

Econ 7010 –Econometrics I

Course Notes

Ryan Godwin

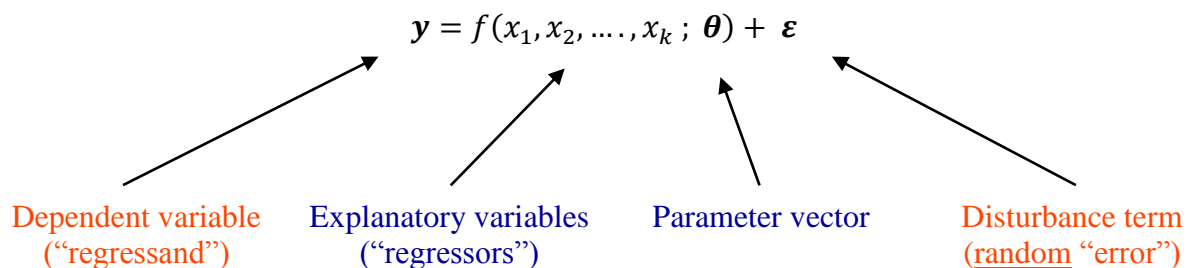
(Original version of notes by David Giles)

Table of Contents

| | |
|---|-----|
| Topic 1: Basic Multiple Regression | 3 |
| Partitioned and Partial Regression | 14 |
| R Introduction | 21 |
| Matrices: Concepts, Definitions & Some Basic Results | 24 |
| Topic 1 Continued: Finite-Sample Properties of the LS Estimator | 34 |
| Introduction to the Monte Carlo Method | 54 |
| Topic 2: Asymptotic Properties of Various Regression Estimators | 59 |
| Topic 3: Inference and Prediction | 77 |
| Topic 4: Model Stability & Specification Analysis | 95 |
| Topic 5: Non-Linear Regression | 106 |
| Topic 6: Non-Spherical Disturbances | 114 |
| Topic 7: Heteroskedasticity | 127 |
| Topic 7 continued: Heteroskedasticity | 136 |
| Topic 8: Autocorrelated Errors | 142 |
| Topic 9: Maximum Likelihood Estimation | 151 |

Topic 1: Basic Multiple Regression

Population “model” –



Note:

- The function, “ f ”, may be linear or non-linear in the variables.
- The function, “ f ”, may be linear or non-linear in the parameters.
- The function, “ f ”, may be non-parametric, but we won’t consider this.
- We’ll focus on models that are parametric, and *usually* linear in the parameters.

Questions:

- Why is the error term needed?
- What is **random**, and what is **deterministic**?

What is **observable**, and what is **unobservable**?

Examples:

1) Keynes’ consumption function:

$$C = \beta_1 + \beta_2 Y + \varepsilon \quad (1)$$

2) Cobb-Douglas production function:

$$Y = AK^{\beta_2}L^{\beta_3}e^{\varepsilon} \quad (2)$$

By taking logs, the Cobb-Douglas production function can be rewritten as:

$$\log Y = \beta_1 + \beta_2 \log K + \beta_3 \log L + \varepsilon, \text{ where } \beta_1 = \log A$$

3) CES production function

$$Y = \varphi(aK^r + (1 - a)L^r)^{1/r} e^{\varepsilon} \quad (3)$$

Taking logs, the CES production function is written as:

$$\log Y = \log \varphi + \frac{1}{r} \log(aK^r + (1 - a)L^r) + \varepsilon$$

Sample Information

- Have a *sample* of “ n ” observations: $\{y_i; x_{i1}, x_{i2}, \dots, x_{ik}\}; i = 1, 2, \dots, n$
- We assume that these observed values are generated by the population model.

Let’s take the case where the model is *linear in the parameters*:

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \varepsilon_i; i = 1, \dots, n \quad (4)$$

Recall that the β ’s and ε are unobservable. So, y_i is generated by 2 components:

1. Deterministic component: $\sum_{j=1}^k \beta_j x_{ij}$.
2. Stochastic component: ε_i .

So, the y_i ’s must be “realized values” of a random variable.

Objectives:

- (i) Estimate unknown parameters
- (ii) Test hypotheses about parameters
- (iii) Predict values of y outside sample

Interpreting the Parameters in a Model

Note that the β 's in equation (4) have an important economics interpretation:

$$\frac{\partial y_i}{\partial x_{1i}} = \beta_1; \text{ etc.}$$

The parameters are the *marginal effects* of the x 's on y , with other factors held constant (*ceteris paribus*). For example, from equation (1):

$$\partial C / \partial Y = \beta_2 = M.P.C.$$

We might wish to test the hypothesis that $\beta_2 = 0.9$, for example.

Depending on how the population model is specified, however, the β 's may *not* be interpreted as marginal effects. For example, after taking logs of the Cobb-Douglas production function in (2), we get the following population model:

$$\log Y = \beta_1 + \beta_2 \log K + \beta_3 \log L + \varepsilon,$$

and

$$\beta_2 = \frac{\partial \log Y}{\partial \log K} = \frac{\partial \log Y}{\partial Y} \times \frac{\partial Y}{\partial K} \times \frac{\partial K}{\partial \log K} = \frac{1}{Y} \times \frac{\partial Y}{\partial K} \times K = \frac{\partial Y/Y}{\partial K/K},$$

so that β_2 is the elasticity of output with respect to capital. The point is that we need to be careful about how the parameters of the model are interpreted.

How could we test the hypothesis of constant returns to scale in the above Cobb-Douglas model?

So, we have a stochastic model that might be useful as a starting point to represent economics relationships. We need to be especially careful about the way in which we specify both parts of the model (the deterministic and stochastic parts).

Assumptions of the Classical Linear Regression Model

All “models” are simplifications of reality. Presumably we want our model to be simple but “realistic” – able to explain actual data in a reliable and robust way.

To begin with we'll make a set of simplifying assumptions for our model. In fact, one of the main objectives of Econometrics is to re-consider these assumptions – are they realistic; can they

be tested; what if they are wrong; can they be “relaxed”? The assumptions relate to: (1) functional form (parameters); (2) regressors; (3) disturbances.

A.1: Linearity

The model is linear in the parameters:

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i \quad ; \quad i = 1, \dots, n.$$

Linearity in the parameters allows the model to be written in matrix notation. Let,

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} ; \quad \mathbf{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} ; \quad X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} ; \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} .$$

$(n \times 1)$
 $(k \times 1)$
 $(n \times k)$
 $(n \times 1)$

Then, we can write the model, for the full sample, as:

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

If we take the i th row (observation) of this model we have:

$$y_i = \mathbf{x}_i \boldsymbol{\beta} + \varepsilon_i \quad (\text{scalar})$$

Notational points

- i. Vectors are in bold.
- ii. The dimensions of vectors/matrices are written (*rows* \times *columns*).
- iii. The first subscript denotes the row, the second subscript the column.
- iv. Some texts (including Greene, 2011), use the convention that vectors are columns. Hence, when an observation (row) is extracted from the X matrix, it is transformed into a column. Hence, the above equation would be expressed as $y_i = \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i$.

A.2: Full Rank

We assume that there are no exact linear dependencies among the columns of X (if there were, then one or more regressor is redundant). Note that X is $(n \times k)$ and $\text{Rank}(X) = k$. So we are also implicitly assuming that $n > k$, since $\text{Rank}(A) \leq \min. \{\#rows, \#cols\}$.

What does this assumption really mean? Suppose we had:

$$y_i = \beta_1 x_{i1} + \beta_2 (2x_{i1}) + \varepsilon_i$$

We can only identify, and estimate, the one function, $(\beta_1 + 2\beta_2)$. In this model, $Rank(X) = k - 1 = 1$. An example which is commonly found in undergraduate textbooks, of where A.2 is violated, is the dummy variable trap.

A.3: Errors Have a Zero Mean

Assume that, *in the population*, $E(\varepsilon_i) = 0$; $i = 1, 2, \dots, n$. So,

$$E(\boldsymbol{\varepsilon}) = E \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} = \mathbf{0} .$$

A.4: Spherical Errors

Assume that, in the population, the disturbances are generated by a process whose variance is constant (σ^2), and that these disturbances are uncorrelated with each other:

$$var(\varepsilon_i) = \sigma^2 ; i = 1, 2, \dots, n \quad (\text{Homoskedasticity})$$

$$cov(\varepsilon_i, \varepsilon_j) = 0 ; \forall i \neq j \quad (\text{no Autocorrelation})$$

Putting these assumptions together we can determine the form of the “covariance matrix” for the random vector, $\boldsymbol{\varepsilon}$.

$$V(\boldsymbol{\varepsilon}) = E [(\boldsymbol{\varepsilon} - E(\boldsymbol{\varepsilon}))(\boldsymbol{\varepsilon} - E(\boldsymbol{\varepsilon}))'] = E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] = \begin{bmatrix} E(\varepsilon_1\varepsilon_1) & \cdots & E(\varepsilon_1\varepsilon_n) \\ \vdots & \ddots & \vdots \\ E(\varepsilon_n\varepsilon_1) & \cdots & E(\varepsilon_n\varepsilon_n) \end{bmatrix}$$

but...

$$E(\varepsilon_i\varepsilon_i) = E(\varepsilon_i^2) = E[(\varepsilon_i - 0)^2] = var(\varepsilon_i) = \sigma^2$$

and

$$E(\varepsilon_i\varepsilon_j) = E[(\varepsilon_i - 0)(\varepsilon_j - 0)] = cov(\varepsilon_i, \varepsilon_j) = 0.$$

So:

$$V(\boldsymbol{\varepsilon}) = \begin{bmatrix} \sigma^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 I_n$$

a scalar matrix.

A.5: Generating Process for \mathbf{X}

The classical regression model assumes that the regressors are “fixed in repeated samples” (laboratory situation). We can assume this – very strong, though.

Alternatively, allow \mathbf{x} 's to be random, but restrict the form of their randomness – assume that the regressors are uncorrelated with the disturbances. The process that generates \mathbf{X} is unrelated to the process that generates $\boldsymbol{\varepsilon}$ in the population.

A.6: Normality of Errors

$$(\boldsymbol{\varepsilon}|\mathbf{X}) \sim N[0, \sigma^2 I_n]$$

This assumption is not as strong as it seems:

- often reasonable due to the Central Limit Theorem (C.L.T.)
- often not needed
- when some distributional assumption is needed, often a more general one is ok

Summary

The classical linear regression model is:

- $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
- $(\boldsymbol{\varepsilon}|\mathbf{X}) \sim N[0, \sigma^2 I_n]$
- $\text{Rank}(\mathbf{X}) = k$
- Data generating processes (D.G.P.s) of \mathbf{X} and $\boldsymbol{\varepsilon}$ are unrelated.

Implications for \mathbf{y} (if \mathbf{X} is non-random; *or* conditional on \mathbf{X}):

$$E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} + E(\boldsymbol{\varepsilon}) = \mathbf{X}\boldsymbol{\beta}$$

$$V(\mathbf{y}) = V(\boldsymbol{\varepsilon}) = \sigma^2 I_n$$

Because *linear* transformations of a Normal random variable are themselves Normal, we also have: $\mathbf{y} \sim N[\mathbf{X}\boldsymbol{\beta}, \sigma^2 I_n]$.

Some Questions

- How reasonable are the assumptions associated with the classical linear regression model?
- How do these assumptions affect the estimation of the model's parameters?
- How do these assumptions affect the way we test hypotheses about the model's parameters?
- Which of these assumptions are used to establish the various results we'll be concerned with?
- Which assumptions can be “relaxed” without affecting these results?

Least Squares Regression

Our first task is to estimate the parameters of our model,

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \boldsymbol{\varepsilon} \sim N[\mathbf{0}, \sigma^2 I_n] .$$

Note that there are $(k + 1)$ parameters, including σ^2 .

- Many possible procedures for estimating parameters.
- Choice should be based not only on computational convenience, but also on the “[sampling properties](#)” of the resulting estimator.
- To begin with, consider *one possible* estimation strategy – **Least Squares**.

For the i^{th} data-point, we have:

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i \quad ,$$

and the population regression is:

$$E(y_i | \mathbf{x}'_i) = \mathbf{x}'_i \boldsymbol{\beta} \quad .$$

We'll estimate $E(y_i | \mathbf{x}'_i)$ by

$$\hat{y}_i = \mathbf{x}'_i \mathbf{b} .$$

In the population, the true (unobserved) disturbance is ε_i [= $y_i - \mathbf{x}'_i \boldsymbol{\beta}$].

When we use \mathbf{b} to estimate $\boldsymbol{\beta}$, there will be some “estimation error”, and the value, $e_i = y_i - \mathbf{x}'_i \mathbf{b}$ will be called the i^{th} “**residual**”.

So,

$$y_i = (\mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i) = (\mathbf{x}'_i \mathbf{b} + e_i) = (\hat{y}_i + e_i)$$

↑ ↑
↑ ↑ ↑

unobserved
observed

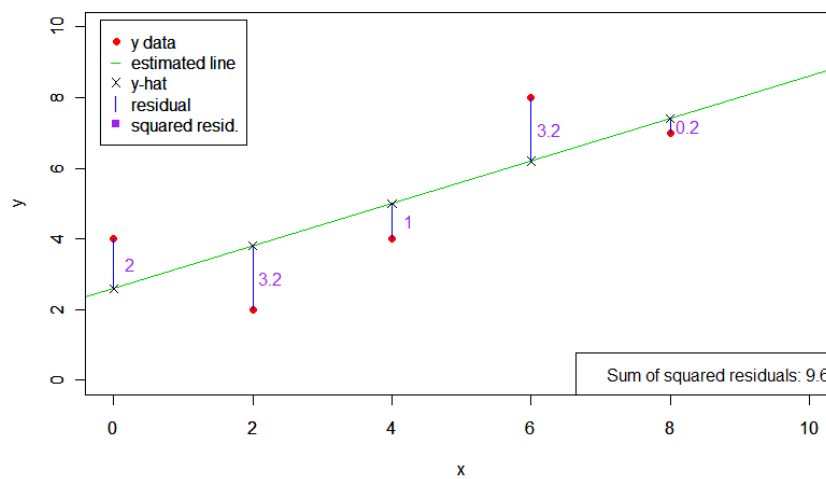
[Population]
[Sample]

The Least Squares Criterion:

“Choose \mathbf{b} so as to minimize the sum of the squared residuals.”

- Why *squared* residuals?
- Why not *absolute values* of residuals?
- Why not use a “minimum distance” criterion?

Fig 1.1. Minimizing the sum of squared residuals, for $y = \{4, 2, 4, 8, 7\}$; $x = \{0, 2, 4, 6, 8\}$.



Minimizing the Sum of Squared Residuals: An Optimization Problem

$$\begin{aligned} \text{Min.}_{(b)} \sum_{i=1}^n e_i^2 &\Leftrightarrow \text{Min.}_{(b)} (\mathbf{e}'\mathbf{e}) \\ &\Leftrightarrow \text{Min.}_{(b)} [(\mathbf{y} - X\mathbf{b})'(\mathbf{y} - X\mathbf{b})]. \end{aligned}$$

Now, let:

$$S = (\mathbf{y} - X\mathbf{b})'(\mathbf{y} - X\mathbf{b}) = \mathbf{y}'\mathbf{y} - \mathbf{b}'X'\mathbf{y} - \mathbf{y}'X\mathbf{b} + \mathbf{b}'X'X\mathbf{b}.$$

Note that,

$$\mathbf{b}'X'\mathbf{y} = \mathbf{y}'X\mathbf{b}.$$

$$(1 \times k)(k \times n)(n \times 1) \quad (1 \times 1)$$

So, $S = \mathbf{y}'\mathbf{y} - 2(\mathbf{y}'X)\mathbf{b} + \mathbf{b}'(X'X)\mathbf{b}.$

Note:

(i) $\partial(\mathbf{a}'\mathbf{x})/\partial\mathbf{x} = \mathbf{a}$

(ii) $\partial(\mathbf{x}'A\mathbf{x})/\partial\mathbf{x} = 2A\mathbf{x}$; if A is *symmetric*

Applying these 2 results –

$$\partial S/\partial\mathbf{b} = \mathbf{0} - 2(\mathbf{y}'X)' + 2(X'X)\mathbf{b} = 2[X'X\mathbf{b} - X'\mathbf{y}].$$

Set this to zero (for a turning point):

$$X'X\mathbf{b} = X'\mathbf{y}, \quad (k \text{ equations in } k \text{ unknowns})$$

$$(k \times n)(n \times k)(k \times 1) \quad (k \times n)(n \times 1) \quad (\text{the “normal equations”})$$

so:

$$\mathbf{b} = (X'X)^{-1}X'\mathbf{y} \quad ; \quad \text{provided that } (X'X)^{-1} \text{ exists}$$

Notice that $X'X$ is $(k \times k)$, and $\text{rank}(X'X) = \text{rank}(X) = k$ (assumption).

This implies that $(X'X)^{-1}$ exists.

We need the “full rank” assumption for the Least Squares estimator, \mathbf{b} , to *exist*.

None of our other assumptions have been used so far.

Check – have we *minimized* S ?

$$\left(\frac{\partial^2 S}{\partial \mathbf{b} \partial \mathbf{b}'}\right) = \partial / \partial \mathbf{b}' [2X'X\mathbf{b} - 2X'\mathbf{y}] = 2(X'X) \quad ; \quad \text{a } (k \times k) \text{ matrix.}$$

Note that $X'X$ is at least positive *semi-definite* –

$$\eta'(X'X)\eta = (X\eta)'(X\eta) = (\mathbf{u}'\mathbf{u}) = \sum_{i=1}^n u_i^2 \geq 0 \quad ;$$

and so if $X'X$ has full rank, it will be *positive-definite*, not negative-definite.

So, our assumption that X has full rank has two implications –

1. The Least Squares estimator, \mathbf{b} , *exists*.
2. Our optimization problem leads to the *minimization* of S , not its maximization!

Aside – OLS formula in scalar form

For a population model with an intercept and a single regressor, you may have seen the following formulas used in undergraduate textbooks:

$$b_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{s_{X,Y}}{s_X^2} \quad ,$$

$$b_0 = \bar{Y} - b_1 \bar{X} \quad ,$$

where $s_{X,Y}$ is the sample covariance between X_i and Y_i , and s_X^2 is the sample variance of X_i .

Some Basic Properties of Least Squares

First, note that the LS residuals are “orthogonal” to the regressors –

$$X'X\mathbf{b} - X'\mathbf{y} = \mathbf{0} \quad \text{ (“normal equations”; } (k \times 1) \text{)}$$

So,

$$-X'(\mathbf{y} - X\mathbf{b}) = -X'\mathbf{e} = \mathbf{0} \quad ;$$

or,

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

If the model includes an intercept term, then one regressor (say, the first column of X) is a unit vector.

In this case we get some further results:

1. The LS residuals sum to zero

$$\begin{aligned}
 X' \mathbf{e} &= \begin{pmatrix} 1 & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ 1 & \cdots & x_{nk} \end{pmatrix}' \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} = \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ x_{1k} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} \\
 &= \begin{pmatrix} \sum_i e_i \\ ? \\ ? \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}
 \end{aligned}$$

From the first element:

$$\sum_{i=1}^n e_i = 0$$

2. Fitted regression passes through sample mean

$$X' \mathbf{y} = X' X \mathbf{b} ,$$

$$\text{or, } \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ x_{1k} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ x_{1k} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} 1 & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ 1 & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} b_1 \\ \vdots \\ b_k \end{pmatrix} .$$

$$\text{So, } \begin{pmatrix} \sum_i y_i \\ ? \\ ? \end{pmatrix} = \begin{pmatrix} n & \sum_i x_{i2} & \cdots \\ ? & \cdots & ? \\ ? & \cdots & ? \end{pmatrix} \begin{pmatrix} b_1 \\ \vdots \\ b_k \end{pmatrix} .$$

From the first row of this vector equation –

$$\sum_i y_i = n b_1 + b_2 \sum_i x_{i2} + \cdots + b_k \sum_i x_{ik}$$

or,

$$\bar{y} = b_1 + b_2 \bar{x}_2 + \cdots + b_k \bar{x}_k$$

3. Sample mean of the fitted y-values equals sample mean of actual y-values

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i = \mathbf{x}_i' \mathbf{b} + e_i = \hat{y}_i + e_i .$$

So,

$$\frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{n} \sum_{i=1}^n \hat{y}_i + \frac{1}{n} \sum_{i=1}^n e_i ,$$

or,

$$\bar{y} = \bar{\hat{y}} + 0 = \bar{\hat{y}}$$

Note: These last 3 results use the fact that the model *includes an intercept*.

Partitioned & Partial Regression

Suppose the regressor matrix can be partitioned into 2 blocks –

$$\mathbf{y} = X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$$

$(n \times 1) \quad (n \times k_1)(k_1 \times 1) \quad (n \times k_2)(k_2 \times 1) \quad (n \times 1)$

The algebra (geometry) of LS estimation provides us with some important results that we'll be able to use to help us at various stages.

The model is:

$$\mathbf{y} = [X_1 : X_2] \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} + \boldsymbol{\varepsilon} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

$(n \times 1) \quad (n \times (k_1 + k_2)) \quad ((k_1 + k_2) \times 1) \quad (n \times 1)$

and $\mathbf{b} = (X'X)^{-1}X'\mathbf{y} \quad ; \quad k = (k_1 + k_2)$

We can write this LS estimator as:

$$\mathbf{b} = \{[X_1 : X_2]'[X_1 : X_2]\}^{-1}[X_1 : X_2]'\mathbf{y}$$

$$= \left\{ \begin{bmatrix} [X_1]' \\ \dots \\ [X_2]' \end{bmatrix} [X_1 : X_2] \right\}^{-1} \begin{bmatrix} [X_1]' \\ \dots \\ [X_2]' \end{bmatrix} \mathbf{y}$$

So,

$$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix}^{-1} \begin{pmatrix} X_1'\mathbf{y} \\ X_2'\mathbf{y} \end{pmatrix}.$$

The “normal equations” underlying this are –

$$(X'X)\mathbf{b} = X'\mathbf{y},$$

or:

$$\begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix} \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{pmatrix} = \begin{pmatrix} X_1'\mathbf{y} \\ X_2'\mathbf{y} \end{pmatrix}.$$

Let's solve these “normal equations” for \mathbf{b}_1 and \mathbf{b}_2 :

$$X_1'X_1\mathbf{b}_1 + X_1'X_2\mathbf{b}_2 = X_1'\mathbf{y} \tag{1}$$

$$X_2'X_1\mathbf{b}_1 + X_2'X_2\mathbf{b}_2 = X_2'\mathbf{y} \tag{2}$$

From [1]:

$$(X_1'X_1)\mathbf{b}_1 = X_1'\mathbf{y} - X_1'X_2\mathbf{b}_2 ,$$

or,
$$\mathbf{b}_1 = (X_1'X_1)^{-1}X_1'\mathbf{y} - (X_1'X_1)^{-1}X_1'X_2\mathbf{b}_2$$

$$= (X_1'X_1)^{-1}[X_1'\mathbf{y} - X_1'X_2\mathbf{b}_2] \quad [3]$$

Note: If $X_1'X_2 = 0$, then $\mathbf{b}_1 = (X_1'X_1)^{-1}X_1'\mathbf{y}$.

(Why do the “partial” and “full” regression estimators coincide in this case?)

Now substitute [3] into [2]:

$$(X_2'X_1)[(X_1'X_1)^{-1}X_1'\mathbf{y} - (X_1'X_1)^{-1}X_1'X_2\mathbf{b}_2] + (X_2'X_2)\mathbf{b}_2 = X_2'\mathbf{y} ,$$

or,

$$[(X_2'X_2) - (X_2'X_1)(X_1'X_1)^{-1}(X_1'X_2)]\mathbf{b}_2 = X_2'\mathbf{y} - (X_2'X_1)(X_1'X_1)^{-1}X_1'\mathbf{y} ,$$

and so:

$$\mathbf{b}_2 = [(X_2'X_2) - (X_2'X_1)(X_1'X_1)^{-1}(X_1'X_2)]^{-1}[X_2'(I - X_1(X_1'X_1)^{-1}X_1')\mathbf{y}] .$$

Define:

$$M_1 = (I - X_1(X_1'X_1)^{-1}X_1') .$$

Then, we can write –

$$\mathbf{b}_2 = (X_2'M_1X_2)^{-1}X_2'M_1\mathbf{y}$$

If we repeat the whole exercise, with X_1 and X_2 interchanged, we get:

$$\mathbf{b}_1 = (X_1'M_2X_1)^{-1}X_1'M_2\mathbf{y}$$

where: $M_2 = (I - X_2(X_2'X_2)^{-1}X_2') .$

- M_1 and M_2 are “*idempotent*” matrices
- $M_iM_i = M_iM_i' = M_i = M_i'M_i$; $i = 1, 2.$

So, finally, we can write:

$$\mathbf{b}_1 = (X_1^*X_1^*)^{-1}X_1^{*\prime}\mathbf{y}_1^*$$

$$\mathbf{b}_2 = (X_2^*X_2^*)^{-1}X_2^{*\prime}\mathbf{y}_2^*$$

where:

$$X_1^* = M_2 X_1 ; X_2^* = M_1 X_2 ; \mathbf{y}_1^* = M_2 \mathbf{y} ; \mathbf{y}_2^* = M_1 \mathbf{y}$$

Why are these results useful?

“Frisch-Waugh-Lovell Theorem”

(Greene, 7th ed., p.33)

Goodness-of-Fit

- One way of measuring the “quality” of fitted regression model is by the extent to which the model “explains” the *sample variation* for \mathbf{y} .
- Sample variance of \mathbf{y} is $\frac{1}{(n-1)} \sum_{i=1}^n (y_i - \bar{y})^2$.
- Or, we could just use $\sum_{i=1}^n (y_i - \bar{y})^2$ to measure *variability*.
- Our “fitted” regression model, using LS, gives us

$$\mathbf{y} = X\mathbf{b} + \mathbf{e} = \hat{\mathbf{y}} + \mathbf{e}$$

where $\hat{\mathbf{y}} = X\mathbf{b} = X(X'X)^{-1}X'\mathbf{y}$

- Recall that *if the model includes an intercept*, then the residuals sum to zero, and $\bar{y} = \bar{\hat{y}}$.

To simplify things, introduce the following matrix:

$$M^0 = [I_n - \frac{1}{n} \mathbf{i}\mathbf{i}']$$

where: $\mathbf{i} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$; $(n \times 1)$

Note that:

- M^0 is an idempotent matrix.
- $M^0 \mathbf{i} = \mathbf{0}$.
- M^0 transforms elements of a vector into deviations from sample mean.
- $\mathbf{y}' M^0 \mathbf{y} = \mathbf{y}' M^0 M^0 \mathbf{y} = \sum_{i=1}^n (y_i - \bar{y})^2$.

Let's check the third of these results:

$$M^0 \mathbf{y} = \left\{ \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix} - \begin{bmatrix} 1/n & \cdots & 1/n \\ \vdots & \ddots & \vdots \\ 1/n & \cdots & 1/n \end{bmatrix} \right\} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

$$= \begin{bmatrix} y_1 - \frac{1}{n}y_1 - \frac{1}{n}y_2 - \cdots - \frac{1}{n}y_n \\ \vdots \\ y_n - \frac{1}{n}y_1 - \frac{1}{n}y_2 - \cdots - \frac{1}{n}y_n \end{bmatrix} = \begin{pmatrix} y_1 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{pmatrix}.$$

Returning to our "fitted" model:

$$\mathbf{y} = X\mathbf{b} + \mathbf{e} = \hat{\mathbf{y}} + \mathbf{e}$$

So, we have:

$$M^0 \mathbf{y} = M^0 \hat{\mathbf{y}} + M^0 \mathbf{e} = M^0 \hat{\mathbf{y}} + \mathbf{e}.$$

[$M^0 \mathbf{e} = \mathbf{e}$; because the residuals sum to zero.]

Then –

$$\begin{aligned} \mathbf{y}' M^0 \mathbf{y} &= \mathbf{y}' M^0 M^0 \mathbf{y} = (M^0 \hat{\mathbf{y}} + \mathbf{e})' (M^0 \hat{\mathbf{y}} + \mathbf{e}) \\ &= \hat{\mathbf{y}}' M^0 \hat{\mathbf{y}} + \mathbf{e}' \mathbf{e} + 2\mathbf{e}' M^0 \hat{\mathbf{y}} \end{aligned}$$

However,

$$\mathbf{e}' M^0 \hat{\mathbf{y}} = \mathbf{e}' M^0 \hat{\mathbf{y}} = (M^0 \mathbf{e})' \hat{\mathbf{y}} = \mathbf{e}' \hat{\mathbf{y}} = \mathbf{e}' X (X'X)^{-1} X' \mathbf{y} = 0.$$

So, we have –

$$\begin{aligned} \mathbf{y}' M^0 \mathbf{y} &= \hat{\mathbf{y}}' M^0 \hat{\mathbf{y}} + \mathbf{e}' \mathbf{e} \\ \sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n e_i^2 \\ \mathbf{SST} &= \mathbf{SSR} + \mathbf{SSE} \end{aligned}$$

Recall: $\bar{\hat{y}} = \bar{y}$.

This lets us define the “**Coefficient of Determination**” –

$$R^2 = \left(\frac{SSR}{SST} \right) = 1 - \left(\frac{SSE}{SST} \right)$$

Note:

- The second equality in definition of R^2 holds only if model *includes an intercept*.
- $R^2 = \left(\frac{SSR}{SST} \right) \geq 0$
- $R^2 = 1 - \left(\frac{SSE}{SST} \right) \leq 1$
- So, $0 \leq R^2 \leq 1$
- Interpretation of “0” and “1” ?
- R^2 is *unitless*.

What happens if we add *any* regressor(s) to the model?

$$\mathbf{y} = X_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon} \quad ; \quad [1]$$

Then:

$$\mathbf{y} = X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + \mathbf{u} \quad ; \quad [2]$$

(A) Applying LS to [2]:

$$\min. (\hat{\mathbf{u}}'\hat{\mathbf{u}}) \quad ; \quad \hat{\mathbf{u}} = \mathbf{y} - X_1\mathbf{b}_1 - X_2\mathbf{b}_2$$

(B) Applying LS to [1]:

$$\min. (\mathbf{e}'\mathbf{e}) \quad ; \quad \mathbf{e} = \mathbf{y} - X_1\hat{\boldsymbol{\beta}}_1$$

Problem (B) is just Problem (A), subject to restriction: $\boldsymbol{\beta}_2 = 0$. Minimized value in (A) must be \leq minimized value in (B). So, $\hat{\mathbf{u}}'\hat{\mathbf{u}} \leq \mathbf{e}'\mathbf{e}$.

What does this imply?

- Adding *any* regressor(s) to the model *cannot increase* (and typically will *decrease*) the sum of squared residuals.
- So, adding *any* regressor(s) to the model *cannot decrease* (and typically will *increase*) the value of R^2 .

- Means that R^2 is not really a very interesting measure of the “quality” of the regression model, in terms of explaining sample variability of the dependent variable.
- For these reasons, we usually use the “adjusted” Coefficient of Determination.

We modify $R^2 = [1 - \frac{e'e}{y'M^0y}]$ to become:

$$\bar{R}^2 = [1 - \frac{e'e/(n-k)}{y'M^0y/(n-1)}].$$

- What are we doing here?

We’re adjusting for “degrees of freedom” in numerator and denominator.

- “Degrees of freedom” = number of independent pieces of information.
- $e = y - Xb$. We estimate k parameters from the n data-points. We have $(n - k)$ “degrees of freedom” associated with the fitted model.
- In denominator – have constructed \bar{y} from sample. “Lost” one degree of freedom.
- Possible for $\bar{R}^2 < 0$ (even with intercept in the model).
- \bar{R}^2 can *increase or decrease* when we add regressors.
- When will it increase (decrease)?

In multiple regression, \bar{R}^2 will *increase* (decrease) if a variable is deleted, if and only if the associated t-statistic has *absolute value less than* (greater than) unity.

- If model *doesn't* include an intercept, then $SST \neq SSR + SSE$, and in this case no longer any guarantee that $0 \leq R^2 \leq 1$.
- Must be careful comparing R^2 and \bar{R}^2 values across models.

Example –

$$(1) \quad \hat{C}_i = 0.5 + 0.8Y_i \quad ; \quad R^2 = 0.90$$

$$(2) \quad \log(\hat{C}_i) = 0.2 + 0.75Y_i \quad ; \quad R^2 = 0.80$$

Sample variation is in *different units*.

Topic 1 Appendix

R code for Fig 1.1

```

#Input the data
y = c(4,2,4,8,7)
x = c(0,2,4,6,8)

### Two ways to get the OLS estimates:
# Calculate slope coefficient using sample covariance and variance
b1 = cov(x,y)/var(x)
b0 = mean(y) - b1*mean(x)

### OR
#Calculate slope and intercept using an R function
summary(lm(y~x))
b0 = lm(y~x)$coeff[1]
b1 = lm(y~x)$coeff[2]

#Get the estimated/fitted/predicted y-values
yhat = b0 + b1*x

#Get the ols residuals
resids = y - yhat

###Graphics###
#Plot the data
plot(x,y,xlim=c(0,10),ylim=c(0,10),pch = 16,col = 2)
#Draw the estimated line
abline(b0,b1,col=3)
#Plot the predicted values (yhat)
par(new=TRUE)
plot(x,yhat,xlim=c(0,10),ylim=c(0,10),pch = 4,col = 1,ylab="")
#Draw the residuals
for(ii in 1:length(y)){
  segments(x[ii],y[ii],x[ii],b0+b1*x[ii],col=4)
}
#Display the squared residuals
for(ii in 1:length(y)){
  text(x[ii]+.25, (b0+b1*x[ii]+y[ii])/2,round((y[ii]-b0-
    b1*x[ii])^2,1),col="purple")
}
#Label the graph
legend("topleft", c("y data", "estimated line","y-
  hat","residual","squared resid."), pch = c(16,NA,4,NA,15)
  ,col=c(2,3,1,4,"purple"), inset = .02)
legend("topleft", c("y data", "estimated line","y-
  hat","residual","squared resid."), pch = c(NA,"_",NA,"|",NA)
  ,col=c(2,3,1,4,"purple"), inset = .02)
legend("bottomright", paste("Sum of squared residuals:",sum((y-b0-
  b1*x)^2)))

```

R is open-source and free, and has a large online user-support base. If you have a problem, Google-ing it will likely provide ample solutions.

For PC: <http://cran.r-project.org/bin/windows/base/>

For Mac: <http://cran.r-project.org/bin/macosx/>

Instructions for installation and download can be found on the above pages, but installation is simple. Download the file, and double-click it.

