  58.9735  .465976  1.73310  12.4771    1937 17.3000  18.8319  8.85474 -1.70455  61.6135
1938 57.5000  61.6210  7.16703 -2.04429  4.48932    1938 15.3000  19.7851  29.3142 -11.5607  5.06162
1939 61.6000  60.4109 -1.93033  7.13043 -1.96382    1939 19.0000  18.0957 -4.75950  24.1830 -8.53863
1940 65.0000  65.0920  .141601  5.51948  7.74881    1940 21.1000  20.2771 -3.90017  11.0526  12.0546
1941 69.7000  76.1503  9.25439  7.23077  16.9887    1941 23.5000  29.7621  26.6471  11.3744  46.7770

......................other variables ...................

           RMSE          RMSE         MAPE     H.Theil: Inequality Coefficients
      (dimensionless)                          Applied Economic Forecasting (1966), p.59
                                                  Eq.(4.5)       Eq.(4.6)

Y(1)=C    .515210e-1     2.80319      3.72349      .964430        1.07951
Y(2)=I    .569947        2.10341
Y(3)=Wp   .561012e-1     2.06894      4.31788      .716739        .791919
Y(4)=X    .787624e-1     4.80013      5.46199      .997709        1.08002
Y(5)=P    .168088        2.92227      11.5514      1.08204        1.11706
Y(6)=K    .104148e-1     2.10341      .730163      .728021        .785039
Y(7)=W    .491070e-1     2.06894      3.81188      .705439        .811302
```

```
            ************************************************
            * Dynamic Solution - OLS structural coefficients *
            ************************************************
             Year 1921                                  Year 1922

Variable  Observed  Computed  % Error        Variable  Observed  Computed  % Error
Y(1)=C     41.9000   43.9284  4.84101        Y(1)=C     45.0000   48.2969  7.32655
Y(2)=I    -.200000  -.211785  5.89235        Y(2)=I     1.90000   3.10527  63.4355
Y(3)=Wp    25.5000   27.6804  8.55070        Y(3)=Wp    29.3000   31.2776  6.74936
Y(4)=X     45.6000   47.6166  4.42236        Y(4)=X     50.1000   54.6022  8.98647
Y(5)=P     12.4000   12.2362 -1.32121        Y(5)=P     16.9000   19.4247  14.9388
Y(6)=K     182.600   182.588 -.645383e-2     Y(6)=K     184.500   185.693  .646878
Y(7)=W     28.2000   30.3804  7.73202        Y(7)=W     32.2000   34.1776  6.14150

.....................other years .......................

Output for Variable Y(1)=C for Years 1921 - 1941    Output for Variable Y(5)=P for Years 1921 - 1941

Year Observed Comput.  % Error  Observed Comput.    Year Observed Comput.  % Error  Observed Comput.
     Value    Value             %Change  %Change         Value    Value             %Change  %Change
1921 41.9000  43.9284  4.84101                      1921 12.4000  12.2362 -1.32121
1922 45.0000  48.2969  7.32655  7.39857  9.94474    1922 16.9000  19.4247  14.9388  36.2903  58.7479
1923 49.2000  52.6653  7.04338  9.33333  9.04487    1923 18.4000  21.3681  16.1308  8.87574  10.0049
1924 50.6000  56.7956  12.2442  2.84553  7.84243    1924 19.4000  24.7105  27.3735  5.43478  15.6419
1925 52.6000  56.5272  7.46618  3.95257 -.472521    1925 20.1000  20.7666  3.31666  3.60825 -15.9601
1926 55.1000  50.3343 -8.64922  4.75285 -10.9557    1926 19.6000  12.6865 -35.2730 -2.48756 -38.9093
1927 56.2000  44.7342 -20.4017  1.99637 -11.1257    1927 19.8000  9.49425 -52.0493  1.02041 -25.1626
1928 57.3000  45.8225 -20.0305  1.95730  2.43285    1928 21.1000  15.0840 -28.5118  6.56566  58.8753
1929 57.8000  51.9065 -10.1963  .872600  13.2773    1929 21.7000  20.6943 -4.63478  2.84360  37.1933
1930 55.0000  54.6348 -.663984 -4.84429  5.25615    1930 15.6000  17.4354  11.7655 -28.1106 -15.7476
1931 50.9000  54.7874  7.63742 -7.45455  .279377    1931 11.4000  16.3514  43.4330 -26.9231 -6.21752
1932 45.6000  52.0730  14.1951 -10.4126 -4.95458    1932 7.00000  12.0939  72.7697 -38.5965 -26.0375
1933 46.5000  50.8066  9.26144  1.97368 -2.43195    1933 11.2000  14.2868  27.5607  60.0000  18.1324
1934 48.7000  52.2007  7.18824  4.73118  2.74394    1934 12.3000  14.7384  19.8247  9.82143  3.16130
1935 51.3000  53.4870  4.26324  5.33881  2.46428    1935 14.0000  14.9109  6.50634  13.8211  1.17003
1936 57.7000  52.8380 -8.42628  12.4756 -1.21340    1936 17.6000  11.2578 -36.0355  25.7143 -24.4997
1937 58.7000  52.9224 -9.84254  1.73310  .159720    1937 17.3000  14.4063 -16.7265 -1.70455  27.9679
1938 57.5000  58.9481  2.51836 -2.04429  11.3858    1938 15.3000  19.1891  25.4189 -11.5607  33.1992
1939 61.6000  64.1598  4.15560  7.13043  8.84133    1939 19.0000  20.8954  9.97560  24.1830  8.89185
1940 65.0000  66.7163  2.64050  5.51948  3.98454    1940 21.1000  20.6711 -2.03266  11.0526 -1.07323
1941 69.7000  75.4129  8.19646  7.23077  13.0352    1941 23.5000  28.2460  20.1958  11.3744  36.6449

.....................other variables ...................

            RMSE          RMSE         MAPE     H.Theil: Inequality Coefficients
       (dimensionless)                              Eq.(4.5)       Eq.(4.6)

Y(1)=C    .978663e-1     5.32480      8.43754      1.30720        1.46318
Y(2)=I    .974583        3.59673
Y(3)=Wp   .130368        4.80780      11.3273      1.12286        1.24064
Y(4)=X    .143506        8.74590      12.7101      1.27135        1.37625
Y(5)=P    .249534        4.33823      22.6569      1.26513        1.30607
Y(6)=K    .295700e-1     5.97202      2.22084      .963788        1.03927
Y(7)=W    .114115        4.80780      9.99802      1.12412        1.29282
```

Impact and Delay (Interim) Multipliers computed from OLS structural coefficients

	Year 1938		from Year 1938 to 1939		from Year 1938 to 1940		from Year 1938 to 1941	
Variable	Multiplier	Elasticity	Multiplier	Elasticity	Multiplier	Elasticity	Multiplier	Elasticity
Exogenous Z(2)=Wg			..... Dynamic solution					
Y(1)=C	2.13175	.266378	1.50454	.180569	.705218	.833886e-1	-.124064	-.130379e-1
Y(2)=I	.783850	2.14272	.898356	2.03129	.191302	.558909	-.349000	-.473862
Y(3)=Wp	1.28134	.231860	1.48196	.251788	.745038	.123286	-.769281e-1	-.108615e-1
Y(4)=X	2.91560	.321922	2.40289	.249480	.896520	.918535e-1	-.473064	-.392769e-1
Y(5)=P	1.63426	.636025	.920937	.355570	.151481	.613186e-1	-.396136	-.114650
Y(6)=K	.783850	.294973e-1	1.68221	.622673e-1	1.87351	.684808e-1	1.52451	.542633e-1
Y(7)=W	2.28134	.349559	1.48196	.214816	.745038	.105200	-.769281e-1	-.939688e-2
Exogenous Z(3)=T			..... Dynamic solution					
Y(1)=C	-1.32106	-.158645	-1.98022	-.228400	-1.15120	-.130820	.644746e-1	.651168e-2
Y(2)=I	-1.14176	-2.99949	-1.40183	-3.04620	-.409872	-1.15083	.451388	.589003
Y(3)=Wp	-1.08235	-.188223	-1.84613	-.301441	-1.18014	-.187676	-.134738e-2	-.182825e-3
Y(4)=X	-2.46282	-.261334	-3.38205	-.337460	-1.56107	-.153709	.515862	.411616e-1
Y(5)=P	-2.38047	-.890341	-1.53592	-.569909	-.380933	-.148191	.517210	.143860
Y(6)=K	-1.14176	-.412918e-1	-2.54359	-.904833e-1	-2.95346	-.103749	-2.50207	-.855888e-1
Y(7)=W	-1.08235	-.159383	-1.84613	-.257179	-1.18014	-.160144	-.134738e-2	-.158172e-3
Exogenous Z(5)=G			..... Dynamic solution					
Y(1)=C	1.67734	.144267	1.88960	.156098	.885708	.720874e-1	-.155816	-.112710e-1
Y(2)=I	.984465	1.85233	1.12828	1.75600	.240263	.483163	-.438321	-.409642
Y(3)=Wp	1.60928	.200438	1.86124	.217665	.935720	.106578	-.966167e-1	-.938948e-2
Y(4)=X	3.66181	.278293	3.01788	.215670	1.12597	.794051e-1	-.594138	-.339539e-1
Y(5)=P	2.05253	.549829	1.15664	.307381	.190251	.530085e-1	-.497521	-.991123e-1
