Econometrics Lecture Notes (I)

Montserrat Farell*

Universitat Autònoma de Barcelona

10th March 2003

Contents

*This notes are slight modifications of part of the book *Lecture Notes in Internet* at http://pareto.uab.es/omega/Project_001 by Professor Michael Creel

1 General Information, Grading, Etc.

Prof. M. Farell

Office: B3-192

Tel. 581-2932

1.1 Suggested texts

Berndt, E., *The Practice of Econometrics: Classic and Contemporary* (1991), Ed. Addison Wesley. This book focuses on applications and empirical implementations of econometrics. It is especially designed for first-year graduate econometrics as a suplementary textbook.

Creel, M., *Lecture Notes in internet* (2002). http://pareto.uab.es/omega/Project_001/.

Green, W., *Econometric Analysis*, Third Edition (1997) Ed. Prentice Hall. This is a book that covers a wide range of topics at an intermediate level.

Hayashi, F., *Econometrics* (2000). This book covers all topics from OLS through cointegration from a modern perspective. It contains also empirical applications and simulation exercises.

More advanced references are Davidson, R. and J. Mackinnon, *Estimation and Inference in Econometrics* (1993), Oxford Univ. Press, and Hamilton, J., *Time Series Analysis* (1994), Princeton Univ. Press.

1.2 Requirements

There will be 7 or 8 problem sets that will provide 30% of the grade and a final exam 70%.

1.3 Web page

There is a class web page which will serve as the distribution point for the class notes and materials for the problem sets. To access it, point your favorite browser at ***pareto.uab.es/mfarell/graduate***. In addition to the department's computer room, the Facultat has a computer room in Aules 21-22-23 where you can find plenty of computers connected to the internet.

2 Economic and econometric models

Economic theory postulates a mathematical relashionship between some economic variables

$$y = f(X),$$

say x is a $kx1$vector. It is a parametric relationship which describes the economic agents' behaviour, and when applying economic analysis to policy questions one can get estimations of the parameters of interest using econometric technics. For example, we can be interested on:

- the input elasticities of substitution in a production or a cost function
- the direct eleasticity quantity-price in a demand function
- the returns to scale a technology exhibits in a production or a cost function
- the marginal propensity to consume in a consumtion function, among others.

Economic theory tells us some characteristics that f should have.
Economic theory tells us which are the vaiables in x.

The econometrician will try to "guess" f and, the parameters of the function from the data gathered of y and x.

The econometric model will be the representation of the economic model. We will consider first that this relation is linear (or that we can make it linear).

$y = f(X, \beta) + \varepsilon$

- y is a $(nx1)$ vector
- x is (nxk)vector
- β is a $(kx1)$ vector

- ε is a $(nx1)$ vector. It is the error term and is interpreted as the sum of several independent effects (variables) not controled by the econometrician.

For the economist the most interesting cases will be the ones where y and X are highly quality data and $y = f(X)$ has a solid theoretical base. Examples can come from two big pieces in the economic theory literature:

a) the theory of the consumer

b) the theory of the firm

a) From utility maximization we obtain the demand functions $x_j^* = (p, M)$, the demand function should satisfy some regularity conditions, such as homogeneity of degree zero on p and M, symetry effects etc.

b) From cost minimization we obtain the indirect cost function $C_i = f(p, Q)$, it is a representation of the technology, from duality theory it recovers all the parameters of the production function. The cost function also should satisfy some conditions, such as homogeneity of degree one on input prices, p, non-decreasing on p, continuous with respect to p etc.

When doing econometric modelling: $y = f(X, \beta) + \varepsilon$ the econometrician needs to specify:

- Which variables X belong to the model

  economic theory

  common sense

- Which is the functional form f

  economic theory gives us the regularity conditions but there are a lot of functional forms that satisfy those conditions

- Which is the error specification $\varepsilon \sim$

  analysis of the context

Origin of the word regression

The linear model presented nowadays, what is understood by the *Gauss linear model* or *the linear regression model* has its main contributions on the work of Halley (1656-1742), De Moivre (1667-1754), Bernoulli (1700-1782), Bayes (1702-1761), Lagrange (1736-1813), Laplace (1749-1827), Legendre (1752-1833), Gauss (1789-1857). Afterwards, with Galton (1822-1911), Edgeworth (1845-1926), Pearson (1857-1936) and Yule (1871-1951) the convergence of descriptuve statistics and the calculus of probabilities in the context of the Gauss linear model was a reality (Spanos 1986; Statistical foundations of Econometric Modelling. *A brief historical overview*. Cambridge University Press).

But the use of the word regression in this context has its origins in the experiments of Galton, when trying to see the relationship between hight of parents and hight of their children. It was clear that tall parents had tall children and short parents had short children. Nevertheless, he discovered that the hight of children from unusually tall or unusually short parents would have the tendency to go to the average of the polpulation hight. He did the experiment fitting the line through the data points that would minimize the sum of the squared of the distances between the points and the line. He called the phenomenon explained above ”regression to the mean” in his words ”regression to mediocrity” (1886) that is the unconditional mean.

75
70
inches

3 Ordinary Least Squares: small samples

3.1 The classical linear model

The econometric model is a representation of the economic model.

$y = f(X, \beta) + \varepsilon$

- the variable under study is the left-hand side, *the dependent variable y*.
- this variable is explained by (related to) several other variables, the right-hand side, *the independent, explanatory, variables, the regressors X*.
- y_i is the $i-th$ value of the dependent variable.
- $x_i = (x_{i1}, x_{i2}, ..., x_{ik})$ is the $i-th$ observation of the k regressors.

Draw a two variable case to motivate the graphical representation of the E(y/x_i).

Data in economics is not experimental so, y and x can be both treated as random variables or y is random and x are fixed values.

- the model is a linear function of the parameter vector β :

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i,$$

$$i = 1, 2, ..., n$$

- $\beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}$ is the regression line, the $\beta's$ are the regression coefficients
- $\beta_j = \frac{\partial y_i}{\partial x_{ij}}$; $j = 1, 2..., k$ because of linearity the marginal effect does not depend on the level of the regressors.

- the error term is the part unexplained by the regressors. It is the sum of the variables not under the control of the econometrician.

Example 1: The consumption function

$$con_i = \beta_1 + \beta_2 DI_i + \varepsilon_i$$

consumption is a linear function of disposable income. In the general notation we used $x_{i1} = 1$for every i. Question: why is it important to introduce a column of ones?

Example 2: A semi-log or log-linear wage equation

$$wage_i = e^{\beta_1} e^{\beta_2 S_i} e^{\beta_3 TEN_i} e^{\beta_4 EXP_i} e^{\varepsilon_i}$$

- $wage_i$ is the wage rate for individual i
- S_i eduction in years for individual i
- TEN_i is tenure, number of years in the current job.
- EXP_i is experience, number of years on all previous and current job.

Taking logs in both sides leads to the semi-log model:

$$log(wage_i) = \beta_1 + \beta_2 S_i + \beta_3 TEN_i + \beta_4 EXP_4 + \varepsilon_i$$

the coefficients in this model are percentatge changes, in a general two variable case we have:

$$log(y_i) = \beta_1 + \beta_2 x_{i2} + \varepsilon_i$$

$\beta_2 = \frac{\partial log(y_i)}{\partial x_{i2}} * \frac{1}{y_i} = \frac{\frac{\partial y_i}{y_i}}{\partial x_{i2}}$ =*relative change in* y_i */ absolute change in* x_{i2}

Example 3: A Cobb-Douglas cost function

By duality a cost function is a dual of a production function and it represents the technology. The indirect cost function gives us, for every level of output and a set of prices, the minimum cost you can produce that output with the represented technology.

$$c = f(p_k, p_l, Q)$$

the Cobb-Douglas econometric model representation is:

$$c = A * p_k^{\beta_k} * p_l^{\beta_l} * Q^{\beta_Q} * e^{\varepsilon}$$

for a set of firms $i = 1, 2..., n$ and taking logarithms in both sides we obtain the following log. model

$$log(c_i) = logA + \beta_k log(p_{ik}) + \beta_l log(p_{il}) + \beta_Q log(Q_i) + \varepsilon_i$$

in a general two varaible case model

$$log(y_i) = \beta_1 + \beta_2 log(x_{i2}) + \varepsilon_i$$

$\beta_2 = \frac{\partial log(y_i)}{\partial log(x_{i2})} = \frac{\frac{\partial y_i}{y_i}}{\frac{\partial x_{i2}}{x_{i2}}} = \frac{\partial y_i}{\partial x_{i2}} * \frac{x_{i2}}{y_i} = \xi$ an elasticity. In the particular case above, it can be proven that the invers of the direct elasticity

$$\xi_{c,Q} = \frac{\partial log(c_i)}{\partial log(Q_i)} = \frac{\frac{\partial c_i}{c_i}}{\frac{\partial Q_i}{Q_i}} = \beta_Q$$

are the returns to scale this technology exhibits,

$$\frac{1}{\xi_{c,Q}} = RTS = \frac{1}{\beta_Q}$$

Observe that RTS donot depend on the level of Q. Is that a restriction that comes from Economic Theory? Think of the commonly assumed U-shape average and marginal cost curves. Does that assumption lead to RTS in this technology to be independent of the level of output? *Draw a cost function, as a function of out put to see the U-shape curves.*

Example 4: Consider an extension to the Cobb-Douglas cost model

$$log(c_i) = logA + \beta_k log(p_{ik}) + \beta_l log(p_{il}) + \beta_Q log(Q_i) + \beta_{QQ}(logQ_i)^2 + \varepsilon_i$$

in this case the direct elasticity $cost - output$ is

$$\xi_{c,Q} = \frac{\partial log(c_i)}{\partial log(Q_i)} = \beta_Q + 2\beta_{QQ} logQ_i$$

in this representation RTS will be a function of the level of output

All these are examples were the linearity assumption holds after a certain transformation of the model. Also non-linearities on the regressors can be accomodated. Other non-linearities on the parameters, for example, would require non-linear estmation.

Matrix Notation

- x_i and β are two $k-dimensional$ vectors:

$$x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ . \\ . \\ . \\ x_{ik} \end{bmatrix} ; \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_k \end{bmatrix}$$

- the inner product is the linear combination,

$$x_i'\beta = \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}$$

The classical linear model is based upon several assumptions:

ASSUMPTION 1: linearity.

The linear model can be written

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

$$i = 1, 2, ..., n$$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix}; \varepsilon = \begin{bmatrix} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix} \end{bmatrix};$$

$$X = \begin{bmatrix} x_1' \\ x_2' \\ . \\ . \\ . \\ x_n' \end{bmatrix} = \begin{bmatrix} x_{11} & . & . & . & x_{1k} \\ x_{21} & . & . & . & x_{2k} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ x_{n1} & . & . & . & x_{nk} \end{bmatrix}$$

$$y = \mathbf{X}\beta + \varepsilon$$

ASSUMPTION 2: X has rank k.

That means there is no perfect multicollinearity, $\rho(X) = k$, X is full rank.

ASSUMPTION 3: strict exogeneity of X.

$$E(\varepsilon_i/X) = 0$$

this states that no observations in X can give information about the expected value of the disturbance. And this conditional expectation implies

$$\begin{aligned} E(\varepsilon_i) &= 0 \\ i &= 1,2,...,n \end{aligned}$$

because of the Law of Total Expectations $E[E(\varepsilon_i/X)] = E(\varepsilon_i)$

$$E(x_{jk}\varepsilon_i) = 0$$

$$j = 1,2,...,n;$$

$$i = 1,2,...,n$$

and those imply

$$E(X'\varepsilon) = 0$$

$$E(y/X) = X\beta$$

strict exogeneity implies orthogonality between the regressors and the error term for

all observations[1].

ASSUMPTION 4: espherical disturbances.