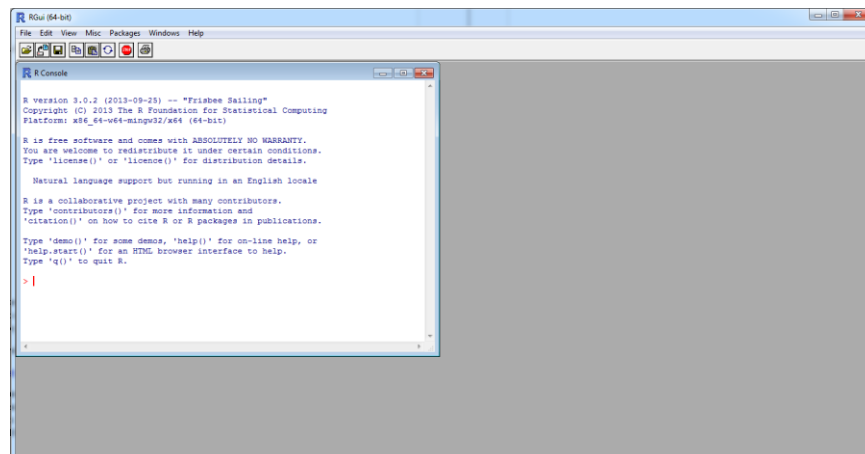
For Windows:



For Mac:

Students have successfully run R on their Macs, however, many students have had problems. I will be unable to help you to get R running on a Mac, as I do not own a Mac.

When you first run R, the window should look something like this:

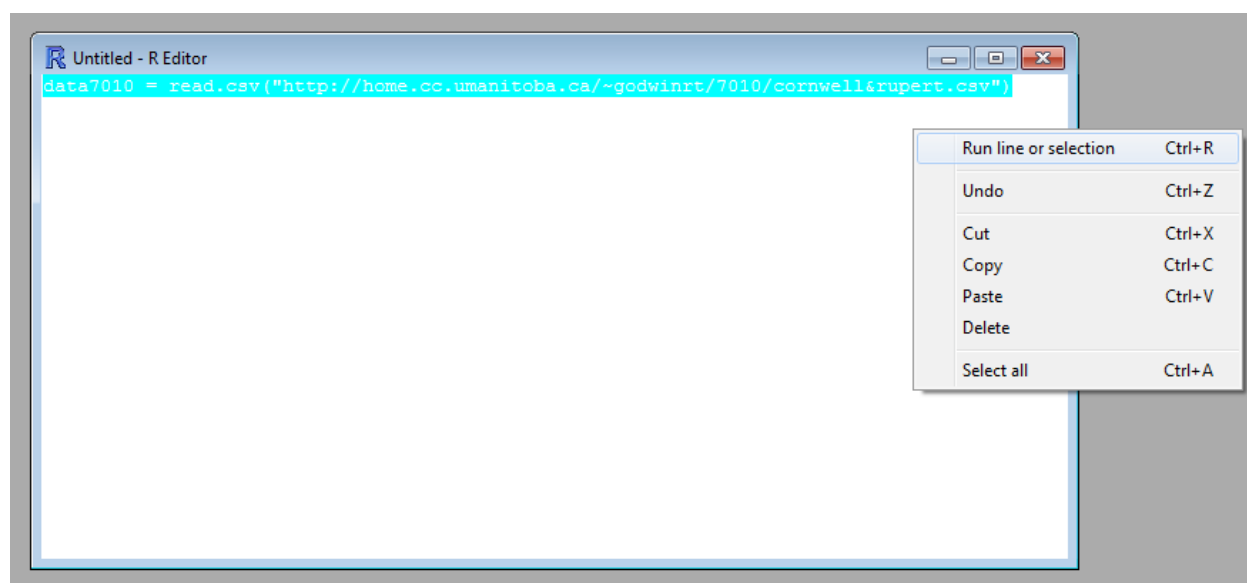


The red cursor is the command prompt, where you can enter R commands. A good way to keep track of your work is to create a script, which you can save, and run commands from. Do this by clicking *File, New script*.

Our first task is to get some data into R. In the script window, type:

```
data7010=read.csv("http://home.cc.umanitoba.ca/~godwinrt/7010/cornwell
&rupert.csv")
```

To run a command from the script window, highlight it, right-click, and select “Run line or selection”.



This data is from Cornwell and Rupert (1988), where the main interest of the study is the effect of education (ED) on the log of wages (LWAGE). A full description of the data is in Greene (2011, Example 8.5, pg. 232). To take a look at the data, you can type:

```
summary(data7010)
```

To see what the first six rows of the data looks like:

```
head(data7010)
```

There are several variables in the dataframe. To load all of these variables into memory, so that each may be referred to easily:

```
attach(data7010)
```

To look at an individual variable, ED for example (years of education), simply type `ED`. This will print out the variable. Not very helpful since there are 4165 observations!

Try typing `summary(ED)`. Some other useful commands, besides the `summary` command, are:

`sum` `mean` `var` `sd` `range` `min` `max` `length`

What do these commands do? You can always type `?length` to get help with a command, but Googling is your best bet.

A good place to start is by visualizing the data. For example, type:

`hist(ED)`

It seems most people in the sample have a high-school education.

To visualize two variables at once, type:

`plot(ED, LWAGE)`

Do you see a positive relationship? You could always verify what you see with:

`cov(ED, LWAGE)`

or

`cor(ED, LWAGE)`

If you want to visualize the relationship between more than one variable, try:

`pairs(~LWAGE+EXP+WKS+ED)`

Finally, run an OLS regression by typing:

`summary(lm(LWAGE ~ EXP + EXP^2 + WKS + OCC + IND + SOUTH + SMSA + MS + UNION + ED + FEM + BLK))`

What are the estimated returns to schooling? Is this estimate statistically significant?

To save your work, make sure the script window is active, then click *File, Save*.

REFERENCES

Cornwell, C., & Rupert, P. (1988). Efficient estimation with panel data: An empirical comparison of instrumental variables estimators. *Journal of Applied Econometrics*, 3(2), 149-155.

Greene, W. H. (2011). *Econometric analysis* 7th edition. Prentice Hall, Upper Saddle River.

ECON 7010: Econometrics I**Matrices: Concepts, Definitions &
Some Basic Results****1. Concepts and Definitions****Vector**

A “vector” is a set of scalar values, or “elements”, placed in a particular order, and then displayed either as a column of values, or a row of values. The number of elements in the vector gives us the vector’s “dimension”.

So, the vector $v_1 = (2 \ 6 \ 3 \ 8)$ is a row vector with 4 elements – it is a (1×4) vector, because it has 1 row with 4 elements. We can also think of these elements as being located in “column” positions, so the vector essentially has one row and 4 columns.

Similarly, the vector $v_2 = \begin{pmatrix} 2 \\ 5 \\ 8 \\ 2 \end{pmatrix}$ is a column vector with 4 elements – it is a (4×1) vector, because

it has 1 column with 4 elements. We can think of these elements as being located in “row” positions, so the vector essentially has one column and 4 rows.

Matrix

A “matrix” is rectangular array of values, or “elements”, obtained by taking several column vectors (of the same dimension) and placing them side-by-side in a specific order. Alternatively, we can think of a matrix as being formed by taking several row vectors (of the same dimension) and placing them one above the other, in a particular order.

For example, if we take the vectors $v_2 = \begin{pmatrix} 2 \\ 5 \\ 8 \\ 2 \end{pmatrix}$ and $v_3 = \begin{pmatrix} 1 \\ 6 \\ 2 \\ 7 \end{pmatrix}$ we can form the matrix

$$V_1 = \begin{bmatrix} 2 & 1 \\ 5 & 6 \\ 8 & 2 \\ 2 & 7 \end{bmatrix}. \text{ If we place the vectors side-by-side in the opposite order, we get a}$$

different matrix, of course, namely:

$$V_2 = \begin{bmatrix} 1 & 2 \\ 6 & 5 \\ 2 & 8 \\ 7 & 2 \end{bmatrix}.$$

Dimension of a Matrix

The “dimension” of a matrix is the number of rows and the number of columns. If there are “ m ” rows and “ n ” columns, the dimension of the matrix is $(m \times n)$. You can see how the way in which the dimension of a vector was defined above is just a special case of this concept.

For example, the matrix $A = \begin{bmatrix} 7 & 3 & 1 \\ 6 & 4 & 3 \\ 5 & 8 & 9 \end{bmatrix}$ is a (3×3) matrix, while the dimension of the matrix

$$D = \begin{bmatrix} 1 & 8 \\ 6 & 5 \\ 8 & 9 \end{bmatrix} \text{ is } (3 \times 2).$$

Square Matrix

A matrix is “square” if it has the same number of rows as columns.

The matrix $A = \begin{bmatrix} 7 & 3 & 1 \\ 6 & 4 & 3 \\ 5 & 8 & 9 \end{bmatrix}$ is square, as it has 3 rows and 3 columns. The matrices

$$D = \begin{bmatrix} 1 & 8 \\ 6 & 5 \\ 8 & 9 \end{bmatrix} \text{ and } E = \begin{bmatrix} 1 & 0 & 8 \\ 8 & 2 & 9 \end{bmatrix} \text{ are not square – they are “rectangular”}.$$

Rectangular Matrix

A rectangular matrix is one whose number of columns is different from its number of rows.

The matrices $D = \begin{bmatrix} 1 & 8 \\ 6 & 5 \\ 8 & 9 \end{bmatrix}$ and $E = \begin{bmatrix} 1 & 0 & 8 \\ 8 & 2 & 9 \end{bmatrix}$ are “rectangular”. The matrix D has 3 rows and 2 columns – it is (3×2) . The matrix E has 2 rows and 3 columns – it is (2×3) .

Leading Diagonal

If the matrix is square, the “leading diagonal” is the string of elements from the top left corner of the matrix to the bottom right corner.

If $A = \begin{bmatrix} 7 & 3 & 1 \\ 6 & 4 & 3 \\ 5 & 8 & 9 \end{bmatrix}$, its leading diagonal contains the elements $(7, 4, 9)$.

Diagonal Matrix

A square matrix is said to be “diagonal” if the only non-zero elements in the matrix occur along the leading diagonal.

The matrix $C = \begin{bmatrix} 7 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 9 \end{bmatrix}$ is a diagonal matrix.

Scalar Matrix

A square matrix is said to be “scalar” if it is diagonal, and all of the elements of its leading diagonal are the same.

The matrix $B = \begin{bmatrix} 7 & 0 & 0 \\ 0 & 7 & 0 \\ 0 & 0 & 7 \end{bmatrix}$ is “scalar”, but the matrix $C = \begin{bmatrix} 7 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 9 \end{bmatrix}$ is not.

Identity Matrix

An “identity” matrix is one which is scalar, with the value “1” for each element on the leading diagonal. (Because this matrix is scalar, it is also a square and diagonal matrix.)

The matrix $I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ is an identity matrix. (We might also name it I_3 to indicate that it is a (3×3) identity matrix.)

An identity matrix serves the same purpose as the number “1” for scalars – if we pre-multiply or post-multiply a matrix by the identity matrix (of the right dimensions), the original matrix is unchanged.

So, if $I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ and $D = \begin{bmatrix} 1 & 8 \\ 6 & 5 \\ 8 & 9 \end{bmatrix}$, then $ID = D = DI$.

Null Matrix

A “null matrix” is one which has the value zero for all of its elements. The matrices

$Z = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ and $N = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$ are both null matrices.

A null matrix serves the same purpose as the number “0” for scalars – if we pre-multiply or post-multiply a matrix by the identity matrix (of the right dimensions), the result is a null matrix.

So, if $Z = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ and $D = \begin{bmatrix} 1 & 8 \\ 6 & 5 \\ 8 & 9 \end{bmatrix}$, then $ZD = N$. [Note that Z is (3×3) , and D

is (3×2) , so ZD must be (3×2) .]

Trace

The “trace” of a square matrix is the sum of the elements on its leading diagonal.

For example, if $A = \begin{bmatrix} 7 & 3 & 1 \\ 6 & 4 & 3 \\ 5 & 8 & 9 \end{bmatrix}$, then $\text{trace}(A) = (7 + 4 + 9) = 20$.

Transpose

The “transpose” of a matrix is obtained by exchanging all of the rows for all of the columns. That is, the first row becomes the first column; the second row becomes the second column; and so on.

If $D = \begin{bmatrix} 1 & 8 \\ 6 & 5 \\ 8 & 9 \end{bmatrix}$, then the transpose of D is $D' = \begin{bmatrix} 1 & 6 & 8 \\ 8 & 5 & 9 \end{bmatrix}$. Sometimes we write D^T

rather than D' to denote the transpose of a matrix. Note that if the original matrix is an $(m \times n)$ matrix, then its transpose will be an $(n \times m)$.

Recall that a vector is just a special type of matrix – a matrix with either just one row, or just one column. So, when we transpose a row vector we just get a column vector with the elements in the same order; and when we transpose a column vector we just get a row vector, with the order of the elements unaltered.

For example, when we transpose the (1×4) row vector, $v_1 = (2 \ 6 \ 3 \ 8)$, we get a column vector which is (4×1) :

$$v_1' = \begin{pmatrix} 2 \\ 6 \\ 3 \\ 8 \end{pmatrix}.$$

Symmetric Matrix

A square matrix is “symmetric” if it is equal to its own transpose – that is, transposing the rows and columns of the matrix leaves it unchanged. In other words, as we look at elements above and below the leading diagonal, we see the same values in corresponding positions – the (i, j) 'th element equals the (j, i) 'th element, for all $i \neq j$.

For example, let $F = \begin{bmatrix} 1 & 5 & 6 \\ 5 & 2 & 4 \\ 6 & 4 & 9 \end{bmatrix}$. Here the $(1, 3)$ element and the $(3, 1)$ element are both 6, *etc.*

Note that $F' = F$, so F is symmetric.

Linear Dependency

Two vectors (and hence two rows, or two columns of a matrix) are “linearly independent” if one vector *cannot* be written as a multiple of the other. So, for example, the vectors $x_1 = (1, 3, 4, 6)$ and $x_2 = (5, 4, 1, 8)$ are linearly independent, but the vectors $x_3 = (1, 2, 4, 8)$ and $x_4 = (2, 4, 8, 16)$ are “linearly dependent”, because $x_4 = 2x_3$.

More generally, a collection of (say) n vectors is linearly independent if no one of the vectors can be written as a linear combination (weighted sum) of the remaining $(n - 1)$ vectors. Consider the

vectors x_1 and x_2 above, together with the vector $x_5 = (4, 1, -3, 2)$. These three vectors are *not* linearly independent, because $x_5 = x_2 - x_1$.

Rank of a Matrix

The “rank” of a matrix is the (smaller of the) number of linearly independent rows or columns in the matrix.

For example, the matrix $D = \begin{bmatrix} 1 & 8 \\ 6 & 5 \\ 8 & 9 \end{bmatrix}$ has a rank of “2”. It has 2 columns, and the first

column is *not* a multiple of the second column. The columns are linearly independent. It has 3 rows – these three rows make up a group of 3 linearly independent vectors, but by convention we define “rank” in terms of the smaller of the number of rows and columns. So this matrix has “full rank”.

On the other hand, the matrix $G = \begin{bmatrix} 1 & 5 & 6 \\ 5 & 2 & 7 \\ 6 & 4 & 10 \end{bmatrix}$ has a rank of “2”, because the third

column is the sum of the first two columns. In this case the matrix has “less than full rank”, because potentially it could have had a rank of “3”, but the one linear dependency reduces the rank below this potential value.

Determinant of a Matrix

The determinant of a (square) matrix is a particular polynomial in the elements of the matrix, and is a scalar quantity. We usually denote the determinant of a matrix A by $|A|$, or $\det.(A)$.

The determinant of a scalar is just the scalar itself.

The determinant of a (2×2) matrix is obtained as follows:

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = (a_{11}a_{22}) - (a_{21}a_{12}).$$

If the matrix is (3×3) , then

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12} \begin{vmatrix} a_{11} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13} \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}$$

$$= a_{11}(a_{22}a_{33} - a_{32}a_{23}) - a_{12}(a_{11}a_{33} - a_{31}a_{23}) + a_{13}(a_{21}a_{32} - a_{31}a_{22})$$

which can then be expanded out completely, and we see that it is just a polynomial in the a_{ij} elements.

Principal Minor Matrices

Let A be an $(n \times n)$ matrix. Then the “principal minor matrices” of A are the sub-matrices formed by deleting the last $(n - 1)$ rows and columns (which leaves only first diagonal element); then deleting the last $(n - 2)$ rows and columns (which leaves the leading (2×2) block of A); then deleting the last $(n - 3)$ rows and columns; *etc.*

If $A = \begin{bmatrix} 7 & 3 & 1 \\ 6 & 4 & 3 \\ 5 & 8 & 9 \end{bmatrix}$, its first principal minor matrix is $A_{(1)} = 7$; the second principal minor

matrix is $A_{(2)} = \begin{bmatrix} 7 & 3 \\ 6 & 4 \end{bmatrix}$; and the third is just A itself.

Note: The term “principal minor” is often used as an abbreviation for “determinant of the principal minor matrix”, so you need to be careful.

Inverse Matrix

Suppose that we have a square matrix, A . If we can find a matrix B , with the same dimension as A , such that $AB = BA = I$ (an identity matrix), then B is called the “inverse matrix” for A , and we denote it as $B = A^{-1}$.

Clearly, the inverse matrix corresponds to the reciprocal when we are dealing with scalar numbers. Note, however, that many square matrices *do not* have an inverse.

Singular Matrix

A square matrix that does *not* have an inverse is said to be a “singular matrix”. On the other hand, if the inverse matrix *does* exist, the matrix is said to be “non-singular”.

For example, every null matrix is singular. Similarly every identity matrix is non-singular, and equal to its own inverse (just as $1/1 = 1$ in the case of scalars).

Computing an Inverse Matrix

You will not have to construct inverse matrices by hand, except in very simple cases – a computer can be used instead. It is worth knowing how to obtain the inverse of a (non-singular) matrix when the matrix is just (2×2) . In this case we first obtain the determinant of the matrix. We then interchange the 2 elements on the leading diagonal of the matrix, and change the signs of the 2 off-diagonal elements. Finally, we divide this transformed matrix by the determinant. Of course, this can only be done if the determinant is non-zero! So, a necessary (but not sufficient) condition for a matrix to be non-singular is that its determinant is non-zero.

To illustrate these calculations, consider the matrix

$R = \begin{bmatrix} 4 & -1 \\ 1 & -2 \end{bmatrix}$. Its determinant is $\Delta = [(4)(-2) - (1)(-1)] = [-8 + 1] = -7$. So, the inverse of

R is the matrix

$$R^{-1} = \left(\frac{1}{\Delta} \right) \begin{bmatrix} -2 & 1 \\ -1 & 4 \end{bmatrix} = \begin{bmatrix} 2/7 & -1/7 \\ 1/7 & -4/7 \end{bmatrix}. \text{ You can check that } RR^{-1} = R^{-1}R = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Definiteness of a Matrix

Suppose that A is any square $(n \times n)$ matrix. The A is “positive definite” if the (scalar) quadratic form, $x'Ax > 0$, for *all* non-zero $(n \times 1)$ vectors, x ; A is “positive semi-definite” if the (scalar) quadratic form, $x'Ax \geq 0$, for *all* non-zero $(n \times 1)$ vectors, x ; A is “negative definite” if the (scalar) quadratic form, $x'Ax < 0$, for *all* non-zero $(n \times 1)$ vectors, x ; and A is “negative semi-definite” if the (scalar) quadratic form, $x'Ax \leq 0$, for *all* non-zero $(n \times 1)$ vectors, x . If the sign of $x'Ax$ varies with the choice of x , then A is said to be “indefinite”.

For example, let $A = \begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix}$. Then

$$x'Ax = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} 4x_1 \\ 2x_2 \end{pmatrix} = 4x_1^2 + 2x_2^2 > 0, \text{ unless}$$

both x_1 and x_2 are zero. So, A is positive definite in this case.

Idempotent Matrix

Suppose that we have a square and symmetric matrix, Q , which has the property that $Q^2 = Q$. Because Q is symmetric, this means that $Q'Q = QQ' = QQ = Q^2 = Q$. Any matrix with this property is called an “idempotent matrix”.

Clearly, the identity matrix, and the null matrix are idempotent. This corresponds with the fact that the only two idempotent scalar numbers are unity and zero. However, other matrices can also be idempotent.

Let X be an $(T \times k)$ matrix, with $T > k$, and such that the square, $(k \times k)$ matrix $(X'X)$ has an inverse (i.e., it is non-singular). Let $P = X(X'X)^{-1}X'$. Note that P is an $(T \times T)$ matrix, so it is square; and also note that

$$P' = [X(X'X)^{-1}X']' = (X')'[(X'X)^{-1}]'X' = X[(X'X)^{-1}]'X' = X(X'X)^{-1}X' = P.$$

That is, P is symmetric. Now, observe that

$$\begin{aligned} P'P &= [X(X'X)^{-1}X']'X(X'X)^{-1}X' = X[(X'X)^{-1}]'X'X(X'X)^{-1}X' = XI(X'X)^{-1}X' \\ &= X(X'X)^{-1}X' = P \end{aligned}$$

and so P is idempotent. You can also check that the matrix $M = (I_T - P)$ is another example of an idempotent matrix.

2. Some Basic Matrix Results

Let A be a square $(n \times n)$ matrix. Then:

1. Let X be an $(m \times n)$ matrix with full rank. Then (XAX') is positive definite if A is positive definite.
2. If A is non-singular (that is, it has an inverse) then it is either positive definite, or negative definite, and its determinant is non-zero.
3. If A is positive semi-definite or negative semi-definite, then its determinant is zero, and it is singular (it does not have an inverse).
4. If A is positive definite then the determinant of A is positive.
5. If A is positive (semi-) definite then all of the leading diagonal elements of A are positive (non-negative).
6. If A is negative (semi-) definite then all of the leading diagonal elements of A are negative (non-positive).
7. A is positive definite *if and only if* the determinants of all of its principal minor matrices are positive.
8. A is negative definite *if and only if* the determinants of the principal minor matrices of order k have sign $(-1)^k$, $k = 1, 2, \dots, n$. (That is, -, +, -, +,.....)
9. Suppose that B is also $(n \times n)$, and that both A and B are non-singular. Then the definiteness of $(A - B)^{-1}$ is the same as the definiteness of $(B^{-1} - A^{-1})$.

10. If A is either positive definite or negative definite, then $\text{rank}(A) = n$.
11. If A is positive semi-definite or negative semi-definite, then $\text{rank}(A) = r < n$.
12. If A is idempotent then it is positive semi-definite.
13. If A is idempotent then $\text{rank}(A) = \text{trace}(A)$, where the trace is the sum of the leading diagonal elements.
14. If C is an $(m \times n)$ matrix, then the rank of C cannot exceed $\min(m, n)$.
15. If A is positive semi-definite or negative semi-definite, then $\text{rank}(A) = r < n$, and it has " r " non-zero eigenvalues
16. If A is either positive definite or negative definite then all of its eigenvalues are non-zero.
17. Suppose that A and B are both $(n \times n)$ matrices. Then $\text{trace}(A + B) = \text{trace}(A) + \text{trace}(B)$.
18. Suppose that A and B are both $(n \times n)$ matrices. Then $(A + B)' = (A' + B')$.
19. Suppose that A is a *non-singular* $(n \times n)$ matrix, then $(A^{-1})' = (A')^{-1}$.
20. Suppose that A and B have dimensions such that AB is defined. Then $(AB)' = (B'A')$.
21. Suppose that A and B are *non-singular* $(n \times n)$ matrices such that both AB and BA are defined. Then $(AB)^{-1} = (B^{-1}A^{-1})$.
22. If D is a square diagonal matrix which is non-singular, then D^{-1} is also diagonal, and the elements of the leading diagonal are the reciprocals of those on the diagonal of D itself.

Topic 1 – Continued.....

Finite-Sample Properties of the LS Estimator

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad ; \quad \boldsymbol{\varepsilon} \sim N[0, \sigma^2 \mathbf{I}_n]$$

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = f(\mathbf{y})$$

$\boldsymbol{\varepsilon}$ is random \Rightarrow \mathbf{y} is random \Rightarrow \mathbf{b} is random

- \mathbf{b} is an *estimator* of $\boldsymbol{\beta}$. It is a function of the *random* sample data.
- \mathbf{b} is a “statistic”.
- \mathbf{b} has a probability distribution – called its *Sampling Distribution*.
- Interpretation of *sampling distribution* –

Repeatedly draw all possible samples of size n .

Calculate values of \mathbf{b} each time.

Construct relative frequency distribution for the \mathbf{b} values and probability of occurrence.

It is a *hypothetical* construct. Why?

- Sampling distribution offers *one* basis for answering the question:

“How good is \mathbf{b} as an estimator of $\boldsymbol{\beta}$?”

Note:

Quality of estimator is being assessed in terms of performance in *repeated samples*. Tells us nothing about quality of estimator for *one particular sample*.

- Let’s explore some of the properties of the LS estimator, \mathbf{b} , and build up its sampling distribution.
- Introduce some general results, and apply them to our problem.

Definition: An estimator, $\hat{\theta}$ is an *unbiased* estimator of the parameter vector, θ , if $E[\hat{\theta}] = \theta$.

That is, $E[\hat{\theta}(\mathbf{y})] = \theta$.

That is, $\int \hat{\theta}(\mathbf{y})p(\mathbf{y} | \theta)d\mathbf{y} = \theta$.

The quantity, $\mathbf{B}(\theta, \mathbf{y}) = E[\hat{\theta}(\mathbf{y}) - \theta]$, is called the “Bias” of $\hat{\theta}$.

Example: $\{y_1, y_2, \dots, y_n\}$ is a random sample from population with a finite mean, μ , and a finite variance, σ^2 .

Consider the *statistic* $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$.

$$\begin{aligned} \text{Then, } E[\bar{y}] &= E\left[\frac{1}{n} \sum_{i=1}^n y_i\right] = \frac{1}{n} \sum_{i=1}^n E(y_i) \\ &= \frac{1}{n} \sum_{i=1}^n \mu = \left(\frac{1}{n} n\mu\right) = \mu. \end{aligned}$$

So, \bar{y} is an *unbiased estimator* of the parameter, μ .

- Here, there are lots of possible unbiased estimators of μ .
- So, need to consider additional characteristics of estimators to help choose.

Return to our LS problem –

$$\mathbf{b} = (X'X)^{-1}X'\mathbf{y}$$

- Recall – either assume that X is *non-random*, or condition on X .
- We’ll assume X is non-random – get same result if we condition on X .

Then: $E(\mathbf{b}) = E[(X'X)^{-1}X'\mathbf{y}] = (X'X)^{-1}X'E(\mathbf{y})$

So,

$$\begin{aligned} E(\mathbf{b}) &= (X'X)^{-1}X'E[X\boldsymbol{\beta} + \boldsymbol{\varepsilon}] = (X'X)^{-1}X'[X\boldsymbol{\beta} + E(\boldsymbol{\varepsilon})] \\ &= (X'X)^{-1}X'[X\boldsymbol{\beta} + \mathbf{0}] = (X'X)^{-1}X'X\boldsymbol{\beta} \\ &= \boldsymbol{\beta}. \end{aligned}$$

The LS estimator of $\boldsymbol{\beta}$ is Unbiased

Definition: Any estimator that is a *linear function* of the random sample data is called a *Linear Estimator*.

Example: $\{y_1, y_2, \dots, y_n\}$ is a random sample from population with a finite mean, μ , and a finite variance, σ^2 .

Consider the *statistic* $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{n} [y_1 + y_2 + \dots + y_n]$.

This statistic is a *linear estimator* of μ .

(Note that the “weights” are non-random.)

Return to our LS problem –

$$\mathbf{b} = (X'X)^{-1}X'\mathbf{y} = A\mathbf{y}$$

$$\begin{matrix} (k \times 1) & & (k \times n)(n \times 1) \end{matrix}$$

Note that, under our assumptions, A is a *non-random* matrix.

So,

$$\begin{pmatrix} b_1 \\ \vdots \\ b_k \end{pmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \cdots & \vdots \\ a_{k1} & \cdots & a_{kn} \end{bmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}.$$

For example, $b_1 = [a_{11}y_1 + a_{12}y_2 + \dots + a_{1n}y_n]$; etc.

The LS estimator, \mathbf{b} , is a linear (& unbiased) estimator of $\boldsymbol{\beta}$

Now let's consider the dispersion (variability) of \mathbf{b} , as an estimator of $\boldsymbol{\beta}$.

Definition: Suppose we have an $(n \times 1)$ random vector, \mathbf{x} . Then the *Covariance Matrix* of \mathbf{x} is defined as the $(n \times n)$ matrix:

$$V(\mathbf{x}) = E[(\mathbf{x} - E(\mathbf{x}))(\mathbf{x} - E(\mathbf{x}))'].$$

- Diagonal elements of $V(\mathbf{x})$ are $var.(x_1), \dots, var.(x_n)$.
- Off-diagonal elements are $covar.(x_i, x_j)$; $i, j = 1, \dots, n$; $i \neq j$.

Return to our LS problem –

We have a $(k \times 1)$ random vector, \mathbf{b} , and we know that $E(\mathbf{b}) = \boldsymbol{\beta}$.

$$V(\mathbf{b}) = E[(\mathbf{b} - E(\mathbf{b}))(\mathbf{b} - E(\mathbf{b}))']$$

Now,

$$\begin{aligned} \mathbf{b} &= (X'X)^{-1}X'\mathbf{y} = (X'X)^{-1}X'(X\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\ &= (X'X)^{-1}(X'X)\boldsymbol{\beta} + (X'X)^{-1}X'\boldsymbol{\varepsilon} \\ &= I\boldsymbol{\beta} + (X'X)^{-1}X'\boldsymbol{\varepsilon}. \end{aligned}$$

So,

$$(\mathbf{b} - \boldsymbol{\beta}) = (X'X)^{-1}X'\boldsymbol{\varepsilon}. \quad [*]$$

Using the result, [*], in $V(\mathbf{b})$, we have:

$$\begin{aligned} V(\mathbf{b}) &= E\{[(X'X)^{-1}X'\boldsymbol{\varepsilon}][(X'X)^{-1}X'\boldsymbol{\varepsilon}']\} \\ &= (X'X)^{-1}X'E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}']X(X'X)^{-1}. \end{aligned}$$

We showed, earlier, that because $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, $V(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2 I_n$.

(What other assumptions did we use to get this result?)

So, we have:

$$\begin{aligned} V(\mathbf{b}) &= (X'X)^{-1}X'E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}']X(X'X)^{-1} = (X'X)^{-1}X'\sigma^2IX(X'X)^{-1} = \sigma^2(X'X)^{-1}(X'X)(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}. \end{aligned}$$

$$\begin{aligned} V(\mathbf{b}) &= \sigma^2(X'X)^{-1} \\ &\quad (k \times k) \end{aligned}$$

Interpret diagonal and off-diagonal elements of this matrix.

Finally, because the error term, $\boldsymbol{\varepsilon}$ is assumed to be Normally distributed,

1. $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$: this implies that \mathbf{y} is also Normally distributed. (Why?)
2. $\mathbf{b} = (X'X)^{-1}X'\mathbf{y} = A\mathbf{y}$: this implies that \mathbf{b} is also Normally distributed.

So, we now have the full **Sampling Distribution** of the LS estimator, \mathbf{b} :

$$\mathbf{b} \sim N[\boldsymbol{\beta}, \sigma^2(X'X)^{-1}]$$

Note:

- This result depends on our various, *rigid*, assumptions about the various components of the regression model.
- The Normal distribution here is a “*multivariate Normal*” distribution.
(See handout on “*Spherical Distributions*”.)
- As with estimation of population mean, $\boldsymbol{\mu}$, in previous example, there are lots of other *unbiased* estimators of $\boldsymbol{\beta}$ in the model $= X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.
- How might we choose between these possibilities? Is *linearity* desirable?

- We need to consider other *desirable* properties that these unbiased estimators may have.
- *One option* is to take account of estimators' *precisions*.

Definition: Suppose we have two *unbiased* estimators, $\widehat{\theta}_1$ and $\widehat{\theta}_2$, of the (scalar) parameter, θ . Then we say that $\widehat{\theta}_1$ is **at least as efficient** as $\widehat{\theta}_2$ if $var.(\widehat{\theta}_1) \leq var.(\widehat{\theta}_2)$.

Note:

1. The variance of an estimator is just the variance of its sampling distribution.
 2. "Efficiency" is a *relative* concept.
 3. What if there are 3 or more unbiased estimators being compared?
- What if one or more of the estimators being compared is *biased* ?
 - In this case we can take account of both variance, and any bias, at the same time by using "*mean squared error*" (MSE) of the estimators.

Definition: Suppose that $\widehat{\theta}$ is an estimator of the (scalar) parameter, θ . Then the MSE of $\widehat{\theta}$ is defined as:

$$MSE(\widehat{\theta}) = E[(\widehat{\theta} - \theta)^2].$$

Note that:

$$MSE(\widehat{\theta}) = var.(\widehat{\theta}) + [Bias(\widehat{\theta})]^2$$

To prove this, write:

$$MSE(\widehat{\theta}) = E[(\widehat{\theta} - \theta)^2] = E\{[(\widehat{\theta}) - E(\widehat{\theta})] + (E(\widehat{\theta}) - \theta)\}^2,$$

expand out, and note that

$$E[E(\widehat{\theta})] = E(\widehat{\theta});$$

and

$$E[\widehat{\theta} - E(\widehat{\theta})] = 0.$$

Definition: Suppose we have two (possibly) *biased* estimators, $\hat{\theta}_1$ and $\hat{\theta}_2$, of the (scalar) parameter, θ . Then we say $\hat{\theta}_1$ is **at least as efficient** as $\hat{\theta}_2$ if $MSE(\hat{\theta}_1) \leq MSE(\hat{\theta}_2)$.

If we extend all of this to the case where we have a vector of parameters, $\boldsymbol{\theta}$, then we have the following definitions:

Definition: Suppose we have two *unbiased* estimators, $\hat{\theta}_1$ and $\hat{\theta}_2$, of the parameter vector, $\boldsymbol{\theta}$. Then we say that $\hat{\theta}_1$ is **at least as efficient** as $\hat{\theta}_2$ if $\Delta = V(\hat{\theta}_2) - V(\hat{\theta}_1)$ is *at least positive semi-definite*.

Definition: Suppose we have two (possibly) *biased* estimators, $\hat{\theta}_1$ and $\hat{\theta}_2$, of the parameter vector, $\boldsymbol{\theta}$. Then we say that $\hat{\theta}_1$ is **at least as efficient** as $\hat{\theta}_2$ if $\Delta = MMSE(\hat{\theta}_2) - MMSE(\hat{\theta}_1)$ is *at least positive semi-definite*.

Note: $MMSE(\hat{\theta}) = E[(\hat{\theta} - \boldsymbol{\theta})(\hat{\theta} - \boldsymbol{\theta})'] = V[\hat{\theta}] + Bias(\hat{\theta})Bias(\hat{\theta})'$.

Taking account of its *linearity*, *unbiasedness*, and its *precision*, in what sense is the LS estimator, \mathbf{b} , of $\boldsymbol{\beta}$ *optimal*?