Y(6)=K	.984465	.254997e-1	2.11274	.538286e-1	2.35301	.592000e-1	1.91468	.469093e-1
Y(7)=W	1.60928	.169726	1.86124	.185703	.935720	.909426e-1	-.966167e-1	-.812338e-2

Dynamic reduced form matrix. Modulus of Eigenvalue Period

Eig(1)=	0.788362	11.7601
Eig(2)=	0.788362	-11.7601
Eig(3)=	0.355372	

Alternative Estimates of structural form coefficients

OLS

1	P	P-1	W
16.2366	.192934	.089885	.796219

1	P	P-1	K-1
10.1258	.479636	.333039	-.111795

1	X	X-1	A
1.49704	.439477	.146090	.130245

2SLS

1	P	P-1	W
16.5548	.017302	.216234	.810183

1	P	P-1	K-1
20.2782	.150222	.615944	-.157788

1	X	X-1	A
1.50030	.438859	.146674	.130396

LIVE (2 iter. from OLS)

1	P	P-1	W
16.8014	-.115631	.312132	.820507

1	P	P-1	K-1
21.6005	.107318	.652790	-.163778

1	X	X-1	A
1.60111	.419711	.164768	.135058

I.I.V. (10 iter. from OLS)

1	P	P-1	W
16.7858	-.117503	.312609	.821456

1	P	P-1	K-1
21.6064	.107126	.652955	-.163805

1	X	X-1	A
1.59635	.420614	.163914	.134838

3SLS

1	P	P-1	W
16.4408	.124890	.163144	.790081

1	P	P-1	K-1
28.1778	-.013079	.755724	-.194848

1	X	X-1	A
1.79722	.400492	.181291	.149674

Iterative 3SLS

1	P	P-1	W
16.5590	.164510	.176564	.765801

1	P	P-1	K-1
42.8963	-.356532	1.01130	-.260200

1	X	X-1	A
2.62477	.374779	.193651	.167926

FIVE (2 iter. from OLS)

1	P	P-1	W
16.4570	.091267	.198055	.789598

1	P	P-1	K-1
24.7860	.029683	.717039	-.178374

1	X	X-1	A
1.93827	.383522	.196435	.157667

FIML

1	P	P-1	W
18.3433	-.232389	.385673	.801844

1	P	P-1	K-1
27.2639	-.801006	1.05185	-.148099

1	X	X-1	A
5.79429	.234118	.284677	.234835

Variance-covariance matrices estimated from structural form residuals (without degrees of freedom correction)

OLS

.851402		
.049497	.824891	
-.380815	.121170	.476417

2SLS

1.04406		
.437848	1.38318	
-.385228	.192606	.476427

LIVE (2 iter. from OLS)

1.44650		
.726600	1.53808	
-.339635	.273549	.486842

I.I.V. (10 iter. from OLS)

1.45233		
.729292	1.53882	
-.341982	.270234	.485911

3SLS

.891760		
.411319	2.09305	
-.393615	.403046	.520027

Iterative 3SLS

.914909		
.641739	4.55536	
-.434985	.734498	.605649

FIVE (2 iter. from OLS)

.937914		
.442160	1.86804	
-.379888	.452247	.565780

FIML

2.10415		
3.87902	12.7715	
.481696	3.85748	1.80112

Identification: Matrices of ‘‘Exclusion Restrictions’’ for the three stochastic equations

```
    Phi1(15,10)           Phi2(15,10)           Phi3(15,10)
0 0 0 0 0 0 0 0 0 0   1 0 0 0 0 0 0 0 0 0 0   1 0 0 0 0 0 0 0 0 0 0
1 0 0 0 0 0 0 0 0 0   0 0 0 0 0 0 0 0 0 0 0   0 1 0 0 0 0 0 0 0 0 0
0 1 0 0 0 0 0 0 0 0   0 1 0 0 0 0 0 0 0 0 0   0 0 0 0 0 0 0 0 0 0 0
0 0 1 0 0 0 0 0 0 0   0 0 1 0 0 0 0 0 0 0 0   0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0   0 0 0 0 0 0 0 0 0 0 0   0 0 1 0 0 0 0 0 0 0 0
0 0 0 1 0 0 0 0 0 0   0 0 0 1 0 0 0 0 0 0 0   0 0 0 1 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0   0 0 0 0 1 0 0 0 0 0 0   0 0 0 0 1 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0   0 0 0 0 0 0 0 0 0 0 0   0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 1 0 0 0 0 0   0 0 0 0 0 1 0 0 0 0 0   0 0 0 0 0 1 0 0 0 0 0
0 0 0 0 0 1 0 0 0 0   0 0 0 0 0 0 1 0 0 0 0   0 0 0 0 0 0 1 0 0 0 0
0 0 0 0 0 0 1 0 0 0   0 0 0 0 0 0 0 1 0 0 0   0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 1 0 0   0 0 0 0 0 0 0 0 1 0 0   0 0 0 0 0 0 0 0 1 0 0
0 0 0 0 0 0 0 0 1 0   0 0 0 0 0 0 0 0 0 1 0   0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0   0 0 0 0 0 0 0 0 0 0 0   0 0 0 0 0 0 0 0 0 1 0
0 0 0 0 0 0 0 0 0 1   0 0 0 0 0 0 0 0 0 0 0   0 0 0 0 0 0 0 0 0 0 1
```

A non-linear version of the Klein-I model (Belsley, 1980): the first structural equation (consumption) is replaced by

$$lnC_t = \alpha_1 + \alpha_2\ lnP_t + \alpha_3\ lnP_{t-1} + \alpha_4\ lnW_t + u_{1,t} \qquad \text{Consumption} \tag{14.70}$$

while all the other equations are the same as (2.1).

Alternative Estimates of structural form coefficients

OLS				I.I.V. (9 iter. from OLS)				NL-FIML			
1	P	P-1	W	1	P	P-1	W	1	P	P-1	W
1.42867	.054133	.017128	.634552	1.43554	.002594	.051195	.646111	1.42365	.048579	.031093	.629689
1	P	P-1	K-1	1	P	P-1	K-1	1	P	P-1	K-1
10.1258	.479636	.333039	-.111795	22.2134	.087431	.669869	-.166555	33.0054	-.248046	.884421	-.209644
1	X	X-1	A	1	X	X-1	A	1	X	X-1	A
1.49704	.439477	.146090	.130245	1.58788	.422223	.162394	.134446	2.85672	.331213	.234773	.163017

```
            ****************************************************************
            * One-Step-Ahead (Static) Solution - OLS structural coefficients *
            ****************************************************************
```

	RMSE (dimensionless)	RMSE	MAPE	H.Theil: Inequality Coefficients Eq.(4.5)	Eq.(4.6)
Y(1)=C	.466356e-1	2.53716	3.67048	.937936	1.04986
Y(2)=I	.559567	2.06510			
Y(3)=Wp	.532651e-1	1.96435	4.34570	.710552	.785083
Y(4)=X	.742997e-1	4.52815	5.56905	.993299	1.07525
Y(5)=P	.159150	2.76688	12.9225	1.10505	1.14081
Y(6)=K	.102252e-1	2.06510	.771039	.722178	.778739
Y(7)=W	.466245e-1	1.96435	3.82079	.700030	.805081
Y(8)=lnC	.115287e-1	.045921	.925162	.911714	1.00951

```
            ************************************************
            * Dynamic Solution - OLS structural coefficients *
            ************************************************
```

	RMSE (dimensionless)	RMSE	MAPE	H.Theil: Inequality Coefficients Eq.(4.5)	Eq.(4.6)
Y(1)=C	.978162e-1	5.32206	8.24905	1.32330	1.48121
Y(2)=I	.976774	3.60481			
Y(3)=Wp	.131097	4.83458	11.1280	1.14688	1.26718
Y(4)=X	.143958	8.77346	12.6106	1.29005	1.39649
Y(5)=P	.250216	4.35009	24.0321	1.29237	1.33419
Y(6)=K	.314611e-1	6.35397	2.35760	.964979	1.04057
Y(7)=W	.114750	4.83458	9.77070	1.14692	1.31904
Y(8)=lnC	.258048e-1	.102785	2.09962	1.33017	1.47285

15 APPENDIX. Multivariate normal distribution (with explict p.d.f.)

1. Transformation of univariate random variables: if the random variable x, with probability density function $f(x)$, is transformed into the random variable $y = y(x)$, if the transformation is continuously differentiable and invertible $[x = x(y)]$, then the p.d.f. of y is $g(y) = f[x(y)]\ |dx(y)/dy|$.

2. Transformation of multivariate random variables: if the $(k \times 1)$ random vector x, with joint p.d.f. $f(x)$, is transformed into the $(k \times 1)$ random vector $y = y(x)$, if the transformation is continuously differentiable and invertible $[x = x(y)]$ with non-singular square Jacobian matrix $\partial x(y)/\partial y'$, then the joint p.d.f. of y is $g(y) = f[x(y)]\ \|\partial x(y)/\partial y'\|$ (where $\|\cdots\|$ means absolute value of determinant).

3. Let μ be a $(k \times 1)$ constant vector and Σ a symmetric $(k \times k)$ positive definite constant matrix, that can be decomposed as $\Sigma = P'P$, with P a non-singular $(k \times k)$ constant matrix. The determinant of P is the square root of the determinant of Σ: $|P| = |\Sigma|^{1/2}$. Let z be a $(k \times 1)$ random vector whose expectation is zero and the variance-covariance matrix is the identity matrix: $E(z) = 0$, $Var(z) = I_k$. Whatever the probability distribution of z, the $(k \times 1)$ random vector $x = P'z + \mu$ has expectation μ and variance-covariance matrix $\Sigma = P'P$. Being P non-singular, the transformation from z to x is continuously differentiable and invertible $[z = P'^{-1}(x - \mu)]$ with non-singular square Jacobian matrix $\partial z(x)/\partial x' = P'^{-1}$.