$E(\varepsilon\varepsilon'/X) = \sigma^2 I_n$.

ASSUMPTION 5: Independent Identically Distributed (IID) with Normality.

$\varepsilon/X \sim N(0, \sigma^2 I_n)$

Assuming that the error term is the sum of independent effects not under the control of the analist it is considered that the conditions of the Central Limit Theorem will generally apply. So the normality assumption will be reasonable in most settings. A usefull implication of normality is that $\varepsilon_i's$ are not only uncorrelated but also independent.

It is very common to assume that $X's$ are non-stochastic. That would be the case in experimental data, where the analyst chooses the values of x_i and observes y_i. It is an unrealistic assumption in some cases, for example with time series data but it is reasonable in cross-section analysis. If they are stochastic then we need strict exogeneity ($X's$ uncorrelated with any ε) or predetermination ($X's$ uncorrelated with contemporaneous errors) in order to get nice properties of the estimator like unbiassness and consistency.

An alternative interpretation is that observations x_i are the same in repeated samples, that is "fixed". And this is equivalent to doing the analysis conditional on the sample observed, thus as if X was non-stochastic.

For now, we will make the assumption of fixed $X's$ and we will derive the properties of the estimator from there. If one is interested on unconditional results, a convenient method of obtaining the statistical properties of $\widehat{\beta}$ unconditioned on the sample is to obtain the results conditioned on X first (see Green or Hayashi for formal results) take the conditioned distribution and average over X (integrate over X)(for results see Green pg. 260-270).

[1]For formal proofs and discussion see Hayashi (Chapter 1)

With fixed regressors the assupmtions can be stated as follows:

A1: linearity

A2: full rank of X

A3: $E(X'\varepsilon) = 0$

A4: $E(\varepsilon\varepsilon') = \sigma^2 I_n$

A5: $\varepsilon \sim N(0, \sigma^2 I_n)$

A6: X are fixed

3.2 Estimation by least squares

Definitions:

- β the vector of unknown parameters
- $\widetilde{\beta}$ any value for the parameters
- $\hat{\beta}$ the OLS estimate for β
- $\widetilde{\varepsilon}_i = y_i - x_i'\widetilde{\beta}$ residual for observation i

$$\begin{aligned} \hat{\beta} &= \arg\min SSR(\widetilde{\beta}) = \sum_{i=1}^{n} \left(y_i - x_i'\widetilde{\beta}\right)^2 \\ s(\widetilde{\beta}) = SSR(\widetilde{\beta}) &= \left(y - X\widetilde{\beta}\right)' \left(y - X\widetilde{\beta}\right) \\ &= y'y - 2y'X\widetilde{\beta} + \beta' X' X\widetilde{\beta} \end{aligned}$$

To minimize the criterion $SSR(\widetilde{\beta})$, take the f.o.n.c. and set them to zero:

$$D_{\widetilde{\beta}} s(\hat{\beta}) = \frac{\partial SSR(\hat{\beta})}{\partial \widetilde{\beta}} = -2X'y + 2X'X\hat{\beta} = 0$$

properties used

$$i)\ \frac{\partial a'x}{\partial x} = a$$

$$ii)\ \frac{\partial x'Ax}{\partial x} = 2Ax;$$

if A symetric,

so

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

To verify that this is a minimum, check the s.o.s.c.:

$$D^2_{\underset{\sim}{\beta}} s(\widehat{\beta}) = \frac{\partial^2_{\underset{\sim}{\beta}} SSR(\widehat{\beta})}{\partial\widehat{\beta}\partial\widehat{\beta}'} = 2X'X$$

Since $\rho(X) = k$, this matrix is positive definite, since it's a quadratic form in a p.d. matrix (identity matrix of order n), so $\hat{\beta}$ is in fact a minimizer.

- The *fitted values* are in the vector $\hat{y} = X\hat{\beta}$.
- The *residuals* are in the vector $\hat{\varepsilon} = y - X\hat{\beta}$
- Note that

$$\begin{aligned} y &= X\beta + \varepsilon \\ &= X\hat{\beta} + \hat{\varepsilon} \end{aligned}$$

3.3 Finite sample properties of the OLS estimator (H-W # 1: simulation)

3.3.1 Unbiasedness

Under assumptions A1-A6:

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

You can see in Hayashi pg. 29 the treatment of this proof when considering the conditional distributions:

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

by linearity of conditional expectations and the fact that conditional on X, $E((X'X)^{-1}X') = (X'X)^{-1}X' = A$; a matrix of constants

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

by strict exogeneity (A3)

$$E(\widehat{\beta}/X) = \beta.$$

3.3.2 Efficiency (Gauss-Markov theorem)

The variance-covariance matrix of the OLS estimator

$$V(\widehat{\beta}) = E((\widehat{\beta}-\beta)(\widehat{\beta}-\beta)')$$

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

therefore the distribution of $\widehat{\beta}$ is $N(\beta, \sigma^2(X'X)^{-1})$.

Asside: A usufull way to derive the distributions of vectors of random variables which are linear combinations of a normaly distributed random vector is aplying the following proposition:

If *x* is a *nx*1random vector where

$$
\begin{aligned}
x &\sim N(\mu, \Sigma) \\
\theta = a + Ax &\sim N(a + A\mu, A\Sigma A')
\end{aligned}
$$

Again, in the case of considering the conditional distributions we would have

(Hayashi, chapter 1):

$$\begin{aligned} V(\widehat{\beta}/X) &= V((\widehat{\beta}-\beta)/X) \\ &= V((X'X)^{-1}X'\varepsilon/X) \\ &= V(A\varepsilon/X) \\ &= AE(\varepsilon\varepsilon'/X)A' \\ &= A\sigma^2 I_n A' \\ &= \sigma^2(X'X)^{-1} \end{aligned}$$

Efficiency:

We say that $\widehat{\beta}$ is efficient for β if $\widehat{\beta}$ is unbiassed for β and

$$V(\widehat{\beta}) \leqslant V(\widetilde{\beta})$$

where $\widetilde{\beta}$ is any other unbiassed estimator of β.

The Gauss-Markov Theorem:

The OLS estimator is a *linear estimator*, which means that it is a linear function of the dependent variable, y.

$$\begin{aligned} \hat{\beta} &= \left[(X'X)^{-1}X'\right]y \\ &= Cy \end{aligned}$$

It is also *unbiased*, as we proved above. One could consider other weights W in place of the OLS weights. We'll still insist upon unbiasedness. Consider $\tilde{\beta} = Wy$. If the

estimator is unbiased

$$\begin{aligned} \mathcal{E}(Wy) &= \mathcal{E}(WX\beta + W\varepsilon) \\ &= WX\beta \\ &= \beta \\ &\Rightarrow \\ WX &= I_K \end{aligned}$$

The variance of $\tilde{\beta}$ is

$$V(\tilde{\beta}) = WW'\sigma^2.$$

Define

$$D = W - (X'X)^{-1}X'$$

so

$$W = D + (X'X)^{-1}X'$$

Since $WX = I_K$, $DX = 0$, so

$$\begin{aligned} V(\tilde{\beta}) &= \left(D + (X'X)^{-1}X'\right)\left(D + (X'X)^{-1}X'\right)'\sigma^2 \\ &= \left(DD' + (X'X)^{-1}\right)\sigma^2 \end{aligned}$$

So

$$V(\tilde{\beta}) \geq V(\hat{\beta}).$$

This is a proof of the Gauss-Markov Theorem.

Theorem 1 (Gauss-Markov) *Under the classical assumptions, the variance of any linear unbiased estimator minus the variance of the OLS estimator is a positive semidef-*

inite matrix.

- It is worth noting that we have not used the normality assumption in any way to prove the Gauss-Markov theorem, so it is valid if the errors are not normally distributed, as long as the other assumptions hold.

For the version of the Gauss-Markov theorem considering conditional expectations see Hayashi pg. 29.

Estimation of σ^2 :

For $\hat{\sigma}^2$ we have

$$\begin{aligned}
\widehat{\sigma^2} &= \frac{1}{n-K}\hat{\varepsilon}'\hat{\varepsilon} \\
\hat{\varepsilon} &= y - X\hat{\beta} \\
&= y - X(X'X)^{-1}X'y \\
&= M_X y \\
&= M_X(X\beta+\varepsilon) \\
&= M_X\varepsilon \\
\mathcal{E}(\widehat{\sigma^2}) &= \frac{1}{n-K}\mathcal{E}(\varepsilon' M\varepsilon) \\
&= \frac{1}{n-K}\mathcal{E}(TrM\varepsilon\varepsilon') \\
&= \frac{1}{n-K}Tr\mathcal{E}(M\varepsilon\varepsilon') \\
&= \frac{1}{n-K}\sigma^2 TrMI_n \\
&= \frac{1}{n-K}\sigma^2\left(n - TrX(X'X)^{-1}X'\right) \\
&= \frac{1}{n-K}\sigma^2\left(n - Tr(X'X)^{-1}X'X\right) \\
&= \sigma^2
\end{aligned}$$

$\hat{\sigma}^2$ is an unbiassed estimator for σ^2.

3.4 Geometric interpretation of least squares estimation

3.4.1 The least squares problem in the context of matrix algebra

(see Green chapter 2) or (Davidson and Mackinon chapter1).

- Given a vector y and a matrix X we are interested on expressing y as a linear combination of X. Say $X\widehat{\beta}$.
- if y lies in the column space of X, we just find $\widehat{\beta}$.
- if y does not lie in that space we can still write

$$y = X\widehat{\beta} + \widehat{\varepsilon}$$

  $\widehat{\varepsilon}$ is the difference (the distance) between y and $X\widehat{\beta}$
- we want to find $\widehat{\beta}$ such that y is the "closest" as possible to $X\widehat{\beta}$. In this case we choose to minimize the Euclidian norm or L^2 norm.

$$\|\widehat{\varepsilon}\| = \sqrt{\widehat{\varepsilon}'\widehat{\varepsilon}}$$

  the problem is to find $\widehat{\beta}$ for which:

$$\|\widehat{\varepsilon}\| = \|y - X\widehat{\beta}\|$$

  is as small as possible. The solution is $\widehat{\beta}$ such that $\widehat{\varepsilon} \perp X\widehat{\beta}$ **this is a mistake in Greene pg. 27** what needs to be orthogonal is $\widehat{\varepsilon} \perp X$ since the orthogonality condition *is to find* $\widehat{\beta}$ *such that there is a minimum distance between the span of* X *and* y

- by the orthogonality condition **this is not the way to find** $\widehat{\beta}$

$$\begin{aligned}
(X\widehat{\beta})'\widehat{\varepsilon} &= 0 \\
\widehat{\beta}'X'(Y - X\widehat{\beta}) &= 0 \\
&= \widehat{\beta}'X'Y - \widehat{\beta}'X'X\widehat{\beta} \\
&= \widehat{\beta}'(X'Y - {}'X'X\widehat{\beta}) \\
since\,\widehat{\beta} &\neq 0 \\
X'Y - {}'X'X\widehat{\beta} &= 0 \\
\hat{\beta} &= (X'X)^{-1}X'y
\end{aligned}$$

let's do it right:

$$\begin{aligned}
(X)'\widehat{\varepsilon} &= 0 \\
(X)'(Y - X\widehat{\beta}) &= 0 \\
X'Y - {}'X'X\widehat{\beta} &= 0 \\
\hat{\beta} &= (X'X)^{-1}X'y
\end{aligned}$$

3.4.2 Projection Matrices

- We have that $X\hat{\beta}$ is the projection of y on the span of X, or

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

Therefore, the matrix that projects y onto the span of X is

$$P_X = X(X'X)^{-1}X'$$

$$X\hat{\beta} = P_X y.$$

- $\hat{\varepsilon}$ is the projection of y *off* the space spanned by X (that is onto the space that is orthogonal to the span of X). We have that

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

  So the matrix that projects y off the span of X is

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

  We have

$$\hat{\varepsilon} = M_X y.$$

- Therefore

$$\begin{aligned} y &= P_X y + M_X y \\ &= X\hat{\beta} + \hat{\varepsilon}. \end{aligned}$$

- Note that both P_X and M_X are *symmetric* and *idempotent*.
  - A symmetric matrix A is one such that $A = A'$.
  - An idempotent matrix A is one such that $A = AA$.
  - The only nonsingular idempotent matrix is the identity matrix. It is easy to

see it from the property of idempotent matrices that their trace equals their rank.