Theorem (Gauss-Markhov):

In the "standard" linear regression model, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, the LS estimator, \mathbf{b} , of $\boldsymbol{\beta}$ is **Best Linear Unbiased** (BLU). That is, it is **Efficient** in the class of all linear and unbiased estimators of $\boldsymbol{\beta}$.

1. Is this an *interesting* result?
2. What *assumptions* about the "standard" model are we going to exploit?

Proof

Let \mathbf{b}_0 be any other *linear* estimator of $\boldsymbol{\beta}$:

$$\mathbf{b}_0 = C\mathbf{y} \quad ; \quad \text{for some non-random } C .$$

$$(k \times 1) \quad (k \times n)(n \times 1)$$

Now, $V(\mathbf{b}_0) = CV(\mathbf{y})C' = C(\sigma^2 I_n)C' = \sigma^2 CC'$

$$(k \times k)$$

Define: $D = C - (X'X)^{-1}X'$

so that $D\mathbf{y} = C\mathbf{y} - (X'X)^{-1}X'\mathbf{y} = \mathbf{b}_0 - \mathbf{b}$.

Now restrict \mathbf{b}_0 to be *unbiased*, so that $E(\mathbf{b}_0) = E(C\mathbf{y}) = CX\boldsymbol{\beta} = \boldsymbol{\beta}$.

This requires that $CX = I$, which in turn implies that

$$DX = [C - (X'X)^{-1}X']X = CX - I = \mathbf{0} \quad (\text{and } D'X' = 0)$$

(What assumptions have we used so far?)

Now, focus on covariance matrix of \mathbf{b}_0 :

$$\begin{aligned} V(\mathbf{b}_0) &= \sigma^2 [D + (X'X)^{-1}X'] [D + (X'X)^{-1}X']' \\ &= \sigma^2 [DD' + (X'X)^{-1}X'X(X'X)^{-1}] \quad ; \quad DX = \mathbf{0} \\ &= \sigma^2 DD' + \sigma^2 (X'X)^{-1} \\ &= \sigma^2 DD' + V(\mathbf{b}), \end{aligned}$$

or, $[V(\mathbf{b}_0) - V(\mathbf{b})] = \sigma^2 DD' \quad ; \quad \sigma^2 > 0$

Now we just have to "sign" this (matrix) difference:

$$\boldsymbol{\eta}'(DD')\boldsymbol{\eta} = (D'\boldsymbol{\eta})'(D'\boldsymbol{\eta}) = \mathbf{v}'\mathbf{v} = \sum_{i=1}^n v_i^2 \geq 0 .$$

So, $\Delta = [V(\mathbf{b}_0) - V(\mathbf{b})]$ is a p.s.d. matrix, implying that \mathbf{b}_0 is *relatively less efficient* than \mathbf{b} .

Result:

The LS estimator is the Best Linear Unbiased estimator of $\boldsymbol{\beta}$.

- What assumptions did we use, and where?
- Were there any standard assumptions that we *didn't* use?
- What does this suggest?

Estimating σ^2

- We now know a lot about estimating $\boldsymbol{\beta}$.
- There's another parameter in the regression model - σ^2 - the variance of each ε_i .
- Note that $\sigma^2 = \text{var.}(\varepsilon_i) = E[(\varepsilon_i - E(\varepsilon_i))^2] = E(\varepsilon_i^2)$.
- The *sample* counterpart to this *population* parameter is the *sample* average of the "residuals": $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n e_i^2 = \frac{1}{n} \mathbf{e}'\mathbf{e}$.
- However, there is a *distortion* in this estimator of σ^2 .
- Although mean of e_i 's is zero (if intercept in model), not all of e_i 's are independent of each other - only $(n - k)$ of them are.
- Why does this distort our potential estimator, $\hat{\sigma}^2$?

Note that: $e_i = (y_i - \hat{y}_i) = (y_i - \mathbf{x}_i'\mathbf{b})$

$$= (\mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i) - \mathbf{x}_i'\mathbf{b}$$

$$= \varepsilon_i + \mathbf{x}_i'(\boldsymbol{\beta} - \mathbf{b})$$

Let's see what properties $\hat{\sigma}^2$ has as an estimator of σ^2 :

$$\mathbf{e} = (\mathbf{y} - \hat{\mathbf{y}}) = (\mathbf{y} - \mathbf{X}\mathbf{b}) = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{M}\mathbf{y},$$

where

$$M = I_n - X(X'X)^{-1}X' \quad ; \quad \textit{idempotent}, \text{ and } MX = 0 .$$

So, $\mathbf{e} = M\mathbf{y} = M(X\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = M\boldsymbol{\varepsilon}$,

and $\mathbf{e}'\mathbf{e} = (M\boldsymbol{\varepsilon})'(M\boldsymbol{\varepsilon}) = \boldsymbol{\varepsilon}'M\boldsymbol{\varepsilon}$; *scalar*

From this, we see that:

$$\begin{aligned} E(\mathbf{e}'\mathbf{e}) &= E[\boldsymbol{\varepsilon}'M\boldsymbol{\varepsilon}] = E[\textit{tr}.(\boldsymbol{\varepsilon}'M\boldsymbol{\varepsilon})] = E[\textit{tr}.(M\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')] \\ &= \textit{tr}.[ME(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')] = \textit{tr}.[M\sigma^2I_n] = \sigma^2\textit{tr}.(M) \\ &= \sigma^2(n - k) \end{aligned}$$

So:

$$E(\hat{\sigma}^2) = E\left(\frac{1}{n}\mathbf{e}'\mathbf{e}\right) = \frac{1}{n}(n - k)\sigma^2 < \sigma^2 \quad ; \quad \mathbf{BIASED}$$

Easy to convert this to an *Unbiased estimator* –

$$s^2 = \frac{1}{(n - k)} \mathbf{e}'\mathbf{e}$$

- “ $(n - k)$ ” is the “*degrees of freedom*” – number of independent sources of information in the “ n ” residuals (e_i 's).
- We can use “ s ” as an estimator of σ , but it is a *biased estimator*.
- Call “ s ” the “*standard error of the regression*”, or the “*standard error of estimate*”.
- s^2 is a *statistic* – has its own sampling distribution, *etc.* More on this to come.
- Let's see one immediate *application* of s^2 and s .
- Recall sampling distribution for LS estimator, \mathbf{b} :

$$\mathbf{b} \sim N[\boldsymbol{\beta} , \sigma^2(X'X)^{-1}]$$
- So, $\textit{var.}(b_i) = \sigma^2[(X'X)^{-1}]_{ii}$; σ^2 is *unobservable*.

- If we want to report variability associated with b_i as an estimator of β_i , we need to use estimator of σ^2 .
- $est. var. (b_i) = s^2[(X'X)^{-1}]_{ii}$.
- $\sqrt{est. var. (b_i)} = s.d. (b_i) = s\{[(X'X)^{-1}]_{ii}\}^{1/2}$.
- We call this the “*standard error*” of b_i .
- This quantity will be very important when it comes to constructing *interval estimates* of our regression coefficients; and when we construct *tests of hypotheses* about these coefficients.

Confidence Intervals & Hypothesis Testing

- So far, we’ve concentrated on “*point*” estimation.
- Need to move on – to do this we’ll need the full sampling distributions of **both** \mathbf{b} and s^2 .
- We will make use of the assumption of *Normally distributed* errors.
- Recall that:

$$\mathbf{b} \sim N[\boldsymbol{\beta}, \sigma^2(X'X)^{-1}]$$

$$b_i \sim N[\beta_i, \sigma^2((X'X)^{-1})_{ii}] \quad ; \quad \text{why still } \textit{Normal}?$$

- So, we can *standardize*:

$$z_i = (b_i - \beta_i) / \sqrt{\sigma^2[(X'X)^{-1}]_{ii}}$$

- But σ^2 is *unknown*, so we can’t use z_i directly to draw inferences about b_i .

Need some preliminary results in order to proceed from here –

Definition: Let $z \sim N[0, 1]$. Then z^2 has a “*Chi-Square*” distribution with one “degree of freedom”.

Definition: Let $z_1, z_2, z_3, \dots, z_m$ be independent $N[0, 1]$ variates. Then the quantity $\sum_{i=1}^m (z_i^2)$ has a Chi-Square distribution with “ m ” d.o.f.

Theorem: Let $\mathbf{x} \sim N[\mathbf{0}, V]$, and let A be a fixed matrix. Then the *quadratic form*, $'A\mathbf{x}$, follows a Chi-Square distribution with r ($= rank(A)$) degrees of freedom, iff AV is an *idempotent matrix*.

Definition: Let $z \sim N[0, 1]$, and let $x \sim \chi_{(v)}^2$, where z and x are *independent*. Then the statistic, $t = z/\sqrt{x/v}$ follows *Student's t distribution*, with “ v ” degrees of freedom.

Now let's consider the sampling distribution of s^2 :

We have
$$s^2 = \frac{1}{(n-k)} \mathbf{e}'\mathbf{e} .$$

So,

$$(n - k)s^2 = (\mathbf{e}'\mathbf{e}) = (\boldsymbol{\varepsilon}'M\boldsymbol{\varepsilon}) .$$

Define the random variable

$$C = \frac{(n-k)s^2}{\sigma^2} = \left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right)' M \left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right) ,$$

where $\boldsymbol{\varepsilon} \sim N[\mathbf{0}, \sigma^2 I_n]$; and so $\left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right) \sim N[\mathbf{0}, I_n]$.

Using the Theorem from last slide, we get the following result for C :

$$C = \left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right)' M \left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right) \sim \chi_{(n-k)}^2 ,$$

because $AV = MI = M$, is *idempotent*, and $r = d.o.f. = rank(A) = rank(M) = tr.(M) = (n - k)$. (Why?)

So, we have the result:

$$\frac{(n - k)s^2}{\sigma^2} \sim \chi_{(n-k)}^2$$

Next, we need to show that b and s^2 are *statistically independent*.

Theorem: Let \mathbf{x} be a normally distributed random vector, and L and A are non-random matrices. Then, the “Linear Form”, $L\mathbf{x}$, and the “Quadratic Form”, $\mathbf{x}'A\mathbf{x}$, are independent if $LA = \mathbf{0}$.

How does this result help us?

- We have $C = \frac{(n-k)s^2}{\sigma^2} = \left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right)'M\left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right)$.
- Also, $\mathbf{b} = (X'X)^{-1}X'\mathbf{y} = (X'X)^{-1}X'(X\boldsymbol{\beta} + \boldsymbol{\varepsilon})$
 $= \boldsymbol{\beta} + (X'X)^{-1}X'\boldsymbol{\varepsilon}$.
- So, $\left[\frac{\mathbf{b}-\boldsymbol{\beta}}{\sigma}\right] = (X'X)^{-1}X'\left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right)$.
- Let $L = (X'X)^{-1}X'$; $A = M$; $\mathbf{x} = \left(\frac{\boldsymbol{\varepsilon}}{\sigma}\right)$
- So, $LA = (X'X)^{-1}X'M = \mathbf{0}$
- This implies that $C = \frac{(n-k)s^2}{\sigma^2}$ and $\left[\frac{\mathbf{b}-\boldsymbol{\beta}}{\sigma}\right]$ are *independent*, and so \mathbf{b} and s^2 are also *statistically independent*.
- C is $\chi^2_{(n-k)}$, and $\left[\frac{\mathbf{b}-\boldsymbol{\beta}}{\sigma}\right] \sim N[\mathbf{0}, (X'X)^{-1}]$, so we immediately get:

Theorem: $t_i = (b_i - \beta_i) / s.e.(b_i)$

has a Student's t distribution with $(n - k)$ d.o.f.

Proof: $\left[\frac{\mathbf{b}-\boldsymbol{\beta}}{\sigma}\right] \sim N[\mathbf{0}, (X'X)^{-1}]$, $\left[\frac{b_i-\beta_i}{\sigma}\right] \sim N[0, ((X'X)^{-1})_{ii}]$

so, $\left[\frac{b_i-\beta_i}{\sigma\sqrt{((X'X)^{-1})_{ii}}}\right] \sim N[0, 1]$.

Also, $C = \frac{(n-k)s^2}{\sigma^2} \sim \chi^2_{(n-k)}$; and we have *independence*.

So, $t_v = N[0, 1] / \sqrt{\chi^2_{(v)}/v}$
 $= \left[\frac{b_i-\beta_i}{\sigma\sqrt{((X'X)^{-1})_{ii}}}\right] / \left[\frac{(n-k)s^2}{\sigma^2} / (n - k)\right]^{1/2}$

$$= \left[\frac{b_i - \beta_i}{s\sqrt{((X'X)^{-1})_{ii}}} \right] = \left[\frac{b_i - \beta_i}{s.e.(b_i)} \right].$$

In this case, $v = (n - k)$, and so:

$$\left[\frac{b_i - \beta_i}{s.e.(b_i)} \right] \sim t_{(n-k)}$$

We can use this to construct *confidence intervals* and *test hypotheses* about β_i .

Note: This last result used all of our assumptions about the linear regression model – including the assumption of *Normality for the errors*.

Example 1:

$$\hat{y} = 1.4 + 0.2x_2 + 0.6x_3$$

(0.7) (0.05) (1.4)

$$H_0: \beta_2 = 0 \quad vs. \quad H_A: \beta_2 > 0$$

$$t = \left[\frac{b_2 - \beta_2}{s.e.(b_2)} \right] = \left[\frac{0.2 - 0}{0.05} \right] = 4 \quad ; \quad \text{suppose } n = 20$$

$$t_c(5\%) = 1.74 \quad ; \quad t_c(1\%) = 2.567 \quad ; \quad \text{d.o.f.} = 17$$

$$t > t_c \Rightarrow \text{Reject } H_0.$$

| Degrees of Freedom | 90th Percentile | 95th Percentile | 97.5th Percentile | 99th Percentile | 99.5th Percentile |
|--------------------|-----------------|-----------------|-------------------|-----------------|-------------------|
| 1 | 3.078 | 6.314 | 12.706 | 31.821 | 63.657 |
| 2 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 |
| : | : | : | : | : | : |
| 15 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 |
| 16 | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 |
| 17 | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 |

Example 2:

$$\hat{y} = 1.4 + 0.2x_2 + 0.6x_3$$

(0.7) (0.05) (1.4)

$$H_0: \beta_1 = 1.5 \quad \text{vs.} \quad H_A: \beta_1 \neq 1.5$$

$$t = \left[\frac{b_1 - \beta_1}{s.e.(b_1)} \right] = \left[\frac{1.4 - 1.5}{0.7} \right] = -0.1429 \quad ; \text{ d.o.f.} = 17$$

$$t_c(5\%) = \pm 2.11$$

$$|t| < t_c \Rightarrow \text{Do Not Reject } H_0$$

(Against H_A , at the 5% significance level.)

Example 3:

$$\hat{y} = 1.4 + 0.2x_2 + 0.6x_3$$

(0.7) (0.05) (1.4)

$$H_0: \beta_1 = 1.5 \quad \text{vs.} \quad H_A: \beta_1 < 1.5$$

$$t = \left[\frac{b_1 - \beta_1}{s.e.(b_1)} \right] = \left[\frac{1.4 - 1.5}{0.7} \right] = -0.1429 \quad ; \text{ d.o.f.} = 17$$

$$p\text{-value} = Pr. [t < -0.1429 | H_0 \text{ is True}]$$

$$\text{in R:} \quad \text{pt}(-0.1429, 17)$$

$$p = 0.444$$

What do you conclude?

Some Properties of Tests:

Null Hypothesis (H_0) Alternative Hypothesis (H_A)

Classical hypothesis testing –

- Assume that H_0 is *TRUE*
- Compute value of test statistic using random sample of data
- Determine *distribution* of the test statistic (*when H_0 is true*)
- Check if observed value of test statistic is likely to occur, *if H_0 is true*
- If this event is sufficiently *unlikely*, then **REJECT H_0** (in favour of H_A)

Note:

1. Can never **accept** H_0 . Why not?
2. What constitutes “*unlikely*” – subjective?
3. Two types of errors we might incur with this process

Type I Error: **Reject H_0** when in fact it is **True**

Type II Error: **Do Not Reject H_0** when in fact it is **False**

- $\text{Pr.}[I] = \alpha =$ Significance level of test = “size” of test
- $\text{Pr.}[II] = \beta$; say
- Value of β will depend on *how* H_0 is **False**. Usually, many ways.
- In classical testing, decide in advance on max. acceptable value of α and then try and design test so as to *minimize* β .
- As β can take different values, may be difficult to design test optimally.
- Why not minimize both? A trade-off for fixed value of n .
- Consider some desirable properties for a test.

Definition:

The “**Power**” of a test is $\Pr.$ [**Reject** H_0 when it is **False**].

So, $\text{Power} = 1 - \Pr.$ [**Do Not Reject** H_0 | H_0 is **False**] = $1 - \beta$.

- As β typically changes, depending on the *way* that H_0 is false, we usually have a **Power Curve**.
- For a fixed value of α , this curve plots Power against parameter value(s).
- We want our tests to have *high power*.
- We want the power of our tests to *increase* as H_0 becomes *increasingly false*.

Property 1

Consider a fixed sample size, n , and a fixed significance level, α .

Then, a test is “**Uniformly Most Powerful**” if its power exceeds (or is no less than) that of *any other test*, for all possible ways that H_0 could be False.

Property 2

Consider a fixed significance level, α .

Then, a test is “**Consistent**” if its power $\rightarrow 1$, as $n \rightarrow \infty$, for all possible ways that H_0 is false.

Property 3

Consider a fixed sample size, n , and a fixed significance level, α .

Then, a test is said to be “Unbiased” its power *never* falls below the significance level.

Property 4

Consider a fixed sample size, n , and a fixed significance level, α .

Then, a test is said to be “**Locally Most Powerful**” if the *slope* of its power curve is greater than the slope of the power curves of all other size – α tests, in a neighbourhood of H_0 .

Note:

- For many testing problems, no UMP test exists. This is why LMP tests are important.
- Why do we use our “t-test” in the regression model –
 1. It is UMP, against 1 –sided alternatives.
 2. It is Unbiased.
 3. It is Consistent.
 4. It is LMP, against both 1-sided and 2-sided alternatives.

Confidence Intervals

We can also use our t-statistic to construct a confidence interval for β_i .

$$Pr. [-t_c \leq t \leq t_c] = (1 - \alpha)$$

$$\Rightarrow Pr. \left[-t_c \leq \left[\frac{b_i - \beta_i}{s.e.(b_i)} \right] \leq t_c \right] = (1 - \alpha)$$

$$\Rightarrow Pr. [-t_c s.e.(b_i) \leq (b_i - \beta_i) \leq t_c s.e.(b_i)] = (1 - \alpha)$$

$$\Rightarrow Pr. [-b_i - t_c s.e.(b_i) \leq (-\beta_i) \leq -b_i + t_c s.e.(b_i)] \\ = (1 - \alpha)$$

$$\Rightarrow Pr. [b_i + t_c s.e.(b_i) \geq \beta_i \geq b_i - t_c s.e.(b_i)] = (1 - \alpha)$$

$$\Rightarrow Pr. [b_i - t_c s.e.(b_i) \leq \beta_i \leq b_i + t_c s.e.(b_i)] = (1 - \alpha)$$

Interpretation –

The interval, $[b_i - t_c s.e.(b_i) , b_i + t_c s.e.(b_i)]$ is *random*.

The parameter, β_i , is *fixed* (but unknown).

If we were to take a sample of n observations, and construct such an interval, and then repeat this exercise many, many, times, then $100(1 - \alpha)\%$ of such intervals would cover the true value of β_i .

If we just construct an interval, for our *given* sample of data, we’ll never know if *this particular* interval covers β_i , or not.

Example 1

$$\hat{y} = 0.3 - 1.4x_2 + 0.7x_3$$

$$(0.1) \quad (1.1) \quad (0.2)$$

Construct a 95% confidence interval for β_1 when $n = 30$.

$$\text{d.o.f.} = (n - k) = 27 \quad ; \quad (\alpha/2) = 0.025$$

$$t_c = \pm 2.052 \quad ; \quad b_1 = 0.3 \quad ; \quad \text{s.e.}(b_1) = 0.1$$

The 95% Confidence Interval is:

$$[b_1 - t_c \text{ s.e.}(b_1) \quad , \quad b_1 + t_c \text{ s.e.}(b_1)]$$

$$\Rightarrow [0.3 - (2.052)(0.1) \quad , \quad 0.3 + (2.052)(0.1)]$$

$$\Rightarrow [0.0948 \quad , \quad 0.5052]$$

Don't forget the units of measurement!

Example 2

$$\hat{y} = 0.3 - 1.4x_2 + 0.7x_3$$

$$(0.1) \quad (1.1) \quad (0.2)$$

Construct a 90% confidence interval for β_2 when $n = 16$.

$$\text{d.o.f.} = (n - k) = 13 \quad ; \quad (\alpha/2) = 0.05$$

$$t_c = \pm 1.771 \quad ; \quad b_2 = -1.4 \quad ; \quad \text{s.e.}(b_2) = 1.1$$

The 95% Confidence Interval is:

$$[b_2 - t_c \text{ s.e.}(b_2) \quad , \quad b_2 + t_c \text{ s.e.}(b_2)]$$

$$\Rightarrow [-1.4 - (1.771)(1.1) \quad , \quad -1.4 + (1.771)(1.1)]$$

$$\Rightarrow [-3.3481 \quad , \quad 0.5481]$$

Don't forget the units of measurement!

Questions:

- Why do we construct the interval *symmetrically* about point estimate, b_i ?
- How can we use a Confidence Interval to test hypotheses?
- For instance, in the last Example, can we reject $H_0: \beta_2 = 0$, against a 2-sided alternative hypothesis?

Introduction to the Monte Carlo Method

Ryan Godwin

ECON 7010

The Monte Carlo method provides a laboratory in which the properties of estimators and tests can be explored. Although the Monte Carlo method is older than the computer, it is associated with repetitive calculations and random number generation, which is greatly assisted by computers.

The Monte Carlo method was used as early as 1933 by Enrico Fermi, and likely contributed to the work that won him the Nobel Prize in 1938 (Anderson, 1986, p. 99). The term “Monte Carlo” was coined in 1947 by Stanislaw Ulam, Nicolas Metropolis, and John von Neumann, and refers to Stanislaw Ulam’s gambling uncle (Metropolis, 1987). The spirit of the method is well captured in the sentiments of Stanislaw Ulam, as he recalls his first thoughts and attempts at practising the method. He was trying to determine the chances that a hand of solitaire would come out successfully. He wondered if the most practical method would be to deal one-hundred hands, and simply observe the outcome (Eckhardt, 1987).

The use of random generation and repetitive calculation are the two central tenets to Monte Carlo experimentation, a method which has flourished since the first electronic computer was built in 1945, and a method which has had a profound impact on mathematics and statistics.

“At long last, mathematics achieved a certain parity - the twofold aspect of experiment and theory - that all other sciences enjoy” (Metropolis, 1987, p.130).

A Simple Monte Carlo Experiment

We have seen the derivation of some of the properties of the OLS estimator \mathbf{b} for $\boldsymbol{\beta}$, in the simple linear regression model. Namely, we have seen that OLS is unbiased and efficient. What if we could *observe* these properties? One of the uses of the Monte Carlo method is to guess at the properties of statistics; properties which may be difficult to derive theoretically.

Let's consider the *unbiasedness* property of OLS, which says that the mean of the *sampling distribution* of \mathbf{b} is $\boldsymbol{\beta}$. If we could mimic the sampling distribution, we should be able to observe this property. Recall how we interpreted the sampling distribution:

1. Repeatedly draw all possible samples of size n .
2. Calculate values of \mathbf{b} each time.
3. Construct a relative frequency distribution for \mathbf{b} .

If we replace “all possible” in Step #1 with “10,000” or “100,000”, then the sampling distribution may easily be synthesized using a computer.

Our Monte Carlo experiment will begin by pretending that the true unobservable population model is *known*. Then, when we calculate the OLS estimates, we pretend that the population model is *unknown*. This way, we can compare \mathbf{b} to $\boldsymbol{\beta}$. Let's begin with an overview of the experiment before writing computer code:

1. Specify the (unobservable) population model: $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. This involves choosing values for $\boldsymbol{\beta}$, choosing the distribution for $\boldsymbol{\varepsilon}$, and creating some arbitrary X data (which will be fixed in repeated samples, in accordance with assumption A.5).
2. “Draw” a sample of \mathbf{y} from the population model. This involves using a *random number generator* to create the $\boldsymbol{\varepsilon}$ values.
3. Calculate \mathbf{b} .
4. Repeat the above steps many (10,000) times, storing each \mathbf{b} .
5. Take the average of all 10,000 \mathbf{b} . If \mathbf{b} is unbiased, this average should be close to $\boldsymbol{\beta}$.

R Code

First we need to determine some parameters for our experiment, such as sample size and the number of Monte Carlo repetitions we would like to use.

```
n = 50
rep = 10000
```

Next, choose values for β . We'll just use an intercept and single x variable.

```
beta1 = 0.5
beta2 = 2
```

We also need to create the x variable. To do this, we will create n random numbers from the uniform (0,1) distribution.

```
x = runif(n)
```

Take a look at x :

```
> x
[1] 0.16113175 0.95362471 0.47450286 0.89152313 0.12962974 0.01736195
[7] 0.45421529 0.75819744 0.09663753 0.51222232 0.47904268 0.93851048
[13] 0.57261715 0.22855245 0.42623832 0.69128449 0.91723239 0.86308324
[19] 0.83708109 0.70848409 0.02601843 0.38442663 0.23403509 0.80584167
[25] 0.70558551 0.54727753 0.98413499 0.63819489 0.21897050 0.98055095
[31] 0.69164831 0.32517447 0.36495332 0.90024951 0.54707758 0.92455957
[37] 0.41021164 0.99205363 0.40688771 0.11455678 0.98368243 0.06997619
[43] 0.85802275 0.58978543 0.13004962 0.45697634 0.12341920 0.62945295
[49] 0.67256565 0.63599985
```

Let's create vectors of zeros, that will later be used to store the OLS estimates for the intercept and slope:


```
b1 = b2 = rep(0,n)
```

Start the Monte Carlo loop:

```
for(j in 1:rep){
```

Create the disturbances vector (ϵ) by randomly generating numbers from the $N(0,1)$ distribution.

```
eps = rnorm(n)
```

Now "draw" a random sample of y values:

```
y = beta1 + beta2*x + eps
```

Calculate and record OLS estimates, and end the Monte Carlo loop:

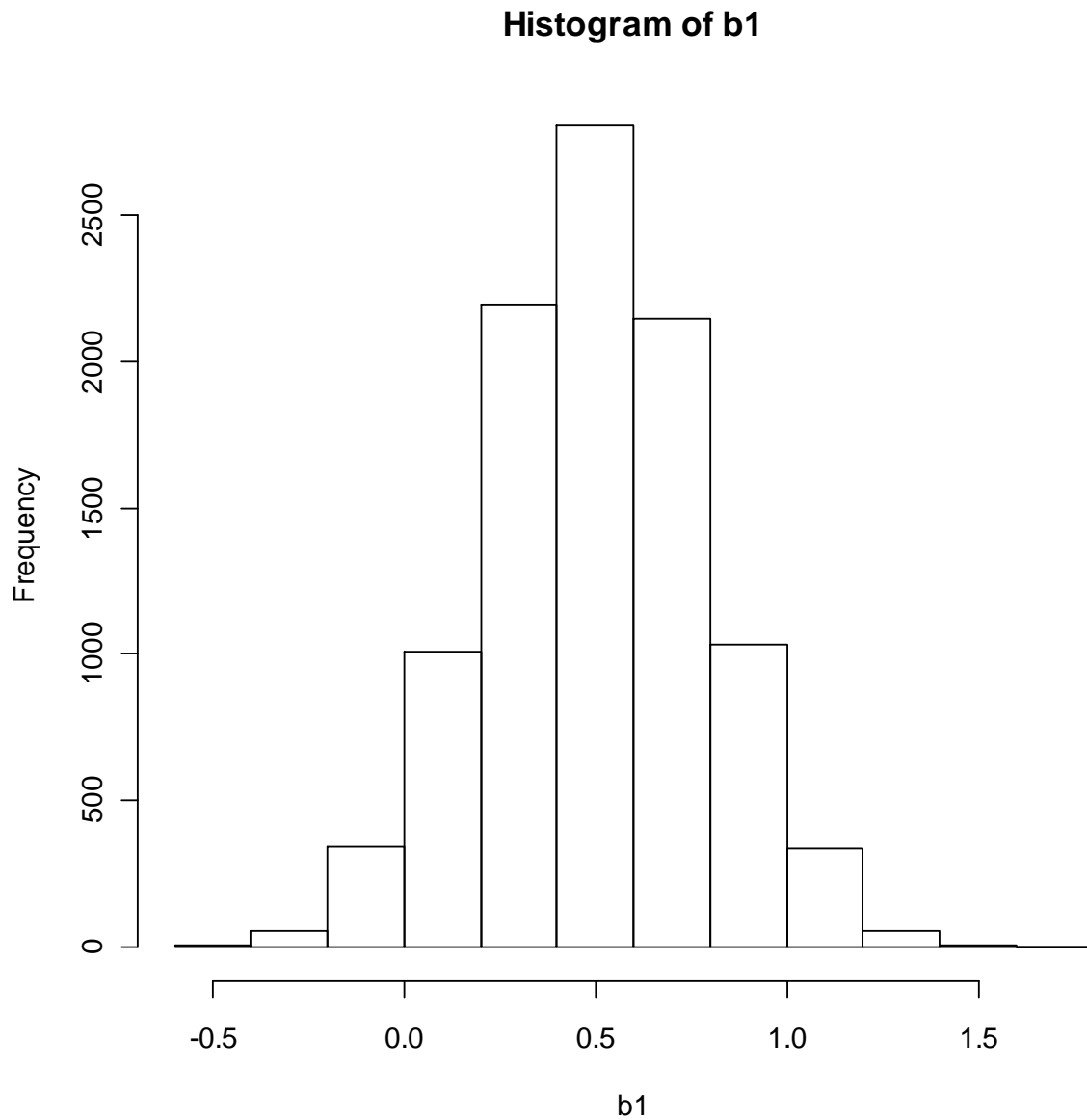
```
b1[j] = lm(y~x)$coefficient[1]  
b2[j] = lm(y~x)$coefficient[2]  
}
```

Now, we let's see if b_2 appears to be unbiased:

```
> mean(b2)  
[1] 2.003847
```

We can also take a look at the simulated sampling distribution:

```
> hist(b1)
```



References

Anderson, H.L. (1986). Metropolis, Monte Carlo, and the MANIAC. *Los Alamos Science* 14: 96-107.

Metropolis, N. (1987). The beginning of the Monte Carlo method. *Los Alamos Science* 15: 125-130.

Eckhardt, R. (1987). Stan Ulam, John von Neumann, and the Monte Carlo method. *Los Alamos Science* 15: 131-137.

Topic 2: Asymptotic Properties of Various Regression Estimators

- Our results to date apply for any *finite* sample size (n).
- In more general models we often can't obtain *exact* results for estimators' properties.
- In this case, we might consider their properties as $n \rightarrow \infty$.
- A way of “approximating” results.
- Also of interest in own right – inferential procedures should “work well” when we have lots of data
- Previous example – hypothesis tests that are “consistent”.

Definition: An estimator, $\hat{\theta}$, for θ , is said to be (weakly) *consistent* if

$$\lim_{n \rightarrow \infty} \{Pr. [|\hat{\theta}_n - \theta| < \epsilon]\} = 1.$$

Note: A *sufficient* condition for this to hold is that *both*

$$(i) \quad Bias(\hat{\theta}_n) \rightarrow \mathbf{0} \quad ; \text{ as } n \rightarrow \infty.$$

$$(ii) \quad V(\hat{\theta}_n) \rightarrow 0 \quad ; \text{ as } n \rightarrow \infty.$$

We call this “**Mean Square Consistency**”. (Often useful for checking.)

If $\hat{\theta}$ is weakly consistent for θ , we say that “the probability limit of $\hat{\theta}$ equals θ .”

We denote this by using “*plim*” operator, and we write

$$plim(\hat{\theta}_n) = \theta \quad \text{or,} \quad \hat{\theta}_n \xrightarrow{p} \theta$$

Example $x_i \sim [\mu, \sigma^2]$ (*i.i.d*)

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$E[\bar{x}] = \frac{1}{n} \sum_{i=1}^n E(x_i) = \frac{1}{n} (n\mu) = \mu \quad (\text{unbiased, for all } n)$$

$$var. [\bar{x}] = \frac{1}{n^2} var. [\sum_{i=1}^n x_i] = \frac{1}{n^2} \sum_{i=1}^n var. (x_i)$$

$$= \frac{1}{n^2} (n\sigma^2) = \sigma^2/n$$

So, \bar{x} is an unbiased estimator of μ , and $\lim_{n \rightarrow \infty} \{var. [\bar{x}]\} = 0$.

This *implies* that \bar{x} is both a **mean-square consistent**, and **weakly consistent** estimator of μ .

Note:

- If an estimator is *inconsistent*, then it is a pretty useless estimator!
- There are many situations in which our LS estimator is *inconsistent*!
- For example –

$$(i) \quad y_t = \beta_1 + \beta_2 x_t + \beta_3 y_{t-1} + \varepsilon_t$$

$$\text{and} \quad \varepsilon_t = \rho \varepsilon_{t-1} + u_t$$

$$(ii) \quad y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \varepsilon_{1t}$$

$$\text{and} \quad x_{2t} = \gamma_1 y_t + \gamma_3 x_{3t} + \gamma_4 x_{4t} + \varepsilon_{2t}$$

Slutsky's Theorem

Let $plim(\hat{\theta}_n) = \mathbf{c}$, and let $f(\cdot)$ be any *continuous* function.

Then, $plim[f(\hat{\theta}_n)] = f(\mathbf{c})$.

For example –

$$plim\left(\frac{1}{\hat{\theta}}\right) = \frac{1}{c} \quad ; \quad \text{scalars}$$

$$plim(e^{\hat{\theta}}) = e^c \quad ; \quad \text{vectors}$$

$$plim(\hat{\theta}^{-1}) = C^{-1} \quad ; \quad \text{matrices}$$

A very useful result – the “**plim**” operator can be used very flexibly.

Asymptotic Properties of LS Estimator(s)

- Consider LS estimator of β under our standard assumptions, in the “large n ” *asymptotic* case.
- Can relax some assumptions:
 - (i) Don't need Normality assumption for the error term of our model
 - (ii) Columns of X can be random – just assume that $\{x'_i, \varepsilon_i\}$ is a random and *independent* sequence; $i = 1, 2, 3, \dots$
 - (iii) Last assumption implies $plim[n^{-1}X'\varepsilon] = \mathbf{0}$. (Greene, pp. 64-65.)
- *Amend* (extend) our assumption about X having full column rank – assume instead that $plim[n^{-1}X'X] = Q$; **positive-definite & finite**
- Note that Q is $(k \times k)$, symmetric, and *unobservable*.
- What are we assuming about the elements of X , which is $(n \times k)$, as n increases without limit?