4. If the elements of the $(k \times 1)$ random vector z are independent standard normal variables [z_i are *i.i.d.* $N(0,1)$], we say that the vector z has a multivariate normal distribution: $z \sim N(0, I_k)$. The joint p.d.f. of the elements of z is the product of the univariate density functions: $f(z) = \prod_{i=1}^{k} \frac{1}{(2\pi)^{1/2}} exp\left[-\frac{z_i^2}{2}\right] = \frac{1}{(2\pi)^{k/2}} exp\left[-\frac{1}{2}\sum_{i=1}^{k} z_i^2\right]$. It is positive everywhere, and its integral is 1, being the product of k integrals, each $= 1$. With vector notation, the same p.d.f. can be written $f(z) = \frac{1}{(2\pi)^{k/2}} exp\left[-\frac{1}{2}z'z\right]$.

5. Using the result on the transformation of random vectors (2), the $(k \times 1)$ random vector $x = P'z + \mu$ has p.d.f. $f(x) = \frac{1}{(2\pi)^{k/2}\|P\|} exp\left[-\frac{1}{2}(x-\mu)'P^{-1}P'^{-1}(x-\mu)\right] = \frac{1}{(2\pi)^{k/2}\|\Sigma\|^{1/2}} exp\left[-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right]$. We say that the $(k \times 1)$ random vector x has a multivariate normal distribution $N(\mu, \Sigma)$. Using the result in (3), μ is the expectation and Σ is the variance-covariance matrix of x.

   Warning. This definition requires Σ to be positive definite. It is possible to define multivariate normal distributions also when Σ is positive semi definite and singular, but these distributions do not admit an explicit p.d.f.

6. If $x \sim N(\mu, \Sigma)$ and Σ is block-diagonal, the corresponding sub-vectors of x are independent multivariate normal vectors. Thus, uncorrelated sub-vectors are independent; moreover, each sub-vector has a marginal distribution multivariate normal. The proof follows considering that Σ^{-1} is also block-diagonal, and $|\Sigma|$ is obtained multiplying the determinants of the diagonal blocks. Thus, the joint p.d.f. $f(x)$ is the product of functions that are exactly the marginal densities of the sub-vectors. In particular, if Σ is diagonal, all the elements of x are independent normal variables.

   Warning. The above properties depend on the *joint* density of the elements of x being normal. If it is only known that the *marginal* densities of the elements are normal, then the joint density needs not be normal and may even not exist. Thus, uncorrelated normal variables need not be independent if they are not *jointly* multivariate normal.

7. If the $(k \times 1)$ random vector $x \sim N(\mu, \Sigma)$ and A is a $(k \times k)$ non-singular constant matrix, the $(k \times 1)$ random vector $y = Ax$ has a multivariate normal distribution $y \sim N(A\mu, A\Sigma A')$. The proof is simply based on the explicit p.d.f. expression of the transformed vector y, considering that A is the Jacobian matrix of the linear transformation.

8. Particular case: decomposing the positive definite matrix $\Sigma = P'P$, with P square and non-singular, the linear transformation $z = P'^{-1}(x - \mu) \sim N(0, I_k)$ (vector of independent standard normal variables).

9. If the $(k \times 1)$ random vector $x \sim N(\mu, \Sigma)$, any sub-vector of x has a marginal distribution multivariate normal, with means, variances and covariances obtained by taking the corresponding elements of μ and Σ. To prove it, let x be *arbitrarily* decomposed into two sub-vectors x_1 and x_2; let μ be correspondingly decomposed into μ_1 and μ_2; let Σ be correspondingly decomposed into $\Sigma_{1,1}$, $\Sigma_{1,2}$, $\Sigma_{2,1} = \Sigma'_{1,2}$, $\Sigma_{2,2}$ (where $\Sigma_{1,1}$ and $\Sigma_{2,2}$ are square blocks), and A a $(k \times k)$ matrix, correspondingly decomposed into blocks $A_{1,1} = I$, $A_{1,2} = -\Sigma_{1,2}\Sigma_{2,2}^{-1}$, $A_{2,1} = 0$, $A_{2,2} = I$. Thus, $A\mu$ has two sub-vectors $\mu_1 - \Sigma_{1,2}\Sigma_{2,2}^{-1}\mu_2$ and μ_2, while $A\Sigma A'$ is block-diagonal, the two square diagonal blocks being $\Sigma_{1,1} - \Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1}$ and $\Sigma_{2,2}$. The linear transformation $y = Ax$ produces the multivariate normal vector $y \sim N(A\mu, A\Sigma A')$ whose variance-covariance matrix is block-diagonal. Thus the two sub-vectors: y_1 and y_2 are independent multivariate normal vectors: $y_1 \sim N[\mu_1 - \Sigma_{1,2}\Sigma_{2,2}^{-1}\mu_2,\ \Sigma_{1,1} - \Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1}]$ and $y_2 \sim N[\mu_2, \Sigma_{2,2}]$. But $x_2 = y_2$; thus the *arbitrary* sub-vector x_2 has the multivariate normal distribution $N(\mu_2, \Sigma_{2,2})$.

10. Particular case: any element of a multivariate normal vector has univariate normal distribution, whose mean and variance are, respectively, the corresponding element of μ and the corresponding diagonal element of Σ.

11. If the $(k \times 1)$ random vector $x \sim N(\mu, \Sigma)$ and D is a $(p \times k)$ constant matrix of rank $p \leq k$, then the $(p \times 1)$ vector $y = Dx \sim N(D\mu, D\Sigma D')$. To prove it, one should add $k - p$ rows to the matrix D, producing a full-rank $(k \times k)$ matrix. Multiplying such a matrix by x produces a multivariate normal vector, whose first sub-vector is Dx.

Warning. The result (11) is a particular case of a more general result, that holds also when $p > k$ or when the rank of D is $< p$, and that we state without proof. *Any linear transformation of a multivariate normal is a multivariate normal*. This property, however, requires to deal also with multivariate normal distributions that do not admit an explicit p.d.f., due to singularity of the variance-covariance matrix (e.g. Rao, 1973, 8a).

12. The conditional distribution of $x_1|x_2$ is multivariate normal $x_1|x_2 \sim N[\mu_1 + \Sigma_{1,2}\Sigma_{2,2}^{-1}(x_2 - \mu_2),\ \Sigma_{1,1} - \Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1}]$. From (9) it follows that, being the Jacobian of the linear transformation $|A| = 1$, the expression of the p.d.f. of x is equal to the expression of the p.d.f. of y, when y in the expression is replaced with Ax. Thus $f(x) = g(y)$, thus $f(x_2)f(x_1|x_2) = g(y_2)g(y_1|y_2) = g(y_2)g(y_1) = f(x_2)g(y_1)$, being $x_2 = y_2$ and being y_1 and y_2 independent. Thus $f(x_1|x_2) = g(y_1)$, when y_1 in the expression of g is replaced with the first sub-vector of Ax. Writing explicitly the expression of the p.d.f. of $y_1 \sim N[\mu_1 - \Sigma_{1,2}\Sigma_{2,2}^{-1}\mu_2,\ \Sigma_{1,1} - \Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1}]$, and replacing in the expression of the p.d.f. y_1 with $x_1 - \Sigma_{1,2}\Sigma_{2,2}^{-1}x_2$ (first sub-vector of Ax), produces the explicit expression of the p.d.f. of a multivariate normal with mean $\mu_1 + \Sigma_{1,2}\Sigma_{2,2}^{-1}(x_2 - \mu_2)$ and variance-covariance matrix $\Sigma_{1,1} - \Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma_{2,1}$.

Remark. The conditional mean of $x_1|x_2$ is a linear function of x_2; the conditional variance is independent of x_2.

16 APPENDIX. Some useful asymptotic results

1. If *plim* $\hat{\theta}_n = \theta_0$ and g is a continuous function, then *plim* $g(\hat{\theta}_n) = g(\theta_0)$. This result extends to continuous functions a result proved by Slutsky (1925) for rational functions (e.g. Rao, 1973, 2c.4.xiii). It holds either in univariate or in multivariate cases.

2. *δ-method* (univariate case e.g. Rao, 1973, 6a.2.i): if $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically $\sim N[0, \sigma^2]$, and g is a continuously differentiable function with nonzero first derivative in θ_0, then $\sqrt{n}[g(\hat{\theta}_n) - g(\theta_0)]$ is asymptotically $\sim N[0, g'^2(\theta_0)\sigma^2]$. The proof follows from a first order Taylor expansion of $g(\hat{\theta}_n)$ with origin θ_0, recalling that the residual is $o_p(\hat{\theta}_n - \theta_0)$.

3. *δ-method* (multivariate case e.g. Rao, 1973, 6a.2.iii): if the vector $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically multivariate normal $\sim N[0, \Sigma]$, and g is a vector of continuously differentiable functions whose first derivatives are not all $= 0$ in θ_0, then $\sqrt{n}[g(\hat{\theta}_n) - g(\theta_0)]$ has an asymptotic multivariate normal distribution $\sim N[0, G\Sigma G']$, where G is the Jacobian matrix $\partial g/\partial\theta'$ computed in θ_0.

Let x_i be a random variable or vector (r.v.), whose probability density function (continuous) is characterized by a parameter (vector) θ: $f(x_i, \theta)$. For any θ, f is a function whose integral is $\equiv 1$.

$$\int_{-\infty}^{+\infty} f(x_i, \theta)\ dx_i \equiv 1 \qquad \forall\theta \tag{17.71}$$

Thus, differentiating (17.71) w.r.t. θ we get

$$\frac{\partial}{\partial\theta}\left[\int_{-\infty}^{+\infty} f(x_i, \theta)\ dx_i\right] \equiv 0 \qquad \forall\theta \tag{17.72}$$

We assume that f satisfies some *regularity conditions* that permit differentiation under integral (for instance, it is twice differentiable w.r.t. θ and the limits of integration are not functions of θ). So, (17.72) can be written

$$\int_{-\infty}^{+\infty} \frac{\partial f(x_i, \theta)}{\partial\theta}\ dx_i \equiv 0 \qquad \forall\theta \tag{17.73}$$

Integration will be confined to the region where f assumes nonzero (positive) values. Thus (17.73) can be written

$$\int \frac{\partial lnf(x_i, \theta)}{\partial\theta} f(x_i, \theta) dx_i \equiv 0 \qquad \forall\theta \tag{17.74}$$

Remark. The proofs of this chapter are based on a *double interpretation* of the function $f(x_i, \theta)$. It must be considered a *probability density function*, but at the same time, being a *transformation* of the r.v. x_i, it is a random variable itself, with expectation and variance. The same *double interpretation* holds for the logarithm of $f(x_i, \theta)$, as well as its derivatives.