3.5 Goodness of fit

The fitted model is

$$y = X\hat{\beta} + \hat{\varepsilon}$$

Take the inner product:

$$y'y = \hat{\beta}'X'X\hat{\beta} + 2\hat{\beta}'X'\hat{\varepsilon} + \hat{\varepsilon}'\hat{\varepsilon}$$

But the middle term of the RHS is zero since $X'\hat{\varepsilon} = 0$, so

$$y'y = \hat{\beta}'X'X\hat{\beta} + \hat{\varepsilon}'\hat{\varepsilon}$$

The *uncentered* R_u^2 is defined as

$$\begin{aligned} R_u^2 &= 1 - \frac{\hat{\varepsilon}'\hat{\varepsilon}}{y'y} \\ &= \frac{\hat{\beta}'X'X\hat{\beta}}{y'y} \\ &= \frac{\| P_X y \|^2}{\| y \|^2} \\ &= \cos^2(\phi), \end{aligned}$$

where ϕ is the angle between y and the span of X *(show with the one regressor, two observation example).*

- The uncentered R^2 changes if we add a constant to y, since this changes ϕ. Another, more common definition measures the contribution of the variables, other than the constant term, to explaining the variation in y.

- Let $\iota = (1, 1, ..., 1)'$, a n -vector. So

$$\begin{aligned} M_\iota &= I_n - \iota(\iota'\iota)^{-1}\iota' \\ &= I_n - \iota\iota'/n \end{aligned}$$

$M_\iota y$ just returns the vector of deviations from the mean.

The *centered* R_c^2 is defined as

$$R_c^2 = 1 - \frac{\hat{\varepsilon}'\hat{\varepsilon}}{y' M_\iota y} = 1 - \frac{RSS}{TSS}$$

Supposing that X contains a column of ones (*i.e.,* there is a constant term),

$$X'\hat{\varepsilon} = 0 \Rightarrow \sum_t \hat{\varepsilon}_t = 0$$

so $M_\iota \hat{\varepsilon} = \hat{\varepsilon}$. In this case

$$\begin{aligned} y' M_\iota y &= \hat{\beta}' X' M_\iota X \hat{\beta} + \hat{\varepsilon}'\hat{\varepsilon} \\ &= \widehat{y}' M_\iota \widehat{y} + \hat{\varepsilon}'\hat{\varepsilon} \end{aligned}$$

So

$$R_c^2 = 1 - \frac{\hat{\varepsilon}'\hat{\varepsilon}}{y' M_\iota y} = \frac{\widehat{y}' M_\iota \widehat{y}}{y' M_\iota y}$$

$$\begin{aligned} y' M_\iota y &= TSS \\ \widehat{y}' M_\iota \widehat{y} &= ESS \\ \hat{\varepsilon}' \hat{M}_\iota \varepsilon &= RSS \end{aligned}$$

3.6 Linear Hypothesis testing under normality (H-W # 2: Hedonic prices for computers, Berndt Chap. 4)

We have now, one or more linear equations on the parameters of the model. These equations are often motivated by the same economic theory on which the model is based or they are based on statements about the explanatory capacity of the variables in the model.

- the restriction to be tested is called the null hypothesis, H_0.
- the model is "the mantained hypothesis", a set of assumptions.
- the model together with the null produces some test-statistic with a known distribution.
- too large of a value of the test statistic is interpreted as a failure of the null. This interpretation is only valid if the model is correctly specified.
- the test statistic may not have the supposed distribution when the null is true but the model is false.

3.6.1 Exact linear restrictions

In many cases, economic theory suggests restrictions on the parameters of a model. For example, a demand function is supposed to be homogeneous of degree zero in prices and income. If we have a Cobb-Douglas (log-linear) model,

$$\ln q = \beta_0 + \beta_1 \ln p_1 + \beta_2 \ln p_2 + \beta_3 \ln m + \varepsilon,$$

then we need that

$$k^0 \ln q = \beta_0 + \beta_1 \ln kp_1 + \beta_2 \ln kp_2 + \beta_3 \ln km + \varepsilon,$$

so

$$\begin{aligned} \beta_1 \ln p_1 + \beta_2 \ln p_2 + \beta_3 \ln m &= \beta_1 \ln kp_1 + \beta_2 \ln kp_2 + \beta_3 \ln km \\ &= (\ln k)(\beta_1 + \beta_2 + \beta_3) + \beta_1 \ln p_1 + \beta_2 \ln p_2 + \beta_3 \ln m. \end{aligned}$$

The only way to guarantee this for arbitrary k is to set

$$\beta_1 + \beta_2 + \beta_3 = 0,$$

which is a *parameter restriction.* In particular, this is a linear equality restriction, which is probably the most commonly encountered case.

3.6.2 Imposition

The general formulation of linear equality restrictions is the model

$$\begin{aligned} y &= X\beta + \varepsilon \\ R\beta &= r \end{aligned}$$

where R is a $Q \times K$ matrix, $Q < K$ and r is a $Q \times 1$ vector of constants.

- We assume R is of rank Q, so that there are no redundant restrictions.
- We also assume that $\exists\beta$ that satisfies the restrictions: they aren't infeasible.

  Draw a picture for two var. model with $\beta = 1$ *as a restriction to motivate the*

relevant distances for an statistic to test the restrictions.

Let's consider how to estimate β subject to the restrictions $R\beta = r$. The most obvious approach is to set up the Lagrangean

$$\min_{\widetilde{\beta}} s(\widetilde{\beta}) = \left(y - X\widetilde{\beta}\right)' \left(y - X\widetilde{\beta}\right) + 2\lambda'(R\widetilde{\beta} - r).$$

The Lagrange multipliers are scaled by 2, which makes thing less messy. The f.o.n.c. are

$$\begin{aligned} D_{\widetilde{\beta}} s(\hat{\beta}, \hat{\lambda}) &= -2X'y + 2X'X\hat{\beta}_R + 2R'\hat{\lambda} \equiv 0 \\ D_{\lambda} s(\hat{\beta}, \hat{\lambda}) &= R\hat{\beta}_R - r \equiv 0, \end{aligned}$$

which can be written as

$$\begin{bmatrix} X'X & R' \\ R & 0 \end{bmatrix} \begin{bmatrix} \hat{\beta}_R \\ \hat{\lambda} \end{bmatrix} = \begin{bmatrix} X'y \\ r \end{bmatrix}.$$

We get

$$\begin{bmatrix} \hat{\beta}_R \\ \hat{\lambda} \end{bmatrix} = \begin{bmatrix} X'X & R' \\ R & 0 \end{bmatrix}^{-1} \begin{bmatrix} X'y \\ r \end{bmatrix}.$$

Aside:**Partition inverse matrices**

Consider the following partitioned matrix: $\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}$

Let's define pivot elements the elements in the main diagonal:

i) Construct the first tableau, pivot element is 1,1 (a):

- element a becomes a^{-1}

- in pivot row do: $(pivot-element)^{-1}*(pivot-row-element):\begin{bmatrix} a^{-1} & a^{-1}b & a^{-1}c \end{bmatrix}$

-in pivot column: $-(pivot-column-element)*(pivot-element)^{-1}:$

$$\begin{bmatrix} a^{-1} & a^{-1}b & a^{-1}c \\ -da^{-1} & & \\ -ga^{-1} & & \end{bmatrix}$$

-off elements: $(off-elements)-[(element-in-same-pivot-row)*(pivot-element)^{-1}$

$*(element-same-pivot-col)]$:

$$\begin{bmatrix} a^{-1} & a^{-1}b & a^{-1}c \\ -da^{-1} & e-da^{-1}b & f-da^{-1c} \\ -ga^{-1} & h-ga^{-1b} & i-ga^{-1}c \end{bmatrix}$$

ii) Construct the next tableau with element 2,2 (e) as the pivot element.

iii) Construct the last tableau with element 3,3 (i) as the pivot element. This will be the final partitioned inverse.

Another way to do it is the following:

Stepwise Inversion:

note that

$$\begin{bmatrix} (X'X)^{-1} & 0 \\ -R(X'X)^{-1} & I_Q \end{bmatrix}\begin{bmatrix} X'X & R' \\ R & 0 \end{bmatrix} \equiv AB$$

$$= \begin{bmatrix} I_K & (X'X)^{-1}R' \\ 0 & -R(X'X)^{-1}R' \end{bmatrix}$$

$$\equiv \begin{bmatrix} I_K & (X'X)^{-1}R' \\ 0 & -P \end{bmatrix}$$

$$\equiv C,$$

and

$$\begin{bmatrix} I_K & (X'X)^{-1}R'P^{-1} \\ 0 & -P^{-1} \end{bmatrix} \begin{bmatrix} I_K & (X'X)^{-1}R' \\ 0 & -P \end{bmatrix} \equiv DC$$
$$= I_{K+Q},$$

so

$$\begin{aligned} DAB &= I_{K+Q} \\ DA &= B^{-1} \\ B^{-1} &= \begin{bmatrix} I_K & (X'X)^{-1}R'P^{-1} \\ 0 & -P^{-1} \end{bmatrix} \begin{bmatrix} (X'X)^{-1} & 0 \\ -R(X'X)^{-1} & I_Q \end{bmatrix} \\ &= \begin{bmatrix} (X'X)^{-1} - (X'X)^{-1}R'P^{-1}R(X'X)^{-1} & (X'X)^{-1}R'P^{-1} \\ P^{-1}R(X'X)^{-1} & -P^{-1} \end{bmatrix}, \end{aligned}$$

so

$$\begin{aligned} \begin{bmatrix} \hat{\beta}_R \\ \hat{\lambda} \end{bmatrix} &= \begin{bmatrix} (X'X)^{-1} - (X'X)^{-1}R'P^{-1}R(X'X)^{-1} & (X'X)^{-1}R'P^{-1} \\ P^{-1}R(X'X)^{-1} & -P^{-1} \end{bmatrix} \begin{bmatrix} X'y \\ r \end{bmatrix} \\ &= \begin{bmatrix} \hat{\beta} - (X'X)^{-1}R'P^{-1}\left(R\hat{\beta} - r\right) \\ P^{-1}\left(R\hat{\beta} - r\right) \end{bmatrix} \\ &= \begin{bmatrix} \left(I_K - (X'X)^{-1}R'P^{-1}R\right) \\ P^{-1}R \end{bmatrix} \hat{\beta} + \begin{bmatrix} (X'X)^{-1}R'P^{-1}r \\ -P^{-1}r \end{bmatrix} \end{aligned}$$

The fact that $\hat{\beta}_R$ and $\hat{\lambda}$ are linear functions of $\hat{\beta}$ makes it easy to determine their distributions, since the distribution of $\hat{\beta}$ is already known. Recall that for x a random vector, and for A and b a matrix and vector of constants, respectively, $Var(Ax+b) =$

$AVar(x)A'$.