Theorem: The LS estimator of β is *weakly consistent*.

Proof:

$$\begin{aligned} \mathbf{b} &= (X'X)^{-1}X'\mathbf{y} = (X'X)^{-1}X'(X\beta + \varepsilon) \\ &= \beta + (X'X)^{-1}X'\varepsilon \\ &= \beta + \left[\frac{1}{n}(X'X)\right]^{-1} \left[\frac{1}{n}X'\varepsilon\right]. \end{aligned}$$

If we now apply Slutsky's Theorem repeatedly, we have:

$$plim(\mathbf{b}) = \beta + Q^{-1} \cdot \mathbf{0} = \beta.$$

- We can also show that s^2 is a consistent estimator for σ^2 .
- Do this in two ways (different assumptions).
- First, assume the errors are Normally distributed – get a strong result.
- Then, relax this assumption and get a weaker result.

Theorem: If the regression model errors are Normally distributed, then s^2 is a *mean-square consistent* estimator for σ^2 .

Proof:

If the errors are Normal, then we know that

$$\frac{(n-k)s^2}{\sigma^2} \sim \chi_{(n-k)}^2$$

Now, (1) $E[\chi_{(n-k)}^2] = (n - k)$

(2) $var. [\chi_{(n-k)}^2] = 2(n - k)$

So, $E(s^2) = \frac{\sigma^2 E[\chi_{(n-k)}^2]}{n-k} = \sigma^2$; *unbiased*

$$var. \left[\frac{(n-k)s^2}{\sigma^2} \right] = 2(n - k)$$

$$\Rightarrow \left[\frac{(n-k)^2}{\sigma^4} \right] var. (s^2) = 2(n - k)$$

$$\Rightarrow var. (s^2) = 2\sigma^4 / (n - k)$$

So, $var. (s^2) \rightarrow 0$, as $n \rightarrow \infty$ (and *unbiased*)

This implies that s^2 is a *mean-square consistent* estimator for σ^2 .

(Implies, in turn, that it is also a *weakly consistent estimator*.)

- With the addition of the (relatively) strong assumption of Normally distributed errors, we get the (relatively) strong result.
- Note that $\hat{\sigma}^2 = (e'e)/n$ is also a *consistent* estimator, even though it is *biased*.
- What other assumptions did we use in the above proof?
- What can we say if we relax the assumption of Normality?
- We need a preliminary result to help us.

Theorem (Khinchine ; WLLN):

Suppose that $\{x_i\}_{i=1}^n$ is a sequence of random variables that are *uncorrelated*, and all drawn from the same distribution with a *finite* mean, μ , and a *finite* variance, σ^2 .

Then, $plim(\bar{x}) = \mu$.

Theorem: In our regression model, s^2 is a *weakly consistent* estimator for σ^2 .

(Notice that this also means that $\hat{\sigma}^2$ is also a weakly consistent estimator, so start with the latter estimator.)

Proof:

$$\begin{aligned}\hat{\sigma}^2 &= \left(\frac{e'e}{n}\right) = \frac{1}{n} \sum_{i=1}^n e_i^2 \\ &= \frac{1}{n} (M\boldsymbol{\varepsilon})'(M\boldsymbol{\varepsilon}) = \frac{1}{n} \boldsymbol{\varepsilon}' M \boldsymbol{\varepsilon} \\ &= \frac{1}{n} [\boldsymbol{\varepsilon}' \boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}' X (X'X)^{-1} X' \boldsymbol{\varepsilon}] \\ &= \left[\left(\frac{1}{n} \boldsymbol{\varepsilon}' \boldsymbol{\varepsilon}\right) - \left(\frac{1}{n} \boldsymbol{\varepsilon}' X\right) \left(\frac{1}{n} X' X\right)^{-1} \left(\frac{1}{n} X' \boldsymbol{\varepsilon}\right) \right].\end{aligned}$$

So, $plim(\hat{\sigma}^2) = plim\left(\frac{1}{n} \boldsymbol{\varepsilon}' \boldsymbol{\varepsilon}\right) - \mathbf{0}' Q^{-1} \mathbf{0} = plim\left[\frac{1}{n} \sum_{i=1}^n \varepsilon_i^2\right]$.

Now, if the errors are pair-wise uncorrelated, so are their squared values.

Also, $E[\varepsilon_i^2] = var.(\varepsilon_i) = \sigma^2$.

By Khintchine's Theorem, we immediately have the result:

$$plim(\hat{\sigma}^2) = \sigma^2,$$

and so $plim(s^2) = \sigma^2$.

- Relaxing the assumption of Normally distributed errors led to a *weaker result* for the consistent estimation of the error variance.
- What other assumptions were used, and where?

An Issue

- Suppose we want to compare the (large n) asymptotic behaviour of our LS estimators with those of other potential estimators.
- These other estimators will presumably also be *consistent*.
- This means that *in each case* the sampling distributions of the estimators collapse to a “spike”, located exactly at the true parameter values.
- So, how can we compare such estimators when n is very large – aren’t they *indistinguishable*?
- If the limiting density of any consistent estimator is a degenerate “spike”, it will have zero variance, in the limit.
- Can we still compare large-sample variances of consistent estimators?
In other words, is it meaningful to think about the concept of asymptotic efficiency?

Asymptotic Efficiency

- **The key to asymptotic efficiency is to “control” for the fact that the distribution of any consistent estimator is “collapsing”, as $n \rightarrow \infty$.**
- The *rate* at which the distribution collapses is crucially important.
- This is probably best understood by considering an example.
- $\{x_i\}_{i=1}^n$; *random sampling* from $[\mu, \sigma^2]$.
- $E[\bar{x}] = \mu$; $var. [\bar{x}] = \sigma^2/n$
- Now construct: $y = \sqrt{n}(\bar{x} - \mu)$.
- Note that $E(y) = \sqrt{n}(E(\bar{x}) - \mu) = 0$.
- Also, $var. [y] = (\sqrt{n})^2 var. (\bar{x} - \mu) = n var. (\bar{x}) = \sigma^2$.
- The scaling we’ve used results in a finite, non-zero, variance.
- $E(y) = 0$, and $var. [y] = \sigma^2$; *unchanged* as $n \rightarrow \infty$.
- So, $y = \sqrt{n}(\bar{x} - \mu)$ has a well-defined “limiting” (asymptotic) distribution.
- The *asymptotic mean* of y is zero, and the *asymptotic variance* of y is σ^2 .
- Question – Why did we scale by \sqrt{n} , and not (say), by n itself ?

- In fact, because we had *independent* x_i 's (*random sampling*), we have the additional result that $y = \sqrt{n}(\bar{x} - \mu) \xrightarrow{d} N[0, \sigma^2]$, the *Lindeberg-Lévy Central Limit Theorem*.
- Now we can define “Asymptotic Efficiency” in a meaningful way.

Definition: Let $\hat{\theta}$ and $\tilde{\theta}$ be two *consistent* estimator of θ ; and suppose that

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} [0, \sigma^2] \text{ , and } \sqrt{n}(\tilde{\theta} - \theta) \xrightarrow{d} [0, \varphi^2] \text{ .}$$

Then $\hat{\theta}$ is “*asymptotically efficient*” relative to $\tilde{\theta}$ if $\sigma^2 < \varphi^2$.

In the case where θ is a *vector*, $\hat{\theta}$ is “*asymptotically efficient*” relative to $\tilde{\theta}$ if

$\Delta = \text{asy. } V(\tilde{\theta}) - \text{asy. } V(\hat{\theta})$ is positive definite.

Asymptotic Distribution of the LS Estimator:

Let's consider the full asymptotic distribution of the LS estimator, \mathbf{b} , for $\boldsymbol{\beta}$ in our linear regression model.

We'll actually have to consider the behaviour of $\sqrt{n}(\mathbf{b} - \boldsymbol{\beta})$:

$$\begin{aligned} \sqrt{n}(\mathbf{b} - \boldsymbol{\beta}) &= \sqrt{n}[(X'X)^{-1}X'\boldsymbol{\varepsilon}] \\ &= \left[\frac{1}{n}(X'X) \right]^{-1} \left(\frac{1}{\sqrt{n}}X'\boldsymbol{\varepsilon} \right). \end{aligned}$$

It can be shown, by the Lindeberg-Feller Central Limit Theorem, that

$$\left(\frac{1}{\sqrt{n}}X'\boldsymbol{\varepsilon} \right) \xrightarrow{d} N[0, \sigma^2 Q],$$

where $Q = \text{plim} \left[\frac{1}{n}(X'X) \right]$.

So, the asymptotic covariance matrix of $\sqrt{n}(\mathbf{b} - \boldsymbol{\beta})$ is

$$plim \left[\frac{1}{n} (X'X) \right]^{-1} (\sigma^2 Q) plim \left[\frac{1}{n} (X'X) \right]^{-1} = \sigma^2 Q^{-1}.$$

In full, the asymptotic distribution of \mathbf{b} is correctly stated by saying that:

$$\sqrt{n}(\mathbf{b} - \boldsymbol{\beta}) \xrightarrow{d} N[\mathbf{0}, \sigma^2 Q^{-1}]$$

The asymptotic covariance matrix is *unobservable*, for **two reasons**:

1. σ^2 is typically *unknown*.
2. Q is *unobservable*.
 - We can estimate σ^2 *consistently*, using s^2 .
 - To estimate $\sigma^2 Q^{-1}$ consistently, we can use $ns^2(X'X)^{-1}$:

$$plim[ns^2(X'X)^{-1}] = plim(s^2)plim \left[\frac{1}{n} (X'X) \right]^{-1} = \sigma^2 Q^{-1}.$$

The square roots of the diagonal elements of $ns^2(X'X)^{-1}$ are the *asymptotic std. errors* for the elements of $\sqrt{n}(\mathbf{b} - \boldsymbol{\beta})$.

Loosely speaking, the asymptotic covariance matrix for \mathbf{b} itself is $s^2(X'X)^{-1}$; and the square roots of the diagonal elements of this matrix are the *asymptotic std. errors* for the b_i 's themselves.

Instrumental Variables

- We have been assuming *either* that the columns of X are *non-random*; or that the sequence $\{\mathbf{x}'_i, \varepsilon_i\}$ is *independent*. Often, neither of these assumptions is tenable.
- This implies that $plim \left(\frac{1}{n} X' \boldsymbol{\varepsilon} \right) \neq \mathbf{0}$, and then the LS estimator is *inconsistent* (prove this).
- In order to motivate a situation where $\{\mathbf{x}'_i, \varepsilon_i\}$ are *dependent*, consider an omitted, or unobservable variable.

We will consider a situation where the unobservable variable is correlated with one of the regressors, and correlated with the dependent variable.

Consider the population model:

$$\mathbf{y} = X_1\beta_1 + X_2\beta_2 + \boldsymbol{\varepsilon}_1. \quad [1]$$

Consider that $cov(X_1, X_2) \neq 0$. For example, X_2 causes X_1 :

$$X_1 = X_2\gamma + \boldsymbol{\varepsilon}_2. \quad [2]$$

Now consider that X_2 is unobservable, so that the observable model is:

$$\mathbf{y} = X_1\beta_1 + \boldsymbol{\varepsilon}_3. \quad [3]$$

- Notice that in [3], $\boldsymbol{\varepsilon}_3$ contains $\beta_2 X_2$, so that X_1 and $\boldsymbol{\varepsilon}_3$ are not independent (X_1 is *endogenous*)
- OLS will be biased, since $E[\boldsymbol{\varepsilon}_3|X_1] \neq \mathbf{0}$
- Note that when estimating from [3], $E[b_1] = \beta_1 + \gamma^{-1}\beta_2$
- OLS will be inconsistent, since $plim\left(\frac{1}{n}X_1'\boldsymbol{\varepsilon}_3\right) \neq \mathbf{0}$
- In such cases we want a safe way of estimating β_1 .
- We just want to ensure that we have an estimator that is (at least) *consistent*.
- One general family of such estimators is the family of **Instrumental Variables (I.V.) Estimators**.

An instrumental variable, Z , must be:

1. Correlated with the endogenous variable(s) X_1
 - Sometimes called the “relevance” of an I.V.
 - This condition can be tested
2. Uncorrelated with the error term, or equivalently, uncorrelated with the dependent variable other than through its correlation with X_1
 - Sometimes called the “exclusion” restriction
 - This restriction cannot be tested directly

Suppose now that we have a variable Z which is

- Relevant: $cov(Z, X_1) \neq 0$
- Satisfies exclusion restriction: $cov(Z, \varepsilon) = 0$. In the above D.G.P.s ([1]- [3]), it is sufficient for the instrument to be uncorrelated with the unobservable variable: $cov(Z, X_2) = 0$.

Validity means that [2] becomes:

$$X_1 = Z\delta + X_2\gamma + \varepsilon_4 \quad [4]$$

Substituting [4] into [1]:

$$\mathbf{y} = X_2\gamma\beta_1 + Z\delta\beta_1 + X_2\beta_2 + \varepsilon_5. \quad [5]$$

X_2 is still unobservable, but is uncorrelated with Z ! The observable population model is now:

$$\mathbf{y} = Z\delta\beta_1 + \varepsilon_6. \quad [6]$$

Now, we have a population model involving β_1 , and where $cov(Z, \varepsilon_6) = 0$. So, $(\delta\beta_1)$ can be estimated by OLS. But we need β_1 !

By Slutsky's Theorem, if $plim(\widehat{\delta\beta_1}) = \delta\beta_1$, and if $plim(\widehat{\delta}) = \delta$, then $plim(\widehat{\delta}^{-1}\widehat{\delta\beta_1}) = \beta_1$. So if we can find a consistent estimator for δ , we can find one for β_1 . How to estimate δ ?

Recall [4]. Since X_2 and Z are uncorrelated, we can estimate δ by an OLS regression of X_1 on Z :

$$\widehat{\delta} = (Z'Z)^{-1}Z'X_1$$

Now solve for $\widehat{\beta_1}$:

$$\widehat{\beta_1} = \widehat{\delta}^{-1}\widehat{\delta\beta_1} = [(Z'Z)^{-1}Z'X_1]^{-1}(Z'Z)^{-1}Z'\mathbf{y}$$

If Z and X_1 have the same number of columns, then:

$$\widehat{\beta_1} = (Z'X_1)^{-1}Z'Z(Z'Z)^{-1}Z'\mathbf{y} = (Z'X_1)^{-1}Z'\mathbf{y}$$

In this example we had one endogenous variable (X_1) and one instrument (Z). In this case, the I.V. estimate may be found by the OLS estimate from a regression of \mathbf{y} on Z by the OLS estimates of a regression of X_1 on Z .

In more general models, we will have more explanatory variables. As long as there is one instrument per endogenous variable, I.V. is possible and the simple I.V. estimator is:

$$b_{IV} = (Z'X)^{-1}Z'\mathbf{y}$$

In general, this estimator is biased. We can show it's consistent, however:

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

$$plim\left(\frac{1}{n}X'X\right) = Q \quad ; \quad \text{p.d. and finite}$$

$$plim\left(\frac{1}{n}X'\boldsymbol{\varepsilon}\right) = \boldsymbol{\gamma} \neq \mathbf{0}$$

Find a (random) $(n \times k)$ matrix, Z , such that:

1. $plim\left(\frac{1}{n}Z'Z\right) = Q_{ZZ} \quad ; \quad \text{p.d. and finite.}$
2. $plim\left(\frac{1}{n}Z'X\right) = Q_{ZX} \quad ; \quad \text{p.d. and finite.}$
3. $plim\left(\frac{1}{n}Z'\boldsymbol{\varepsilon}\right) = \mathbf{0} \quad .$

Then, consider the estimator: $\mathbf{b}_{IV} = (Z'X)^{-1}Z'\mathbf{y}$. This is a *consistent* estimator of $\boldsymbol{\beta}$.

$$\begin{aligned} \mathbf{b}_{IV} &= (Z'X)^{-1}Z'\mathbf{y} = (Z'X)^{-1}Z'(X\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\ &= (Z'X)^{-1}Z'X\boldsymbol{\beta} + (Z'X)^{-1}Z'\boldsymbol{\varepsilon} \\ &= \boldsymbol{\beta} + (Z'X)^{-1}Z'\boldsymbol{\varepsilon} \\ &= \boldsymbol{\beta} + \left(\frac{1}{n}Z'X\right)^{-1} \left(\frac{1}{n}Z'\boldsymbol{\varepsilon}\right) . \end{aligned}$$

$$\text{So, } plim(\mathbf{b}_{IV}) = \boldsymbol{\beta} + [plim\left(\frac{1}{n}Z'X\right)]^{-1}plim\left(\frac{1}{n}Z'\boldsymbol{\varepsilon}\right)$$

$$= \boldsymbol{\beta} + Q_{ZX}^{-1}\mathbf{0} = \boldsymbol{\beta} \quad (\text{consistent})$$

Choosing different Z matrices generates different members of I.V. family.

Although we won't *derive* the full asymptotic distribution of the I.V. estimator, note that it can be expressed as:

$$\sqrt{n}(\mathbf{b}_{IV} - \boldsymbol{\beta}) \xrightarrow{d} N[\mathbf{0}, \sigma^2 Q_{ZX}^{-1} Q_{ZZ} Q_{XZ}^{-1}]$$

where $Q_{XZ} = Q_{ZX}'$. [How would you estimate Asy. Covar. Matrix?]

Interpreting I.V. as two-stage least squares (2SLS)

1st stage: Regress X on Z , get \hat{X} .

- \hat{X} contains the variation in X due to Z *only*
- \hat{X} is not correlated with $\boldsymbol{\varepsilon}$

2nd stage: Estimate the model $\mathbf{y} = \hat{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

From 1st stage: $\hat{X} = Z(Z'Z)^{-1}Z'X$

From 2nd stage: $\mathbf{b}_{IV} = [X'Z(Z'Z)^{-1}Z'X]^{-1}X'Z(Z'Z)^{-1}Z'\mathbf{y}$

In fact, this is the *Generalized* I.V. estimator of $\boldsymbol{\beta}$. We can actually use more instruments than regressors (the "Over-Identified" case).

Note that if X and Z have the same dimensions, then the generalized estimator collapses to the simple one.

Let's check the consistency of the I.V. estimator. Let $M_Z = Z(Z'Z)^{-1}Z'$. Then the generalized I.V. estimator is:

$$\mathbf{b}_{IV} = [X'M_Z X]^{-1} X'M_Z \mathbf{y}$$

$$\begin{aligned} \mathbf{b}_{IV} &= [X'M_Z X]^{-1} X'M_Z \mathbf{y} = [X'M_Z X]^{-1} X'M_Z (X\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\ &= [X'M_Z X]^{-1} X'M_Z X\boldsymbol{\beta} + [X'M_Z X]^{-1} X'M_Z \boldsymbol{\varepsilon} \\ &= \boldsymbol{\beta} + [X'Z(Z'Z)^{-1}Z'X]^{-1} X'Z(Z'Z)^{-1}Z'\boldsymbol{\varepsilon} \end{aligned}$$

So,

$$\mathbf{b}_{IV} = \boldsymbol{\beta} + \left[\left(\frac{1}{n} X'Z \right) \left(\frac{1}{n} Z'Z \right)^{-1} \left(\frac{1}{n} Z'X \right) \right]^{-1} \left(\frac{1}{n} X'Z \right) \left(\frac{1}{n} Z'Z \right)^{-1} \left(\frac{1}{n} Z'\boldsymbol{\varepsilon} \right).$$

Modify our assumptions:

We have a (random) $(n \times L)$ matrix, Z , such that:

1. $plim\left(\frac{1}{n}Z'Z\right) = Q_{ZZ}$; $(L \times L)$, p.d.s. and finite.
2. $plim\left(\frac{1}{n}Z'X\right) = Q_{ZX}$; $(L \times k)$, rank = k , and finite.
3. $plim\left(\frac{1}{n}Z'\boldsymbol{\varepsilon}\right) = \mathbf{0}$; $(L \times 1)$

So,

$$plim(\mathbf{b}_{IV}) = \boldsymbol{\beta} + [Q_{XZ}Q_{ZZ}^{-1}Q_{ZX}]^{-1}Q_{XZ}Q_{ZZ}^{-1}\mathbf{0} = \boldsymbol{\beta} ; \text{ consistent}$$

Similarly, a *consistent estimator* of σ^2 is

$$s_{IV}^2 = (\mathbf{y} - X\mathbf{b}_{IV})'(\mathbf{y} - X\mathbf{b}_{IV})/n$$

residual vector 

- Recall that each choice of Z leads to a *different* I.V. estimator.
- Z must be chosen in way that ensures consistency of the I.V. estimator.
- How might we choose a suitable set of instruments, *in practice*?
- If we have several “valid” sets of instruments, how might we choose between them?

For the “simple” regression model, recall that:

$$\sqrt{n}(\mathbf{b}_{IV} - \boldsymbol{\beta}) \xrightarrow{d} N[\mathbf{0}, \sigma^2 Q_{ZX}^{-1} Q_{ZZ} Q_{XZ}^{-1}]$$

so if $k = 1$,

$$Q_{ZZ} = plim\left(n^{-1} \sum_{i=1}^n z_i^2\right)$$

$$Q_{ZX} = \text{plim} \left(n^{-1} \sum_{i=1}^n z_i x_i \right) = Q_{XZ}$$

The *asymptotic efficiency* of \mathbf{b}_{IV} will be higher, the more highly correlated are Z and X , *asymptotically*.

We need to find instruments that are uncorrelated with the errors, but highly correlated with the regressors – *asymptotically*.

This is not easy to do!

- **Time –series data** -
 1. Often, we can use lagged values of the regressors as suitable instruments.
 2. This will be fine as long as the errors are serially uncorrelated.
- **Cross-section data** –
 1. Geography, weather, biology.
 2. Various “old” tricks – *e.g.*, using “ranks” of the data as instruments.

Testing if I.V. estimation is needed

- Why does LS fail, and when do we need I.V.?
- If $\text{plim} \left(\frac{1}{n} X' \boldsymbol{\varepsilon} \right) \neq \mathbf{0}$.
- We can *test* to see if this is a problem, & decide if we should use LS or I.V.

The Hausman Test

We want to test $H_0 : \text{plim} \left(\frac{1}{n} X' \boldsymbol{\varepsilon} \right) = \mathbf{0}$ vs. $H_A : \text{plim} \left(\frac{1}{n} X' \boldsymbol{\varepsilon} \right) \neq \mathbf{0}$

- If we reject H_0 , we will use I.V. estimation.
- If we cannot reject H_0 , we'll use LS estimation.
- Hausman test is a general “testing strategy” that can be applied in many situations – not just for this particular situation!
- Basic idea – construct 2 estimators of $\boldsymbol{\beta}$:

1. \mathbf{b}_E : estimator is both *consistent and asymptotically efficient* if H_0 true.
 2. \mathbf{b}_I : estimator is at least *consistent*, even if H_0 false.
- In our case here, \mathbf{b}_E is the LS estimator; and \mathbf{b}_I is the I.V. estimator.
 - If H_0 is true, we'd expect $(\mathbf{b}_I - \mathbf{b}_E)$ to be “small”, at least for large n , as both estimators are consistent in that case.
 - Hausman shows that $\hat{V}(\mathbf{b}_I - \mathbf{b}_E) = \hat{V}(\mathbf{b}_I) - \hat{V}(\mathbf{b}_E)$, if H_0 is true.
 - So, the test statistic is, $H = (\mathbf{b}_I - \mathbf{b}_E)' [\hat{V}(\mathbf{b}_I) - \hat{V}(\mathbf{b}_E)]^{-1} (\mathbf{b}_I - \mathbf{b}_E)$.
 - $H \xrightarrow{d} \chi_J^2$, if H_0 is true.
 - Here, J is the number of columns in X which *may* be correlated with the errors, & for which we need instruments.
 - Problem – often, $[\hat{V}(\mathbf{b}_I) - \hat{V}(\mathbf{b}_E)]$ is *singular*, so H is *not defined*.
 - One option is to replace the “regular inverse” with a “generalized inverse”.
 - Another option is to modify H so that it becomes:

$$H^* = (\mathbf{b}_I^* - \mathbf{b}_E^*)' [\hat{V}(\mathbf{b}_I^*) - \hat{V}(\mathbf{b}_E^*)]^{-1} (\mathbf{b}_I^* - \mathbf{b}_E^*) \xrightarrow{d} \chi_J^2 ; \text{ if } H_0 \text{ true.}$$
 - Here, \mathbf{b}_I^* and \mathbf{b}_E^* are the $(J \times 1)$ vectors formed by using only the elements of \mathbf{b}_I and \mathbf{b}_E that correspond to the “problematic” regressors.
 - Constructing H^* is not very convenient unless $J = 1$.

The Durbin-Wu Test

This test is *specific* to testing

$$H_0 : \text{plim} \left(\frac{1}{n} X' \boldsymbol{\varepsilon} \right) = \mathbf{0} \quad \text{vs.} \quad H_A : \text{plim} \left(\frac{1}{n} X' \boldsymbol{\varepsilon} \right) \neq \mathbf{0}$$

Again, an asymptotic test.

Testing the exogeneity of Instruments

The key assumption that ensures the consistency of I.V. estimators is that

$$plim \left(\frac{1}{n} Z' \boldsymbol{\varepsilon} \right) = \mathbf{0} .$$

This condition involves the *unobservable* $\boldsymbol{\varepsilon}$. In general, it cannot be tested.

“Weak Instruments” – Problems arise if the instruments are *not* well correlated with the regressors (not relevant).

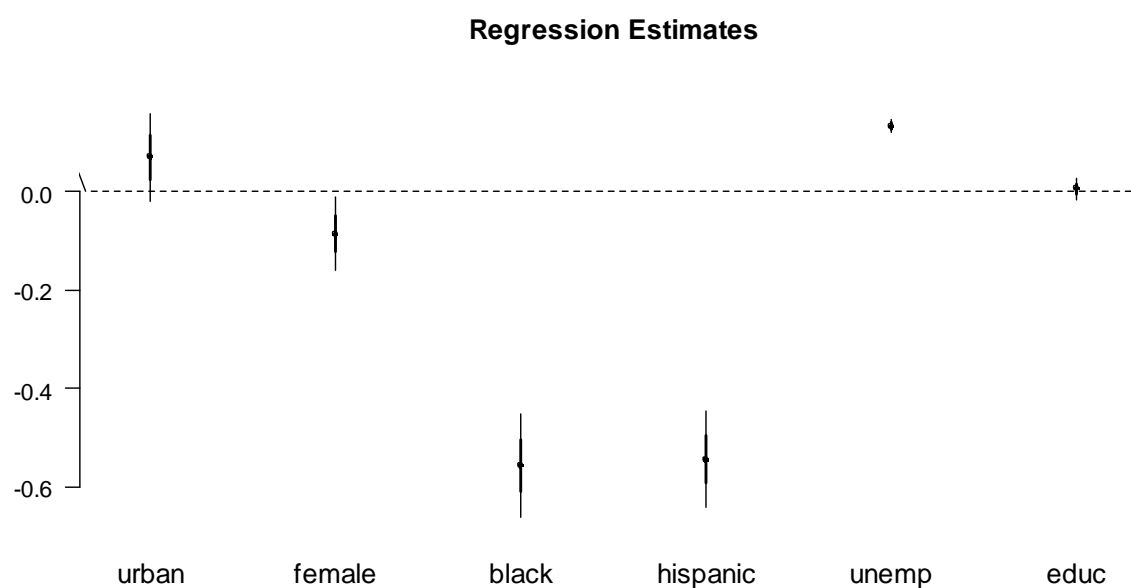
- These problems go beyond loss of asymptotic efficiency.
- Small-sample bias of I.V. estimator can be greater than that of LS!
- Sampling distribution of I.V. estimator can be bi-modal!
- Fortunately, we can again *test* to see if we have these problems.

Empirical Example: Using geographic variation in college proximity to estimate the return to schooling¹

- Have data on **wage**, **years of education**, and demographic variables
- Want to estimate the return to education
- Problem: **ability** (intelligence) may be correlated with (cause) both **wage** and **education**
- Since **ability** is unobservable, it is contained in the error term
- The **education** variable is then correlated with the error term (endogenous)
- OLS estimation of the returns to **education** may be inconsistent

First, let's try OLS.

```
library(AER)
attach(CollegeDistance)
lm(wage ~ urban + gender + ethnicity + unemp + education)
```



Note that the returns to education are not statistically significant.

Now let's try using **distance from college** (while attending high school) as an instrument for **education**. For the instrument to be valid, we require that **distance** and **education** be correlated:

```
summary(lm(education ~ distance))
```

¹ Card, David. *Using geographic variation in college proximity to estimate the return to schooling*. No. w4483. National Bureau of Economic Research, 1993.

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|----------|------------|---------|-------------|
| (Intercept) | 13.93861 | 0.03290 | 423.683 | < 2e-16 *** |
| distance | -0.07258 | 0.01127 | -6.441 | 1.3e-10 *** |

While `distance` appears to be statistically significant, this isn't quite enough to test for validity (a testing problem we won't address here).

From the 2SLS interpretation, we know that we can get the IV estimator by:

1.) getting the predicted values from a regression of `education` on `distance`

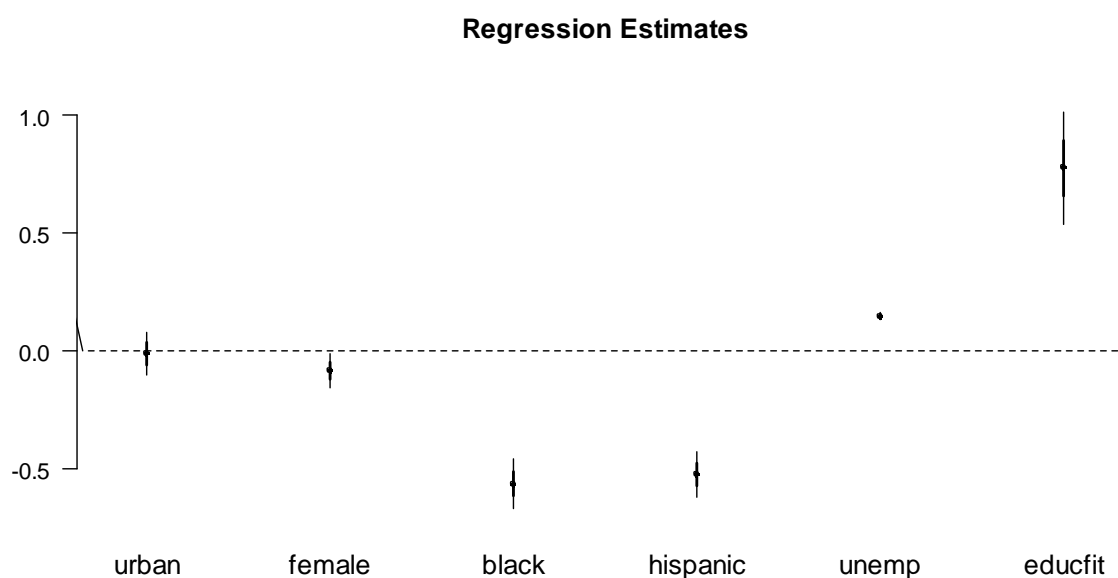
```
educfit = predict(lm(education ~ distance))
```

2.) regressing `wage` on the same variables, but using `educfit` instead of `education`

```
lm(wage ~ urban + gender + ethnicity + unemp + educfit)
```

Note that `educfit` is the variation in `education` as it can be explained by `distance`. These fitted values are uncorrelated with `ability`, since `distance` is uncorrelated with `ability` (by assumption).

Results of IV estimation:



The estimate for the return to education is now positive, and significant.

Topic 3: Inference and Prediction

We'll be concerned here with testing more general hypotheses than those seen to date. Also concerned with constructing interval predictions from our regression model.

Examples

- $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$; $H_0: \boldsymbol{\beta} = \mathbf{0}$ vs. $H_A: \boldsymbol{\beta} \neq \mathbf{0}$
- $\log(Q) = \beta_1 + \beta_2 \log(K) + \beta_3 \log(L) + \varepsilon$
 $H_0: \beta_2 + \beta_3 = 1$ vs. $H_A: \beta_2 + \beta_3 \neq 1$
- $\log(q) = \beta_1 + \beta_2 \log(p) + \beta_3 \log(y) + \varepsilon$
 $H_0: \beta_2 + \beta_3 = 0$ vs. $H_A: \beta_2 + \beta_3 \neq 0$

If we can obtain one model from another by imposing restrictions on the parameters of the first model, we say that the 2 models are “*Nested*”.

We'll be concerned with (several) possible restrictions on $\boldsymbol{\beta}$, in the usual model:

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad ; \quad \boldsymbol{\varepsilon} \sim N[0, \sigma^2 I_n]$$

(X is non-random ; $\text{rank}(X) = k$)

Let's focus on *linear restrictions*:

$$r_{11}\beta_1 + r_{12}\beta_2 + \cdots + r_{1k}\beta_k = q_1$$

$$r_{21}\beta_1 + r_{22}\beta_2 + \cdots + r_{2k}\beta_k = q_2$$

•

(J restrictions)

•

$$r_{J1}\beta_1 + r_{J2}\beta_2 + \cdots + r_{Jk}\beta_k = q_J$$

Some (many?) of the r_{ij} 's may be zero.

- Combine these J restrictions:

$$R\boldsymbol{\beta} = \mathbf{q} \quad ; \quad R \text{ and } \mathbf{q} \text{ are } \textit{known, \& non-random}$$

$$(J \times k)(k \times 1) \quad (J \times 1)$$

- We'll assume that $\text{rank}(R) = J (< k)$.

- No conflicting or redundant restrictions.
- What if $J = k$?

Examples

1. $\beta_2 = \beta_3 = \dots = \beta_k = 0$

$$R = \begin{bmatrix} 0 & 1 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad ; \quad \mathbf{q} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$$

2. $\beta_2 + \beta_3 = 1$

$$R = [0 \quad 1 \quad 1 \quad 0 \quad \dots \quad 0] \quad ; \quad q = 1$$

3. $\beta_3 = \beta_4$; and $\beta_1 = 2\beta_2$

$$R = \begin{bmatrix} 0 & 0 & 1 & -1 & 0 & \dots & 0 \\ 1 & -2 & 0 & 0 & 0 & \dots & 0 \end{bmatrix} \quad ; \quad \mathbf{q} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$$

- Suppose that we just estimate the model by LS, and get $\mathbf{b} = (X'X)^{-1}X'\mathbf{y}$.
- It is very unlikely that $R\mathbf{b} = \mathbf{q}$!
- Denote $\mathbf{m} = R\mathbf{b} - \mathbf{q}$.
- Clearly, \mathbf{m} is a $(J \times 1)$ *random vector*.
- Let's consider the sampling distribution of \mathbf{m} :

$$\mathbf{m} = R\mathbf{b} - \mathbf{q} \quad ; \quad \text{it is a *linear* function of } \mathbf{b}.$$

If the errors in the model are Normal, then \mathbf{b} is Normally distributed, & hence \mathbf{m} is Normally distributed.

$$E[\mathbf{m}] = RE[\mathbf{b}] - \mathbf{q} = R\boldsymbol{\beta} - \mathbf{q} \quad (\text{What assumptions used?})$$

$$\text{So, } E[\mathbf{m}] = \mathbf{0} \quad ; \quad \text{iff } R\boldsymbol{\beta} = \mathbf{q}$$

$$\text{Also, } V[\mathbf{m}] = V[R\mathbf{b} - \mathbf{q}] = V[R\mathbf{b}] = RV[\mathbf{b}]R'$$

$$= R\sigma^2(X'X)^{-1}R' = \sigma^2R(X'X)^{-1}R'$$

(What assumptions used?)