The derivative (vector of derivatives) $\partial lnf(x_i, \theta)/\partial\theta$ (*gradient* of the log-density) is usually called the *score*. If derivative (vector) is computed at the *true* parameter value, so that $f(x_i, \theta)$ is the probability density of the r.v. x_i, equation (17.74) is the expectation of the r.v. *score*

$$E\left[\frac{\partial lnf(x_i, \theta)}{\partial\theta}\right] = \int \frac{\partial lnf(x_i, \theta)}{\partial\theta} f(x_i, \theta) dx_i = 0 \tag{17.75}$$

Thus the expectation of the score is zero. The variance-covariance matrix of the score, $\Im(\theta)$, is called *information matrix* (more precisely, Fisher's information measure on θ contained in the r.v. x_i)

$$\Im(\theta) = Var\left[\frac{\partial lnf(x_i, \theta)}{\partial\theta}\right] = E\left[\frac{\partial lnf(x_i, \theta)}{\partial\theta}\frac{\partial lnf(x_i, \theta)}{\partial\theta'}\right] \tag{17.76}$$

Further differentiation of (17.74) gives

$$\int \left[\frac{\partial^2 lnf(x_i, \theta)}{\partial\theta\partial\theta'} f(x_i, \theta) + \frac{\partial lnf(x_i, \theta)}{\partial\theta}\frac{\partial f(x_i, \theta)}{\partial\theta'}\right] dx_i \equiv 0 \qquad \forall\theta$$

that is

$$\int \frac{\partial^2 lnf(x_i, \theta)}{\partial\theta\partial\theta'} f(x_i, \theta) dx_i + \int \frac{\partial lnf(x_i, \theta)}{\partial\theta}\frac{\partial lnf(x_i, \theta)}{\partial\theta'} f(x_i, \theta) dx_i \equiv 0 \qquad \forall\theta \tag{17.77}$$

Again, if derivatives are computed at the *true* parameter value, so that $f(x_i, \theta)$ is the probability density of the r.v. x_i, the two terms in equation (17.77) are expectations, so

$$E\left[\frac{\partial^2 lnf(x_i, \theta)}{\partial\theta\partial\theta'}\right] + E\left[\frac{\partial lnf(x_i, \theta)}{\partial\theta}\frac{\partial lnf(x_i, \theta)}{\partial\theta'}\right] \equiv 0 \qquad \forall\theta \tag{17.78}$$

The second term of the sum is the information matrix (17.76). Thus, from (17.78) we get an alternative expression for the information matrix

$$\Im(\theta) = E\left[-\frac{\partial^2 lnf(x_i, \theta)}{\partial\theta\partial\theta'}\right] \tag{17.79}$$

that is the expected Hessian of the log-density, with the opposite sign.

If x_1, x_2, ..., x_n are independent draws from the same distribution (random sample), the joint density of the sample is $f(x_1, x_2, ..., x_n, \theta) = f(x_1, \theta)f(x_2, \theta)...f(x_n, \theta)$; to simplify notations, it will simply be indicated as $f(x, \theta)$. The log-density of the sample will be therefore the sum of the log-densities, while its first and second derivatives as well as their expectations will be sums of the corresponding derivatives or expectations. As a straightforward consequence, the expectation of the score of the sample will be zero, while the information in the whole sample will be $n\Im(\theta)$

$$E\left[\frac{\partial lnf(x, \theta)}{\partial\theta}\right] = 0$$

$$n\Im(\theta) = Var\left[\frac{\partial lnf(x,\theta)}{\partial\theta}\right] = E\left[\frac{\partial lnf(x,\theta)}{\partial\theta}\frac{\partial lnf(x,\theta)}{\partial\theta'}\right] = E\left[-\frac{\partial^2 lnf(x,\theta)}{\partial\theta\partial\theta'}\right] \tag{17.80}$$

Remark. If $f(x_i,\theta)$ (and therefore $f(x,\theta)$) is a family of strictly positive functions whose integral is identically = 1 for any θ, but for *no value* of θ it is the probability density function of the r.v. x_i, all the above identities involving integrals (eqs. 17.71 - 17.74 and 17.78) are still valid, but they *cannot* be interpreted as *expected values*. This remark will be important when dealing with *quasi*-likelihood (or *pseudo*-likelihood).

17.1 Cramér-Rao inequality

Let θ be a single parameter, x_1, x_2, ..., x_n independent draws (random sample; each x_i can be a single variable or a vector), and let $f(x,\theta)$ indicate the joint probability density of the whole sample. Let $t(x)$ be an estimator of the parameter θ (of course, t will be a function of the sample, and *not* of the parameter itself). Its expectation

$$E\left[t(x)\right] = \int t(x)f(x,\theta)dx \tag{17.81}$$

of course, will be a function of the parameter θ and *not* of the sample.
Under the usual regularity conditions, differentiating (17.81) we get

$$\frac{\partial E\left[t(x)\right]}{\partial\theta} = \int t(x)\frac{\partial f(x,\theta)}{\partial\theta}dx = \int t(x)\frac{\partial lnf(x,\theta)}{\partial\theta}f(x,\theta)dx = E\left[t(x)\frac{\partial lnf(x,\theta)}{\partial\theta}\right] = Cov\left[t(x),\frac{\partial lnf(x,\theta)}{\partial\theta}\right] \tag{17.82}$$

Since the squared covariance cannot exceed the product of the two variances, we have

$$\left[\frac{\partial E\left[t(x)\right]}{\partial\theta}\right]^2 \leq Var\left[t(x)\right]Var\left[\frac{\partial lnf(x,\theta)}{\partial\theta}\right] = Var\left[t(x)\right]\left[n\Im(\theta)\right] \tag{17.83}$$

If the expected value of the estimator is a *regular* function of θ

$$E\left[t(x)\right] = h(\theta) \qquad \text{so that} \qquad \frac{\partial E\left[t(x)\right]}{\partial\theta} = \frac{\partial h(\theta)}{\partial\theta}$$

the Cramér-Rao inequality follows from (17.83)

$$Var\left[t(x)\right] \geq \left[\frac{\partial h(\theta)}{\partial\theta}\right]^2 \left[n\Im(\theta)\right]^{-1} \tag{17.84}$$

In the particular case of an *unbiased* estimator, $E[t(x)] = h(\theta) = \theta$; thus $\partial h(\theta)/\partial\theta = 1$, thus the Cramér-Rao inequality becomes

$$Var\left[t(x)\right] \geq \left[n\Im(\theta)\right]^{-1} \tag{17.85}$$

An unbiased estimator is *efficient* if its variance is the lower bound of the inequality: $\left[n\Im(\theta)\right]^{-1}$.

17.1.1 Multidimensional Cramér-Rao inequality

When θ is a $(k\times 1)$ vector of parameters, analogously to (17.82) we have

$$\frac{\partial E\left[t(x)\right]}{\partial\theta'} = E\left[t(x)\frac{\partial lnf(x,\theta)}{\partial\theta'}\right] = Cov\left[t(x),\frac{\partial lnf(x,\theta)}{\partial\theta'}\right]$$

thus we can write as follows the variance-covariance matrix of the $(2k\times 1)$ vector

$$Var\begin{bmatrix} t(x) \\ \frac{\partial lnf(x,\theta)}{\partial\theta} \end{bmatrix} = \begin{bmatrix} Var[t(x)] & Cov\left[t(x),\frac{\partial lnf(x,\theta)}{\partial\theta'}\right] \\ Cov\left[t(x)',\frac{\partial lnf(x,\theta)}{\partial\theta}\right] & n\Im(\theta) \end{bmatrix} = \begin{bmatrix} Var[t(x)] & \frac{\partial E[t(x)]}{\partial\theta'} \\ \frac{\partial E\left[t(x)'\right]}{\partial\theta} & n\Im(\theta) \end{bmatrix} \tag{17.86}$$

which is positive semi definite, being a variance-covariance matrix $(2k\times 2k)$. Thus, pre- and post-multiplication by a matrix and its transpose still provides a positive semi definite matrix. In particular, if the information matrix is non-singular (i.e. the derivatives of the log-density are not linearly dependent), pre-multiplication by the $(k\times 2k)$ matrix $\left[I_k\ ;\ -\frac{\partial E[t(x)]}{\partial\theta'}[n\Im(\theta)]^{-1}\right]$ and post-multiplication by its transpose produces

$$\left[I_k\ ;\ -\frac{\partial E[t(x)]}{\partial\theta'}[n\Im(\theta)]^{-1}\right]\begin{bmatrix} Var[t(x)] & \frac{\partial E[t(x)]}{\partial\theta'} \\ \frac{\partial E\left[t(x)'\right]}{\partial\theta} & n\Im(\theta) \end{bmatrix}\begin{bmatrix} I_k \\ -[n\Im(\theta)]^{-1}\frac{\partial E[t(x)']}{\partial\theta} \end{bmatrix} = Var\left[t(x)\right] - \left[\frac{\partial E\left[t(x)\right]}{\partial\theta'}\right]\left[n\Im(\theta)\right]^{-1}\left[\frac{\partial E\left[t(x)'\right]}{\partial\theta}\right]$$

which is a positive semi definite matrix, implying the Cramér-Rao inequality

$$Var\left[t(x)\right] \geq \left[\frac{\partial E\left[t(x)\right]}{\partial\theta'}\right]\left[n\Im(\theta)\right]^{-1}\left[\frac{\partial E\left[t(x)'\right]}{\partial\theta}\right] \tag{17.87}$$

where Var must be interpreted as the variance-covariance matrix of the estimator $t(x)$. The inequality is valid if the estimator is biased, but its expected value is a regular function of θ (analogously to eq.17.84).