Though this is the obvious way to go about finding the restricted estimator, an easier way, if the number of restrictions is small, is to impose them by substitution. Write

$$\begin{aligned} y &= X_1\beta_1 + X_2\beta_2 + \varepsilon \\ \begin{bmatrix} R_1 & R_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} &= r \end{aligned}$$

where R_1 is $Q \times Q$ nonsingular. Supposing the Q restrictions are linearly independent, one can always make R_1 nonsingular by reorganizing the columns of X. Then

$$\beta_1 = R_1^{-1}r - R_1^{-1}R_2\beta_2.$$

Substitute this into the model

$$\begin{aligned} y &= X_1R_1^{-1}r - X_1R_1^{-1}R_2\beta_2 + X_2\beta_2 + \varepsilon \\ y - X_1R_1^{-1}r &= \left[X_2 - X_1R_1^{-1}R_2\right]\beta_2 + \varepsilon \end{aligned}$$

or with the appropriate definitions,

$$y_R = X_R\beta_2 + \varepsilon.$$

This model satisfies the classical assumptions, *supposing the restriction is true*. One can estimate by OLS. The variance of $\hat{\beta}_2$ is as before

$$V(\hat{\beta}_2) = \left(X_R'X_R\right)^{-1}\sigma^2$$

and the estimator is

$$\hat{V}(\hat{\beta}_2) = \left(X_R' X_R\right)^{-1} \hat{\sigma}^2$$

where one estimates σ_0^2 in the normal way, using the restricted model, *i.e.*,

$$\widehat{\sigma^2} = \frac{\left(y_R - X_R \widehat{\beta_2}\right)' \left(y_R - X_R \widehat{\beta_2}\right)}{n - (K - Q)}$$

To recover $\hat{\beta}_1$, use the restriction. To find the variance of $\hat{\beta}_1$, use the fact that it is a linear function of $\hat{\beta}_2$, so

$$\begin{aligned} V(\hat{\beta}_1) &= R_1^{-1} R_2 V(\hat{\beta}_2) R_2' \left(R_1^{-1}\right)' \\ &= R_1^{-1} R_2 \left(X_2' X_2\right)^{-1} R_2' \left(R_1^{-1}\right)' \sigma_0^2 \end{aligned}$$

3.6.3 Properties of the restricted estimator

We have that

$$\begin{aligned} \hat{\beta}_R &= \hat{\beta} - (X'X)^{-1} R' P^{-1} \left(R\hat{\beta} - r\right) \\ &= \hat{\beta} + (X'X)^{-1} R' P^{-1} r - (X'X)^{-1} R' P^{-1} R (X'X)^{-1} X' y \\ &= \beta + (X'X)^{-1} X'\varepsilon + (X'X)^{-1} R' P^{-1} [r - R\beta] - (X'X)^{-1} R' P^{-1} R (X'X)^{-1} X'\varepsilon \\ \hat{\beta}_R - \beta &= (X'X)^{-1} X'\varepsilon \\ &+ (X'X)^{-1} R' P^{-1} [r - R\beta] \\ &- (X'X)^{-1} R' P^{-1} R (X'X)^{-1} X'\varepsilon \end{aligned}$$

Mean squared error is

$$MSE(\hat{\beta}_R) = \mathcal{E}[(\hat{\beta}_R - \beta)(\hat{\beta}_R - \beta)']$$

Noting that the crosses between the second term and the other terms expect to zero, and that the cross of the first and third has a cancellation with the square of the third, we obtain

$$\begin{aligned} MSE(\hat{\beta}_R) &= (X'X)^{-1}\sigma^2 \\ &+ (X'X)^{-1}R'P^{-1}[r-R\beta][r-R\beta]'P^{-1}R(X'X)^{-1} \\ &- (X'X)^{-1}R'P^{-1}R(X'X)^{-1}\sigma^2 \end{aligned}$$

So, the first term is the OLS covariance. The second term is PSD, and the third term is NSD.

- If the restriction is true, the second term is 0, so we are better off. *True restrictions improve efficiency of estimation.*
- If the restriction is false, we may be better or worse off, in terms of MSE, depending on the magnitudes of $r - R\beta$ and σ^2.

Observe that the MSE allows the compareson between biased and unbiased estimators. Consider a scalar

$$MSE(\widehat{\theta}_j) = V(\widehat{\theta}_j) + (E(\widehat{\theta}_j) - \theta_j))^2 = V(\widehat{\theta}_j) + bias^2$$

make a picture of the two suposed distributions.

3.6.4 Testing

In many cases, one wishes to test economic theories. If theory suggests parameter restrictions, as in the above homogeneity example, one can test theory by testing parameter restrictions. A number of tests are available.

3.6.5 t-test

Suppose one has the model

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

and one wishes to test the *single restriction* $H_0 : R\beta = r$ vs. $H_A : R\beta \neq r$. Under H_0, with normality of the errors,

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

so

$$\frac{R\hat{\beta} - r}{\sqrt{R(X'X)^{-1}R'\sigma^2}} = \frac{R\hat{\beta} - r}{\sigma^2\sqrt{R(X'X)^{-1}R'}} \sim N(0,1).$$

The problem is that σ^2 is unknown. One could use the consistent estimator $\widehat{\sigma^2}$ in place of σ^2, but the test would only be valid asymptotically in this case.

Proposition 2

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

as long as the $N(0,1)$ *and the* $\chi^2(q)$ *are independent.*

We need a few results on the χ^2 distribution.

Proposition 3 *If the n dimensional random vector* $x \sim N(0,V)$*, then* $x'V^{-1}x \sim \chi^2(n)$.

We'll prove this one as an indication of how the following unproven propositions could be proved.

Proof. Factor V^{-1} as PP' (this is the Cholesky factorization). Then consider $y = P'x$. We have

$$y \sim N(0, P'VP)$$

but

$$\begin{aligned} VPP' &= I_n \\ P'VPP' &= P' \end{aligned}$$

so $PVP' = I_n$.

$$\begin{aligned} y &\sim N(0, I_n) \\ y'y &= x'PP'x \\ &= xV^{-1}x \\ &\sim \chi^2(n) \end{aligned}$$

A more general proposition which implies this result is

Proposition 4 *If the n dimensional random vector* $x \sim N(0,V)$, *then*

$$x'Bx \sim \chi^2(\rho(B)) \tag{2}$$

if and only if BV *is idempotent.*

An immediate consequence is

Proposition 5 *If the random vector (of dimension n)* $x \sim N(0,I)$, *and B is idempotent with rank r, then*

$$x'Bx \sim \chi^2(r). \tag{3}$$

Consider the random variable

$$\begin{aligned}\frac{\hat{\varepsilon}'\hat{\varepsilon}}{\sigma^2} &= \frac{\varepsilon' M_X \varepsilon}{\sigma^2} \\ &= \left(\frac{\varepsilon}{\sigma}\right)' M_X \left(\frac{\varepsilon}{\sigma}\right) \\ &\sim \chi^2(n-K)\end{aligned}$$

Proposition 6 *If the random vector (of dimension n)* $x \sim N(0,I)$*, then* Ax *and* $x'Bx$ *are independent if* $AB = 0$.

Now consider (remember that we have only one restriction in this case)

$$\frac{\frac{R\hat{\beta}-r}{\sigma\sqrt{R(X'X)^{-1}R'}}}{\sqrt{\frac{\hat{\varepsilon}'\hat{\varepsilon}}{(n-K)\sigma^2}}} = \frac{R\hat{\beta}-r}{\hat{\sigma}\sqrt{R(X'X)^{-1}R'}}$$

This will have the $t(n-K)$ distribution if $\hat{\beta}$ and $\hat{\varepsilon}'\hat{\varepsilon}$ are independent. But $\hat{\beta} = \beta + (X'X)^{-1}X'\varepsilon$ and

$$(X'X)^{-1}X'M_X = 0,$$

so

$$\frac{R\hat{\beta}-r}{\hat{\sigma}\sqrt{R(X'X)^{-1}R'}} = \frac{R\hat{\beta}-r}{\hat{\sigma}_{R\hat{\beta}}} \sim t(n-K)$$

In particular, for the commonly encountered *test of significance* of an individual coefficient, for which $H_0 : \beta_j = 0$ vs. $H_0 : \beta_j \neq 0$, the test statistic is

$$\frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim t(n-K)$$

Note: the $t-$ test is strictly valid only if the errors are actually normally distributed. If one has nonnormal errors, one needs asymptotic results to justify taking critical values from the $N(0,1)$ distribution, since $t(n-K) \xrightarrow{d} N(0,1)$ as $n \to \infty$. In practice, a

conservative procedure is to take critical values from the t distribution if nonnormality is suspected. This will reject H_0 less often since the t distribution is fatter-tailed than is the normal.

3.6.6 p-value (as a decision rule)

Instead of finding the t-statistic you can calculate

$$\begin{aligned} p-value &= Pr(t > |t_j|) * 2 \\ Pr(-|t_j| < t < |t_j|\ ) &= 1-p \end{aligned}$$

3.6.7 *F* test

The F test allows testing multiple restrictions jointly.

Interpretation of test statistics

Now that we have a menu of test statistics, we need to know how to use them.

Proposition 7 *If $x \sim \chi^2(r)$ and $y \sim \chi^2(s)$, then*

$$\frac{x/r}{y/s} \sim F(r,s) \tag{4}$$

provided that x and y are independent.

Proposition 8 *If the random vector (of dimension n) $x \sim N(0,I)$, then $x'Ax$ and $x'Bx$ are independent if $AB = 0$.*

Using these results, and previous results on the χ^2 distribution, it is simple to show that the following statistic has the F distribution:

$$F = \frac{\left(R\hat{\beta}-r\right)'\left(R(X'X)^{-1}R'\right)^{-1}\left(R\hat{\beta}-r\right)}{q\hat{\sigma}^2} \sim F(q,n-K).$$

A numerically equivalent expression is

$$\frac{(RSS_R - RSS_U)/q}{RSS_U/(n-K)} \sim F(q, n-K).$$

3.6.8 Confidence intervals

Confidence intervals for single coefficients are generated in the normal manner. Given the t statistic

$$t - statistic = \frac{\hat{\beta}_j - a}{\widehat{\sigma_{\hat{\beta}_j}}}$$

a $100(1-\alpha)\%$ confidence interval for β is defined by the bounds of the set of a such that the $t - statistic$ does not reject $H_0 : \beta_j = a$, using a α significance level:

$$\begin{aligned} C(\alpha) &= \{\beta_j : -t_{\alpha/2,(n-k)} < \frac{\hat{\beta}_j - a}{\widehat{\sigma_{\hat{\beta}_j}}} < t_{\alpha/2,(n-k)}\} \\ Pr(-t_{\alpha/2,(n-k)} < \frac{\hat{\beta}_j - a}{\widehat{\sigma_{\hat{\beta}_j}}} < t_{\alpha/2,(n-k)}) &= 1-\alpha \end{aligned}$$

inside the expression of the probability, multiply by $\widehat{\sigma_{\hat{\beta}_j}}$, subtract $\hat{\beta}_j$, multiply by -1 and reorder.

The set of such β_j is the interval

$$\hat{\beta}_j \pm \widehat{\sigma_{\hat{\beta}_j}} t_{\alpha/2}$$

A confidence ellipse for two coefficients jointly would be, analogously, the set of $\{\beta_1, \beta_2\}$ such that the F (or some other test statistic) doesn't reject at the specified critical value. This generates an ellipse, if the estimators are correlated. *Draw a picture here.*

- The region is an ellipse, since the CI for an individual coefficient defines a (in-

finitely long) rectangle with total prob. mass $1-\alpha$, since the other coefficient is marginalized (e.g., can take on any value). Since the ellipse is bounded in both dimensions but also contains mass $1-\alpha$, it must extend beyond the bounds of the individual CI.

- From the pictue we can see that:
  - Rejection of hypotheses individually does not imply that the joint test will reject.
  - Joint rejection does not imply individal tests will reject.