So, $\mathbf{m} \sim N[\mathbf{0}, \sigma^2 R(X'X)^{-1}R']$.

Let's see how we can use this information to *test* if $R\boldsymbol{\beta} = \mathbf{q}$. (Intuition?)

Definition: The *Wald Test Statistic* for testing $H_0: R\boldsymbol{\beta} = \mathbf{q}$ vs. $H_A: R\boldsymbol{\beta} \neq \mathbf{q}$ is:
 $W = \mathbf{m}'[V(\mathbf{m})]^{-1}\mathbf{m}$.

So, *if H_0 is true:*

$$\begin{aligned} W &= (R\mathbf{b} - \mathbf{q})'[\sigma^2 R(X'X)^{-1}R']^{-1}(R\mathbf{b} - \mathbf{q}) \\ &= (R\mathbf{b} - \mathbf{q})'[R(X'X)^{-1}R']^{-1}(R\mathbf{b} - \mathbf{q})/\sigma^2 . \end{aligned}$$

Because $\mathbf{m} \sim N[\mathbf{0}, \sigma^2 R(X'X)^{-1}R']$, then *if H_0 is true:*

$$W \sim \chi_{(J)}^2 \quad ; \quad \text{provided that } \sigma^2 \text{ is known.}$$

Notice that:

- This result is valid only *asymptotically* if σ^2 is unobservable, and we replace it with *any consistent estimator*.
- We would reject H_0 if $W > \text{critical value}$. (i.e., when $\mathbf{m} = R\mathbf{b} - \mathbf{q}$ is sufficiently “large”.)
- The Wald test is a *very general* testing procedure – other testing problems.
- Wald test statistic always constructed using an estimator that *ignores* the restrictions being tested.
- As we'll see, *for this particular* testing problem, we can modify the Wald test slightly, and obtain a test that is **exact in finite samples**, and has excellent power properties.

What is the F-statistic?

To derive this test statistic, we need a *preliminary result*.

Definition:

Let $x_1 \sim \chi^2_{(v_1)}$ and $x_2 \sim \chi^2_{(v_2)}$ *and independent*

Then

$$F = \frac{\frac{x_1}{v_1}}{\frac{x_2}{v_2}} \sim F_{(v_1, v_2)} \quad ; \quad \text{Snedecor's F-Distribution}$$

Note:

- $(t_{(v)})^2 = F_{(1, v)}$; *Why does this make sense?*
- $v_1 F_{(v_1, v_2)} \xrightarrow{d} \chi^2_{(v_1)}$; *Explanation?*

Let's proceed to our main result, which involves the statistic, $F = \left(\frac{W}{J}\right) \left(\frac{\sigma^2}{s^2}\right)$.

Theorem:

$F = \left(\frac{W}{J}\right) \left(\frac{\sigma^2}{s^2}\right) \sim F_{(J, (n-k))}$, if the Null Hypothesis $H_0: R\boldsymbol{\beta} = \mathbf{q}$ is true.

Proof:

$$\begin{aligned} F &= \frac{(\mathbf{Rb} - \mathbf{q})' [R(X'X)^{-1}R']^{-1} (\mathbf{Rb} - \mathbf{q})}{\sigma^2} \left(\frac{1}{J}\right) \left(\frac{\sigma^2}{s^2}\right) \\ &= \frac{(\mathbf{Rb} - \mathbf{q})' [\sigma^2 R(X'X)^{-1}R']^{-1} (\mathbf{Rb} - \mathbf{q}) / J}{\left[\frac{(n-k)s^2}{\sigma^2}\right] / (n-k)} = \left(\frac{N}{D}\right) \end{aligned}$$

where $D = \left[\frac{(n-k)s^2}{\sigma^2}\right] / (n-k) = \chi^2_{(n-k)} / (n-k)$.

Consider the numerator:

$$N = (\mathbf{Rb} - \mathbf{q})' [\sigma^2 R(X'X)^{-1}R']^{-1} (\mathbf{Rb} - \mathbf{q}) / J.$$

Suppose that H_0 is TRUE, so that $\mathbf{Rb} = \mathbf{q}$, and then

$$(R\mathbf{b} - \mathbf{q}) = (R\mathbf{b} - R\boldsymbol{\beta}) = R(\mathbf{b} - \boldsymbol{\beta}) .$$

Now, recall that

$$\mathbf{b} = (X'X)^{-1}X'\mathbf{y} = (X'X)^{-1}X'(X\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{\beta} + (X'X)^{-1}X'\boldsymbol{\varepsilon} .$$

So,
$$R(\mathbf{b} - \boldsymbol{\beta}) = R(X'X)^{-1}X'\boldsymbol{\varepsilon} ,$$

and
$$N = [R(X'X)^{-1}X'\boldsymbol{\varepsilon}]'[\sigma^2 R(X'X)^{-1}R']^{-1}[R(X'X)^{-1}X'\boldsymbol{\varepsilon}]/J$$

$$= \left(\frac{1}{J}\right) (\boldsymbol{\varepsilon}/\sigma)'[Q](\boldsymbol{\varepsilon}/\sigma) ,$$

where
$$Q = X(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}X' ,$$

and
$$(\boldsymbol{\varepsilon}/\sigma) \sim N[\mathbf{0}, I_n] .$$

Now, $(\boldsymbol{\varepsilon}/\sigma)'[Q](\boldsymbol{\varepsilon}/\sigma) \sim \chi_{(r)}^2$ if and only if Q is *idempotent*, where

$$r = \text{rank}(Q) .$$

Easy to check that Q is idempotent.

So,
$$\text{rank}(Q) = \text{tr.}(Q)$$

$$= \text{tr.}\{X(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}X'\}$$

$$= \text{tr.}\{(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}X'\}$$

$$= \text{tr.}\{R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}\}$$

$$= \{[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}R'\}$$

$$= \text{tr.}(I_J) = J .$$

So,
$$N = \left(\frac{1}{J}\right) (\boldsymbol{\varepsilon}/\sigma)'[Q](\boldsymbol{\varepsilon}/\sigma) = \chi_{(J)}^2/J .$$

- In the construction of F we have a ratio of 2 Chi-Square statistics, each divided by their degrees of freedom.
- Are N and D *independent*?

- The Chi-Square statistic in N is: $(\boldsymbol{\varepsilon}/\sigma)'[Q](\boldsymbol{\varepsilon}/\sigma)$.
- The Chi-Square statistic in D is: $\left[\frac{(n-k)s^2}{\sigma^2}\right]$ (see bottom of slide 13)

Re-write this:

$$\begin{aligned} \left[\frac{(n-k)s^2}{\sigma^2}\right] &= \frac{(n-k)}{\sigma^2} (\mathbf{e}'\mathbf{e}/(n-k)) = (\mathbf{e}'\mathbf{e}/\sigma^2) \\ &= (M\boldsymbol{\varepsilon}/\sigma)'(M\boldsymbol{\varepsilon}/\sigma) = (\boldsymbol{\varepsilon}/\sigma)'M(\boldsymbol{\varepsilon}/\sigma). \end{aligned}$$

So, we have

$$(\boldsymbol{\varepsilon}/\sigma)'[Q](\boldsymbol{\varepsilon}/\sigma) \quad \text{and} \quad (\boldsymbol{\varepsilon}/\sigma)'M(\boldsymbol{\varepsilon}/\sigma).$$

These two statistics are *independent* if and only if $MQ = 0$.

$$\begin{aligned} MQ &= [I - X(X'X)^{-1}X'] X(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}X' \\ &= Q - X(X'X)^{-1}X'X(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}X' \\ &= Q - X(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}X' \\ &= Q - Q = 0. \end{aligned}$$

So, if H_0 is **TRUE**, our statistic, F is the ratio of 2 *independent* Chi-Square variates, each divided by their degrees of freedom.

This implies that, if H_0 is **TRUE**,

$$F = \frac{(\mathbf{Rb}-\mathbf{q})'[R(X'X)^{-1}R']^{-1}(\mathbf{Rb}-\mathbf{q})/J}{s^2} \sim F_{(J,(n-k))}$$

What assumptions have been used ? *What if H_0 is FALSE ?*

Implementing the test –

- Calculate F .
- Reject $H_0: R\boldsymbol{\beta} = \mathbf{q}$ in favour of $H_A: R\boldsymbol{\beta} \neq \mathbf{q}$ if $> c_\alpha$.

Why do we use *this particular test* for linear restrictions?

This F -test is **Uniformly Most Powerful**.

Another point to note –

$$(t_{(v)})^2 = F_{(1,v)}$$

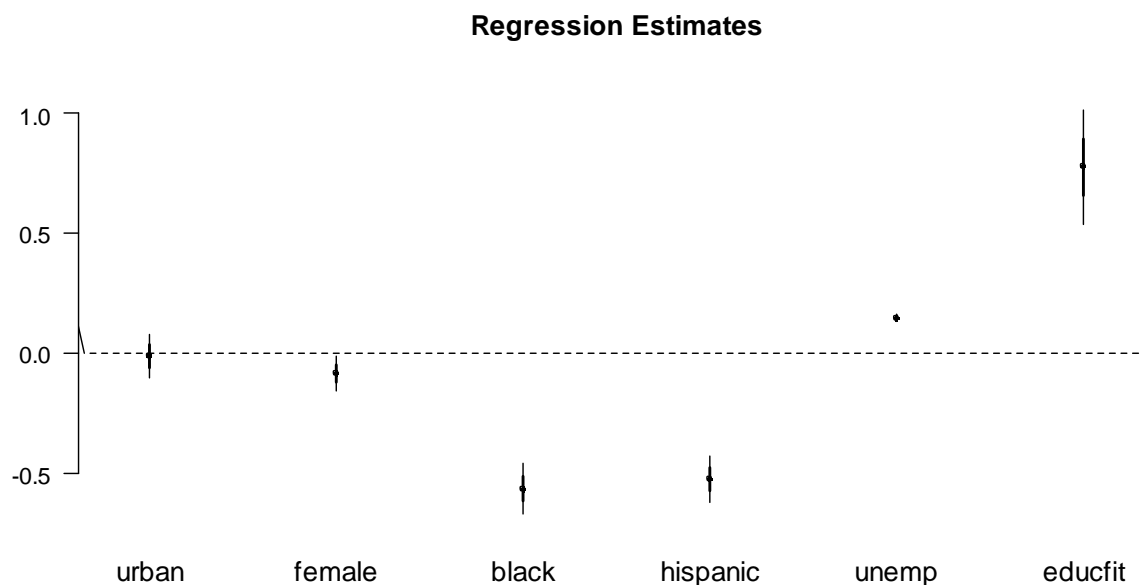
Consider $t_{(n-k)} = (b_i - \beta_i) / (s.e. (b_i))$

Then, $(t_{(n-k)})^2 \sim F_{(1,(n-k))}$; t-test is **UMP** against **1-sided alternatives**

Example

Let's return to the Card (1993) data, used as an example of I.V.

Recall the results of the IV estimation:



```
resiv = lm(wage ~ urban + gender + ethnicity + unemp + educfit)
summary(resiv)
```

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------------|-----------|------------|---------|--------------|
| (Intercept) | -2.053604 | 1.675314 | -1.226 | 0.2203 |
| urbanyes | -0.013588 | 0.046403 | -0.293 | 0.7697 |
| genderfemale | -0.086700 | 0.036909 | -2.349 | 0.0189 * |
| ethnicityafam | -0.566524 | 0.051686 | -10.961 | < 2e-16 *** |
| ethnicityhispanic | -0.529088 | 0.048429 | -10.925 | < 2e-16 *** |
| unemp | 0.145806 | 0.006969 | 20.922 | < 2e-16 *** |
| educfit | 0.774340 | 0.120372 | 6.433 | 1.38e-10 *** |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

$$(n - k) = (4739 - 7) = 4732$$

Residual standard error: 1.263 on 4732 degrees of freedom

Multiple R-squared: 0.1175, Adjusted R-squared: 0.1163

F-statistic: 105 on 6 and 4732 DF, p-value: < 2.2e-16

Let's test the hypothesis that *urban* and *gender* are jointly insignificant.

$H_0: \beta_2 = \beta_3 = 0$ vs. $H_A: \text{At least one of these coeffs.} \neq 0. (J=2)$

Let's see R-code for calculating the F-stat from the formula:

$$F = \frac{(Rb - q)'[R(X'X)^{-1}R']^{-1}(Rb - q)/J}{s^2} = (Rb - q)'[Rs^2(X'X)^{-1}R']^{-1}(Rb - q)/J$$

```
R = matrix(c(0,0,1,0,0,1,0,0,0,0,0,0,0,0),2,7)
```

```
> R
```

```
      [,1] [,2] [,3] [,4] [,5] [,6] [,7]
[1,]    0    1    0    0    0    0    0
[2,]    0    0    1    0    0    0    0
```

```
b = matrix(resiv$coef, 7, 1)
```

```
> b
```

```
      [,1]
[1,] -2.05360353
[2,] -0.01358775
[3,] -0.08670020
[4,] -0.56652448
[5,] -0.52908814
[6,]  0.14580613
[7,]  0.77433967
```

```
q = matrix(c(0,0), 2, 1)
```

```
> q
```

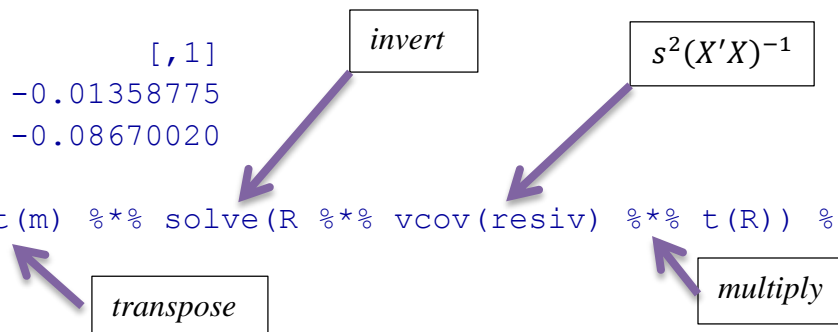
```
      [,1]
[1,]    0
[2,]    0
```

```
m = R%%b - q
```

```
> m
```

```
      [,1]
[1,] -0.01358775
[2,] -0.08670020
```

```
F = t(m) %% solve(R %% vcov(resiv) %% t(R)) %% m
```



```
> F
```

```
      [,1]
[1,] 5.583774
```

Is this F-stat “large”?

```
> 1 - pf(F, 2, 4732)
```

```
      [,1]
[1,] 0.003783159
```

Should we be using the F-test?

```
Wald = 2*F
```

```
> 1 - pchisq(Wald, 2)
```

```
      [,1]
[1,] 0.003758353
```

Why are the p-values from the Wald and F-test so similar?

Restricted Least Squares Estimation:

If we test the validity of certain linear restrictions on the elements of β , and we can't reject them, how might we incorporate the restrictions (*information*) into the estimator?

Definition: The “Restricted Least Squares” (RLS) estimator of β , in the model, $\mathbf{y} = X\beta + \varepsilon$, is the vector, \mathbf{b}_* , which minimizes the sum of the squared residuals, subject to the constraint(s) $R\mathbf{b}_* = \mathbf{q}$.

- Let's obtain the expression for this new estimator, and derive its sampling distribution.
- Set up the Lagrangian: $\mathcal{L} = (\mathbf{y} - X\mathbf{b}_*)'(\mathbf{y} - X\mathbf{b}_*) + 2\lambda'(R\mathbf{b}_* - \mathbf{q})$
- Set $(\partial\mathcal{L}/\partial\mathbf{b}_*) = \mathbf{0}$; $(\partial\mathcal{L}/\partial\lambda) = \mathbf{0}$, and solve

$$\mathcal{L} = \mathbf{y}'\mathbf{y} + \mathbf{b}_*'X'X\mathbf{b}_* - 2\mathbf{y}'X\mathbf{b}_* + 2\lambda'(R\mathbf{b}_* - \mathbf{q})$$

$$(\partial\mathcal{L}/\partial\mathbf{b}_*) = 2X'X\mathbf{b}_* - 2X'\mathbf{y} + 2R'\lambda = \mathbf{0} \quad [1]$$

$$(\partial\mathcal{L}/\partial\lambda) = 2(R\mathbf{b}_* - \mathbf{q}) = \mathbf{0} \quad [2]$$

From [1]:

$$R'\lambda = X'(\mathbf{y} - X\mathbf{b}_*)$$

So, $R(X'X)^{-1}R'\lambda = R(X'X)^{-1}X'(\mathbf{y} - X\mathbf{b}_*)$

or, $\lambda = [R(X'X)^{-1}R']^{-1}R(X'X)^{-1}X'(\mathbf{y} - X\mathbf{b}_*) \quad [3]$

Inserting [3] into [1], and dividing by “2”:

$$(X'X)\mathbf{b}_* = X'\mathbf{y} - R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}X'(\mathbf{y} - X\mathbf{b}_*)$$

So, $(X'X)\mathbf{b}_* = X'\mathbf{y} - R'[R(X'X)^{-1}R']^{-1}R(\mathbf{b} - \mathbf{b}_*)$

or,

$$\mathbf{b}_* = (X'X)^{-1}X'\mathbf{y} - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(R\mathbf{b} - R\mathbf{b}_*)$$

or, using [2]:

$$\mathbf{b}_* = \mathbf{b} - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(R\mathbf{b} - \mathbf{q})$$

- RLS = LS + “Adjustment Factor”.
- What if $R\mathbf{b} = \mathbf{q}$?
- Interpretation of this?
- What are the properties of this RLS estimator of $\boldsymbol{\beta}$?

Theorem: The RLS estimator of $\boldsymbol{\beta}$ is *Unbiased* if $R\boldsymbol{\beta} = \mathbf{q}$ is TRUE.

Otherwise, the RLS estimator is *Biased*.

Proof:

$$\begin{aligned} E(\mathbf{b}_*) &= E(\mathbf{b}) - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(RE(\mathbf{b}) - \mathbf{q}) \\ &= \boldsymbol{\beta} - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(R\boldsymbol{\beta} - \mathbf{q}) . \end{aligned}$$

So, if $R\boldsymbol{\beta} = \mathbf{q}$, then $E(\mathbf{b}_*) = \boldsymbol{\beta}$.

Theorem: The covariance matrix of the RLS estimator of $\boldsymbol{\beta}$ is

$$V(\mathbf{b}_*) = \sigma^2(X'X)^{-1}\{I - R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}\}$$

Proof:

$$\begin{aligned} \mathbf{b}_* &= \mathbf{b} - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(R\mathbf{b} - \mathbf{q}) \\ &= \{I - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R\}\mathbf{b} + \boldsymbol{\alpha} \end{aligned}$$

where

$$\boldsymbol{\alpha} = (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}\mathbf{q} \quad (\text{non-random})$$

So, $V(\mathbf{b}_*) = AV(\mathbf{b})A'$,

where $A = \{I - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R\}$.

That is, $V(\mathbf{b}_*) = AV(\mathbf{b})A' = \sigma^2 A(X'X)^{-1}A'$ (assumptions?)

Now let's look at the matrix, $A(X'X)^{-1}A'$.

$$\begin{aligned} A(X'X)A' &= \{I - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R\} (X'X)^{-1} \\ &\quad \times \{I - R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}\} \\ &= (X'X)^{-1} + (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1} \\ &\quad - 2(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1} \\ &= (X'X)^{-1}\{I - R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}\}. \end{aligned}$$

So,

$$V(\mathbf{b}_*) = \sigma^2 (X'X)^{-1}\{I - R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}\}.$$

(What assumptions have we used to get this result?)

We can use this result immediately to establish the following.....

Theorem: The matrix, $V(\mathbf{b}) - V(\mathbf{b}_*)$, is at least positive semi-definite.

Proof:

$$\begin{aligned} V(\mathbf{b}_*) &= \sigma^2 (X'X)^{-1}\{I - R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}\} \\ &= \sigma^2 (X'X)^{-1} - \sigma^2 (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1} \\ &= V(\mathbf{b}) - \sigma^2 (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1} \end{aligned}$$

So, $V(\mathbf{b}) - V(\mathbf{b}_*) = \sigma^2 \Delta$, say

where $\Delta = (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}$.

This matrix is **square**, **symmetric**, and of **full rank**. So, Δ is at least **p.s.d.**.

- This tells us that the variability of the RLS estimator is no more than that of the LS estimator, *whether or not the restrictions are true*.
- Generally, the RLS estimator will be “more precise” than the LS estimator.
- **When will the RLS and LS estimators have the same variability?**
- In addition, we know that the RLS estimator is unbiased *if the restrictions are true*.
- So, *if the restrictions are true*, the RLS estimator, \mathbf{b}_* , is more efficient than the LS estimator, \mathbf{b} , of the coefficient vector, $\boldsymbol{\beta}$.

Also note the following:

- *If the restrictions are false*, and we consider $\text{MSE}(\mathbf{b})$ and $\text{MSE}(\mathbf{b}_*)$, then the relative efficiency can go either way.
- *If the restrictions are false*, not only is \mathbf{b}_* biased, it's also **inconsistent**.

So, it's a good thing that that we know how to construct the UMP test for the validity of the restrictions on the elements of $\boldsymbol{\beta}$!

In practice:

- Estimate the unrestricted model, using LS.
- Test $H_0: R\boldsymbol{\beta} = \mathbf{q}$ vs. $H_A: R\boldsymbol{\beta} \neq \mathbf{q}$.
- If the null hypothesis can't be rejected, re-estimate the model with RLS.
- Otherwise, retain the LS estimates.

Example: Cobb-Douglas Production Function²

```
>
cobbdata=read.csv("http://home.cc.umanitoba.ca/~godwinrt/7010/cobbd.csv")
> attach(cobbdata)
> res = lm(log(y) ~ log(k) + log(l))
> summary(res)
```

Call:

```
lm(formula = log(y) ~ log(k) + log(l))
```

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) | |
|-------------|----------|------------|---------|----------|-----|
| (Intercept) | 1.8444 | 0.2336 | 7.896 | 7.33e-08 | *** |
| log(k) | 0.2454 | 0.1069 | 2.297 | 0.0315 | * |
| log(l) | 0.8052 | 0.1263 | 6.373 | 2.06e-06 | *** |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2357 on 22 degrees of freedom

Multiple R-squared: 0.9731, Adjusted R-squared: 0.9706

F-statistic: 397.5 on 2 and 22 DF, p-value: < 2.2e-16

What's this?

Let's get the SSE from this regression, for later use:

```
> sum(res$residuals^2)
[1] 1.22226
```

SSE = 1.22226

Test the hypothesis of constant returns to scale:

$$H_0: \beta_2 + \beta_3 = 1 \quad \text{vs.} \quad H_A: \beta_2 + \beta_3 \neq 1$$

```
> R = matrix(c(0,1,1),1,3)
```

```
> R
```

```
      [,1] [,2] [,3]
[1,]    0    1    1
```

² The data are from table F7.2, Greene, 2012

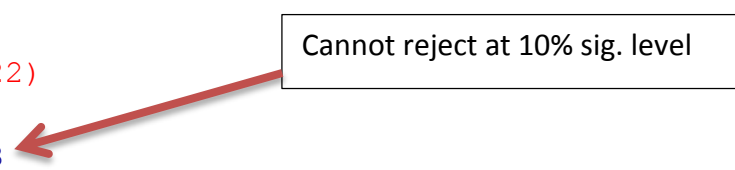
```
> b = matrix(res$coef, 3, 1)
> b
      [,1]
[1,] 1.8444157
[2,] 0.2454281
[3,] 0.8051830

> q = 1

> m = R%%b - q
> m
      [,1]
[1,] 0.05061103

> F = t(m) %% solve(R %% vcov(res) %% t(R)) %% m
> F
      [,1]
[1,] 1.540692

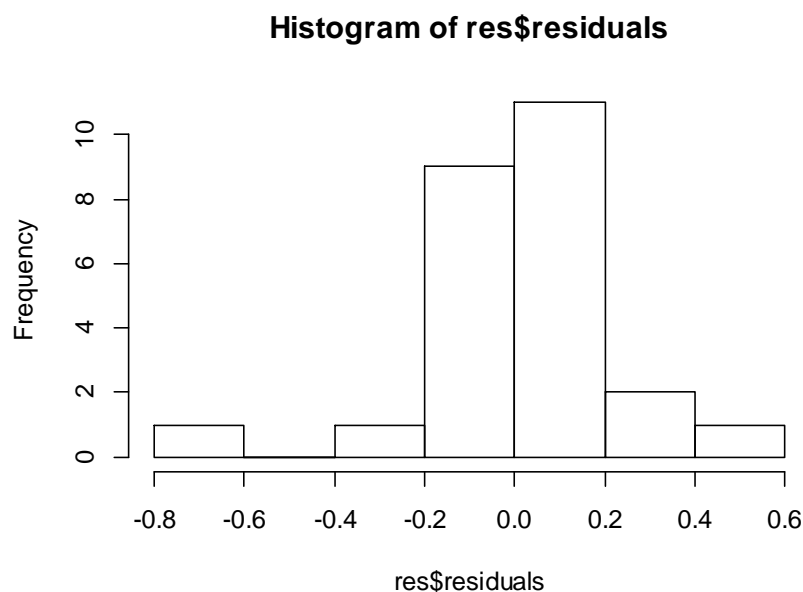
> 1 - pf(F, 1, 22)
      [,1]
[1,] 0.2275873
```



Cannot reject at 10% sig. level

Are the residuals normally distributed?

```
> hist(res$residuals)
```



```
> library(tseries)
> jarque.bera.test(res$residuals)
Jarque Bera Test
```

Might want to use Wald test instead!

```
data: res$residuals
X-squared = 5.5339, df = 2, p-value = 0.06285
F-test "supported" the validity of the restriction on the coefficients, so now impose this restriction of CRTS. Use RLS:
```

$$\log(Q/L) = \beta_1 + \beta_2 \log(K/L) + \varepsilon$$

```
> rlsres = lm(log(y/l) ~ log(k/l))
> summary(rlsres)
```

Call:

```
lm(formula = log(y/l) ~ log(k/l))
```

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) | |
|-------------|----------|------------|---------|----------|-----|
| (Intercept) | 2.0950 | 0.1189 | 17.615 | 7.55e-15 | *** |
| log(k/l) | 0.2893 | 0.1020 | 2.835 | 0.00939 | ** |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2385 on 23 degrees of freedom
 Multiple R-squared: 0.2589, Adjusted R-squared: 0.2267
 F-statistic: 8.036 on 1 and 23 DF, p-value: 0.009387

```
> sum(rlsres$residuals^2)
```

```
[1] 1.307857
```

SSE = 1.307857

From the LS and RLS results for this particular application, note that

$$\mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) = 1.22226$$

$$\mathbf{e}_*'\mathbf{e}_* = (\mathbf{y} - \mathbf{X}\mathbf{b}_*)'(\mathbf{y} - \mathbf{X}\mathbf{b}_*) = 1.307857$$

So, $\mathbf{e}_*'\mathbf{e}_* > \mathbf{e}'\mathbf{e}$.

- In fact this inequality will *always hold*.
- What's the intuition behind this?

Note that:

$$\begin{aligned}\mathbf{e}_* &= (\mathbf{y} - \mathbf{X}\mathbf{b}_*) = \mathbf{y} - \mathbf{X}\mathbf{b} + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{q}) \\ &= \mathbf{e} + \mathbf{X}(\mathbf{X}'\mathbf{X})\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{q})\end{aligned}$$

Now, recall that $\mathbf{X}'\mathbf{e} = \mathbf{0}$.

So,

$$\mathbf{e}_*'\mathbf{e}_* = \mathbf{e}'\mathbf{e} + (\mathbf{R}\mathbf{b} - \mathbf{q})'\mathbf{A}(\mathbf{R}\mathbf{b} - \mathbf{q}),$$

where:

$$\begin{aligned}\mathbf{A} &= [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1} \\ &= [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1} \quad ; \quad \text{this matrix has full rank, and is p.d.s.}\end{aligned}$$

So, $\mathbf{e}_*'\mathbf{e}_* > \mathbf{e}'\mathbf{e}$, because $(\mathbf{R}\mathbf{b} - \mathbf{q})'\mathbf{A}(\mathbf{R}\mathbf{b} - \mathbf{q}) > 0$.

This last result also gives us an alternative (convenient) way of writing the formula for the F-statistic:

$$\begin{aligned}(\mathbf{e}_*'\mathbf{e}_* - \mathbf{e}'\mathbf{e}) &= (\mathbf{Rb} - \mathbf{q})'A(\mathbf{Rb} - \mathbf{q}) \\ &= (\mathbf{Rb} - \mathbf{q})'[R(X'X)^{-1}R']^{-1}(\mathbf{Rb} - \mathbf{q}).\end{aligned}$$

Recall that:

$$F = \frac{(\mathbf{Rb} - \mathbf{q})'[R(X'X)^{-1}R']^{-1}(\mathbf{Rb} - \mathbf{q})/J}{s^2}$$

So, clearly,

$$F = \frac{(\mathbf{e}_*'\mathbf{e}_* - \mathbf{e}'\mathbf{e})/J}{s^2} = \frac{(\mathbf{e}_*'\mathbf{e}_* - \mathbf{e}'\mathbf{e})/J}{\mathbf{e}'\mathbf{e}/(n - k)}$$

For the last example:

$$J = 1 ; (n - k) = (25 - 3) = 22$$

$$(\mathbf{e}_*'\mathbf{e}_*) = 1.307857 \quad ; \quad (\mathbf{e}'\mathbf{e}) = 1.22226$$

$$\text{So, } F = \frac{(1.307857 - 1.22226)/1}{1.22226/22} = 1.54070 \quad \checkmark$$

In Retrospect

- Now we can see why $R^2 \uparrow$ when we add any regressor to our model (and $R^2 \downarrow$ when we delete any regressor).
- Deleting a regressor is equivalent to imposing a zero restriction on one of the coefficients.
- The residual sum of squares \uparrow and so $R^2 \downarrow$.

Exercise: use the R^2 from the unrestricted and restricted model to calculate F .

Estimating the Error Variance

We have considered the RLS estimator of $\boldsymbol{\beta}$. What about the corresponding estimator of the variance of the error term, σ^2 ?

Theorem:

Let \mathbf{b}^* be the RLS estimator of $\boldsymbol{\beta}$ in the model,

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad ; \quad \boldsymbol{\varepsilon} \sim [0, \sigma^2 I_n]$$

and let the corresponding residual vector be $\mathbf{e}_* = (\mathbf{y} - X\mathbf{b}^*)$. Then the following estimator of σ^2 is *unbiased*, if the restrictions, $R\boldsymbol{\beta} = \mathbf{q}$, are satisfied: $s_*^2 = (\mathbf{e}_*' \mathbf{e}_*) / (n - k + J)$.

See if you can prove this result!

Topic 4: Model Stability & Specification Analysis

- Our results to date presume that our regression model holds for the full sample that we are working with.
- Our results also presume that the model is correctly specified, in the following sense:
 - The functional form is correct.
 - All of the relevant regressors have been included.
 - No redundant regressors have been included.
 - The only “unexplained” variation in the dependent variable is purely random “noise”, as represented by a “well-behaved” error term.
- In this section we’ll re-consider item 1, above, and items 2 (b) & (c).
- The other items will be considered later.

Specification Analysis

(Henri Theil, 1957)

We’ll consider various issues associated with the choice of regressors in our linear regression model.

Omission of Relevant Regressors

D.G.P.: $y = X_1\beta_1 + X_2\beta_2 + \varepsilon$; $E[\varepsilon] = \mathbf{0}$

F.M.: $y = X_1\beta_1 + u$

So, $b_1 = (X_1'X_1)^{-1}X_1'y$

$$= (X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2 + \varepsilon)$$

$$= \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 + (X_1'X_1)^{-1}X_1'\varepsilon$$

Let’s consider the bias of this estimator –

$$E[b_1] = \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2$$

$$\neq \beta_1 \quad ; \quad \text{unless } X_1'X_2 = \mathbf{0} \quad ; \quad \text{or } X_2\beta_2 = \mathbf{0}$$

- So, in general, this estimator will be **Biased**.
- This is just an example of imposing false restrictions on some elements of the $\boldsymbol{\beta}$ vector.
- The estimator, b_1 , will also be **inconsistent**.
- However, there will be a reduction in the variance of the estimator, through the imposition of the restrictions, even though they are *false*.

Now consider the converse situation –

Inclusion of Irrelevant Regressors

$$\text{D.G.P.:} \quad \mathbf{y} = X_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon} \quad ; \quad E[\boldsymbol{\varepsilon}] = \mathbf{0}$$

$$\text{F.M.:} \quad \mathbf{y} = X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + \mathbf{u} = X\boldsymbol{\beta} + \mathbf{u}$$

where,

$$X = [X_1, X_2] \quad ; \quad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}$$

$$\begin{aligned} \text{So,} \quad \mathbf{b} &= \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{pmatrix} = (X'X)^{-1}X'\mathbf{y} \\ &= (X'X)^{-1}X'(X_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}) . \end{aligned}$$

Now, we can write: $X_1 = (X_1, X_2) \begin{pmatrix} I \\ 0 \end{pmatrix} = XS$, say.

$$\begin{aligned} \text{So,} \quad \mathbf{b} &= \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{pmatrix} = (X'X)^{-1}X'S\boldsymbol{\beta}_1 + (X'X)^{-1}X'\boldsymbol{\varepsilon} \\ &= S\boldsymbol{\beta}_1 + (X'X)^{-1}X'\boldsymbol{\varepsilon} . \end{aligned}$$

$$\text{Then,} \quad E[\mathbf{b}] = E \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{pmatrix} = S\boldsymbol{\beta}_1 = \begin{pmatrix} I \\ 0 \end{pmatrix} \boldsymbol{\beta}_1 = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \mathbf{0} \end{pmatrix} .$$

That is,

$$E[\mathbf{b}_1] = \boldsymbol{\beta}_1 \quad ; \quad \text{and} \quad E[\mathbf{b}_2] = \mathbf{0} (= \boldsymbol{\beta}_2) .$$

So, in this case the LS estimator is **Unbiased** (and also **Consistent**).