For unbiased estimators, where $E[t(x)] = \theta$, the inequality still has the form of equation (17.85),

17.2 Maximum Likelihood estimator: consistency (simplified proof)

When the sample x_1, x_2, ..., x_n is observed, the function of θ defined by $L(x,\theta) = f(x,\theta)$ is called the *likelihood* of θ given the observations.

Remark. Textbooks in probability and statistics often adopt two different notations: $f(x|\theta) = f(x_1,\theta)f(x_2,\theta)...f(x_n,\theta)$ when considering the probability density function, to put into evidence that the r.v. is x, while θ is a given parameter (vector); $L(\theta|x)$ when considering the likelihood, thus evidencing a function of θ, while x is a realized value of a set of observations. Of course they clearly specify that $L(\theta|x) = f(x|\theta)$. These different notations are unnecessary for the purposes of this chapter, which thus adopts standard *mathematical notations* ($f(x,\theta)$ or $L(x,\theta)$) to indicate functions of x and θ.

Under suitable regularity conditions, maximum likelihood yields an estimator which is consistent, asymptitically normal with mean equal to the *true* parameter value and variance-covariance matrix equal to the inverse of the information matrix.
To simplify the proof, we consider θ a single parameter. The sample x_1, x_2, ..., x_n, ... is made of independent draws from the same population; the density of each x_i belongs to the *family* of density functions $f(x_i,\theta)$ for a particular value $\theta = \theta_0$; θ_0 can be called the *true* parameter value.
Regularity conditions are requested to ensure that

1. differentiation can be done under integral for any θ belonging to an interval (or a compact set) that contains the *true* θ_0 as an interior point;
2. the score, evaluated at θ_0, has positive finite variance $\Im(\theta_0)$;
3. the residual of a first order Taylor expansion of the score is bounded by a function of x_i with finite expectation (for instance, this condition could be satisfied assuming boundedness of the third derivative of the log-likelihood).

Applying first order Taylor expansion to the score, with initial point θ_0 we get

$$\frac{\partial lnf(x_i,\theta)}{\partial\theta} = \left[\frac{\partial lnf(x_i,\theta)}{\partial\theta}\right]_{\theta_0} + \left[\frac{\partial^2 lnf(x_i,\theta)}{\partial\theta^2}\right]_{\theta_0}(\theta-\theta_0) + Res(x_i,\theta,\theta_0)$$

Summing for $i = 1, 2, ..., n$ and dividing by n (averaging) we get

$$\frac{1}{n}\frac{\partial lnL(x,\theta)}{\partial\theta} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial lnf(x_i,\theta)}{\partial\theta} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial lnf(x_i,\theta)}{\partial\theta}\right]_{\theta_0} + \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial^2 lnf(x_i,\theta)}{\partial\theta^2}\right]_{\theta_0}(\theta-\theta_0) + \frac{1}{n}\sum_{i=1}^{n}Res(x_i,\theta,\theta_0) \quad (17.88)$$

Some *suitable* form of the weak law of large numbers (WLLN) ensures that

$$plim\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial lnf(x_i,\theta)}{\partial\theta}\right]_{\theta_0} = 0 \qquad \text{and} \qquad plim\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial^2 lnf(x_i,\theta)}{\partial\theta^2}\right]_{\theta_0} = -\Im(\theta_0) \quad (17.89)$$

thus, for a conveniently large n, the first term on the right hand side of (17.88) will be negligible, while the second term will be negative if $(\theta-\theta_0)$ is positive, and will be positive if $(\theta-\theta_0)$ is negative. Concerning the residual term, for *large* n and *small* $(\theta-\theta_0)$, regularity conditions and Taylor expansion properties ensure that its contribution is negligible with respect to the other terms. The consequence is that, when n is large enough, analysing an arbitrarily small interval around θ_0, the left hand side of (17.88) is positive on the left of θ_0, negative on the right: thus, arbitrarily close to θ_0 there is a point where the log-likelihood has a *local* maximum (and the score is zero). This value will be indicated with $\hat{\theta}$. It is called the *maximum likelihood estimator* of θ.

Remark. The (simplified) proof given above ensures the existence of a consistent root of the likelihood equation; it does not enable us to identify it (for instance, in case of multiple roots). It can be shown that the consistent root corresponds to the supremum of the likelihood with probability 1 (see Rao, 1973, 5f.2, who refers to papers by Wald, 1949, Le Cam, 1953, 1956, and Bahadur, 1958).

17.3 Maximum Likelihood estimator: asymptotic normality

Considering again θ a vector of parameters, if (17.88) is computed at $\hat{\theta}$, the left hand side is zero. Multiplying by $\sqrt{n}$ we get

$$\sqrt{n}(\hat{\theta}-\theta_0) = \left[-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 lnf(x_i,\theta)}{\partial\theta\partial\theta'}\right]^{-1}_{\theta_0}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{\partial lnf(x_i,\theta)}{\partial\theta}\right]_{\theta_0} + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}Res(x_i,\hat{\theta},\theta_0)\right\} \quad (17.90)$$

When $n \to \infty$ (and therefore $\hat{\theta} \to \theta_0$) still the contribution of the residual term becomes negligible. Concerning the term with second order derivatives, it converges in probability to the information matrix $\Im(\theta_0)$ (17.89), while some suitable form of the Central Limit Theorem (CLT) ensures that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{\partial lnf(x_i,\theta)}{\partial\theta}\right]_{\theta_0} \xrightarrow[n\to\infty]{distr} N\left[0,\Im(\theta_0)\right] \quad (17.91)$$

thus, from (17.90) and (17.91) we get

$$\sqrt{n}(\hat{\theta}-\theta_0) \xrightarrow[n\to\infty]{distr} N\left[0,\Im(\theta_0)^{-1}\right] \quad (17.92)$$

17.4 Maximum Likelihood: estimation of the (asymptotic) variance-covariance matrix

The practical consequence of (17.92) is that, when n is large enough, $\sqrt{n}(\hat{\theta} - \theta_0)$ has approximately a normal distribution with zero mean and $\Im(\theta_0)^{-1}$ variance-covariance matrix. Thus $(\hat{\theta} - \theta_0)$ has approximately a normal distribution with zero mean and $\Im(\theta_0)^{-1}/n$ variance-covariance matrix, that is

$$\hat{\theta} \;\; approx. \;\sim N\left[\theta_0, \Im(\theta_0)^{-1}/n\right]$$

Practical estimation of the information matrix can be done in two different ways, using the sample analogues of the *expectations* on the right hand sides of (17.76) or (17.79): each expectation is replaced by the sample average, and derivatives are computed at $\hat{\theta}$

1. Hessian estimator of $\Im = \frac{1}{n}\sum_{i=1}^{n}\left[-\frac{\partial^2 lnf(x_i,\theta)}{\partial\theta\partial\theta'}\right]_{\hat{\theta}} = \frac{1}{n}\left[-\frac{\partial^2 lnL(x,\theta)}{\partial\theta\partial\theta'}\right]_{\hat{\theta}}$
2. Outer Product estimator of $\Im = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial lnf(x_i,\theta)}{\partial\theta}\frac{\partial lnf(x_i,\theta)}{\partial\theta'}\right]_{\hat{\theta}}$

As a consequence, also the practical estimation of the variance-covariance matrix of $\hat{\theta}$ can be done in two different ways: using the Hessian or using the Outer Product matrix

1. $\widehat{Var}(\hat{\theta}) = (H)^{-1} = \left[-\frac{\partial^2 lnL(x,\theta)}{\partial\theta\partial\theta'}\right]^{-1}_{\hat{\theta}}$
2. $\widehat{Var}(\hat{\theta}) = (OP)^{-1} = \left\{\sum_{i=1}^{n}\left[\frac{\partial lnf(x_i,\theta)}{\partial\theta}\frac{\partial lnf(x_i,\theta)}{\partial\theta'}\right]_{\hat{\theta}}\right\}^{-1}$

17.5 Misspecification, Pseudo-Likelihood and "sandwich matrix" (simplified proof)

The sample x_1, x_2, ..., x_n, ... is made of independent draws from the same population. *We believe* that the density of each x_i belongs to the *family* of density functions $f(x_i,\theta)$, so we build the *presumed* likelihood and maximize it, obtaining $\hat{\theta}$. However, it may happen that the density of the r.v. does not belong to to the *family* of density functions $f(x_i,\theta)$, so our *presumed* likelihood is in fact misspecified: it will be called *pseudo-likelihood* or *quasi-likelihood*, and the derivative of the logarithm will be called *pseudo-score*. The *integral* (17.74) is still zero, but for no value of θ such an integral can be considered the expectation of the pseudo-score, as $f(x_i,\theta)$ is not the density of x_i.