Example: (Hayashi pg 45) assume $k=2$ and consider:

$$\begin{aligned} H_0 : \beta_1 &= 1 \\ \beta_2 &= 0 \end{aligned}$$

this can be written as a linear hypothesis $R\beta = r$ for $R = I_2$ and $r = (1,0)\prime$ so the *F test* should be used. It is tempting, however, to conduct the $t-test$ separately for each individual coefficient of the hypothesis. We might accept H_0if both restrictions $\beta_1 = 1$ *and* $\beta_2 = 0$ pass the $t-test$. This amounts to using the confidence region of:

$$\begin{aligned} CI[(\beta_1,\beta_2) \quad &: \quad (\hat{\beta}_1 - \widehat{\sigma_{\hat{\beta}_1}} t_{\alpha/2,(n-k)} < \beta_1 < (\hat{\beta}_1 + \widehat{\sigma_{\hat{\beta}_1}} t_{\alpha/2,(n-k)}) \\ &: \quad (\hat{\beta}_2 - \widehat{\sigma_{\hat{\beta}_2}} t_{\alpha/2,(n-k)} < \beta_2 < (\hat{\beta}_2 + \widehat{\sigma_{\hat{\beta}_2}} t_{\alpha/2,(n-k)})] \end{aligned}$$

which is a rectangular region in the (β_1,β_2) plane.

On the other hand, the confidence region for the $F-test$ is

$$[(\beta_1,\beta_2):(\hat{\beta}_1-\beta_1,\hat{\beta}_2-\beta_2)\widehat{(V(\hat{\beta}))}^{-1}\begin{pmatrix}\hat{\beta}_1-\beta_1\\ \hat{\beta}_2-\beta_2\end{pmatrix}]<2F_{\alpha(q,n-k)}$$

the acceptance region is an elipse in the (β_1,β_2) plane.

3.6.9 Examples

When considering a set of restrictions we have to be careful about not stating redundant or inconsistent equations. Let's consider the wage equation in the example of the introduction

$$log(wage_i)=\beta_1+\beta_2 S_i+\beta_3 TEN_i+\beta_4 EXP_4+\varepsilon_i$$

We might wish to test that education and tenure have equal impact on the wage rate and that there is no experience effect:

$$\begin{aligned}H_0:\beta_2 &= \beta_3 \Rightarrow \beta_2-\beta_3=0\\ \beta_4 &= 0\end{aligned}$$

since the two rows are linearly independent the rank condition is satisfied.

Suppose now that additionally we write $\beta_2-\beta_3=\beta_4$ If you construct the R matrix you will see that is a $3x4$matrix but the rank is 2. This are called *redundant restrictions*.

You can have also *inconsistent restrictions* That happen when there is no β that can

satisfy those restrictions example

$$\begin{aligned} H_0 : \beta_2 &- \beta_3 = 0 \\ \beta_4 &= 0 \\ \beta_4 &= 0.5 \end{aligned}$$

3.7 Maximum Likelihood estimation (brief review)

Motivation:

- We will see that $\widehat{\beta}_{OLS}$ attains the Cramer-Rao lower bound, $\widehat{\beta}_{OLS}$ is the Best Unbiased Estimator (BUE).
- The $F-test$ is a likelihood ratio test when errors are normal.

3.7.1 The log-likelihood for the regression model

Since we have the assumption of N of the errors we can use the maximum likelihood (ML) principle to estimate the parameters of the model

$$\begin{aligned} y &= X\beta + \varepsilon \\ \varepsilon &\sim N(0, \sigma^2 I_n) \end{aligned}$$

- we will see that the ML estimator of the regression model $\widetilde{\beta}_{ML}$is the same as the $\widehat{\beta}_{OLS}$
- and that $\widetilde{\sigma}^2_{ML}$ differs slightly from $\widehat{\sigma}^2_{OLS}$.

The ML primciple:

We want to choose the parameter estimates that maximize the probability of obtaining the observed sample.

Definitions:

- θ is the true parameter vector
- y is a random vector of sample size n, $y = (y_1, y_2, .., y_n)$
- $f(y;\theta)$ is the joint density characterized by a parameter vector θ.
- the likelihood function is just this density evaluated at any other values $\widetilde{\theta}$

$$\mathcal{L}(y;\theta) = f(y;\widetilde{\theta})$$

aside: here again we are considering $X's$ fixed, if $X's$ are stochastic then we have

$$f(y,X;\eta) = f(y/X;\theta) \cdot f(X;\psi)$$

we don't know $f(X;\psi)$, we can find $\widetilde{\theta}_{ML}$ if there is no functional relationship between θ and ψ, that is $X's$ are exogenous.

- the likelihood function gives us the value that the density takes when we introduce our sample observations given any hypothetical value $\widetilde{\theta}$
- the *ML* estimate of the true parameter vector θ is $\widetilde{\theta}_{ML}$ the value for the parameters that maximizes the likelihood given the data.

The log-likelihood

$$\begin{aligned} y &= X\beta+\varepsilon \\ \varepsilon &\sim N(0,\sigma^2 I_n) \\ y &\sim N(X\beta,\sigma^2 I_n) \end{aligned}$$

we have observed

- $y_1, y_2, \ldots, y_n$
- $y's$ are independent
- consider $f(y_1, y_2, \ldots, y_n; \theta)$, where $\theta = \beta_1, \beta_2, \ldots, \beta_k, \sigma^2$, in matrix notation

$$f(y) = (2\pi\sigma^2)^{-n/2} exp - \frac{1}{2\sigma^2}(y - X\beta)\prime(y - X\beta)$$

replacing the true parameters β, σ^2 by the hypothetical parameters $\widetilde{\beta}, \widetilde{\sigma}^2$ and taking logs, we obtain the log-likelihood function

$$logL(\widetilde{\beta}, \widetilde{\sigma}^2) = -\frac{n}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y - X\widetilde{\beta})\prime(y - X\widetilde{\beta})$$

the *ML* estimator of (β, σ^2) is the $(\beta, \widetilde{\sigma}^2)_{ML}$ that maximizes this log-likelihood.

F.O.N.C:

define: $\widetilde{\gamma} = \widetilde{\sigma}^2$

$$\begin{aligned} (1)\ \frac{\partial lnL}{\partial \widetilde{\beta}} &= -\frac{1}{2\widetilde{\sigma}^2}[-2X\prime y - 2X\prime X\widetilde{\beta}]\ |_{\widetilde{\beta}=\widetilde{\beta}_{ML}} = 0 \\ (2)\ \frac{\partial lnL}{\partial \widetilde{\gamma}} &= \frac{-n}{2\widetilde{\gamma}} + \frac{1}{2\widetilde{\gamma}^2}(y - X\beta)\prime(y - X\beta)\ |_{\widetilde{\gamma}=\widetilde{\gamma}_{ML}} = 0 \end{aligned}$$

let's rewrite the log-likelihood

$$\begin{aligned} lnL(\widetilde{\beta},\widetilde{\sigma}^2) &= -\frac{n}{2}ln(2\pi)-\frac{n}{2}ln(\widetilde{\gamma})-\frac{1}{2\widetilde{\gamma}}(y-X\widetilde{\beta})\prime(y-X\widetilde{\beta}) \\ (2)\ \frac{\partial lnL}{\partial\widetilde{\gamma}} &= -\frac{n}{2}\frac{1}{\widetilde{\gamma}}+\frac{2}{2^2\widetilde{\gamma}^2}(y-X\widetilde{\beta})\prime(y-X\widetilde{\beta}) \\ &= -\frac{n}{2}\frac{1}{\widetilde{\sigma}^2}+\frac{1}{2\widetilde{\sigma}^4}(y-X\widetilde{\beta})\prime(y-X\widetilde{\beta}) \end{aligned}$$

from (1)

$$\begin{aligned} X\prime y &= X\prime X\widetilde{\beta}_{ML} \\ \widetilde{\beta}_{ML} &= (X\prime X)^{-1}X\prime y \end{aligned}$$

from (2)

$$\begin{aligned} \frac{n}{\widetilde{\sigma}^2} &= \frac{(y-X\widetilde{\beta})\prime(y-X\widetilde{\beta})}{\widetilde{\sigma}^4} \\ \frac{\widetilde{\sigma}^4}{\widetilde{\sigma}^2} &= \widetilde{\sigma}^2=\frac{\widetilde{\varepsilon}_{ML}\prime\widetilde{\varepsilon}_{ML}}{n} \end{aligned}$$

and since

$$\widehat{\beta}_{OLS}=\widetilde{\beta}_{ML}\Rightarrow\widetilde{\varepsilon}_{ML}\prime\widetilde{\varepsilon}_{ML}=\hat{\varepsilon}'_{OLS}\hat{\varepsilon_{OLS}}$$

$\widetilde{\sigma}^2_{ML}=\frac{n-k}{n}\widehat{\sigma}^2_{OLS}$ it is a biased estimator of σ^2 (not when n goes to infinity latter).

3.7.2 The Cramer-Rao Bound for the Classical Regression Model

The Cramer-Rao Inequality

Let z be a vector of random variables (not necessarily independent) which joint density is given by $f(z;\theta)$ where θ is an m-dimensional vector of parameters in some parameter space Θ. Let $L(\widetilde{\theta})\equiv f(z;\widetilde{\theta})$ be the likelihood function and let $\widehat{\theta}(z)$ be an

unbiased estimator of θ with a finite var-cov matrix. Then under some regularity conditions on $f(z;\theta)$

$$Var[(\widehat{\theta}(z)] \geq I(\theta)^{-1} \equiv Cramer - Rao\ Lower\ Bound$$

where $I(\theta)$ is the information matrix

$$I(\theta) = -E[\frac{\partial^2 lnL(\widetilde{\theta})}{\partial\widetilde{\theta}\partial\widetilde{\theta}\prime}] \mid_{\widetilde{\theta}=\theta} = -E[\frac{\partial^2 lnL(\theta)}{\partial\widetilde{\theta}\partial\widetilde{\theta}\prime}]$$

those regularity conditions are to guarantee that operations of differentiation and taking expectations can be interchanged.

For the classical regression model it can be shown that the regularity conditions are satisfied for the normal density.

The parameter vector θ is $(\beta\prime, \sigma^2)\prime$ therefore $\widetilde{\theta} = (\widetilde{\beta}\prime, \widetilde{\gamma})\prime$

$$\frac{\partial^2 lnL(\theta)}{\partial\widetilde{\theta}\partial\widetilde{\theta}\prime} = \begin{bmatrix} \frac{\partial^2 lnL(\theta)}{\partial\widetilde{\beta}\partial\widetilde{\beta}\prime} & \frac{\partial^2 lnL(\theta)}{\partial\widetilde{\beta}\partial\widetilde{\gamma}} \\ \frac{\partial^2 lnL(\theta)}{\partial\widetilde{\gamma}\partial\widetilde{\beta}\prime} & \frac{\partial^2 lnL(\theta)}{\partial^2\widetilde{\gamma}} \end{bmatrix}$$

$$\begin{aligned}
(1)\ \frac{\partial lnL(\theta)}{\partial\widetilde{\beta}} &= -\frac{1}{\gamma}X\prime(y - X\beta) \\
(2)\ \frac{\partial lnL(\theta)}{\partial\widetilde{\gamma}} &= \frac{-n}{2\gamma} + \frac{1}{2\gamma^2}(y - X\beta)\prime(y - X\beta) \\
(1,1)\ \frac{\partial^2 lnL(\theta)}{\partial\widetilde{\beta}\partial\widetilde{\beta}\prime} &= -\frac{1}{\gamma}(X\prime X) \\
(2,2)\ \frac{\partial^2 lnL(\theta)}{\partial^2\widetilde{\gamma}} &= \frac{n}{2\gamma^2} - \frac{1}{\gamma^3}(y - X\beta)\prime(y - X\beta) \\
(1,2) = (2,1)\ \frac{\partial^2 lnL(\theta)}{\partial\widetilde{\gamma}\partial\widetilde{\beta}\prime} &= -\frac{1}{\gamma^2}X\prime(y - X\beta)
\end{aligned}$$

Since evaluated at the true parameters $y - X\beta = \varepsilon$ and $E(\varepsilon) = 0$ and $E(\varepsilon\prime\varepsilon) = n\sigma^2$

$$I(\theta) = -E[\frac{\partial^2 lnL(\theta)}{\partial\widetilde{\theta}\partial\widetilde{\theta}\prime}] = \begin{bmatrix} \frac{1}{\sigma^2}(X\prime X) & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}$$

this block diagonal matrix can be easily inverted to obtain the Cramer-Rao lower bound

$$C - R\ bound \equiv I(\theta)^{-1} = \begin{bmatrix} \sigma^2(X\prime X)^{-1} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}$$

4 Ordinary Least Squares: large samples

The importance of OLS in econometrics is that it has good properties for a class of models different than the classical. This is useful in economics since often the assumptions of the exact distribution are not satisfied. The finite sample properties cannot be claimed if the errors are not normal or if the model is not linear. We will relax now the assumption of normality of the error term and we will derive an approximation to the distribution of the estimator and its associate statistics suppossing that the sample is sufficiently large. This is the asymptotic or large-sample theory approach.