- In the case where we include irrelevant regressors, we are effectively *ignoring some valid restrictions* on β .
- Although the LS estimator is Unbiased, it is also **Inefficient**.
- The “costs” of wrongly omitting regressors usually exceed those of wrongly including extraneous ones.
- This suggests that a “**General-to-Specific**” model building strategy may be preferable to a “Specific-to-General” one. *(David Hendry)*
- Over-fit the model, then simplify it on the basis of significance and specification testing.
- Generally this involves a *sequence* of “nested” hypotheses – increasingly restrictive. Stop when restrictions are rejected.
- **Issues:** (a) Degrees of freedom; (b) Loss of precision; (c) Dependence of test statistics, and distortions due to “pre-testing”.

Testing for Structural Change

- Suppose that a “shift” in the model occurs due to some exogenous “shock”.
- Define a **Dummy Variable**:

$$D_t = 0 \quad ; \quad \text{before the shock}$$

$$D_t = 1 \quad ; \quad \text{after the shock}$$
- Need not involve “time”. Could be 2 regions, for example.
- Could be more than one “shift”.
- Do the values of the Dummy variable have to be 0 and 1?
- Then, consider a model of the form:

$$y_t = \beta_1 + \beta_2 x_{2t} + \cdots + \beta_k D_t + \varepsilon_t$$

- We could then think of testing

$$H_0: \beta_k = 0 \quad \text{vs.} \quad H_A: \beta_k \neq 0$$
- Rejection of H_0 implies there is a particular type of **structural change** in the model. (A *shift in the level*.)
- Or, more generally, consider a model of the form:

$$y_t = \beta_1 + \beta_2 x_{2t} + \cdots + \beta_{k-1} (D_t \times x_{2t}) + \beta_k D_t + \varepsilon_t$$
- We could then think of testing

$$H_0: \beta_{k-1} = \beta_k = 0 \quad \text{vs.} \quad H_A: \text{Not } H_0$$

- Rejection of H_0 implies there is a different type of **structural change** in the model. (A shift in the level and one of the marginal effects.)
- Using the dummy variable *fully*, in this way (with intercept and *all* slope coefficients) turns out to be equivalent to the following –

The Chow Test

(Gregory Chow, 1960)

- Suppose there is a natural break-point in the sample after n_1 observations, and we have:

$$\mathbf{y}_1 = X_1 \boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}_1 \quad ; \quad \boldsymbol{\varepsilon}_1 \sim N[0, \sigma^2 I_{n_1}] \quad (n_1)$$

$$\mathbf{y}_2 = X_2 \boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}_2 \quad ; \quad \boldsymbol{\varepsilon}_2 \sim N[0, \sigma^2 I_{n_2}] \quad (n_2)$$

- X_1 and X_2 relate to the same regressors, but different sub-samples. Similarly for \mathbf{y}_1 and \mathbf{y}_2 .
Let $n = (n_1 + n_2)$.
- We can write the full model as:

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{pmatrix}$$

$$(n \times 1) \quad (n \times 2k) \quad (2k \times 1) \quad (n \times 1)$$

or,

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad ; \quad \boldsymbol{\varepsilon} \sim N[0, \sigma^2 I_n]$$

- If we estimate each part of the model separately, using LS, we get:

$$\mathbf{b}_1 = (X_1' X_1)^{-1} X_1' \mathbf{y}_1 \quad ; \quad \mathbf{e}_1 = \mathbf{y}_1 - X_1 \mathbf{b}_1$$

$$\mathbf{b}_2 = (X_2' X_2)^{-1} X_2' \mathbf{y}_2 \quad ; \quad \mathbf{e}_2 = \mathbf{y}_2 - X_2 \mathbf{b}_2$$

- The total sum of squared residuals for all $n = (n_1 + n_2)$ observations is then:

$$\mathbf{e}'\mathbf{e} = \mathbf{e}_1' \mathbf{e}_1 + \mathbf{e}_2' \mathbf{e}_2$$

Suppose that we want to test $H_0: \beta_1 = \beta_2$ vs. $H_A: \beta_1 \neq \beta_2$

- That is we want to test the null hypothesis “*There is no structural break*”.
- One way to interpret this problem is as follows:

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

$$H_0: R\beta = q \quad \text{vs.} \quad H_A: R\beta \neq q$$

where: $R = [I_k \quad -I_k]$; $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$; $q = \mathbf{0}$.

If there are k regressors, then q is $(k \times 1)$, and $J = k$.

- Then, we can apply the usual F -test for exact linear restrictions:

$$F = (Rb - q)'[R(X'X)^{-1}R']^{-1}(Rb - q)/(ks^2)$$

$$F \sim F_{k, n-2k} \quad \text{if } H_0 \text{ is True}$$

- Alternatively, recall that we can write the test statistic as:

$$F = \frac{[(e_*'e_*) - (e'e)]/k}{(e'e)/(n_1 + n_2 - 2k)}$$

- Here, e_* is the residual vector associated with the RLS estimator, b_* , of β .
- An easy way to obtain b_* , and hence e_* , is to write:

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}$$

$(n \times 1) \quad (n \times 2k) \quad (2k \times 1) \quad (n \times 1)$

and then restrict $\beta_1 = \beta_2 = \bar{\beta}$ (say), yielding the model:

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \bar{\beta} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}$$

- That is, we just “stack up the y and X data for both sub-samples – that is, estimate the one model for the full sample.
- This will yield b_* , and hence e_* .
- Notice that we assumed that $\sigma_1^2 = \sigma_2^2$.
- Major complications without this restriction: “**Behrens-Fisher Problem**”.

- If we have random regressors, we can still use the **Wald Test**.
- $kF \xrightarrow{d} \chi^2_{(k)}$; if H_0 is True.

Example

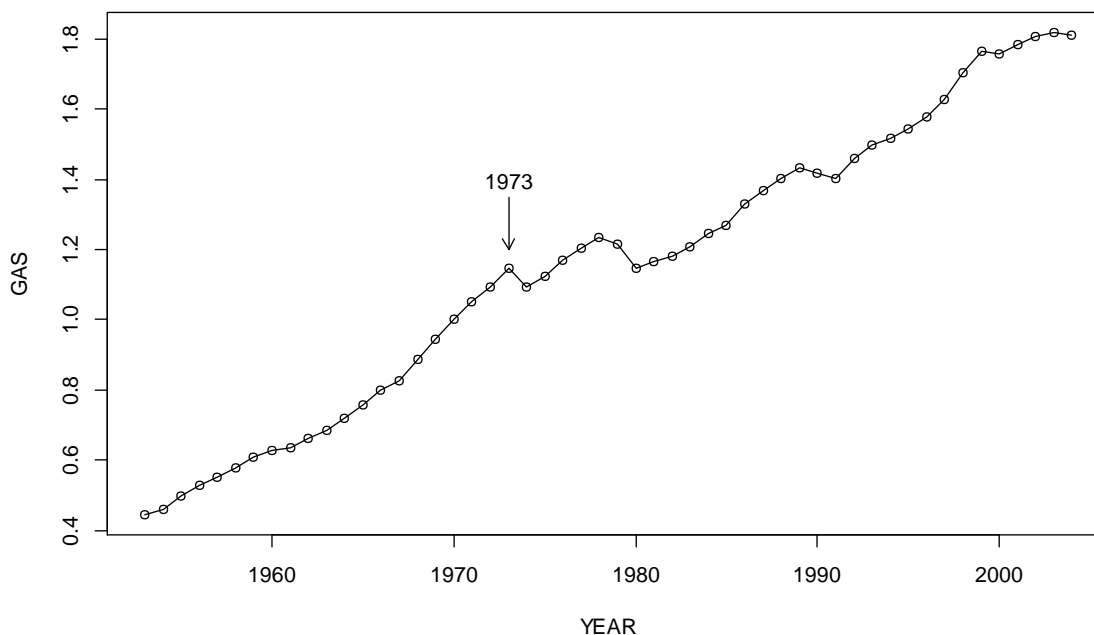
Let's see this illustrated. We'll see *two* equivalent ways of testing for this type of structural change.

Consider the following model for per-capita gasoline consumption³:

$$\ln GAS = \beta_1 + \beta_2 YEAR + \beta_3 \ln Income/Pop + \beta_4 \ln GASP + \beta_5 \ln PNC + \beta_6 \ln PUC + \varepsilon$$

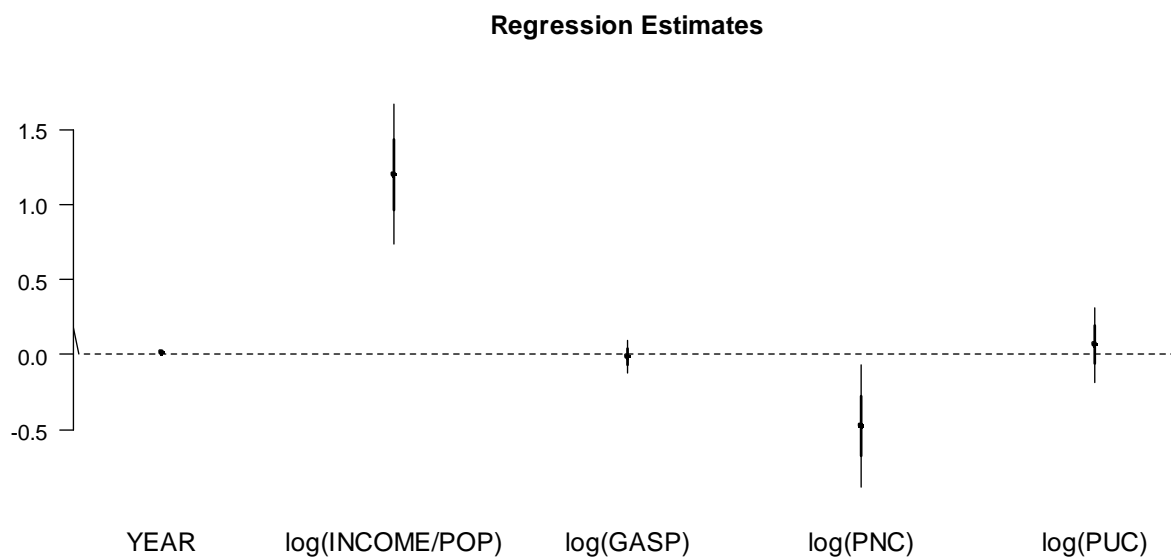
Where *GASP* is the price of gasoline, *PNC* is the price of new cars, and *PUC* the price of used cars. We will consider an exogenous shock for the year 1973.

Per Capita Gasoline Consumption (U.S.A.)



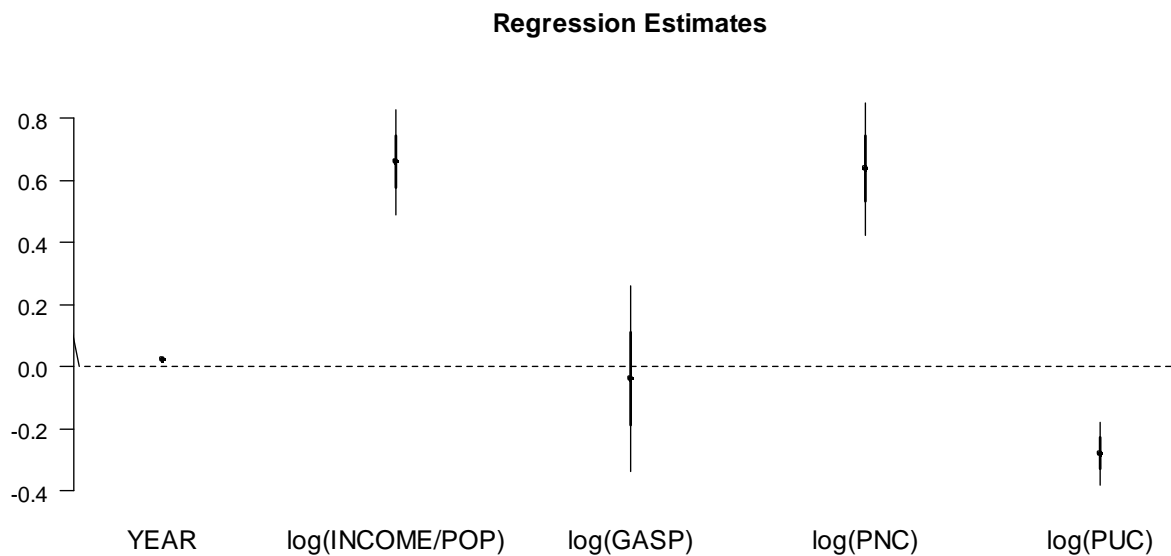
³ Data from Greene (2012), Table F2.2

Estimate the **pooled** model (using **all** observations):



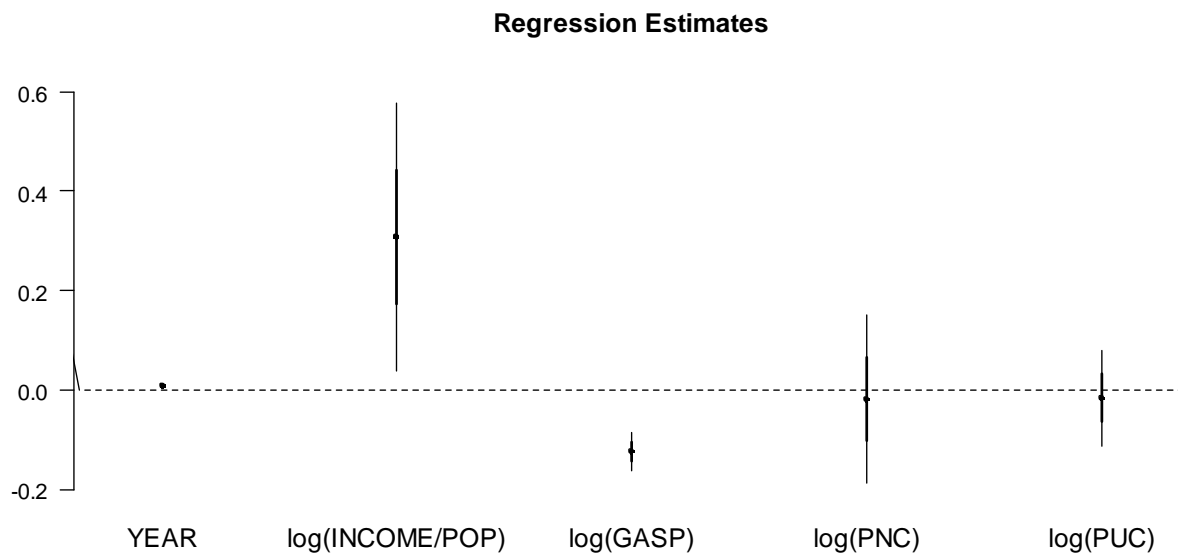
$$e'_*e_* = 0.16302$$

Re-estimate the model using data **up to 1973** only (**pre-shock** data):



$$e'_1e_1 = 0.00184$$

Re-estimate the model using data [after](#) 1973 only (post-shock data):



$$e_2'e_2 = 0.00739$$

Chow test:

$$F = \frac{[(e_*'e_*) - (e_1'e_1 + e_2'e_2)]/k}{(e_1'e_1 + e_2'e_2)/(n_1 + n_2 - 2k)} = \frac{[0.16302 - 0.00184 - 0.00739]/6}{(0.00184 + 0.00739)/(52 - 12)} = 111.267$$

From an F-distribution with 6 and 40 degrees of freedom, the p-value associated with this test statistic is 0.000.

An alternate way to calculate this test statistic is to estimate a model using dummy variables, and perform an F-test for the joint significance of all coefficients associated with a dummy variable.

DUM = 0 (1953 – 1973) ; = 1 (1974 – 2004)

| | Estimate | Std. Error | t value | Pr(> t) | |
|-----------------------|------------|------------|---------|----------|-----|
| (Intercept) | -60.998170 | 5.308283 | -11.491 | 3.03e-14 | *** |
| YEAR | 0.024922 | 0.002960 | 8.420 | 2.15e-10 | *** |
| log (INCOME/POP) | 0.660168 | 0.116328 | 5.675 | 1.35e-06 | *** |
| log (GASP) | -0.036362 | 0.205657 | -0.177 | 0.860553 | |
| log (PNC) | 0.638100 | 0.146745 | 4.348 | 9.18e-05 | *** |
| log (PUC) | -0.279605 | 0.069318 | -4.034 | 0.000240 | *** |
| DUM | 31.954337 | 5.859984 | 5.453 | 2.77e-06 | *** |
| YEAR:DUM | -0.015663 | 0.003184 | -4.920 | 1.53e-05 | *** |
| log (INCOME/POP) :DUM | -0.352420 | 0.166666 | -2.115 | 0.040750 | * |
| log (GASP) :DUM | -0.087200 | 0.206332 | -0.423 | 0.674837 | |
| log (PNC) :DUM | -0.656235 | 0.164627 | -3.986 | 0.000277 | *** |
| log (PUC) :DUM | 0.263556 | 0.081286 | 3.242 | 0.002394 | ** |

$$e'e = 0.00922$$

Note that $e_1'e_1 + e_2'e_2 = e'e$!

Insufficient Degrees of Freedom

- What if either $n_1 < k$, or $n_2 < k$?
- In this case we can't fit one of the sub-sample regressions, so F can't be computed.
- There is a second version of the Chow test, designed for this situation ("Chow Forecast Test").

Also, note:

- Location of break-point(s) assumed known.
- Situation becomes much more complicated if we have to *estimate* break-point locations(s).

Using the Wald Test

- If *any* of the usual assumptions that underly the F-test for exact linear restrictions are violated, then the usual Chow test is *not valid*.
- We can, however, still use the Wald test version of the Chow test.
- It will be valid only asymptotically (large n).
- It may have poor performance in small samples.
- Examples where we would use the Wald version of the Chow test –
 1. Random regressors (*e.g.*, lagged dependent variable).
 2. Non-Normal errors.

Appendix – R Code

```
#Data is from Greene, Table F2.2
#You will have to install the "arm" package if you wish to use "coefplot".
library(arm)
gasdata = read.csv("http://home.cc.umanitoba.ca/~godwinrt/7010/gas.csv")
attach(gasdata)

#View the break-point:
plot(YEAR, GAS)
lines(YEAR, GAS)
text(1973, 1.4, "1973")
arrows(1973, 1.35, 1973, 1.2, length = 0.1)

#Estimate the pooled model:
eq1 = lm(log(GASEXP/GASP/POP) ~ YEAR + log(INCOME/POP) + log(GASP) + log(PNC)
+ log(PUC))
#View the estimated coefficients:
coefplot(eq1, vertical=FALSE, var.las = 1, cex.var=1.2)
#Get the sum of squared residuals from the pooled (restricted) model:
sSER = sum(eq1$residuals^2)

#Use only the first 21 observations (up to 1973):
preshock = gasdata[1:21,]
attach(preshock)
eq2 = lm(log(GASEXP/GASP/POP) ~ YEAR + log(INCOME/POP) + log(GASP) + log(PNC)
+ log(PUC))
coefplot(eq2, vertical=FALSE, var.las = 1, cex.var=1.2)
sSE1 = sum(eq2$residuals^2)

#Use only the last 31 observations (after 1973):
postshock = gasdata[22:52,]
```

```
attach(postshock)
eq3 = lm(log(GASEXP/GASP/POP) ~ YEAR + log(INCOME/POP) + log(GASP) + log(PNC)
  + log(PUC))
coefplot(eq3,vertical=FALSE,var.las = 1,cex.var=1.2)
sseu2 = sum(eq3$residuals^2)

#Calculate Chow test statistic:
chow = ((sSER - SSEU1 - SSEU2)/6)/((SSEU1 + SSEU2)/(52 - 12))
#p-value:
1 - pf(chow,6,40)

#Estimate the model with dummy variables:
DUM = c(rep(0,21),rep(1,31))
attach(gasdata)
eq4 = lm(log(GASEXP/GASP/POP) ~ YEAR + log(INCOME/POP) + log(GASP) + log(PNC)
  + log(PUC) + DUM + DUM*YEAR + DUM*log(INCOME/POP) + DUM*log(GASP) +
  DUM*log(PNC) + DUM*log(PUC))
summary(eq4)
ssedum = sum(eq4$residuals^2)
```

Topic 5: Non-Linear Regression

- The models we've worked with so far have been *linear in the parameters*.
- They've been of the form: $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
- Many models based on economic theory are actually *non-linear* in the parameters.

CES Production function:

$$Y_i = \gamma [\delta K_i^{-\rho} + (1 - \delta)L_i^{-\rho}]^{-v/\rho} \exp(\varepsilon_i)$$

$$\text{or, } \ln(Y_i) = \ln(\gamma) - \left(\frac{v}{\rho}\right) \ln[\delta K_i^{-\rho} + (1 - \delta)L_i^{-\rho}] + \varepsilon_i$$

Linear Expenditure System:

(Stone, 1954)

$$\text{Max. } U(\mathbf{q}) = \sum_i \beta_i \ln(q_i - \gamma_i) \quad (\text{Stone-Geary / Klein-Rubin})$$

$$\text{s.t. } \sum_i p_i q_i = M$$

- Yields the following system of demand equations:

$$p_i q_i = \gamma_i p_i + \beta_i (M - \sum_j \gamma_j p_j) \quad ; \quad i = 1, 2, \dots, n$$

- The β_i 's are the *Marginal Budget Shares*.
- So, we require that $0 < \beta_i < 1$; $i = 1, 2, \dots, n$.
- Engel aggregation implies that

$$1. \quad \sum_i \gamma_i = 0 .$$

$$2. \quad \sum_i \beta_i = 1 .$$

- In general, suppose we have a single non-linear equation:

$$y_i = f(x_{i1}, x_{i2}, \dots, x_{ik}; \theta_1, \theta_2, \dots, \theta_p) + \varepsilon_i$$

- We can still consider a "Least Squares" approach.
- The **Non-Linear Least Squares** estimator is the vector, $\hat{\boldsymbol{\theta}}$, that *minimizes* the quantity:

$$S(X, \boldsymbol{\theta}) = \sum_i [y_i - f_i(X, \boldsymbol{\theta})]^2 .$$

- Clearly the usual LS estimator is just a special case of this.
- To obtain the estimator, we differentiate S with respect to each element of $\hat{\boldsymbol{\theta}}$; set up the "p" first-order conditions and solve.

- Difficulty – usually, the first-order conditions are themselves non-linear in the unknowns (the parameters).
- This means there is (generally) no exact, closed-form, solution.
- Can't write down an explicit formula for the estimators of parameters.

Example

$$y_i = \theta_1 + \theta_2 x_{i2} + \theta_3 x_{i3} + (\theta_2 \theta_3) x_{i4} + \varepsilon_i$$

$$S = \sum_i [y_i - \theta_1 - \theta_2 x_{i2} - \theta_3 x_{i3} - (\theta_2 \theta_3) x_{i4}]^2$$

$$\frac{\partial S}{\partial \theta_1} = -2 \sum_i [y_i - \theta_1 - \theta_2 x_{i2} - \theta_3 x_{i3} - (\theta_2 \theta_3) x_{i4}]$$

$$\frac{\partial S}{\partial \theta_2} = -2 \sum_i [(\theta_3 x_{i4} + x_{i2})(y_i - \theta_1 - \theta_2 x_{i2} - \theta_3 x_{i3} - \theta_2 \theta_3 x_{i4})]$$

$$\frac{\partial S}{\partial \theta_3} = -2 \sum_i [(\theta_2 x_{i4} + x_{i3})(y_i - \theta_1 - \theta_2 x_{i2} - \theta_3 x_{i3} - \theta_2 \theta_3 x_{i4})]$$

Setting these 3 equations to zero, we can't solve analytically for the estimators of the three parameters.

- In situations such as this, we need to use a numerical algorithm to obtain *a solution* to the first-order conditions.
- Lots of methods for doing this – one possibility is Newton's algorithm (the **Newton-Raphson algorithm**).

Methods of Descent

$$\tilde{\theta} = \theta_0 + s \mathbf{d}(\theta_0)$$

θ_0 = initial (vector) value.

s = step-length (positive scalar)

$\mathbf{d}(\cdot)$ = direction vector

- Usually, $d(\cdot)$ Depends on the gradient vector at θ_0 .
- It may also depend on the change in the gradient (the Hessian matrix) at θ_0 .
- Some specific algorithms in the “family” make the step-length a function of the Hessian.
- One very useful, specific member of the family of “Descent Methods” is the **Newton-Raphson algorithm**:

Suppose we want to minimize some function, $f(\theta)$.

Approximate the function using a Taylor’s series expansion about $\tilde{\theta}$, the vector value that minimizes $f(\theta)$:

$$f(\theta) \cong f(\tilde{\theta}) + (\theta - \tilde{\theta})' \left(\frac{\partial f}{\partial \theta} \right)_{\tilde{\theta}} + \frac{1}{2!} (\theta - \tilde{\theta})' \left[\frac{\partial^2 f}{\partial \theta \partial \theta'} \right]_{\tilde{\theta}} (\theta - \tilde{\theta})$$

Or:

$$f(\theta) \cong f(\tilde{\theta}) + (\theta - \tilde{\theta})' g(\tilde{\theta}) + \frac{1}{2!} (\theta - \tilde{\theta})' H(\tilde{\theta})(\theta - \tilde{\theta})$$

So,

$$\frac{\partial f(\theta)}{\partial \theta} \cong 0 + (\theta - \tilde{\theta})' g(\tilde{\theta}) + \frac{1}{2!} 2H(\tilde{\theta})(\theta - \tilde{\theta})$$

However, $g(\tilde{\theta}) = 0$; as $\tilde{\theta}$ locates a minimum.

So,

$$(\theta - \tilde{\theta}) \cong H^{-1}(\tilde{\theta}) \left(\frac{\partial f(\theta)}{\partial \theta} \right);$$

or,
$$\tilde{\theta} \cong \theta - H^{-1}(\tilde{\theta})g(\theta)$$

This suggests a numerical algorithm:

Set $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ to begin, and then iterate –

$$\boldsymbol{\theta}_1 = \boldsymbol{\theta}_0 - H^{-1}(\boldsymbol{\theta}_1)g(\boldsymbol{\theta}_0)$$

$$\boldsymbol{\theta}_2 = \boldsymbol{\theta}_1 - H^{-1}(\boldsymbol{\theta}_2)g(\boldsymbol{\theta}_1)$$

$$\vdots \quad \vdots \quad \quad \quad \vdots$$

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n - H^{-1}(\boldsymbol{\theta}_{n+1})g(\boldsymbol{\theta}_n)$$

or, approximately:

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n - H^{-1}(\boldsymbol{\theta}_n)g(\boldsymbol{\theta}_n)$$

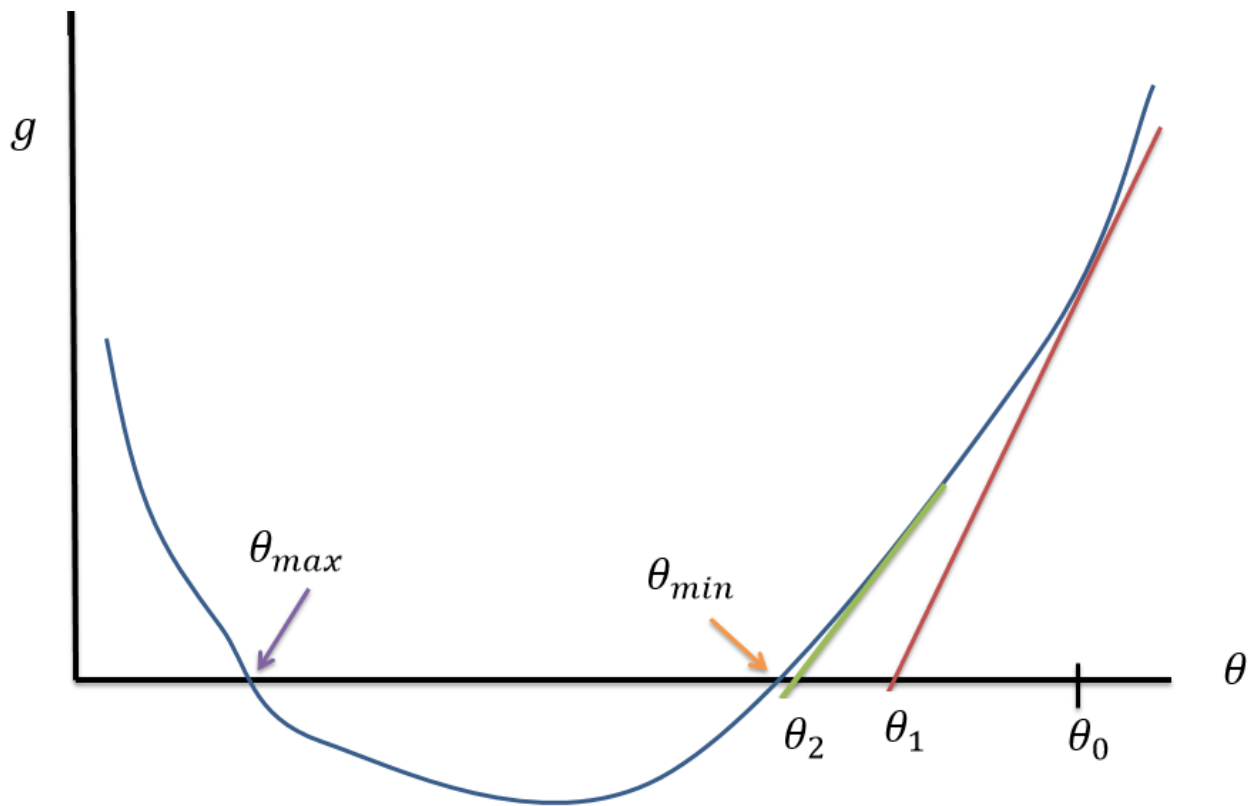
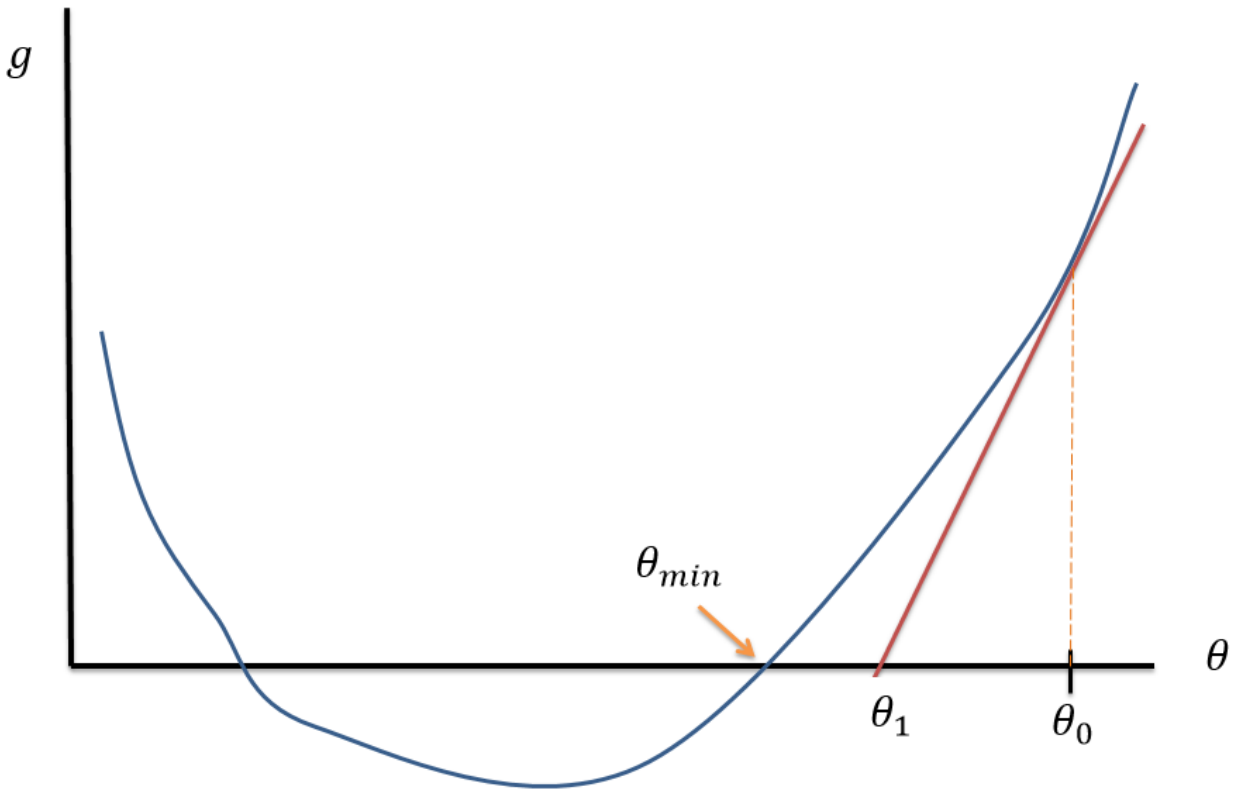
Stop if $\left| \frac{(\theta_{n+1}^{(i)} - \theta_n^{(i)})}{\theta_n^{(i)}} \right| < \varepsilon^{(i)} \quad ; \quad i = 1, 2, \dots, p$

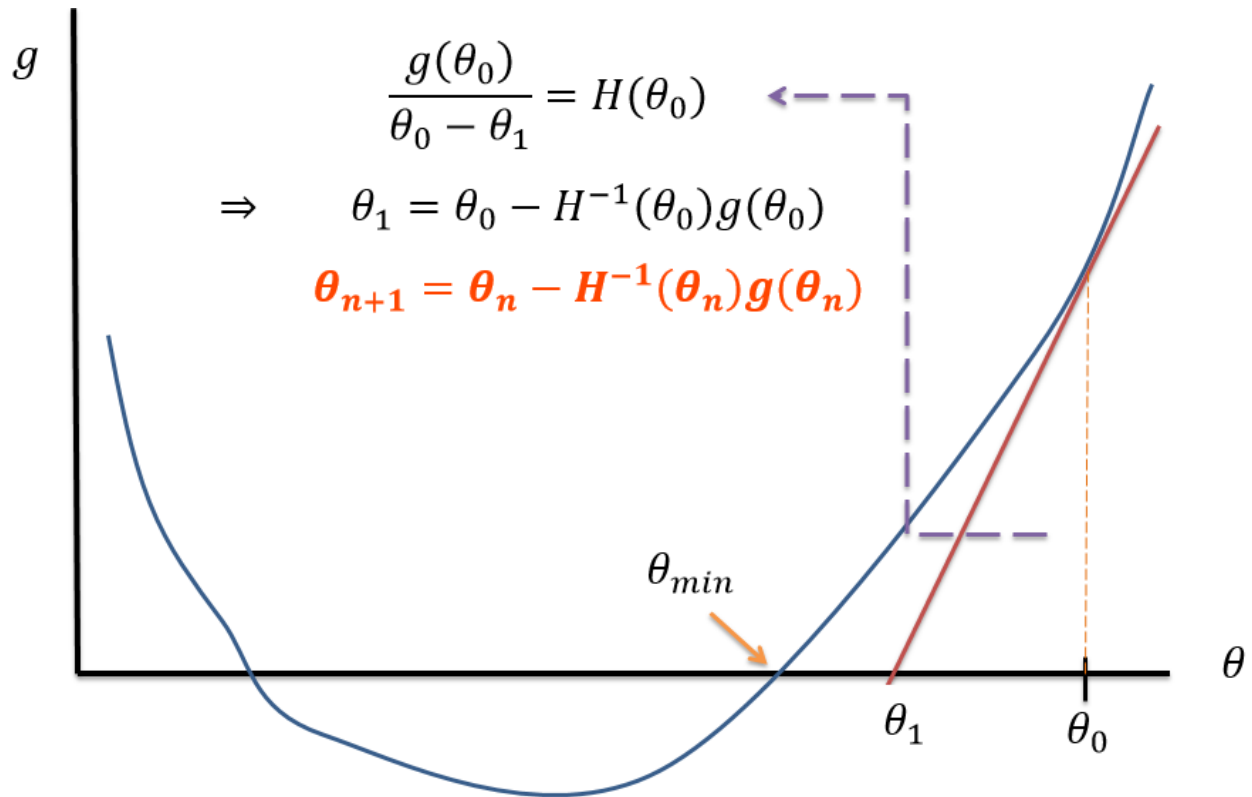
Note:

1. $s = 1$.
2. $\mathbf{d}(\boldsymbol{\theta}_n) = -H^{-1}(\boldsymbol{\theta}_n)g(\boldsymbol{\theta}_n)$.
3. Algorithm *fails* if H ever becomes *singular* at any iteration.
4. Achieve a *minimum* of $f(\cdot)$ if H is *positive definite*.
5. Algorithm may locate only a *local* minimum.
6. Algorithm may *oscillate*.