We assume that the *true* density of x_i belongs to a regular (but *unknown*) family of density functions $g(x_i,\theta)$, and we still call θ_0 the *true* parameter value. Unlike (17.75), the expectation of the pseudo-score in θ_0 can be nonzero

$$E\left[\frac{\partial lnf(x_i,\theta)}{\partial\theta}\right]_{\theta_0} = \int \left.\frac{\partial lnf(x_i,\theta)}{\partial\theta}\right|_{\theta_0} g(x_i,\theta_0)dx \neq 0 \tag{17.93}$$

It may happen that the expectation is zero if evaluated at a different value of the parameter (vector): such a value, $\theta^\star$, is called *pseudo-true-value*

$$E\left[\frac{\partial lnf(x_i,\theta)}{\partial\theta}\right]_{\theta^\star} = 0 \tag{17.94}$$

Performing analogously to (17.88) the Taylor expansion with initial value $\theta^\star$ rather than θ_0

$$\frac{1}{n}\frac{\partial lnL(x,\theta)}{\partial\theta} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial lnf(x_i,\theta)}{\partial\theta} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial lnf(x_i,\theta)}{\partial\theta}\right]_{\theta^\star} + \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial^2 lnf(x_i,\theta)}{\partial\theta\partial\theta'}\right]_{\theta^\star}(\theta-\theta^\star) + \frac{1}{n}\sum_{i=1}^{n}Res(x_i,\theta,\theta^\star) \tag{17.95}$$

and then computing it at $\hat{\theta}$, the left hand side is zero. Multiplying by $\sqrt{n}$ we get

$$\sqrt{n}(\hat{\theta}-\theta^\star) = \left[-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 lnf(x_i,\theta)}{\partial\theta\partial\theta'}\right]^{-1}_{\theta^\star}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{\partial lnf(x_i,\theta)}{\partial\theta}\right]_{\theta^\star} + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}Res(x_i,\hat{\theta},\theta^\star)\right\} \tag{17.96}$$

We still assume (without proof) that, under suitable regularity conditions, when $n\to\infty$ the contribution of the residual term becomes negligible. As in section (17.3), we first apply (some suitable form of) the Law of Large Numbers to the term with second order derivatives, whose probability limit will be a constant matrix (let's call A). Then we apply (some suitable form of) the Central Limit Theorem to the term with first order derivatives, each of which has expected value zero (according to 17.94), and call B the variance-covariance matrix of the asymptotic normal distribution

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{\partial lnf(x_i,\theta)}{\partial\theta}\right]_{\theta^\star} \;\; \underset{n\to\infty}{\overset{distr}{\longrightarrow}} \;\; N\left[0,B\right] \tag{17.97}$$

thus

$$\sqrt{n}(\hat{\theta}-\theta^\star) \;\; \underset{n\to\infty}{\overset{distr}{\longrightarrow}} \;\; N\left[0,A^{-1}BA^{-1}\right]$$

While in section (17.3) A and B were both equal to the information matrix, here they can be different. The asymptotic variance-covariance matrix $A^{-1}BA^{-1}$ can be called *sandwich* matrix. An obvious way of estimating A and B is to use the Hessian and the matrix of outer products, respectively. Thus, practical estimation of the variance-covariance matrix of $\hat{\theta}$ can be done as

$$\widehat{Var}(\hat{\theta}) = H^{-1}\ OP\ H^{-1} \tag{17.98}$$

with the same expressions for H and OP as in section (17.4). This expression provides a *robust* estimator of the variance-covariance matrix of the pseudo-maximum-likelihood parameters $\hat{\theta}$.

17.6 Linear regression model with normal error terms: maximum likelihood and ordinary least squares

With the usual symbols, let the model and the vector of parameters be

$$y = X\beta + u \qquad u \sim N(0, \sigma^2 I_n) \qquad \theta = \begin{bmatrix} \beta \\ \sigma^2 \end{bmatrix}$$

Being the Jacobian of the transformation $\partial u_i/\partial y_i = 1$ (so that the density $f(y_i, \theta) = f(u_i, \theta)$) the log-likelihood is

$$lnL(y, \theta) = -\frac{n}{2} ln(2\pi) - \frac{n}{2} ln(\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)$$

the score is

$$\frac{\partial lnL(y,\theta)}{\partial \theta} = \begin{bmatrix} \frac{\partial lnL}{\partial \beta} \\ \frac{\partial lnL}{\partial \sigma^2} \end{bmatrix} = \begin{bmatrix} -\frac{1}{\sigma^2}(X'X\beta - X'y) \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(y - X\beta)'(y - X\beta) \end{bmatrix}$$

the Hessian matrix is

$$H(\theta) = \frac{\partial^2 lnL(y,\theta)}{\partial\theta\partial\theta'} = \begin{bmatrix} -\frac{1}{\sigma^2}X'X & \frac{1}{\sigma^4}(X'X\beta - X'y) \\ \frac{1}{\sigma^4}(\beta'X'X - y'X) & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}(y - X\beta)'(y - X\beta) \end{bmatrix} \tag{17.99}$$

The expectation of the off-diagonal blocks of the Hessian matrix is zero, and the expectation of the last block is $n/(2\sigma^4) - (1/\sigma^6)n\sigma^2 = -n/(2\sigma^4)$ So, the information matrix is

$$n\Im(\theta) = E[-H(\theta)] = \begin{bmatrix} \frac{1}{\sigma^2}X'X & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}$$

and its inverse (the Cramér-Rao bound for the covariance matrix of any unbiased estimator) is

$$[n\Im(\theta)]^{-1} = \begin{bmatrix} \sigma^2(X'X)^{-1} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}$$

The covariance matrix of coefficients estimated by OLS is $(X'X)^{-1}\sigma^2$, so OLS coefficients attain the Cramér-Rao bound. But the OLS estimator of σ^2 does not attain the bound. In fact, remembering that $\hat{\sigma}^2/\sigma^2$ is a random variable χ^2_{n-k} divided by $n-k$, and that the variance of the χ^2_{n-k} is $2(n-k)$, we get:

$$Var(\hat{\sigma}^2) = \left[\frac{\sigma^2}{n-k}\right]^2 Var\left(\chi^2_{n-k}\right) = \frac{2\sigma^4}{n-k}$$

which is larger than the Cramér-Rao bound (however, it is not possible to find an unbiased estimator of σ^2 with a smaller variance; see Rao, 1973, 5a.2).

Remark. If the Hessian (17.99) is *estimated*, that is it is computed at the OLS estimated parameters $\hat{\beta}$ and $\hat{\sigma}^2$, the off diagonal blocks are zero ($X'X\hat{\beta} - X'y = -X'\hat{u} = 0$).

Remark. Obviously, the *good* properties of the OLS estimator just described are no more valid if some elements of x_i are correlated with u_i. In principle, the likelihood should be re-specified, to take explicitly into account the correlation, and maximum likelihood would be different from the simple OLS estimator.

17.7 Nonlinear regression model: maximum likelihood and nonlinear least squares

Let the model and the vector of parameters be

$$y_i = q(x_i, \beta) + u_i \qquad u_i\ i.i.d.\ N(0, \sigma^2) \qquad i = 1, 2, ..., n \qquad \theta = \begin{bmatrix} \beta \\ \sigma^2 \end{bmatrix} \tag{17.100}$$

where q is a nonlinear function of the explanatory variables x_i and of the coefficients β, satisfying some *regularity* conditions (continuity and differentiability). Almost all properties of the linear regression with normal errors apply to the nonlinear regression as well. The only difference is that, unlike the linear case, estimation of coefficients usually requires the application of a numerical technique (e.g. Newton-Raphson or similar), as it cannot be done in closed form.

Being the Jacobian of the transformation $\partial u_i/\partial y_i = 1$ (so that the density $f(y_i,\theta) = f(u_i,\theta)$) the log-likelihood is

$$lnL(y,\theta) = -\frac{n}{2}ln(2\pi) - \frac{n}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[y_i - q(x_i,\beta)\right]^2 \tag{17.101}$$

the score is

$$\frac{\partial lnL(y,\theta)}{\partial\theta} = \begin{bmatrix} \frac{\partial lnL}{\partial\beta} \\ \frac{\partial lnL}{\partial\sigma^2} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2}\sum_{i=1}^{n}\left[y_i - q(x_i,\beta)\right]\frac{\partial q(x_i,\beta)}{\beta} \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}\left[y_i - q(x_i,\beta)\right]^2 \end{bmatrix} \tag{17.102}$$

thus the system of first order conditions is

$$\begin{bmatrix} \frac{1}{\sigma^2}\sum_{i=1}^{n}\left[y_i - q(x_i,\beta)\right]\frac{\partial q(x_i,\beta)}{\beta} = 0 \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}\left[y_i - q(x_i,\beta)\right]^2 = 0 \end{bmatrix} \tag{17.103}$$

Solution of the last equation gives

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - q(x_i,\beta)\right]^2 \tag{17.104}$$

that can be substituted into (17.101) producing the *concentrated* log-likelihood

$$lnL(y,\beta) = -\frac{n}{2}ln(2\pi) - \frac{n}{2} - \frac{n}{2}ln\left\{\frac{1}{n}\sum_{i=1}^{n}\left[y_i - q(x_i,\beta)\right]^2\right\} \tag{17.105}$$

There is no more the parameter σ^2, so the concentrated log-likelihood has to be maximized only with respect to the coefficients β. From equation (17.105) it is clear that the maximum of the concentrated log-likelihood is the minimum of the sum of squared errors $[y_i - q(x_i,\beta)]^2$; thus maximum likelihood is *nonlinear least squares*.
After β has been estimated minimizing (with some numerical technique) the sum of squared errors, the estimate of σ^2 is obtained from (17.104); it is the average of the squared residuals, analogously to the linear regression case.

Remark. Rather than minimizing the sum of squared residuals, one could minimize "$n/2$ *ln* of the average of the squared residuals" (as in equation 17.105), using the Newton-Raphson procedure (at least in the last iterations). The coefficient estimates would obviously be the same, but there would be no need of any further calculation to estimate the variance-covariance matrix of the coefficients: it would simply be the inverse of the last iteration's Hessian matrix.
However, from a computational viewpoint, convergence of the Newton-Raphson procedure is usually faster when the method is applied to the sum of squared residuals. Thus, it might be more convenient to split the procedure in two parts: first compute coefficients minimizing the sum of squared residuals; then, when convergence has been achieved, compute (and invert) the Hessian of "$n/2$ *ln* of the average of the squared residuals" as an estimate of the coefficients variance-covariance matrix.