We will see that $\widehat{\beta}_{OLS}$ in large samples is:

- consistent
- asymptotically normal
- asymptotically efficient (definition only)

Some results needed:

Lemma 1: (preservation of convergence for continuous transformations)

Supose $a(.)$ is a continuous vector-valued continuous function that does not depend on n

(a)

$$z \to_p \alpha \quad \Rightarrow \quad a(z_n) \to_p a(\alpha)$$
$$plim\, a(z_n) \quad = \quad a(plim(z_n))$$

provided the $plim$exists

(b)

$$z_n \to_d z \Rightarrow a(z_n) \to_d a(z)$$

Lemma 2:

(a)

$$x_n \to_d x,\ y_n \to_p \alpha \quad \Rightarrow \quad x_n + y_n \to_d x + \alpha$$

(b)

$$x_n \to_d x,\ y_n \to_p 0 \quad \Rightarrow \quad y_n \prime x_n \to_p 0$$

(c)

$$x_n \to_d x,\ A_n \to_p A \quad \Rightarrow \quad A_n x_n + y_n \to_d Ax$$

provided $A_n$and $x_n$are comformable, in particular

$$if\ x \sim N(0\Sigma)\ then\ A_n x_n \to_d N(0, A\Sigma A\prime)$$

(d)

$$x_n \to_d x,\ A_n \to_p A \quad \Rightarrow \quad x_n\prime A_n^{-1} x_n \to_d x\prime A^{-1} x$$

provided $A_n$and $x_n$are comformable and Anon-singular.

Law of Large Numbers (a version of the Chebyshev WLLN)

- $\{z_i\}$is a sequence of random scalars (extendable to vectors)
- independent
- $E(z_i) = \mu$
- $V(x_i)$ finite

$$\overline{z_n} = \frac{\sum_i z_i}{n} \to_p \mu$$

Central Limit Theorem (a version of the CLT, the Lindeberg-Feller theorem)

- $\{z_i\}$is a sequence of random vectors
- independent
- $E(z_i) = \mu_i$
- $Var - cov(x_i) = \Sigma_i$

$$\begin{aligned} \sqrt{n}(\overline{z_n} - \mu_i \quad &= \quad \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (z_i - \mu_i) \\ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (z_i - \mu_i) \quad &\to_d \quad N(0, \Sigma) \end{aligned}$$

where $\Sigma = \lim_{n\to\infty}(\overline{\Sigma}_n) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \Sigma_i$

4.1 Asymptotic properties of the OLS estimator (H-W # 3: simulation)

We can view estimators as sequences of random variables (Hayashi chapter 2):

Let $\hat{\theta}_n$be an estimator of a parameter vector θ based on a sample size n. The sequence $\{\hat{\theta}_n\}$is an example of a sequence of random variables:

- $\hat{\theta}_n$is a consistent estimator for θif $plim_{n\to\infty}\hat{\theta}_n = \theta$. And the asymptotic bias is defined as $plim_{n\to\infty}\hat{\theta}_n - \theta$.
- If an estimator is consistent then the asymptotic bias is zero.
- $\hat{\theta}_n$is asymptotically normal if $\sqrt{n}(\hat{\theta}_n - \theta) \to_d N(0,\Sigma)$. This is a definition and the estimator is called $\sqrt{n} - consistent$. The acronim is CAN. The matrix $\Sigma$is called the asymptotic variance $Avar(\hat{\theta}_n)$. Some authors use $Avar(\hat{\theta}_n) = \frac{\Sigma}{n}$ which is zero in the limit (not useful for seing the properties of the limitting distribution before it collapses). We will follow the definition of $Avar(\hat{\theta}_n) = asymptotic\ variance(\sqrt{n}(\hat{\theta}_n - \theta))$.

4.1.1 Consistency of OLS

Define:

$$\begin{aligned} S_{XX} &= \frac{1}{n}X\prime X \\ S_{Xy} &= \frac{1}{n}X\prime y \end{aligned}$$

then

$$\widehat{\beta} = (X\prime X^{-1}(X\prime y) = S_{XX}^{-1}S_{Xy}$$

Assumptions:

A1: Linearity

A2: full rank of X

A3: $E(X\prime\varepsilon) = 0$

A4: $E(\varepsilon\varepsilon\prime) = \sigma^2 I_n$

A5: $\varepsilon \sim IID$

A6: X are fixed

A7: $\lim_{n\to\infty} \frac{X\prime X}{n} = \Sigma_{XX}$ a positive definite (finite) matrix.

See that:

$$S_{XX} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i\prime$$

notation that will be useful latter.

$$\begin{aligned} \widehat{\beta} &= \beta + (X\prime X^{-1}(X\prime y) \\ &= \beta + \left(\frac{X\prime X}{n}\right)^{-1}\frac{X\prime\varepsilon}{n} \end{aligned}$$

we want to find $plim(\widehat{\beta} - \beta)$

$$\begin{aligned} \widehat{\beta} - \beta &= \left(\frac{X\prime X}{n}\right)^{-1}\frac{X\prime\varepsilon}{n} \\ plim(\widehat{\beta} - \beta) &= plim\left(\frac{X\prime X}{n}\right)^{-1} plim\frac{X\prime\varepsilon}{n}. \end{aligned}$$

By assumption $\lim_{n\to\infty} \frac{X\prime X}{n} = \Sigma_{XX}$, by lemma 1.a.iv

$$\begin{aligned} if\ \left(\frac{X\prime X}{n}\right) &\to_p \Sigma_{XX} \\ \left(\frac{X\prime X}{n}\right)^{-1} &\to_p \Sigma_{XX}^{-1} \\ plim(\widehat{\beta} - \beta) &= \Sigma_{XX}^{-1} plim\frac{X\prime\varepsilon}{n} \\ \frac{X\prime\varepsilon}{n} &= \frac{1}{n}\sum_{i=1}^{n} x_i\varepsilon_i \end{aligned}$$

Define $g_i = x_i\varepsilon_i$then $\frac{1}{n}\sum_{i=1}^{n} x_i\varepsilon_i = \overline{g}$. Applying the law of large numbers (LLN) for IID random variables,

$$\overline{g} \rightarrow_p E(g_i)$$

since $X's$ are non-stochastic $E(g_i) = x_i E(\varepsilon_i) = 0$; therefore

$$\begin{aligned} \widehat{\beta} - \beta &= S_{XX}\overline{g} \\ plim(\widehat{\beta} - \beta) &= plim(S_{XX})\, plim(\overline{g}) \\ plim(\widehat{\beta} - \beta) &= \Sigma_{XX}\, 0 = 0 \end{aligned}$$

$\widehat{\beta}$ is consistent

4.1.2 Asymptotic Normality of OLS

Definition: A consistent estimator is asymptotically normal if

$$\sqrt{n}(\widehat{\theta}_n - \theta) \rightarrow_d N(0, \Sigma).$$

Such an estimator is called $\sqrt{n} - consistent$. The acronim used is *CAN* estimator (consistent and asymptotically normal). The variance Σ is by definition the asymptotic variance $Avar(\widehat{\theta}_n)$. It is the variance of the limiting distribution of $\sqrt{n}(\widehat{\theta}_n - \theta)$.

We want to see the limiting distribution of

$$\sqrt{n}(\widehat{\beta} - \beta) \rightarrow_d ?$$

we can write

$$\sqrt{n}(\widehat{\beta} - \beta) = \left(\frac{X\prime X}{n}\right)^{-1} \frac{X\prime\varepsilon}{\sqrt{n}}$$

$$\left(\frac{X\prime X}{n}\right)^{-1} \quad \rightarrow_p \quad \Sigma_{XX}$$
$$\frac{X\prime\varepsilon}{\sqrt{n}} = \sqrt{n} \quad (\overline{g} - 0) \rightarrow_d \quad N(0,\Sigma)$$

by the CLT stated above. And

$$\sqrt{n}(\widehat{\beta} - \beta) \quad \rightarrow_d \quad N(0, \Sigma_{XX}^{-1}\Sigma\Sigma_{XX}^{-1})$$

and by lemma 2(c). Where $\Sigma = \lim_{n\rightarrow\infty} V(\frac{X\prime\varepsilon}{\sqrt{n}})$

$$\begin{aligned}
\frac{X\prime\varepsilon}{n} &= \frac{1}{n}\sum_{i=1}^{n} x_i\varepsilon_i \\
V(x_i\varepsilon_i) &= \Sigma_i = E(x_i\varepsilon_i\varepsilon_i\prime x_i\prime) = E(\varepsilon_i^2 x_i x_i\prime) = \sigma^2(x_i x_i\prime) \\
V(\frac{1}{n}\sum_{i=1}^{n} x_i\varepsilon_i) &= \frac{1}{n}\sum_i \sigma^2(x_i x_i\prime) \\
&= \frac{1}{n}\sigma^2 X\prime X \\
&= \sigma^2\frac{X\prime X}{n} \\
Avar(\overline{g}) &= \lim_{n\rightarrow\infty} V(\sqrt{n}\overline{g}) \\
&= \lim_{n\rightarrow\infty} V(\frac{X\prime\varepsilon}{\sqrt{n}}) \\
&= \sigma^2 plim\frac{X\prime X}{n} \\
&= \sigma^2\Sigma_{XX}
\end{aligned}$$

therefore

$$\begin{aligned}
\sqrt{n}(\widehat{\beta} - \beta) &= \left(\frac{X\prime X}{n}\right)^{-1}\frac{X\prime\varepsilon}{\sqrt{n}} \rightarrow_d N(0, \Sigma_{XX}^{-1}\sigma^2\Sigma_{XX}\Sigma_{XX}^{-1}) \\
\sqrt{n}(\widehat{\beta} - \beta) &\rightarrow_d N(0, \sigma^2\Sigma_{XX}^{-1})
\end{aligned}$$

In summary the OLS estimator is normally distributed in small and large samples if

epsilons are normal. If epsilons are not normal $\widehat{\beta}$ is asyptotically normally distributed.

4.1.3 Asymptotic Efficiency

Definition: An estimator $\widehat{\theta}$ for $\theta$is asyptotically efficient if it is *CAN* and $Avar(\widetilde{\theta}) - Avar(\widehat{\theta})$ is a posotive semidefinite matrix for $\widetilde{\theta}$ any other *CAN* estimator.

- if $\varepsilon$are non-normal we need nonlinear estimation to achieve asymptotic efficiency.