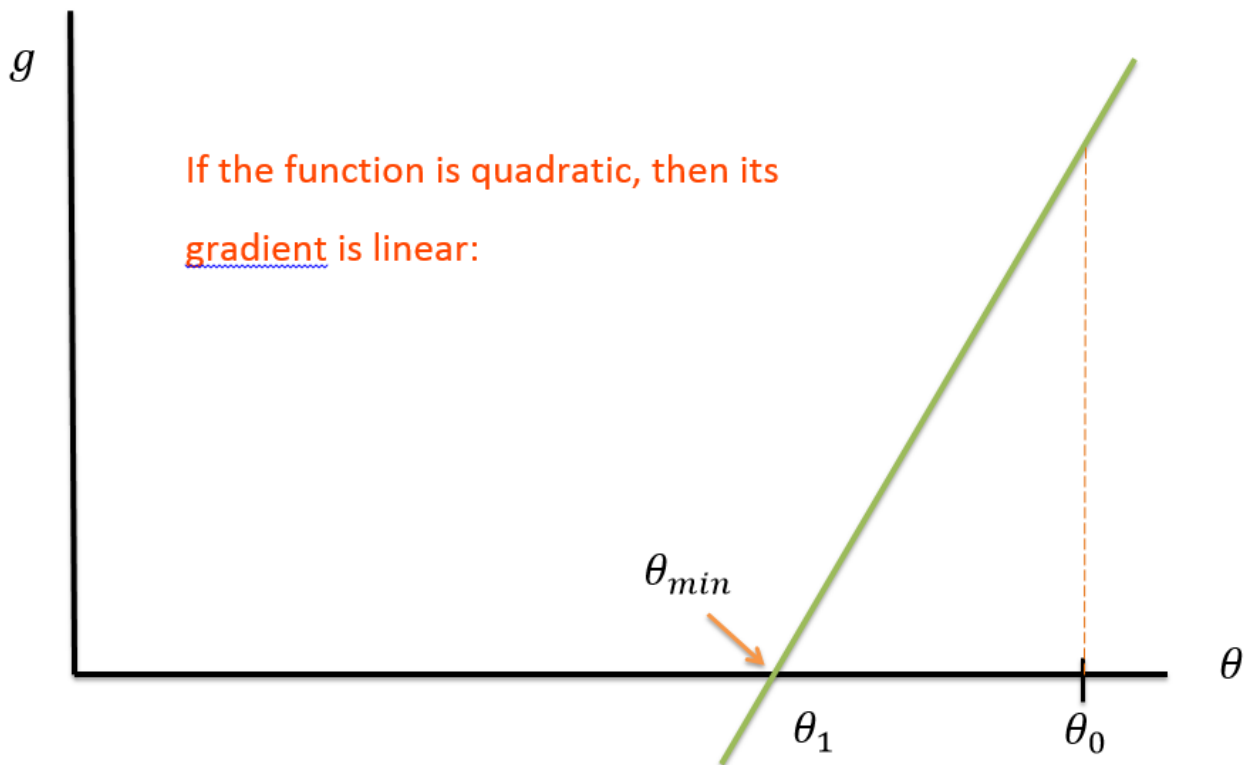
The algorithm can be given a nice *geometric interpretation* – scalar θ .

To find an extremum of $f(\cdot)$, solve $\frac{\partial f(\theta)}{\partial \theta} = g(\theta) = 0$.

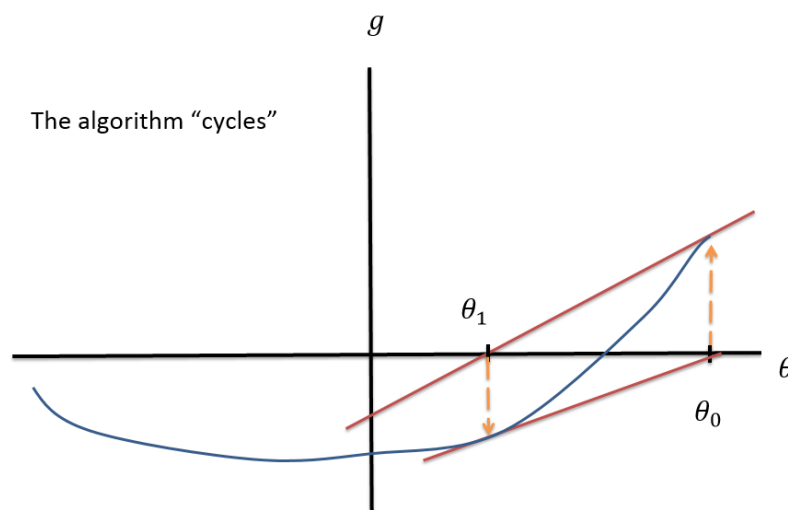
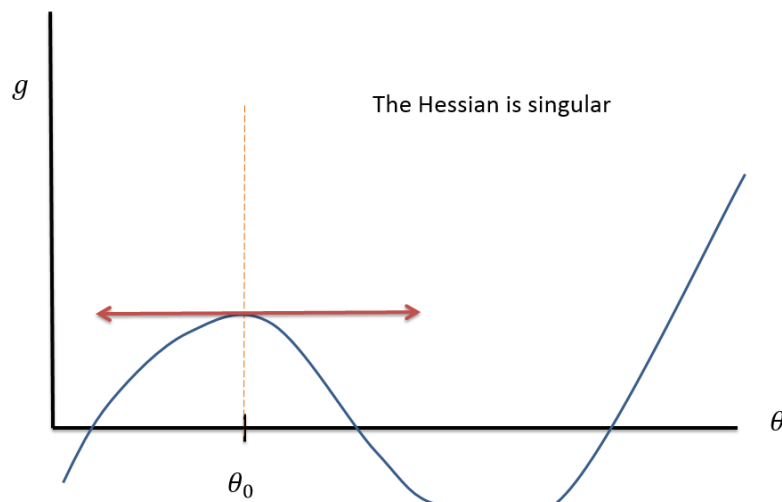
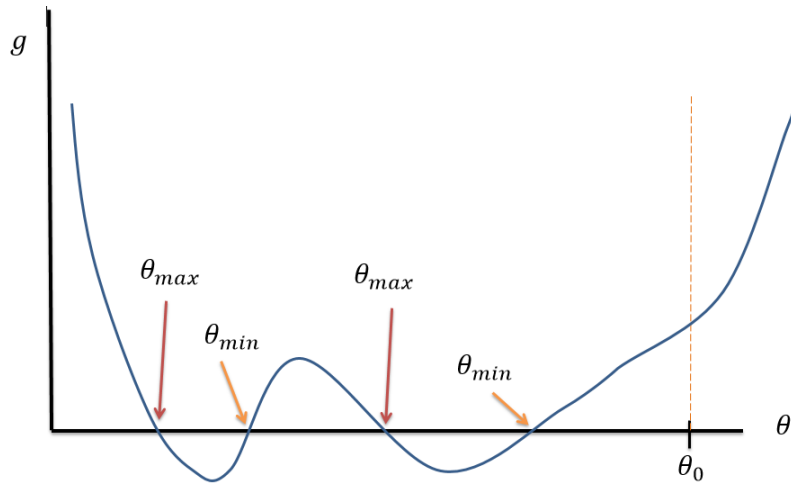




If $f(\theta)$ is *quadratic* in θ , then the algorithm converges in one iteration:



In general, different choices of θ_0 may lead to different solutions, or no solution at all.



Example

(Where we actually know the answer)

$$f(\theta) = 3\theta^4 - 4\theta^3 + 1 \quad \text{locate minimum}$$

Analytically:

$$g(\theta) = 12\theta^3 - 12\theta^2 = 12\theta^2(\theta - 1)$$

$$H(\theta) = 36\theta^2 - 24\theta = 12\theta(3\theta - 2)$$

Turning points at $\theta = 0, 0, 1$.

$$H(0) = 0 \quad \text{saddlepoint}$$

$$H(1) = 12 \quad \text{minimum}$$

Algorithm

$$\theta_{n+1} = \theta_n - H^{-1}(\theta_n)g(\theta_n)$$

$$\theta_0 = 2 \quad (\text{say})$$

$$\theta_1 = 2 - \left(\frac{48}{96}\right) = 1.5$$

$$\theta_2 = 1.5 - \left(\frac{13.5}{45}\right) = 1.2$$

$$\theta_3 = 1.2 - \left(\frac{3.456}{23.040}\right) = 1.05$$

⋮

*etc.*Try: $\theta_0 = -2; \theta_0 = 0.5$

Topic 6: Non-Spherical Disturbances

Our basic linear regression model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad ; \quad \boldsymbol{\varepsilon} \sim N[\mathbf{0}, \sigma^2 \mathbf{I}_n]$$

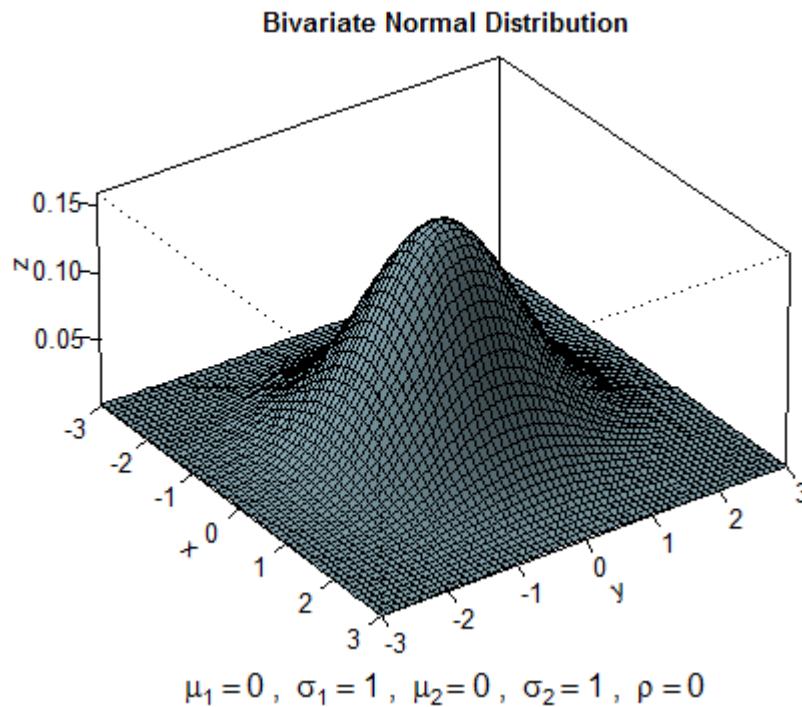
Now we'll generalize the specification of the error term in the model:

$$E[\boldsymbol{\varepsilon}] = \mathbf{0} \quad ; \quad E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] = \boldsymbol{\Sigma} = \sigma^2 \boldsymbol{\Omega} \quad ; \quad (\& \text{Normal})$$

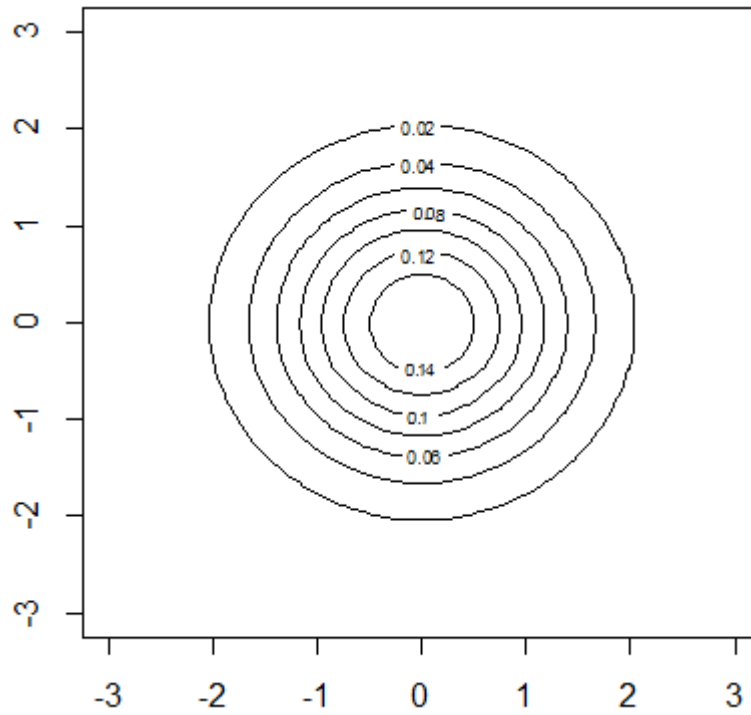
This allows for the possibility of one or both of

- Heteroskedasticity
- Autocorrelation (Cross-section; Time-series; Panel data)

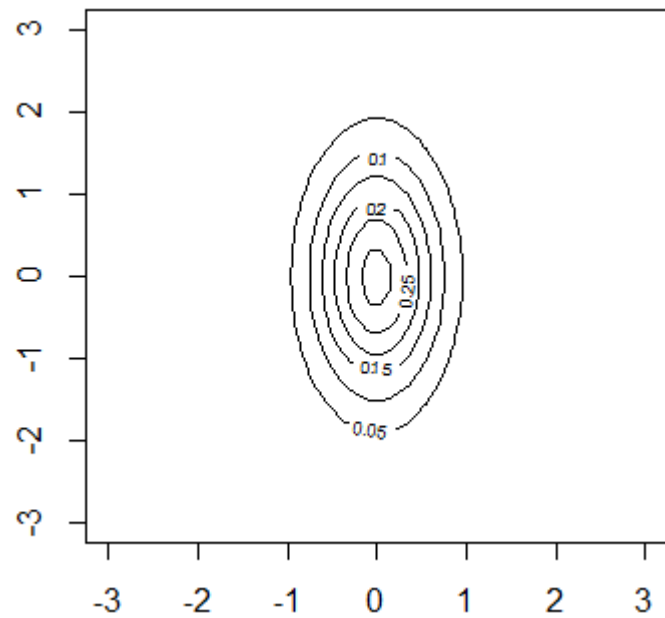
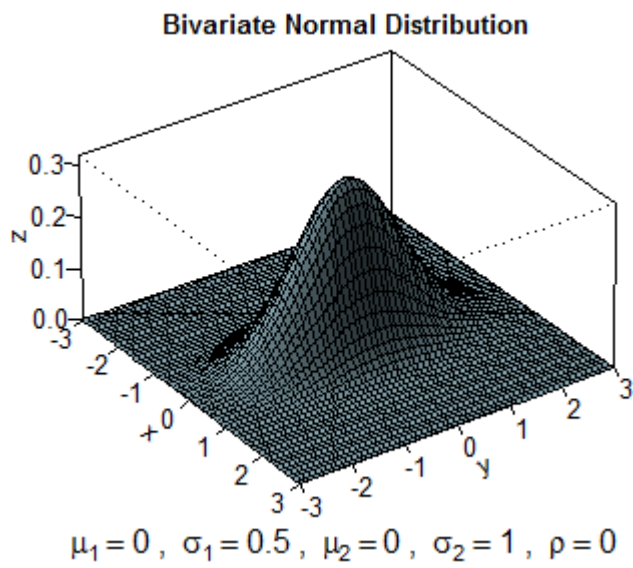
Spherical Disturbances – Homoskedasticity and Non-Autocorrelation



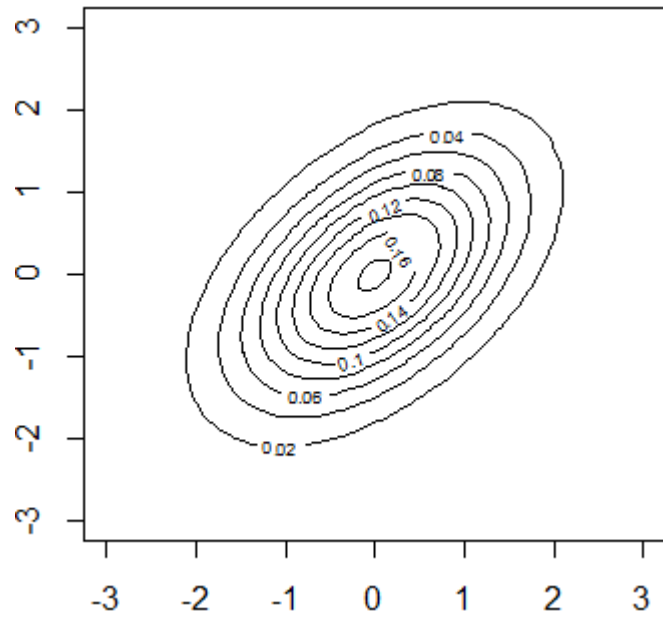
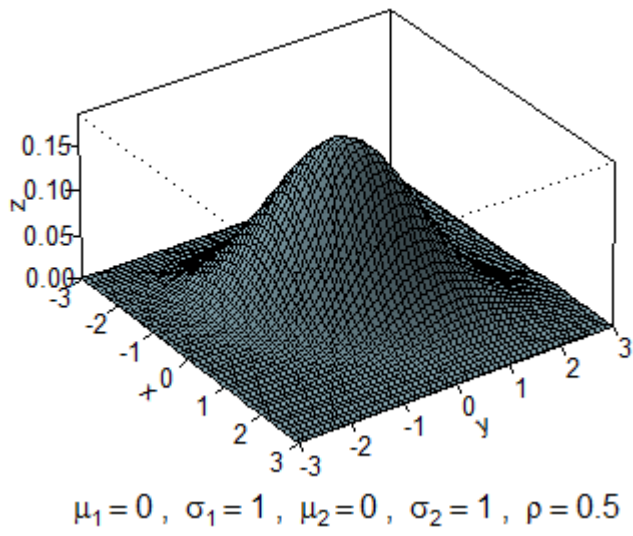
In the above, consider $x = \varepsilon_i$ and $y = \varepsilon_j$. The joint probability density function, $p(\varepsilon_i, \varepsilon_j)$, is in the direction of the z axis. Below is a contour of the above perspective. If we consider the joint distribution of *three* error terms instead of *two*, the circles below would become spheres, hence the terminology “spherical disturbances.”



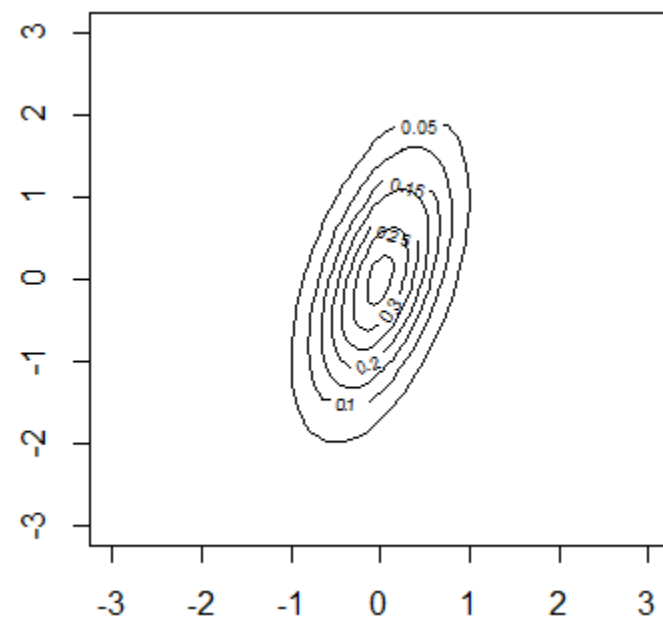
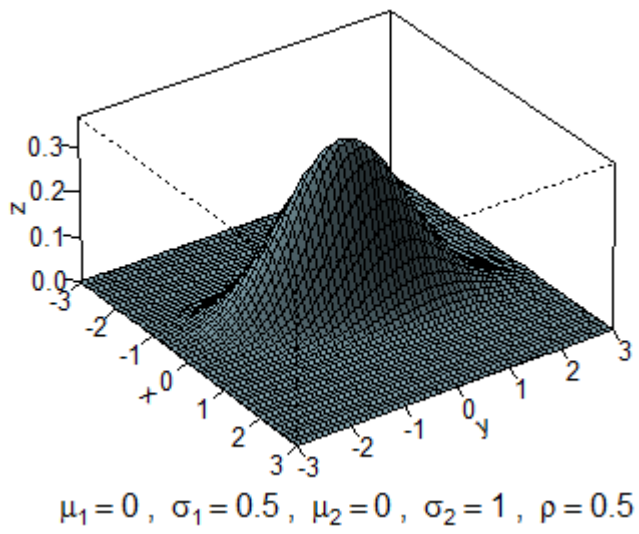
Non-Spherical Disturbances – Heteroskedasticity and Non-Autocorrelation



Non-Spherical Disturbances – Homoskedasticity and Autocorrelation



Non-Spherical Disturbances – Heteroskedasticity and Autocorrelation



- How does the more general situation of non-spherical disturbances affect our (Ordinary) Least Squares estimator?
- In particular, let's first look at the sampling distribution of \mathbf{b} :

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

So,

$$E(\mathbf{b}) = \boldsymbol{\beta} + (X'X)^{-1}X'E(\boldsymbol{\varepsilon}) = \boldsymbol{\beta} .$$

The more general form of the covariance matrix for the error term does not alter the fact that the OLS estimator is *unbiased*.

- Next, consider the covariance matrix of our OLS estimator in this more general situation:

$$\begin{aligned}V(\mathbf{b}) &= V[\boldsymbol{\beta} + (X'X)^{-1}X'\boldsymbol{\varepsilon}] = V[(X'X)^{-1}X'\boldsymbol{\varepsilon}] \\ &= [(X'X)^{-1}X'V(\boldsymbol{\varepsilon})X(X'X)^{-1}] \\ &= [(X'X)^{-1}X'\sigma^2\Omega X(X'X)^{-1}] \\ &\neq [\sigma^2(X'X)^{-1}] .\end{aligned}$$

- So, under our full set of modified assumptions about the error term:

$$\mathbf{b} \sim N[\boldsymbol{\beta}, V^*]$$

where

$$V^* = \sigma^2[(X'X)^{-1}X'\Omega X(X'X)^{-1}] .$$

- So, the usual computer output will be misleading, *numerically*, as it will be based on using the wrong formula, namely $s^2(X'X)^{-1}$.
- The standard errors, t-statistics, *etc.* will all be incorrect.
- As well as being *unbiased*, the OLS point estimator of $\boldsymbol{\beta}$ will still be *weakly consistent*.
- The I.V. estimator of $\boldsymbol{\beta}$ will still be *weakly consistent*.

- The NLLS estimator of the model's parameters will still be *weakly consistent*.
- **However**, the usual estimator for the covariance matrix of \mathbf{b} , namely $s^2(X'X)^{-1}$, will be an *inconsistent estimator* of the true covariance matrix of \mathbf{b} !
- This has serious implications for inferences based on confidence intervals, tests of significance, *etc.*
- So, we need to know how to deal with these issues.
- This will lead us to some *generalized estimators*.
- First, let's deal with the most pressing issue – the inconsistency of the estimator for the covariance matrix of \mathbf{b} .

White's Heteroskedasticity-Consistent Covariance Matrix Estimator

- If we knew $\sigma^2\Omega$, then the “estimator” of the covariance matrix for \mathbf{b} would just be:

$$\begin{aligned} V^* &= [(X'X)^{-1}X'\sigma^2\Omega X(X'X)^{-1}] \\ &= \frac{1}{n} \left[\left(\frac{1}{n} X'X \right)^{-1} \left(\frac{1}{n} X' \sigma^2 \Omega X \right) \left(\frac{1}{n} X'X \right)^{-1} \right] \\ &= \frac{1}{n} \left[\left(\frac{1}{n} X'X \right)^{-1} \left(\frac{1}{n} X' \Sigma X \right) \left(\frac{1}{n} X'X \right)^{-1} \right] \end{aligned}$$

- If Σ is *unknown*, then we need to find a consistent estimator of $\left(\frac{1}{n} X' \Sigma X \right)$.
- (Why not an estimator of just Σ ?)
- Note that at this stage of the discussion, the form of the Σ matrix is quite arbitrary.
- Let $Q^* = \left(\frac{1}{n} X' \Sigma X \right) \quad (k \times k)$

$$= \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n \sigma_{ij} \mathbf{x}_i \mathbf{x}_j'$$

$$(k \times 1) \quad (1 \times k)$$

- In the case of *heteroskedastic errors*, things simplify, because $\sigma_{ij} = 0$, for $i \neq j$.

Then, we have

$$Q^* = \frac{1}{n} \sum_{i=1}^n \sigma_i^2 \mathbf{x}_i \mathbf{x}_i'$$

- White (1980) showed that if we define

$$S_0 = \frac{1}{n} \sum_{i=1}^n e_i^2 \mathbf{x}_i \mathbf{x}_i'$$

Then , $plim(S_0) = Q^*$.

- This means that we can estimate the model by OLS; get the associated residual vector, \mathbf{e} ; and then a consistent estimator of V^* , the covariance matrix of \mathbf{b} , will be:

$$\hat{V}^* = \frac{1}{n} \left[\left(\frac{1}{n} X'X \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n e_i^2 \mathbf{x}_i \mathbf{x}_i' \right) \left(\frac{1}{n} X'X \right)^{-1} \right]$$

or,

$$\hat{V}^* = n[(X'X)^{-1}S_0(X'X)^{-1}] .$$

- \hat{V}^* is a **consistent estimator** of V^* , regardless of the (unknown) form of the heteroskedasticity.
- This includes **no heteroskedasticity** (i.e., homoscedastic errors).
- Newey & West produced a corresponding consistent estimator of V^* for when the errors possibly exhibit autocorrelation (of some unknown form).
- Note that the White and the Newey-West estimators modify just the estimated covariance matrix of \mathbf{b} – not \mathbf{b} itself.
- As a result, the t -statistics, F -statistic, etc., will be modified, but only in a manner that is appropriate **asymptotically**.
- So, if we have heteroskedasticity (or autocorrelation), whether we modify the covariance estimator or not, the usual test statistics will be unreliable **in finite samples**.
- A good practical solution is to use White's (or Newey-West's) adjustment, and then use the Wald test, rather than the F -test for exact linear restrictions.
- This Wald test will incorporate the consistent estimator of the covariance matrix of \mathbf{b} , and so it will still be valid, **asymptotically**.

- Now let's turn to the estimation of β , taking account of the fact that the error term has a non-scalar covariance matrix.
- Using this information should enable us to improve the *efficiency* of the LS estimator of the coefficient vector.

Generalized Least Squares

(Alexander Aitken, 1935)

- In the present context, (Ordinary) LS ignores some important information, and we'd anticipate that this will result in a loss of efficiency when estimating β .
- Let's see how to obtain the fully efficient (linear unbiased) estimator.
- Recall that $V(\boldsymbol{\varepsilon}) = E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] = \Sigma = \sigma^2\Omega$.
- Generally, Ω will be *unknown*. However, to begin with, let's consider the case where it is actually *known*.
- Clearly, Ω must be *symmetric*, as it is a covariance matrix.
- Suppose that Ω is also *positive-definite*.
- Then, Ω^{-1} is also positive-definite, and so there exists a *non-singular* matrix, P , such that $\Omega^{-1} = P'P$.
- In fact, $P' = C\Lambda^{-1/2}$, where the columns of C are the characteristic vectors of Ω , and $\Lambda^{1/2} = \text{diag.}(\sqrt{\lambda_i})$. Here, the $\{\lambda_i\}$ are the characteristic roots of Ω .
- Our model is:

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad ; \quad \boldsymbol{\varepsilon} \sim [0, \sigma^2\Omega]$$

- Pre-multiply the equation by P :

$$P\mathbf{y} = PX\boldsymbol{\beta} + P\boldsymbol{\varepsilon}$$

or,

$$\mathbf{y}^* = X^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}^* \quad ; \quad \text{say}$$

- Now, Ω is non-random, so P is also non-random.
- So, $E[\boldsymbol{\varepsilon}^*] = E[P\boldsymbol{\varepsilon}] = P E[\boldsymbol{\varepsilon}] = \mathbf{0}$.

- And $V[\boldsymbol{\varepsilon}^*] = V[P\boldsymbol{\varepsilon}]$

$$= PV(\boldsymbol{\varepsilon})P'$$

$$= P(\sigma^2\Omega)P' = \sigma^2P\Omega P'$$

- Note that $P\Omega P' = P(\Omega^{-1})^{-1}P'$

$$= P(P'P)^{-1}P'$$

$$= PP^{-1}(P')^{-1}P' = I$$

- (Because P is both **square and non-singular**.)
- So, $E[\boldsymbol{\varepsilon}^*] = \mathbf{0}$ and $V[\boldsymbol{\varepsilon}^*] = \sigma^2 I$.
- The transformed model, $\mathbf{y}^* = X^* \boldsymbol{\beta} + \boldsymbol{\varepsilon}^*$, has an error-term that satisfies the *usual assumptions*. **In particular, it has a scalar covariance matrix.**
- So, if we apply (Ordinary) Least Squares to the model, $\mathbf{y}^* = X^* \boldsymbol{\beta} + \boldsymbol{\varepsilon}^*$, we'll get the BLU estimator of $\boldsymbol{\beta}$, by the Gauss-Markhov Theorem.
- We call this the **Generalized Least Squares Estimator** of $\boldsymbol{\beta}$.
- The formula for this estimator is readily determined:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= [X^{*'}X^*]^{-1}X^{*'}\mathbf{y}^* \\ &= [(PX)'(PX)]^{-1}(PX)'(P\mathbf{y}) \\ &= [X'P'PX]^{-1}X'P'P\mathbf{y} \\ &= [X'\Omega^{-1}X]^{-1}X'\Omega^{-1}\mathbf{y}\end{aligned}$$

- Note that we can also write the GLS estimator as:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= [X'(\sigma^2\Omega)^{-1}X]^{-1}X'(\sigma^2\Omega)^{-1}\mathbf{y} \\ &= [X'\Sigma^{-1}X]^{-1}X'\Sigma^{-1}\mathbf{y} = [X'\Omega^{-1}X]^{-1}X'\Omega^{-1}\mathbf{y}\end{aligned}$$

- Clearly, because $E[\boldsymbol{\varepsilon}^*] = \mathbf{0}$ as long as the regressors are non-random, the GLS estimator, $\hat{\boldsymbol{\beta}}$ is **unbiased**.
- Moreover, the covariance matrix of the GLS estimator is:

$$V(\hat{\boldsymbol{\beta}}) = [X'\Omega^{-1}X]^{-1}X'\Omega^{-1}V(\mathbf{y})\{[X'\Omega^{-1}X]^{-1}X'\Omega^{-1}\}'$$

$$= [X'\Omega^{-1}X]^{-1}X'\Omega^{-1}\sigma^2\Omega^{-1}X[X'\Omega^{-1}X]^{-1}$$

$$= \sigma^2[X'\Omega^{-1}X]^{-1}.$$

- If the errors are Normally distributed, then the full sampling distribution of the GLS estimator of β is:

$$\hat{\beta} \sim N[\beta, \sigma^2[X'\Omega^{-1}X]^{-1},]$$

- The GLS estimator is just the OLS estimator, applied to the transformed model, and the latter model satisfies all of the usual conditions.
- So, the *Gauss-Markhov Theorem* applies to the GLS estimator.
- The GLS estimator is BLU for this more general model (with a non-scalar error covariance matrix).
- Note: OLS must be *inefficient* in the present context.
- Have a more general form of the GMT – the OLS version is a special case.
- Moreover, all of the results that we established with regard to testing for linear restrictions and incorporating them into our estimation, also apply if we make some obvious modifications.
- $\hat{\beta}$ = GLS estimator
 $\hat{\varepsilon} = \mathbf{y}^* - X^*\hat{\beta}$
 $\hat{\sigma}^2 = \hat{\varepsilon}'\hat{\varepsilon}/(n - k)$
- Then, to test $H_0: R\beta = \mathbf{q}$ vs. $H_A: R\beta \neq \mathbf{q}$ we would use the test statistic,

$$F = (R\hat{\beta} - \mathbf{q})'[R(X^{**}X^*)^{-1}R']^{-1}(R\hat{\beta} - \mathbf{q}) / J\hat{\sigma}^2$$
- If H_0 is true, then is distributed as $F_{J, n-k}$.
- We can also construct the **Restricted GLS estimator**, in the same way that we obtained the restricted OLS estimator of β .

- Check for yourself that this restricted estimator is

$$\begin{aligned}\widehat{\beta}_r &= \widehat{\beta} - (X^*{}'X^*)^{-1}R'[R(X^*{}'X^*)^{-1}R']^{-1}(R\widehat{\beta} - \mathbf{q}) \\ &= \widehat{\beta} - (X'\Omega^{-1}X)^{-1}R'[R(X'\Omega^{-1}X)^{-1}R']^{-1}(R\widehat{\beta} - \mathbf{q}) \\ &= \widehat{\beta} - (X'\Sigma^{-1}X)^{-1}R'[R(X'\Sigma^{-1}X)^{-1}R']^{-1}(R\widehat{\beta} - \mathbf{q})\end{aligned}$$

- Then, if the residuals from this restricted GLS estimation are defined as $\widehat{\boldsymbol{\varepsilon}}_r = \mathbf{y} - X\widehat{\beta}_r$, we can also write the F-test statistic as:

$$F = [\widehat{\boldsymbol{\varepsilon}}_r'\widehat{\boldsymbol{\varepsilon}}_r - \widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}}] / (J\widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}} / (n - k))$$

- Recalling our formula for the GLS estimator, we see that it depends on the (usually unknown) covariance matrix of the error term:

$$\widehat{\beta} = [X'\Omega^{-1}X]^{-1}X'\Omega^{-1}\mathbf{y} .$$

“Feasible” GLS Estimator

- In order to be able to implement the GLS estimator, in practice, we’re usually going to have to provide a *suitable estimator* of Ω (or Σ).
- Presumably we’ll want to obtain an estimator that is *at least consistent*, and place this into the formula for the GLS estimator, yielding:

$$\widetilde{\beta} = [X'\widehat{\Omega}^{-1}X]^{-1}X'\widehat{\Omega}^{-1}\mathbf{y}$$

- Problem: The Ω matrix is $(n \times n)$, and it has $n(n + 1)/2$ *distinct* elements. However, we have only n observations on the data. This precludes obtaining a consistent estimator.
- We need to constrain the elements of Ω in some way.
- In practice, this won’t be a big problem, because usually there will be lots of “structure” on the form of Ω .
- Typically, we’ll have $\Omega = \Omega(\boldsymbol{\theta})$, where the vector, $\boldsymbol{\theta}$ has low dimension.