17.8 Nonlinear regression model with autocorrelated errors

We consider the same case as (17.100) when the error terms have a stationary AR(1) structure

$$y_t = q(x_t,\beta) + u_t \qquad u_t = \rho u_{t-1} + \varepsilon_t \qquad \varepsilon_t \ i.i.d.\ N(0,\sigma^2) \qquad t = 2,...,n \tag{17.106}$$

where the vector of parameters is

$$\theta = \begin{bmatrix} \beta \\ \rho \\ \sigma^2 \end{bmatrix} \qquad |\rho| < 1 \qquad Var(u_t) = \frac{\sigma^2}{1-\rho^2} \tag{17.107}$$

Subtracting from (17.106) its lagged value, we get

$$y_t - \rho y_{t-1} = q(x_t,\beta) - \rho q(x_{t-1},\beta) + \varepsilon_t \qquad \varepsilon_t \ i.i.d.\ N(0,\sigma^2) \qquad t = 2,...,n \tag{17.108}$$

that can be treated as a nonlinear regression model with *i.i.d.* normal errors. Thus, to maximize the concentrated log-likelihood with respect to the coefficients (β and ρ), one has to minimize the sum of squared errors of the transformed model (17.108)

$$\sum_{t=1}^{n}\left\{[y_t - \rho y_{t-1}] - [q(x_t,\beta) - \rho q(x_{t-1},\beta)]\right\}^2 \tag{17.109}$$

After $\hat{\beta}$ and $\hat{\rho}$ have been computed from minimization of (17.109), the estimate $\hat{\sigma}^2$ is obtained, as usual, as the average of the squared residuals of (17.108).

Remark. The same argument of the previous section can be applied here as well. Rather than minimizing the sum of squared residuals (17.109), one could minimize "$n/2$ *ln* of the average of the squared residuals" using the Newton-Raphson procedure. This would produce the same estimates of β and ρ, and the estimate of their variance-covariance matrix would be the inverse of the last iteration's Hessian matrix. From a computational viewpoint, however, it is usually more convenient to first compute the estimates of β and ρ; then, after convergence of the iterative maximization procedure has been achieved, compute (and invert) the Hessian of "$n/2$ *ln* of the average of the squared residuals" as an estimate of the coefficients variance-covariance matrix.

18 APPENDIX. Complements of linear algebra: Kronecker product

Let A be an $m \times n$ matrix and B a $p \times q$ matrix; the Kronecker product of the two matrices is a matrix with dimensions $mp \times nq$; using a block-representation, this product can be defined as follows

$$A \otimes B = \left[\begin{array}{c|c|c|c|c} a_{1,1}B & a_{1,2}B & a_{1,3}B & ... & a_{1,n}B \\ a_{2,1}B & a_{2,2}B & a_{2,3}B & ... & a_{2,n}B \\ a_{3,1}B & a_{3,2}B & a_{3,3}B & ... & a_{3,n}B \\ ... & ... & ... & ... & ... \\ a_{m,1}B & a_{m,2}B & a_{m,3}B & ... & a_{m,n}B \end{array} \right] \tag{18.110}$$

The Kronecker product is distributive with respect to the sum; that is
$(A + C) \otimes B = A \otimes B + C \otimes B$, if A and C have the same dimensions;
$A \otimes (B + D) = A \otimes B + A \otimes D$, if B and D have the same dimensions (the proof is straightforward).
The transpose of the Kronecker product is the Kronecker product of the two transposed matrices:
$(A \otimes B)' = A' \otimes B'$
(the proof is straightforward, looking at the block-representation 18.110).
If A and C are conformable for the ordinary multiplication of matrices (rows by columns), and B and D are also conformable for the multiplication, then

$$(A \otimes B)(C \otimes D) = (AC) \otimes (BD) \tag{18.111}$$

To prove it, first of all it must be verified the equality of the dimensions of the matrices on both sides. Then, the equality is proved by writing the explicit expression of a generic element of the matrix on the left hand side and of the corresponding element of the matrix on the right hand side (for simplicity, one can write explicitly the element $(1, 1)$ which is $(A \otimes B)_{1.}(C \otimes D)_{.1}$ for the matrix on the left hand side, and $(AC)_{1,1}(BD)_{1,1} = (A_{1.}C_{.1})(B_{1.}D_{.1})$ for the matrix on the right hand side, easily verifying their equality; equality for all the other corresponding elements could be proved in the same way).
If A and B are two non-singular square matrices, the inverse of the Kronecker product is the Kronecker product of the two inverted matrices:
$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$
The proof follows straightforwardly from the above theorem, observing that if A has dimensions $m \times m$ and B has dimensions $n \times n$,
$(A^{-1} \otimes B^{-1})(A \otimes B) = (A^{-1}A) \otimes (B^{-1}B) = I_m \otimes I_n = I_{mn}$.
In particular, if Σ is a $(k \times k)$ nonsingular matrix, then
$(\Sigma \otimes I_n)^{-1} = \Sigma^{-1} \otimes I_n$.

19 APPENDIX. Two useful derivatives

19.1 Derivatives of a determinant

Let A be a non-singular square matrix $(n \times n)$, $|A|$ its determinant and $||A||$ the absolute value of the determinant. Then

$$\frac{\partial ln||A||}{\partial A} = (A')^{-1}$$

To prove it, first consider the derivative of the determinant with respect to $a_{i,j}$ (the $i, j-th$ element of the matrix), having expanded the determinant according to the cofactors of the $i-th$ row

$$|A| = a_{i,1}A_{i,1} + a_{i,2}A_{i,2} + ... + a_{i,n}A_{i,n}$$

No cofactor depends on $a_{i,j}$. Thus $\partial|A|/\partial a_{i,j} = A_{i,j}$. Then, applying the chain rule for derivatives

$$\frac{\partial ln||A||}{\partial a_{i,j}} = \frac{\partial ln||A||}{\partial |A|} \frac{\partial |A|}{\partial a_{i,j}} = \frac{A_{i,j}}{|A|}$$

which is the $j, i-th$ element of A^{-1}.

19.2 Derivatives of the elements of an inverse

Let A be a non-singular square matrix $(n \times n)$, $a_{i,j}$ its $i, j-th$ element and $a^{h,k}$ the $h, k-th$ element of A^{-1}. Then

$$\frac{\partial a^{h,k}}{\partial a_{i,j}} = -a^{h,i}a^{j,k}$$

(Theil, 1971, p.33 suggests as a simple *mnemonic rule* the familiar traffic sign *"no U turn anytime"*, where the "no" is represented by the minus sign and the "U" indicates the order in which the indices on the left hand side appear on the right: down from h to i, then up from j to k).

The proof follows in a straightforward way from the more general case, where all the elements of the matrix are functions of a scalar variable (say x). Now, considering the matrix $A(x)$ and its inverse $A^{-1}(x)$, the product $A(x)\ A^{-1}(x) = I_n$ (constant) thus all derivatives are zero

$$\frac{\partial A(x)\ A^{-1}(x)}{\partial x} = \frac{\partial A(x)}{\partial x}A^{-1}(x) + A(x)\frac{\partial A^{-1}(x)}{\partial x} = 0 \qquad \text{Thus} \qquad \frac{\partial A^{-1}(x)}{\partial x} = -A^{-1}(x)\frac{\partial A(x)}{\partial x}A^{-1}(x)$$

In the particular case where $x = a_{i,j}$ $(i, j - th$ element of $A)$, then all derivatives of the elements of A are zero, except for the $i, j - th$ element, whose derivative is $= 1$. So the matrix $\partial A(x)/\partial x$ has all elements $= 0$, with the exception of the $i, j - th$ element which is 1. The $h, k - th$ element of product is thus the product of the $h, i - th$ and $j, k - th$ elements of A^{-1}.

20 REFERENCES

20.1 Identification, estimation and simulation of simultaneous equation models - Surveys

1. Amemiya, T. (1983): "Non-linear Regression Models", in *Handbook of Econometrics*, ed. by Z. Griliches and M. D. Intriligator. Amsterdam: North-Holland Publishing Company, Vol. I, 333-389.

2. Fair, R. C. (1986): "Evaluating the Predictive Accuracy of Models", in *Handbook of Econometrics*, ed. by Z. Griliches and M. D. Intriligator. Amsterdam: North-Holland Publishing Company, Vol. III, 1979-1995.

3. Hausman, J. A. (1983): "Specification and Estimation of Simultaneous Equation Models", in *Handbook of Econometrics*, ed. by Z. Griliches and M. D. Intriligator. Amsterdam: North-Holland Publishing Company, Vol. I, 391-448.

4. Hsiao, C. (1983): "Identification", in *Handbook of Econometrics*, ed. by Z. Griliches and M. D. Intriligator. Amsterdam: North-Holland Publishing Company, Vol. I, 223-283.

20.2 Textbooks

1. Greene, W. H. (2012): *Econometric Analysis* (7*th* edition). Boston: Person Education Limited.

2. Hadley, G. (1961): *Linear Algebra*. Reading, MA: Addison-Wesley Publishing Company, Inc.

3. Johnston, J. (1984): *Econometric Methods* (3*rd* edition). New York: McGraw-Hill, Inc.

4. Rao, C. R. (1973): *Linear Statistical Inference and its Applications* (2*nd* edition). New York: John Wiley & Sons, Inc.

5. Schmidt, P. (1976): *Econometrics*. New York: Marcel Dekker, Inc.

6. Stock, J. H., and M. W. Watson (2015): *Introduction to Econometrics* (Updated 3*rd* edition). Boston: Person Education Limited.

7. Theil, H. (1971): *Principles of Econometrics*. New York: John Wiley & Sons, Inc.

20.3 Specific references

20.3.1 *Klein-I model*

1. Klein, L. R. (1950): *Economic Fluctuations in the United States, 1921-1941*. New York: John Wiley & Sons, Inc., Cowles Commission Monograph No. 11.