4.2 Hypthesis testing

4.2.1 Wald-type tests

The Wald principle is based on the idea that if a restriction is true, the unrestricted model should "approximately" satisfy the restriction. Given that the least squares estimator is asymptotically normally distributed:

$$\sqrt{n}\left(\hat{\beta}-\beta\right) \xrightarrow{d} N\left(0,\sigma^2\Sigma_{XX}^{-1}\right)$$

then under $H_0 : R\beta = r$, we have

$$\sqrt{n}\left(R\hat{\beta}-r\right) \xrightarrow{d} N\left(0,\sigma^2 R\Sigma_{XX}^{-1}R'\right)$$

so by Proposition [3]

$$n\left(R\hat{\beta}-r\right)'\left(\sigma_0^2 R\Sigma_{XX}^{-1}R'\right)^{-1}\left(R\hat{\beta}-r\right) \xrightarrow{d} \chi^2(q)$$

Note that Σ_{XX}^{-1} or σ^2 are not observable. The test statistic we use substitutes the consistent estimators. Use $(X'X/n)^{-1}$ as the consistent estimator of Σ_{XX}^{-1}. With this, there

is a cancellation of $n's$, and the statistic to use is

$$\left(R\hat{\beta}-r\right)'\left(\widehat{\sigma_0^2}R(X'X)^{-1}R'\right)^{-1}\left(R\hat{\beta}-r\right)\xrightarrow{d}\chi^2(q)$$

- The Wald test is a simple way to test restrictions without having to estimate the restricted model.
- Note that this formula is similar to one of the formulae provided for the F test.

4.2.2 Score-type tests (Rao tests, Lagrange multiplier tests)

In some cases, an unrestricted model may be nonlinear in the parameters, but the model is linear in the parameters under the null hypothesis. For example, the model

$$y=(X\beta)^{\gamma}+\varepsilon$$

is nonlinear in β and γ, but is linear in β under $H_0:\gamma=1$. Estimation of nonlinear models is a bit more complicated, so one might prefer to have a test based upon the restricted, linear model. The score test is useful in this situation.

- Score-type tests are based upon the general principle that the gradient vector of the unrestricted model, evaluated at the restricted estimate, should be asymptotically normally distributed with mean zero, if the restrictions are true. The original development was for ML estimation, but the principle is valid for a wide variety of estimation methods.

We have seen that

$$\begin{aligned}\hat{\lambda} &= \left(R(X'X)^{-1}R'\right)^{-1}\left(R\hat{\beta}-r\right)\\ &= P^{-1}\left(R\hat{\beta}-r\right)\end{aligned}$$

Given that

$$\sqrt{n}\left(R\hat{\beta}-r\right) \xrightarrow{d} N\left(0,\sigma^2 R\Sigma_{XX}^{-1}R'\right)$$

under the null hypothesis,

$$\sqrt{n}\hat{\lambda} \xrightarrow{d} N\left(0,\sigma^2 P^{-1}R\Sigma_{XX}^{-1}R'P^{-1}\right)$$

Show as an exercise:

$$\hat{\lambda}'\left(\frac{R(X'X)^{-1}R'}{\sigma^2}\right)\hat{\lambda} \xrightarrow{d} \chi^2(q)$$

since the powers of n cancel. To get a usable test statistic substitute a consistent estimator of σ^2.

- This makes it clear why the test is sometimes referred to as a Lagrange multiplier test. It may seem that one needs the actual Lagrange multipliers to calculate this. If we impose the restrictions by substitution, these are not available. Note that the test can be written as

  $$\frac{\left(R'\hat{\lambda}\right)'(X'X)^{-1}R'\hat{\lambda}}{\sigma^2} \xrightarrow{d} \chi^2(q)$$

  However, we can use the fonc for the restricted estimator:

  $$-X'y+X'X\hat{\beta}_R+R'\hat{\lambda}$$

  to get that

  $$\begin{aligned} R'\hat{\lambda} &= X'(y-X\hat{\beta}_R) \\ &= X'\hat{\varepsilon}_R \end{aligned}$$

Substituting this into the above, we get

$$\frac{\hat{\varepsilon}_R' X (X'X)^{-1} X' \hat{\varepsilon}_R}{\sigma^2} \xrightarrow{d} \chi^2(q)$$

but this is simply

$$\hat{\varepsilon}_R' \frac{P_X}{\sigma^2} \hat{\varepsilon}_R \xrightarrow{d} \chi^2(q).$$

To see why the test is also known as a score test, note that the fonc for restricted least squares

$$-X'y + X'X\hat{\beta}_R + R'\hat{\lambda}$$

give us

$$R'\hat{\lambda} = X'y - X'X\hat{\beta}_R$$

and the rhs is simply the gradient (score) of the unrestricted model, evaluated at the restricted estimator. The scores evaluated at the unrestricted estimate are identically zero. The logic behind the score test is that the scores evaluated at the restricted estimate should be approximately zero, if the restriction is true. The test is also known as a Rao test, since P. Rao first proposed it in 1948.

4.2.3 Likelihood ratio-type tests

The Wald test can be calculated using the unrestricted model. The score test can be calculated using only the restricted model. The likelihood ratio test, on the other hand, uses both the restricted and the unrestricted estimators. The test statistic is

$$LR = 2\left(\ln L(\hat{\theta}) - \ln L(\tilde{\theta})\right)$$

where $\hat{\theta}$ is the unrestricted estimate and $\tilde{\theta}$ is the restricted estimate.

Final commments:

It can be shown, for linear regression models subject to linear restrictions, and if $\frac{\hat{\varepsilon}'\hat{\varepsilon}}{n}$ is used to calculate the Wald test and $\frac{\hat{\varepsilon}_R'\hat{\varepsilon}_R}{n}$ is used for the score test, that

$$W > LR > LM.$$

For this reason, the Wald test will always reject if the LR test rejects, and in turn the LR test rejects if the LM test rejects. This is a bit problematic: there is the possibility that by careful choice of the statistic used, one can manipulate reported results to favor or disfavor a hypothesis. A conservative/honest approach would be to report all three test statistics when they are available. In the case of linear models with normal errors the F test is to be preferred, since asymptotic approximations are not an issue.

The small sample behavior of the tests can be quite different. The true size (probability of rejection of the null when the null is true) of the Wald test is often dramatically higher than the nominal size associated with the asymptotic distribution. Likewise, the true size of the score test is often smaller than the nominal size.

4.2.4 Non-linear Hypothesis Testing (the Delta Method)

Testing nonlinear restrictions of a linear model is not much more difficult, at least when the model is linear. Since estimation subject to nonlinear restrictions requires nonlinear estimation methods, which are beyond the score of this course, we'll just consider the Wald test for nonlinear restrictions on a linear model.

Lemma 3: Suppose $\{x_n\}$ is a sequence of $k-dim$random vectors such that

$$x_n \to_p \beta$$

and

$$\sqrt{n}(x_n - \beta) \to_d z$$

Consider $a(.)$ a q-vector valued function. Suppose $a(.)$ has continuous first derivatives that we evaluate at β

$$D_{\beta'} a(\beta)\big|_{\beta} = A(\beta)$$

We suppose that the restrictions are not redundant in a neighborhood of β, so that

$$\rho(A(\beta)) = q$$

then

$$\sqrt{n}(a(x_n) - a(\beta)) \to_d A(\beta) z$$

In particular

$$\begin{aligned} \sqrt{n}(x_n - \beta) &\to_d \quad N(0, \Sigma) \\ \sqrt{n}(a(x_n) - a(\beta)) &\to_d \quad N(0, A(\beta)\Sigma A(\beta)\prime) \end{aligned}$$

Application: consider the set of $q$non-linear restrictions $a(\beta) = 0$

$$\begin{aligned} \sqrt{n}\left(\hat{\beta} - \beta\right) &\to_d N\left(0, \sigma^2 \Sigma_{XX}^{-1}\right) \\ \sqrt{n}(a(\hat{\beta}) - a(\beta)) &\to_d N(0, A(\beta)\sigma^2 \Sigma_{XX}^{-1} A(\beta)\prime) \end{aligned}$$

by proposition 3:

$$na(\hat{\beta})' \left(A(\beta)\sigma^2 \Sigma_{XX}^{-1} A(\beta)'\right)^{-1} a(\hat{\beta}) \xrightarrow{d} \chi^2(q)$$

substitute the unknown parameters for consisitent estimators:

$$\begin{aligned}
\widehat{\beta} &\rightarrow_p \beta \\
\widehat{\sigma^2} &\rightarrow_p \sigma^2 \\
S_{XX} &\rightarrow_p \Sigma_{XX} \\
A(\widehat{\beta}) &\rightarrow_p A(\beta)
\end{aligned}$$

then

$$na(\hat{\beta})' \left(A(\widehat{\beta})\widehat{\sigma^2}\frac{X\prime X}{n}A(\widehat{\beta})' \right)^{-1} a(\hat{\beta}) \xrightarrow{d} \chi^2(q)$$

The above result is based on a first order Taylor's series expansion of $a(\hat{\beta})$ about β:

$$a(\hat{\beta}) = a(\beta) + A(\beta^*)(\hat{\beta} - \beta)$$

where β^* is a convex combination of $\hat{\beta}$ and β. Under the null hypothesis we have

$$a(\hat{\beta}) = A(\beta^*)(\hat{\beta} - \beta)$$

Due to consistency of $\hat{\beta}$ we can replace β^* by β, asymptotically, so

$$\sqrt{n}a(\hat{\beta}) \stackrel{a}{=} \sqrt{n}A(\beta)(\hat{\beta} - \beta)$$

- This is known in the literature as the *Delta method*, or as *Klein's approximation*.
- Since this is a Wald test, it will tend to over-reject in finite samples. The score and LR tests are also possibilities, but they require estimation methods for non-linear models, which aren't in the scope of this course.

Note that this also gives a convenient way to estimate nonlinear functions and associ-

ated asymptotic confidence intervals. If the nonlinear function $a(\beta)$ is not hypothesized to be zero, we just have

$$\sqrt{n}\left(a(\hat{\beta}) - a(\beta)\right) \xrightarrow{d} N\left(0, A(\beta)\Sigma_{XX}^{-1}A(\beta)'\sigma^2\right)$$

so an approximation to the distribution of the function of the estimator is

$$a(\hat{\beta}) \approx N(a(\beta), A(\beta)(X'X)^{-1}A(\beta)'\sigma^2)$$

For example, the vector of elasticities of a function $f(x)$ is

$$\mathcal{E}(x) = \frac{\partial f(x)}{\partial x}\frac{x}{f(x)}$$

where I'm using element-by-element multiplication and division. Suppose we estimate a linear function

$$y = x'\beta + \varepsilon.$$

The elasticities of y w.r.t. x are

$$\eta_j(x) = \frac{\beta_j}{x'\beta}x_j$$

The estimator of the ith elasticity is

$$\widehat{\eta}_i(x) = \frac{\hat{\beta}_i}{x'\hat{\beta}}x_i$$

To calculate the estimated standard errors of all five elasticites, use

$$\begin{aligned} A_j(\beta) &= \frac{\partial \eta_j(x)}{\partial \beta'} \\ &= \frac{[\begin{array}{ccccc} 0 & 0 & 0 & x_j & 0 \end{array}]x'\beta - x(x_j\beta_j)}{(x'\beta)^2} \end{aligned}$$

to obtain the j th row of $A(\beta)$, and apply the above formula. Note that the elasticity and the standard error are functions of x.

4.3 Bootstrapping

When we rely on asymptotic theory to use the normal distribution-based tests and confidence intervals, we're often at serious risk of making important errors. If the sample size is small and errors are highly nonnormal, the small sample distribution of $\sqrt{n}\left(\hat{\beta} - \beta\right)$ may be very different than its large sample distribution. Also, the distributions of test statistics may not resemble their limiting distributions at all. A means of trying to gain information on the small sample distribution of test statistics and estimators is the *bootstrap*. We'll consider a simple example, just to get the main idea.

Suppose that

$$\begin{array}{ccc} y & = & X\beta_0 + \varepsilon \\ \varepsilon & \sim & IID(0, \sigma_0^2) \end{array}$$

X is nonstochastic

Given that the distribution of ε is unknown, the distribution of $\hat{\beta}$ will be unknown in small samples. However, since we have random sampling, we could generate *artificial*

data. The steps are:

1. Draw n observations from $\hat{\varepsilon}$ **with replacement**. Call this vector $\tilde{\varepsilon}^j$ (it's a $n \times 1$).
2. Then generate the data by $\tilde{y}^j = X\hat{\beta} + \tilde{\varepsilon}^j$
3. Now take this and estimate

$$\tilde{\beta}^j = (X'X)^{-1}X'\tilde{y}^j.$$

4. Save $\tilde{\beta}^j$
5. Repeat steps 1-4, until we have a large number, J, of $\tilde{\beta}^j$.

With this, we can use the replications to calculate the *empirical distribution of* $\tilde{\beta}_j$. One way to form a 100(1-α)% confidence interval for β_0 would be to order the $\tilde{\beta}^j$ from smallest to largest, and drop the first and last $J\alpha/2$ of the replications, and use the remaining endpoints as the limits of the CI. Note that this will not give the shortest CI if the empirical distribution is skewed.

- Suppose one was interested in the distribution of some function of $\hat{\beta}$, for example a test statistic. Simple: just calculate the transformation for each j, and work with the empirical distribution of the transformation.
- If the assumption of iid errors is too strong (for example if there is heteroscedasticity or autocorrelation, see below) one can work with a bootstrap defined by sampling from (y,x) with replacement.
- How to choose J: J should be large enough that the results don't change with repetition of the entire bootstrap. This is easy to check. If you find the results change a lot, increase J and try again.

- The bootstrap is based fundamentally on the idea that the empirical distribution of (y, x) converges to the actual sampling distribution as n becomes large, so statistics based on sampling from the empirical distribution should converge in distribution to statistics based on sampling from the actual sampling distribution.

- In finite samples, this doesn't hold. At a minimum, the bootstrap is a good way to check if asymptotic theory results offer a decent approximation to the small sample distribution.

5 Generalized least squares

One of the assumptions we've made up to now is that

$$\varepsilon_t \sim IID(0, \sigma^2),$$

or occasionally

$$\varepsilon_t \sim IIN(0, \sigma^2).$$

Now we'll investigate the consequences of nonidentically and/or dependently distributed errors. The model is

$$\begin{aligned} y &= X\beta + \varepsilon \\ \mathcal{E}(\varepsilon) &= 0 \\ V(\varepsilon) &= \Sigma \\ \mathcal{E}(X'\varepsilon) &= 0 \end{aligned}$$

where Σ is a general symmetric positive definite matrix (we'll write β in place of β_0 to simplify the typing of these notes).

- The case where Σ is a diagonal matrix gives uncorrelated, nonidentically distributed errors. This is known as *heteroscedasticity.*
- The case where Σ has the same number on the main diagonal but nonzero elements off the main diagonal gives identically (assuming higher moments are also the same) dependently distributed errors. This is known as *autocorrelation*.
- The general case combines heteroscedasticity and autocorrelation. This is known as "nonspherical" disturbances, though why this term is used, I have no idea.

Perhaps it's because under the classical assumptions, a joint confidence region for ε would be an $n-$ dimensional hypersphere.

5.1 Effects of nonspherical disturbances on the OLS estimator

The least square estimator is

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

- Conditional on X, or supposing that X is independent of ε, we have unbiasedness, as before.
- The variance of $\hat{\beta}$, supposing X is nonstochastic, is

$$\begin{aligned}\mathcal{E}\left[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'\right] &= \mathcal{E}\left[(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}\right] \\ &= (X'X)^{-1}X'\Sigma X(X'X)^{-1}\end{aligned}$$

  Due to this, any test statistic that is based upon $\widehat{\sigma^2}$ or the probability limit $\widehat{\sigma^2}$ of is invalid. In particular, the formulas for the t, F, χ^2 based tests given above do not lead to statistics with these distributions.
- $\hat{\beta}$ is still consistent, following exactly the same argument given before.
- If ε is normally distributed, then, conditional on X

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

  The problem is that Σ is unknown in general, so this distribution won't be useful

for testing hypotheses.

- Without normality, and unconditional on X we still have

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

we now have to look at the limitting distribution of $\frac{X'\varepsilon}{\sqrt{n}}$. The CLT will still apply if dependencies between the $\varepsilon's$ are not so strong and/or variances are not very big. Therefore

$$\frac{X'\varepsilon}{\sqrt{n}} = \sqrt{n}(\overline{g}-0) \quad \rightarrow_d \quad N(0,\Omega)$$

$$\Omega = \lim_{n\rightarrow\infty} V(\tfrac{X'\varepsilon}{\sqrt{n}}) = \lim_{n\rightarrow\infty} E(\frac{X\prime\varepsilon\varepsilon\prime X}{n})$$

assumming X's non-stochastic

$$\Omega = \lim_{n\rightarrow\infty} E(\frac{X\prime\Sigma X}{n})$$

- by lemma 2(c)

$$\sqrt{n}\left(\hat{\beta}-\beta\right) \rightarrow_d N(0,\Sigma_{XX}^{-1}\Omega\Sigma_{XX}^{-1})$$

Summary: OLS with heteroscedasticity and/or autocorrelation is:

- unbiased in the same circumstances in which the estimator is unbiased with iid errors
- has a different variance than before, so the previous test statistics aren't valid
- is consistent

- is asymptotically normally distributed, but with a different limiting covariance matrix. Previous test statistics aren't valid in this case for this reason.
- is inefficient, as is shown below.

5.2 The GLS estimator

Suppose Σ were known. Then one could form the Cholesky decomposition

$$PP' = \Sigma^{-1}$$

We have

$$PP'\Sigma = I_n$$

so

$$P'\left(P\Sigma P'\right) = P',$$

which implies that

$$P'\Sigma P = I_n$$

Consider the model

$$P'y = P'X\beta + P'\varepsilon,$$

or, making the obvious definitions,

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

This variance of $\varepsilon^* = P'\varepsilon$ is

$$\begin{aligned} \mathcal{E}(P'\varepsilon\varepsilon'P) &= P'\Sigma P \\ &= I_n \end{aligned}$$

Therefore, the model

$$\begin{aligned} y^* &= X^*\beta + \varepsilon^* \\ \mathcal{E}(\varepsilon^*) &= 0 \\ V(\varepsilon^*) &= I_n \\ \mathcal{E}(X^{*\prime}\varepsilon^*) &= 0 \end{aligned}$$

satisfies the classical assumptions (with modifications to allow stochastic regressors and nonnormality of ε). The GLS estimator is simply OLS applied to the transformed model:

$$\begin{aligned} \hat{\beta}_{GLS} &= (X^{*\prime}X^*)^{-1}X^{*\prime}y^* \\ &= (X'PP'X)^{-1}X'PP'y \\ &= (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y \end{aligned}$$

The GLS estimator is unbiased in the same circumstances under which the OLS estimator is unbiased. For example, assuming X is nonstochastic

$$\begin{aligned} \mathcal{E}(\hat{\beta}_{GLS}) &= \mathcal{E}\left\{(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y\right\} \\ &= \mathcal{E}\left\{(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}(X\beta+\varepsilon\right\} \\ &= \beta. \end{aligned}$$

The variance of the estimator, conditional on X can be calculated using

$$\begin{aligned}\hat{\beta}_{GLS} &= (X^{*\prime}X^*)^{-1}X^{*\prime}y^* \\ &= (X^{*\prime}X^*)^{-1}X^{*\prime}(X^*\beta+\varepsilon^*) \\ &= \beta+(X^{*\prime}X^*)^{-1}X^{*\prime}\varepsilon^*\end{aligned}$$

so

$$\begin{aligned}\mathcal{E}\left\{\left(\hat{\beta}_{GLS}-\beta\right)\left(\hat{\beta}_{GLS}-\beta\right)'\right\} &= \mathcal{E}\left\{(X^{*\prime}X^*)^{-1}X^{*\prime}\varepsilon^*\varepsilon^{*\prime}X^*(X^{*\prime}X^*)^{-1}\right\} \\ &= (X^{*\prime}X^*)^{-1}X^{*\prime}X^*(X^{*\prime}X^*)^{-1} \\ &= (X^{*\prime}X^*)^{-1} \\ &= (X'\Sigma^{-1}X)^{-1}\end{aligned}$$

Either of these last formulas can be used.

- All the previous results regarding the desirable properties of the least squares estimator hold, when dealing with the transformed model.
- Tests are valid, using the previous formulas, as long as we substitute X^* in place of X. Furthermore, any test that involves σ^2 can set it to 1. This is preferable to re-deriving the appropriate formulas.
- The GLS estimator is more efficient than the OLS estimator. This is a consequence of the Gauss-Markov theorem, since the GLS estimator is based on a model that satisfies the classical assumptions but the OLS estimator is not. To

see this directly, not that(Hayashi pg. 92)

$$\begin{aligned} Var(\hat{\beta}) - Var(\hat{\beta}_{GLS}) &= (X'X)^{-1}X'\Sigma X(X'X)^{-1} - (X'\Sigma^{-1}X)^{-1} \\ &= \end{aligned}$$

- As one can verify by calculating fonc, the GLS estimator is the solution to the minimization problem

$$\hat{\beta}_{GLS} = \arg\min(y - X\beta)'\Sigma^{-1}(y - X\beta)$$

so the *metric* Σ^{-1} is used to weight the residuals.

5.3 Feasible GLS estimation

The problem is that Σ isn't known usually, so this estimator isn't available.

- Consider the dimension of Σ : it's an $n \times n$ matrix with $\left(n^2 - n\right)/2 + n = \left(n^2 + n\right)/2$ unique elements.
- The number of parameters to estimate is larger than n and increases faster than n. There's no way to devise an estimator that satisfies a LLN without adding restrictions.
- The *feasible GLS estimator* is based upon making sufficient assumptions regarding the form of Σ so that a consistent estimator can be devised.

Suppose that we *parameterize* Σ as a function of X and θ, where θ may include β as well as other parameters, so that

$$\Sigma = \Sigma(X, \theta)$$

where θ is of fixed dimension. If we can consistently estimate θ, we can consistently estimate Σ, as long as $\Sigma(X,\theta)$ is a continuous function of θ (by the Slutsky theorem). In this case,

$$\widehat{\Sigma} = \Sigma(X,\hat{\theta}) \xrightarrow{p} \Sigma(X,\theta)$$

If we replace Σ in the formulas for the GLS estimator with $\widehat{\Sigma}$, we obtain the FGLS estimator. **The FGLS estimator shares the same asymptotic properties as GLS. These are**

1. Consistency
2. Asymptotic normality
3. Asymptotic efficiency *if* the errors are normally distributed. (Cramer-Rao).
4. Test procedures are asymptotically valid.

In practice, the usual way to proceed is

1. Define a consistent estimator of θ. This is a case-by-case proposition, depending on the parameterization $\Sigma(\theta)$. We'll see examples below.
2. Form $\widehat{\Sigma} = \Sigma(X,\hat{\theta})$
3. Calculate the Cholesky factorization $\widehat{P} = Chol(\hat{\Sigma}^{-1})$.
4. Transform the model using

$$\hat{P}'y = \hat{P}'X\beta + \hat{P}'\varepsilon$$

5. Estimate using OLS on the transformed model.