20.3.2 *Identification*

1. Fisher, F. M. (1959): "Generalization of the Rank and Order Conditions for Identifiability", *Econometrica* 27, 431-447.

2. Fisher, F. M. (1966): *The Identification Problem in Econometrics*. New York: McGraw-Hill.

3. Koopmans, T. C. (1949): "Identification Problems in Economic Model Construction", *Econometrica* 17, 125-144.

20.3.3 *Simulation, forecasting, multipliers, dynamic properties, control*

1. Bianchi, C., G. Calzolari, and P. Corsi (1981): "Estimating Asymptotic Standard Errors and Inconsistencies of Impact Multipliers in Nonlinear Econometric Models", *Journal of Econometrics* 16, 277-294.

2. Bianchi, C., and G. Calzolari (1980): "The One-Period Forecast Errors in Nonlinear Econometric Models", *International Economic Review* 21, 201-208. Reprinted in *Macroeconometric Modelling*, ed. by K. F. Wallis (1994). Cheltenham: Edward Elgar Publishing Ltd., *The International Library of Critical Writings in Econometrics*, Vol. 2, 183-190.

3. Chow, G. C. (1975): *Analysis and Control of Dynamic Economic Systems*. New York: John Wiley & Sons, Inc.

4. Duesenberry, J. S., G. Fromm, L. R. Klein, and E. Kuh, eds. (1969): *The Brookings Model: Some Further Results*. Amsterdam: North-Holland Publishing Company.

5. Evans, M. K., L. R. Klein, and G. R. Schink (1968): *The Wharton Econometric Forecasting Model*. Philadelphia: University of Pennsylvania, Economics Research Unit, Studies in Quantitative Economics No. 2.

6. Goldberger, A. S. (1959): *Impact Multipliers and Dynamic Properties of the Klein-Goldberger Model*. Amsterdam: North-Holland Publishing Company,

7. Goldberger, A. S., A. L. Nagar, and H. S. Odeh (1961): "The Covariance Matrices of Reduced-Form Coefficients and of Forecasts for a Structural Econometric Model", *Econometrica* 29, 556-573.

8. Howrey, E. P., and H. H. Kelejian (1971): "Simulation versus Analytical Solutions: the Case of Econometric Models", in *Computer Simulation Experiments with Models of Economic Systems*, ed. by T. H. Naylor. New York: John Wiley & Sons, Inc., 299-319.

9. Howrey, E. P., and L. R. Klein (1972): "Dynamic Properties of Nonlinear Econometric Models", *International Economic Review* 13, 599-618.

10. Kendrick, D. (1981): *Stochastic Control for Economic Models*. New York: McGraw-Hill.

11. Theil, H. (1966): *Applied Economic Forecasting*. Amsterdam: North-Holland Publishing Company.

12. Tinbergen, J. (1952): *On the Theory of Economic Policy*. Amsterdam: North-Holland Publishing Company,

20.3.4 *Instrumental variables*

1. Bowden, R. J., and D. A. Turkington (1984): *Instrumental Variables*. Cambridge University Press, Econometric Society Monographs in Quantitative Economics.

2. Brundy, J. M., and D. W. Jorgenson (1971): "Efficient Estimation of Simultaneous Equations by Instrumental Variables", *The Review of Economics and Statistics* 53, 207-224.

3. Brundy, J. M., and D. W. Jorgenson (1974): "The Relative Efficiency of Instrumental Variables Estimators of Systems of Simultaneous Equations", *Annals of Economic and Social Measurement* 3, 679-700.

4. Dhrymes, P. J. (1971): "A Simplified Structural Estimator for Large-Scale Econometric Models", *Australian Journal of Statistics* 13, 168-175.

5. Dutta, M., and E. Lyttkens (1974): "Iterative Instrumental Variables Method and Estimation of a Large Simultaneous System", *Journal of the American Statistical Association* 69, 977-986.

6. Geary, R. C. (1949): "Determination of Linear Relations between Systematic Parts of Variables with Errors of Observation, the Variances of which are Unknown", *Econometrica* 17, 30-59.

7. Lyttkens, E. (1974): "The Iterative Instrumental Variables Method and the Full Information Maximum Likelihood Method for Estimating Interdependent Systems", *Journal of Multivariate Analysis* 4, 283-307.

8. Reiersøl, O. (1945): "Confluence Analysis by Means of Instrumental Sets of Variables", *Arkiv for Matematik, Astronomi och Fysik* 32, 1-119.

9. Sargan, J. D. (1958): "The Estimation of Economic Relationships Using Instrumental Variables", *Econometrica* 26, 393-415.

20.3.5 *SURE, 2SLS, 3SLS, k-class (linear systems)*

1. Basmann, R. L. (1957): "A Generalized Classical Method of Linear Estimation of Coefficients in a Structural Equation", *Econometrica* 25, 77-83.

2. Nagar, A. L. (1959): "The Bias and Moment Matrix of the General k-class Estimators of the Parameters in Simultaneous Equations", *Econometrica* 27, 575-595.

3. Theil, H. (1958): *Economic Forecasts and Policy*. Amsterdam: North-Holland Publishing Company.

4. 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.

5. Zellner, A., and H. Theil (1962): "Three-Stage Least Squares: Simultaneous Estimation of Simultaneous Equations", *Econometrica* 30, 54-78.

20.3.6 *Limited and Full Information Maximum Likelihood*

1. Anderson, T. W. (2005): "Origins of the Limited Information Maximum Likelihood and Two-Stage Least Squares Eastimators", *Journal of Econometrics* 127, 1-16.

2. Anderson, T. W., and H. Rubin (1949): "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations", *Annals of Mathematical Statistics* 20, 46-63.

3. Anderson, T. W., and H. Rubin (1950): "The Asymptotic Properties of Estimates of the Parameters of a Single Equation in a Complete System of Stochastic Equations", *Annals of Mathematical Statistics* 21, 570-582.

4. Berndt, E. K., B. H. Hall, R. E. Hall, and J. A. Hausman (1974): "Estimation and Inference in Nonlinear Structural Models", *Annals of Economic and Social Measurement* 3, 653-665.

5. Chernoff, H., and N. Divinsky (1953): "The Computation of Maximum-Likelihood Estimates of Linear Structural Equations", in *Studies in Econometric Method*, ed. by W. C. Hood and T. C. Koopmans. New York: John Wiley & Sons, Inc., Cowles Commission Monograph No. 14, 236-302.

6. Koopmans, T. C., H. Rubin, and R. B. Leipnik (1950): "Measuring the Equation Systems of Dynamic Economics", in *Statistical Inference in Dynamic Economic Models*, ed. by T. C. Koopmans. New York: John Wiley & Sons, Inc., Cowles Commission Monograph No. 10, 53-237.

7. Mann, H. B., and A. Wald (1943): "On the Statistical Treatment of Linear Stochastic Difference Equations", *Econometrica* 11, 173-220.

20.3.7 *Maximum Likelihood with covariance restrictions*

1. Rothenberg, T. J. (1973): *Efficient Estimation with A Priori Information*. New Haven: Yale University Press, Cowles Foundation Monograph No. 23.

20.3.8 *Estimation of nonlinear simultaneous equations*

1. Amemiya, T. (1977): "The Maximum Likelihood and the Nonlinear Three-Stage Least Squares Estimator in the General Nonlinear Simultaneous Equation Model", *Econometrica* 45, 955-968.

2. Belsley, D. A. (1980): "On the Efficient Computation of the Nonlinear Full-Information Maximum-Likelihood Estimator", *Journal of Econometrics* 14, 203-225.

3. Calzolari, G., L. Panattoni, and C. Weihs (1987): "Computational Efficiency of FIML Estimation", *Journal of Econometrics* 36, 299-310.

4. Calzolari, G., and L. Panattoni (1988): "Alternative Estimators of FIML Covariance Matrix: A Monte Carlo Study", *Econometrica* 56, 701-714.

5. Gallant, A. R. (1977): "Three-Stage Least-Squares Estimation for a System of Simultaneous, Nonlinear, Implicit Equations", *Journal of Econometrics* 5, 71-88.

6. Hatanaka, M. (1978): "On the Efficient Estimation Methods for the Macro-Economic Models Nonlinear in Variables", *Journal of Econometrics* 8, 323-356.

7. Phillips, P. C. B. (1982): "On the Consistency of Nonlinear FIML", *Econometrica* 50, 1307-1324.

20.3.9 *Instrumental variables $\Longrightarrow$ Full Information Maximum Likelihood*

1. Calzolari, G., and L. Sampoli (1993): "A Curious Result on Exact FIML and Instrumental Variables", *Econometric Theory* 9, 296-309.

2. Dagenais, M. G. (1978): "The Computation of FIML Estimates as Iterative Generalized Least Squares Estimates in Linear and Nonlinear Simultaneous Equations Models", *Econometrica* 46, 1351-1362.

3. Durbin, J. (1963, 1988): "Maximum Likelihood Estimation of the Parameters of a System of Simultaneous Regression Equations". London School of Economics: discussion paper presented at *The European Meeting of the Econometric Society*, Copenhagen, 1963. Published in *Econometric Theory* 4 (1988), 159-170.

4. Hausman, J. A. (1974): "Full Information Instrumental Variables Estimation of Simultaneous Equations Systems", *Annals of Economic and Social Measurement* 3, 641-652.

5. Hausman, J. A. (1975): "An Instrumental Variable Approach to Full Information Estimators for Linear and Certain Nonlinear Econometric Models", *Econometrica* 43: 727-738.

6. Hendry, D. F. (1976): "The Structure of Simultaneous Equations Estimators", *Journal of Econometrics* 4, 51-88.