Example: Heteroskedasticity

Suppose that $var.(\varepsilon_i) \propto (\theta_1 + \theta_2 z_i) = \sigma^2(\theta_1 + \theta_2 z_i)$

Then,

$$\Omega = \begin{pmatrix} \theta_1 + \theta_2 z_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \theta_1 + \theta_2 z_n \end{pmatrix}$$

There are just two parameters that have to be estimated, in order to obtain $\widehat{\Omega}$.

Example: Autocorrelation

Suppose that the errors follow a *first-order autoregressive process*:

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t \quad ; \quad u_t \sim N[0, \sigma_u^2] \quad (\text{i.i.d.})$$

Then (for reasons we'll see later),

$$\Omega = \frac{\sigma_u^2}{1-\rho^2} \begin{bmatrix} 1 & \rho & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \rho^{n-2} \\ \vdots & \rho & \ddots & \vdots \\ \rho^{n-1} & \cdots & \cdots & 1 \end{bmatrix} = \Omega(\rho).$$

- So, typically, we'll just have to estimate a very small number of parameters in order to get an estimator of Ω .
- As long as we use a *consistent estimator* for these parameters – the elements of θ , this will give us a consistent estimator of Ω and of Ω^{-1} , by Slutsky's Theorem.
- This in turn, will ensure that our Feasible GLS estimator of β is also *weakly consistent*:

$$\begin{aligned} plim(\tilde{\beta}) &= plim \left\{ [X' \widehat{\Omega}^{-1} X]^{-1} X' \widehat{\Omega}^{-1} \mathbf{y} \right\} \\ &= plim \left\{ [X' \Omega^{-1} X]^{-1} X' \Omega^{-1} \mathbf{y} \right\} \\ &= plim(\hat{\beta}) = \beta . \end{aligned}$$

- Also, if $\widehat{\Omega}$ is consistent for Ω then $\tilde{\beta}$ will be *asymptotically efficient*.

- In general, we can say little about the *finite-sample* properties of our feasible GLS estimator.
- Usually it will be *biased*, and the nature of the bias will depend on the form of Ω and our choice of $\hat{\Omega}$.
- In order to apply either the GLS estimator, or the feasible GLS estimator, we need to know the form of Ω .
- Typically, this is achieved by postulating various forms, and testing to see if these are supported by the data.

Appendix – R-Code for perspective plots and contours

(see <http://quantcorner.wordpress.com/2012/09/21/bivariate-normal-distribution-with-r/>)

```
# Édouard Tallent @ TaGoMa.Tech
# September 2012
# This code plots simulated bivariate normal distributions

# Some variable definitions
mu1 <- 0 # expected value of x
mu2 <- 0 # expected value of y
sig1 <- 0.5 # variance of x
sig2 <- 1 # variance of y
rho <- 0.5 # corr(x, y)

# Some additional variables for x-axis and y-axis
xm <- -3
xp <- 3
ym <- -3
yp <- 3

x <- seq(xm, xp, length= as.integer((xp + abs(xm)) * 10)) # vector
series x
y <- seq(ym, yp, length= as.integer((yp + abs(ym)) * 10)) # vector
series y

# Core function
bivariate <- function(x,y){
  term1 <- 1 / (2 * pi * sig1 * sig2 * sqrt(1 - rho^2))
  term2 <- (x - mu1)^2 / sig1^2
  term3 <- -(2 * rho * (x - mu1)*(y - mu2))/(sig1 * sig2)
  term4 <- (y - mu2)^2 / sig2^2
  z <- term2 + term3 + term4
  term5 <- term1 * exp((-z / (2 * (1 - rho^2))))
  return (term5)
}
```

```
# Computes the density values
z <- outer(x,y,bivariate)

# Plot
persp(x, y, z, main = "Bivariate Normal Distribution",
      sub = bquote(bold(mu[1])==".(mu1)~", "~sigma[1]==".(sig1)~",
                  "~mu[2]==".(mu2)~", "~sigma[2]==".(sig2)~", "~rho==".(rho)),
      col="lightblue", theta = 55, phi = 30, r = 40, d = 0.1, expand
      = 0.5,ltheta = 90, lphi = 180, shade = 0.4, ticktype =
      "detailed", nticks=5)

#In order to see the contours, use:
#contour(x,y,z)
```

Topic 7: Heteroskedasticity

Consider the linear regression model

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad ; \quad \boldsymbol{\varepsilon} \sim N[\mathbf{0}, \sigma^2\Omega]$$

where

$$\sigma^2\Omega = \sigma^2 \begin{bmatrix} \omega_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \omega_{nn} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix} = \text{diag.}(\sigma_i^2)$$

Then the errors exhibit Heteroskedasticity, but they are still uncorrelated.

- We know, from Topic 6, that in this case the OLS estimator of $\boldsymbol{\beta}$ is unbiased and consistent, but it is *inefficient*.
- We know that we can use White's modified estimator for the covariance matrix of $\boldsymbol{\beta}$ to ensure that the standard errors of the b_i 's are *consistent* estimators for the true s.e. (b_i)'s.
- We also know that we can use GLS to obtain the BLU estimator of $\boldsymbol{\beta}$ if Ω is *known*.

• If
$$\Omega = \begin{bmatrix} \omega_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \omega_{nn} \end{bmatrix},$$

then
$$P = \begin{bmatrix} \omega_{11}^{-1/2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \omega_{nn}^{-1/2} \end{bmatrix},$$

so that
$$P'P = \Omega^{-1}$$

- So, in this particular case, GLS estimation involves transforming the data:

$$\mathbf{y}^* = P\mathbf{y} \quad ; \quad X^* = PX$$

- Just multiply the model by the matrix, P , or simply scale the i^{th} observation of all variables by $\omega_{ii}^{-1/2}$:

$$\omega_{ii}^{-1/2}y_i = \beta_1\omega_{ii}^{-1/2} + \beta_2\left(\omega_{ii}^{-\frac{1}{2}}x_{i2}\right) + \cdots + \left(\omega_{ii}^{-\frac{1}{2}}\varepsilon_i\right)$$

- This particular variant of GLS is often referred to as “**Weighted Least Squares**” estimation. It is just OLS applied using “weighted” data.

Example:

$$y_i = \beta_1 + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \varepsilon_i$$

$$E[\varepsilon_i] = 0 \quad ; \quad \text{var.}[\varepsilon_i] \propto x_{i2}^2$$

So, we can write:

$$\text{var.}[\varepsilon_i] = \sigma^2 x_{i2}^2 \quad ; \quad \omega_{ii} = x_{i2}^2 \quad ; \quad \omega_{ii}^{-1/2} = 1/x_{i2}$$

$$(y_i/x_{i2}) = \beta_1(1/x_{i2}) + \beta_2 + \dots + \beta_k(x_{ik}/x_{i2}) + \varepsilon_i^*$$

where $\varepsilon_i^* = \left(\frac{\varepsilon_i}{x_{i2}}\right)$; $E[\varepsilon_i^*] = 0$ (assumption?)

$$\text{var.}[\varepsilon_i^*] = (1/x_{i2})^2 \text{var.}[\varepsilon_i] = (1/x_{i2})^2 \sigma^2 x_{i2}^2 = \sigma^2$$

Example:

$$y_i = \beta_1 + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \varepsilon_i$$

$$E[\varepsilon_i] = 0 \quad ; \quad \text{var.}[\varepsilon_i] \propto z_i^p$$

$$\text{var.}[\varepsilon_i] = \sigma^2 z_i^p \quad ; \quad \omega_{ii} = z_i^p \quad ; \quad \omega_{ii}^{-1/2} = z_i^{-p/2}$$

$$(y_i z_i^{-p/2}) = \beta_1(z_i^{-p/2}) + \beta_2(x_{i2} z_i^{-p/2}) + \dots + \beta_k(x_{ik} z_i^{-p/2}) + \varepsilon_i^*$$

where $\varepsilon_i^* = (\varepsilon_i z_i^{-p/2})$; $E[\varepsilon_i^*] = 0$ (assumption?)

$$\text{var.}[\varepsilon_i^*] = (z_i^{-p/2})^2 \text{var.}[\varepsilon_i] = z_i^{-p} \sigma^2 z_i^p = \sigma^2$$

Note that in this case we end up with a fitted model with no intercept, but we are still estimating the original parameters of interest.

- In some cases we will actually **know** the form of the heteroskedasticity, so we can apply WLS directly.

Example:

$$y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

$$E[\varepsilon_i] = 0 \quad ; \quad \text{var.}[\varepsilon_i] = \sigma^2 \quad ; \quad \text{i.i.d}$$

However, suppose that we only observe “grouped” data, rather than the observations on the individual agents.

This happens frequently in practice, when data are released in this way to preserve confidentiality.

Suppose there are m groups (*e.g.*, income groups), with n_j observations in the j^{th} group; $j = 1, 2, \dots, m$.

The model we can *actually estimate* is of the form:

$$\bar{y}_j = \beta_1 + \beta_2 \bar{x}_{j2} + \cdots + \beta_k \bar{x}_{jk} + \bar{\varepsilon}_j \quad ; \quad j = 1, 2, \dots, m$$

and clearly,

$$E[\bar{\varepsilon}_j] = E\left[\frac{1}{n_j} \sum_{i=1}^{n_j} \varepsilon_i\right] = \left[\frac{1}{n_j} \sum_{i=1}^{n_j} E(\varepsilon_i)\right] = 0$$

$$\text{var.}[\bar{\varepsilon}_j] = \text{var.}\left[\frac{1}{n_j} \sum_{i=1}^{n_j} \varepsilon_i\right] = \left[\frac{1}{n_j^2} \sum_{i=1}^{n_j} \text{var.}(\varepsilon_i)\right]$$

$$= (n_j \sigma^2 / n_j^2)$$

$$= (\sigma^2 / n_j) .$$

The n_j values are generally reported, so we know the error covariance matrix:

$$\sigma^2 \Omega = \sigma^2 \begin{bmatrix} 1/n_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/n_m \end{bmatrix} .$$

Because Ω is *known*, we can compute the GLS estimator of the coefficient vector immediately:

$$\hat{\beta} = [X' \Omega^{-1} X]^{-1} X' \Omega^{-1} \mathbf{y} .$$

However, in many other applications, we *won't know* the values of the elements of Ω , and we'll have to use **Feasible GLS estimation**.

FGLS Example:

- Estimate β by OLS (b is at least consistent)
- Obtain the OLS residuals, \mathbf{e}
- Estimate Ω by: $\hat{\Omega}_{OLS} = \text{diag}(e_1^2, \dots, e_n^2)$
- Estimate $\hat{\beta}_{FGLS1} = [X' \hat{\Omega}_{OLS}^{-1} X]^{-1} X' \hat{\Omega}_{OLS}^{-1} \mathbf{y}$

The procedure can be iterated, until estimation of $\hat{\Omega}$ converges. Note that the benefit of iterating is questionable, as each estimator for β past the first iteration is consistent.

Testing for Homoskedasticity

- Clearly, it would be very useful to have a test of the hypothesis that the errors in our regression model are *homoscedastic*, against the alternative that they exhibit some sort of *heteroskedasticity*.
- Recall that heteroskedasticity reduces the efficiency of the OLS estimator of β and has serious implications for the properties of the associated standard errors, confidence intervals, and tests.
- Because OLS is still a *consistent* estimator of β even if the errors are heteroskedastic, this means that we can use the OLS residuals to construct tests that will still be (at least) asymptotically valid.
- In particular, we can use these residuals to construct asymptotically valid tests for homoskedasticity.

White's Test

$$y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

$$E[\varepsilon_i] = 0 \quad ; \quad \text{var.}[\varepsilon_i] = \sigma_i^2 \quad ; \quad i.i.d$$

Consider the following null and alternative hypotheses:

$$H_0: \sigma_i^2 = \sigma^2 \quad ; \quad i = 1, 2, \dots, n \quad \quad H_A: \text{Not } H_0$$

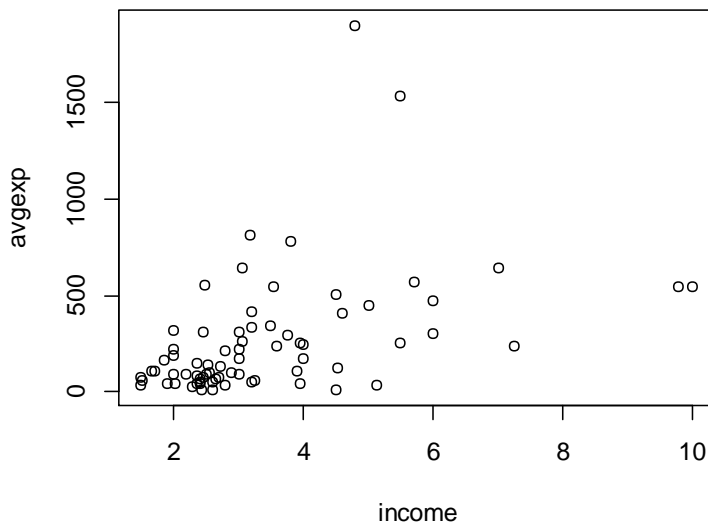
- The Alternative Hypothesis is very general.
- No specific form of heteroskedasticity is declared.
- To implement the test –
 1. Estimate the model by OLS, and get the residuals, $e_i ; i = 1, 2, \dots, n$.
 2. Using OLS, regress the e_i^2 values on each of the x 's in the original model; their squared values; all of the cross-products of the regressors; and an intercept.
 3. nR^2 from the regression in Step 2 is *asymptotically* $\chi_{(p)}^2$ if H_0 is true; where p is the number of parameters that are estimated at Step 2.
 4. Reject H_0 in favour of H_A if $nR^2 > c(\alpha)$.
- Note the *limitations* of this test:
 1. It is valid only asymptotically.
 2. The test is “non-constructive”, in the sense that if we reject H_0 , we don't know what form of heteroskedasticity we may have.
 3. This means that it won't be clear what form the GLS estimator should take.
- However, this may be enough information to alert us to the fact that we should probably use White's “heteroskedasticity-consistent” estimator of $V(\mathbf{b})$.
- In fact, there is little, if anything, to be lost in using this covariance matrix estimator, anyway, as long as the sample is large.

Example

Data is on average monthly credit card expenditure (*avgexp*). The explanatory variables are *age*, *ownrent* (= 1 if homeowner, = 0 if renter), and *income* (in \$10,000). Produce a scatter plot of *avgexp* against *income*.

```
ccard=read.csv("http://home.cc.umanitoba.ca/~godwinrt/7010/creditcard.csv")
attach(ccard)
plot(income, avgexp)
```

Does it look like heteroskedasticity is apparent?



Estimate the following model by OLS:

$$avgexp = \beta_1 + \beta_2 age + \beta_3 ownrent + \beta_4 income + \beta_5 income^2 + \varepsilon$$

```
income2 = income^2
res = lm(avgexp ~ age + ownrent + income + income2)
summary(res)
```

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) | |
|-------------|----------|------------|---------|----------|----|
| (Intercept) | -237.147 | 199.352 | -1.190 | 0.23841 | |
| age | -3.082 | 5.515 | -0.559 | 0.57814 | |
| ownrent | 27.941 | 82.922 | 0.337 | 0.73721 | |
| income | 234.347 | 80.366 | 2.916 | 0.00482 | ** |
| income2 | -14.997 | 7.469 | -2.008 | 0.04870 | * |

White's heteroskedasticity consistent standard errors can be calculated using standard econometric software (e.g. Eviews, Stata). However, we can easily write R code to estimate the appropriate variance-covariance matrix.

Recall that in the presence of heteroskedasticity, White's estimator for the *var-cov* matrix of b is:

$$\hat{V}^* = n[(X'X)^{-1}S_0(X'X)^{-1}]$$

where

$$S_0 = \frac{1}{n} \sum_{i=1}^n e_i^2 \mathbf{x}_i \mathbf{x}_i'$$

To code this into R:

```
resids2 = res$residuals^2
```

Get the squared resid. from 1st regression

```
n = length(avgexp)
```

Read the sample size from the data

```
X = matrix(c(rep(1,n), age, ownrent, income, income2), n, 5)
```

```
S = matrix(0, 5, 5)
```

Create "empty" S matrix

Create X matrix

```
for(i in 1:n){
```

```
S = S + (resids2[i]) * X[i,] %*% t(X[i,])
```

```
}
```

This is a "for" loop. In each iteration, $e_i^2 \mathbf{x}_i \mathbf{x}_i'$ will be added to the S matrix.

```
S = S/n
```

```
diag((n*solve(t(X) %*% X) %*% S %*% solve(t(X) %*% X))^0.5)
```

Finally, this reports the diagonal elements of the \hat{V}^* matrix.

```
212.990530  3.301661  92.187777  88.866352  6.944563
```

How do these compare to the previous standard errors?

White's Heteroskedasticity Test - Example

We'll regress the squared residuals from the OLS regression on all explanatory variables, and squared and cross-products of the explanatory variables. If the R^2 from this auxiliary regression is high enough, we'll reject the null of homoscedasticity.

First, create all the variables needed for the auxiliary regression, then run OLS:

```
age2 = age^2
income4 = income^4
age_own = age*ownrent
age_inc = age*income
age_inc2 = age*income2
own_inc = ownrent*income
own_inc2 = ownrent*income2
inc_inc2 = income^3
summary(lm(resids2 ~ age + ownrent + income + income2 + age2 +
  income4 + age_own + age_inc + age_inc2 + own_inc + own_inc2 +
  inc_inc2))
```

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|------------|------------|---------|----------|
| (Intercept) | 1637390.4 | 1290979.7 | 1.268 | 0.2097 |
| age | 5366.2 | 48893.8 | 0.110 | 0.9130 |
| ownrent | 812036.8 | 991630.2 | 0.819 | 0.4161 |
| income | -2021697.6 | 1053559.1 | -1.919 | 0.0598 . |
| income2 | 669055.3 | 365666.7 | 1.830 | 0.0724 . |
| age2 | -424.1 | 627.5 | -0.676 | 0.5018 |
| income4 | 3762.7 | 2277.4 | 1.652 | 0.1038 |
| age_own | 4661.7 | 14424.6 | 0.323 | 0.7477 |
| age_inc | 11499.9 | 15614.3 | 0.736 | 0.4643 |
| age_inc2 | -1093.3 | 1568.1 | -0.697 | 0.4884 |
| own_inc | -510192.3 | 469792.6 | -1.086 | 0.2819 |
| own_inc2 | 51835.1 | 61799.8 | 0.839 | 0.4050 |
| inc_inc2 | -86805.3 | 51162.6 | -1.697 | 0.0950 . |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 274600 on 59 degrees of freedom

Multiple R-squared: 0.199, Adjusted R-squared: 0.0361

F-statistic: 1.222 on 12 and 59 DF, p-value: 0.2905

- Can variation in $e'e$ be explained?
- Should we use the F-test reported in the regression results?
- ```
> 1 - pchisq(n*0.199,12)
[1] 0.280255
```

So, even though regression seems apparent from the plot of *avgexp* against *income*, we cannot reject the null of homoskedasticity using White's test.

What would be the safe thing to do in this case?



## Topic 7 Continued: Heteroskedasticity

### Goldfeld-Quandt Test

- Suppose that we have two samples of data. That is, we have sampled from two *potentially different* populations.
- We want to test if the variance of the error term for our regression model is the same for both populations.
- We'll assume that we know that the coefficient vector *is the same* for both populations.
- So:

$$\mathbf{y}_1 = X_1\boldsymbol{\beta} + \boldsymbol{\varepsilon}_1 \quad ; \quad \boldsymbol{\varepsilon}_1 \sim N[0, \sigma_1^2 I_{n_1}]$$

$$\mathbf{y}_2 = X_2\boldsymbol{\beta} + \boldsymbol{\varepsilon}_2 \quad ; \quad \boldsymbol{\varepsilon}_2 \sim N[0, \sigma_2^2 I_{n_2}]$$

*(Subscripts denote samples)*

- We want to test  $H_0: \sigma_1^2 = \sigma_2^2$  vs.  $H_A: \sigma_1^2 > \sigma_2^2$  (say)

The Goldfeld-Quandt test for homoscedasticity is constructed as follows:

1. Fit the model, using OLS, over each of the two samples, *separately*.
2. Let the two residual vectors be  $\mathbf{e}_1$  and  $\mathbf{e}_2$ .
3. If the errors are Normally distributed, then the statistics:

$$(\mathbf{e}_i' \mathbf{e}_i) / (\sigma_i^2) \sim \chi_{(n_i - k)}^2 \quad ; \quad i = 1, 2.$$

4. The two regressions are fitted quite separately, so these two statistics are *statistically independent*.
5. Consider the statistic:

$$F = (\mathbf{e}_1' \mathbf{e}_1) / (\sigma_1^2 (n_1 - k)) / (\mathbf{e}_2' \mathbf{e}_2) / (\sigma_2^2 (n_2 - k))$$

6. If  $H_0: \sigma_1^2 = \sigma_2^2$  is true, then  $F = \left( \frac{s_1^2}{s_2^2} \right) \sim F_{(n_1 - k; n_2 - k)}$ .
7. We would **reject  $H_0$**  if  $F > c(\alpha)$ .

- If we *do not reject*  $H_0$ , then we would estimate the (common) coefficient vector,  $\boldsymbol{\beta}$ , by "pooling" both samples together, and applying OLS.
- On the other hand, if we reject  $H_0$ , then we would estimate the (common) coefficient vector,  $\boldsymbol{\beta}$ , by GLS.

- Let's see what form the latter estimator takes in this particular case.
- Recall that we have:

$$\mathbf{y}_1 = X_1\boldsymbol{\beta} + \boldsymbol{\varepsilon}_1 \quad ; \quad \boldsymbol{\varepsilon}_1 \sim N[0, \sigma_1^2 I_{n_1}] \quad (n_1)$$

$$\mathbf{y}_2 = X_2\boldsymbol{\beta} + \boldsymbol{\varepsilon}_2 \quad ; \quad \boldsymbol{\varepsilon}_2 \sim N[0, \sigma_2^2 I_{n_2}] \quad (n_2)$$

- Let  $\phi = (\sigma_1/\sigma_2)$  ; and let  $\hat{\phi} = (s_1/s_2)$  ;  
where  $s_i^2 = (\mathbf{e}_i' \mathbf{e}_i)/(n_i - k)$  ;  $i = 1, 2$ .
- Note that  $\hat{\phi}$  is a *consistent* estimator of  $\phi$ .
- If we knew the value of  $\phi$ , we could use it to scale the model for the second sub-sample, as follows:

$$\phi \mathbf{y}_2 = \phi X_2 \boldsymbol{\beta} + \phi \boldsymbol{\varepsilon}_2 \quad (n_2)$$

where  $E[\phi \boldsymbol{\varepsilon}_2] = 0$  and

$$V[\phi \boldsymbol{\varepsilon}_2] = \phi^2 V[\boldsymbol{\varepsilon}_2] = \left( \frac{\sigma_1^2}{\sigma_2^2} \right) \sigma_2^2 I_{n_2} = \sigma_1^2 I_{n_2}$$

- That is, the full error vector,  $\boldsymbol{\varepsilon}' = (\boldsymbol{\varepsilon}_1', \phi \boldsymbol{\varepsilon}_2')'$ , is *homoscedastic*.
- GLS estimation then amounts to applying OLS to the “pooled” data, but where the data associated with the second sub-sample have been transformed in the above way.
- Typically, we won't know the value of  $\phi = (\sigma_1/\sigma_2)$ , but we can use  $\hat{\phi} = (s_1/s_2)$  instead to implement *feasible* GLS estimation.
- Because  $\hat{\phi}$  is a consistent estimator of  $\phi$ , this feasible GLS estimator will be consistent for  $\boldsymbol{\beta}$ .

### Example

- Investment data for 2 companies – General Electric & Westinghouse
- 20 years of annual data for each company – 1935 to 1954
- I = Gross investment, in 1947 dollars
- V = Market value of company as of 31 December, in 1947 dollars
- K = Stock of plant & equipment, in 1947 dollars
- “Pool” the data – first 20 observations are for General Electric; second 20 observations are for Westinghouse

First, take a look at the data:

```
fglsdata=read.csv("http://home.cc.umanitoba.ca/~godwinrt/7010/fgls.csv")
```

```
attach(fglsdata)
```

```
fglsdata
```

|    | Year | Ige   | Vge    | Kge   | Iw    | Vw     | Kw    |
|----|------|-------|--------|-------|-------|--------|-------|
| 1  | 1935 | 33.1  | 1170.6 | 97.8  | 12.93 | 191.5  | 1.8   |
| 2  | 1936 | 45.0  | 2015.8 | 104.4 | 25.90 | 516.0  | 0.8   |
| 3  | 1937 | 77.2  | 2803.3 | 118.0 | 35.05 | 729.0  | 7.4   |
| 4  | 1938 | 44.6  | 2039.7 | 156.2 | 22.89 | 560.4  | 18.1  |
| 5  | 1939 | 48.1  | 2256.2 | 172.6 | 18.84 | 519.9  | 23.5  |
| 6  | 1940 | 74.4  | 2132.2 | 186.6 | 28.57 | 628.5  | 26.5  |
| 7  | 1941 | 113.0 | 1834.1 | 220.9 | 48.51 | 537.1  | 36.2  |
| 8  | 1942 | 91.9  | 1588.0 | 287.8 | 43.34 | 561.2  | 60.8  |
| 9  | 1943 | 61.3  | 1749.4 | 319.9 | 37.02 | 617.2  | 84.4  |
| 10 | 1944 | 56.8  | 1687.2 | 321.3 | 37.81 | 626.7  | 91.2  |
| 11 | 1945 | 93.6  | 2007.7 | 319.6 | 39.27 | 737.2  | 92.4  |
| 12 | 1946 | 159.9 | 2208.3 | 346.0 | 53.46 | 760.5  | 86.0  |
| 13 | 1947 | 147.2 | 1656.7 | 456.4 | 55.56 | 581.4  | 111.1 |
| 14 | 1947 | 146.3 | 1604.4 | 543.4 | 49.56 | 662.3  | 130.6 |
| 15 | 1949 | 98.3  | 1431.8 | 618.3 | 32.04 | 583.8  | 141.8 |
| 16 | 1950 | 93.5  | 1610.5 | 647.4 | 32.24 | 635.2  | 136.7 |
| 17 | 1951 | 135.2 | 1819.4 | 671.3 | 54.38 | 723.8  | 129.7 |
| 18 | 1952 | 157.3 | 2079.7 | 726.1 | 71.78 | 864.1  | 145.5 |
| 19 | 1953 | 179.5 | 2371.6 | 800.3 | 90.08 | 1193.5 | 174.8 |
| 20 | 1954 | 189.6 | 2759.9 | 888.9 | 68.60 | 1188.9 | 213.5 |

Estimate the “pooled” regression:

$$I = c(I_{ge}, I_w)$$

$$V = c(V_{ge}, V_w)$$

```
K = c(Kge, Kw)
res = lm(I ~ V + K)
summary(res)
```

Coefficients:

|             | Estimate  | Std. Error | t value | Pr(> t )     |
|-------------|-----------|------------|---------|--------------|
| (Intercept) | 17.872001 | 7.024081   | 2.544   | 0.0153 *     |
| V           | 0.015193  | 0.006196   | 2.452   | 0.0191 *     |
| K           | 0.143579  | 0.018601   | 7.719   | 3.19e-09 *** |

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.16 on 37 degrees of freedom  
Multiple R-squared: 0.8098, Adjusted R-squared: 0.7995  
F-statistic: 78.75 on 2 and 37 DF, p-value: 4.641e-14

Perform White's Heteroskedasticity test:

```
resids2 = res$residuals^2
V2 = V^2
K2 = K^2
VK = V*K
summary(lm(resids2 ~ V + K + V2 + K2 + VK))
```

Coefficients:

|             | Estimate   | Std. Error | t value | Pr(> t ) |
|-------------|------------|------------|---------|----------|
| (Intercept) | -1.643e+02 | 4.553e+02  | -0.361  | 0.7204   |
| V           | -1.591e-01 | 1.053e+00  | -0.151  | 0.8808   |
| K           | 5.238e+00  | 2.592e+00  | 2.021   | 0.0512 . |
| V2          | 6.041e-06  | 3.413e-04  | 0.018   | 0.9860   |
| K2          | -8.899e-03 | 3.860e-03  | -2.305  | 0.0274 * |
| VK          | 1.233e-03  | 1.381e-03  | 0.893   | 0.3781   |

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 586.8 on 34 degrees of freedom  
Multiple R-squared: 0.337, Adjusted R-squared: 0.2395  
F-statistic: 3.457 on 5 and 34 DF, p-value: 0.01242

```
1 - pchisq(40*0.337, 5)
0.01927276
```

Now let's try the Goldfeld-Quandt Test:

```
resGE = lm(Ige ~ Vge + Kge)
summary(resGE)
```

Coefficients:

```

 Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.95631 31.37425 -0.317 0.755
Vge 0.02655 0.01557 1.706 0.106
Kge 0.15169 0.02570 5.902 1.74e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Residual standard error: **27.88** on 17 degrees of freedom  
Multiple R-squared: 0.7053, Adjusted R-squared: 0.6706  
F-statistic: 20.34 on 2 and 17 DF, p-value: 3.088e-05

$$\frac{e_1'e_1}{n_1 - k} = 27.88^2 = 777.45$$

```

resW = lm(Iw ~ Vw + Kw)
summary(resW)

```

Coefficients:

```

 Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.50939 8.01529 -0.064 0.95007
Vw 0.05289 0.01571 3.368 0.00365 **
Kw 0.09241 0.05610 1.647 0.11787

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Residual standard error: **10.21** on 17 degrees of freedom  
Multiple R-squared: 0.7444, Adjusted R-squared: 0.7144  
F-statistic: 24.76 on 2 and 17 DF, p-value: 9.196e-06

$$\frac{e_2'e_2}{n_2 - k} = 10.21^2 = 104.31$$

In this example, there is more variability in the error term over the first sub-sample (General Electric) than there is over the second sub-sample (Westinghouse):  $s_1^2 = 777.45$  ;  $s_2^2 = 104.31$

- $H_0: \sigma_1^2 = \sigma_2^2$  vs.  $H_A: \sigma_1^2 > \sigma_2^2$
- $F = (777.45/104.31) = 7.45$
- If  $H_0$  is true,  $F \sim F_{(n_1-k; n_2-k)} = F_{(17; 17)}$
- 5% critical value = 2.4 ; 1% critical value = 3.5
- $1 - \text{pf}(7.45, 17, 17)$   
 $7.172914e-05$
- **Reject  $H_0$**
- So, leave the data for the *first sub-sample unchanged*, but multiply the data (including the intercept) for the *second sub-sample* by  $\hat{\phi} = \frac{s_1}{s_2} = \frac{27.88}{10.21} = 2.73$
- This means that instead of using a constant term in our regression, we must create a vector that consists of 20 1's, followed by 20 values of 2.73 (Cstar), and use this vector as the first term in our regression.

```
Istar = c(Ige, 2.73 * Iw)
Cstar = c(rep(1,20), rep(2.73,20))
Vstar = c(Vge, 2.73 * Vw)
Kstar = c(Kge, 2.73 * Kw)
summary(lm(Istar ~ Cstar + Vstar + Kstar -1))
```

Coefficients:

|       | Estimate  | Std. Error | t value | Pr(> t ) |     |
|-------|-----------|------------|---------|----------|-----|
| Cstar | 16.747017 | 4.785409   | 3.500   | 0.00123  | **  |
| Vstar | 0.020391  | 0.007245   | 2.814   | 0.00778  | **  |
| Kstar | 0.133713  | 0.024144   | 5.538   | 2.65e-06 | *** |

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 29.74 on 37 degrees of freedom  
Multiple R-squared: 0.9436, Adjusted R-squared: 0.939  
F-statistic: 206.3 on 3 and 37 DF, p-value: < 2.2e-16

## Topic 8: Autocorrelated Errors

Consider the standard linear regression model

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad ; \quad \boldsymbol{\varepsilon} \sim N[\mathbf{0}, \sigma^2 I_n]$$

- Among other things, because the off-diagonal elements of  $V(\boldsymbol{\varepsilon})$  are all zero in value, we are assuming that the elements of the error vector are pair-wise *uncorrelated*.
- That is, they do not exhibit any *Autocorrelation*.
- Often, this assumption is unreasonable – especially with *time-series data*.
- Often, current values of the error term are correlated with past values.
- We often say they are “*Serially Correlated*”.
- In this case, the off-diagonal elements of  $V(\boldsymbol{\varepsilon})$  will be non-zero.
- The particular values they take will depend on the *form of autocorrelation*.
- That is, they will depend on the *pattern of the correlations* between the elements of the error vector.

$$V(\boldsymbol{\varepsilon}) = \begin{bmatrix} \sigma^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma^2 \end{bmatrix}$$

- If the *errors* themselves are autocorrelated, often this will be reflected in the regression *residuals* also being autocorrelated.
- That is, the residuals will follow some sort of *pattern*, rather than just being random.
- Typically, this reflects a mis-specification of the model *structure* itself.
- If the errors of our model are autocorrelated, then the OLS estimator of  $\boldsymbol{\beta}$  usually will be unbiased and consistent, but it will be inefficient.
- In addition  $V(\mathbf{b})$  will be computed incorrectly, and the standard errors, *etc.*, will be *inconsistent*.
- So, we need to consider formal methods for
  1. Testing for the presence/absence of autocorrelation.
  2. Estimating models when the errors are autocorrelated.
- It will be helpful to consider various specific forms of autocorrelation that may arise in practice.

- As we'll see, typically we can represent the important forms of autocorrelation with the addition of just a small number of parameters.
- That is,  $V(\boldsymbol{\varepsilon})$  will be a function of  $\sigma^2$ , and just a small number of additional (unknown) parameters.

### Autoregressive Process

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t \quad ; \quad u_t \sim i.i.d.N[0, \sigma_u^2] \quad ; \quad |\rho| < 1$$

This is an AR(1) model for the error process.

More generally:

$$\varepsilon_t = \rho_1\varepsilon_{t-1} + \rho_2\varepsilon_{t-2} + \dots + \rho_p\varepsilon_{t-p} + u_t \quad ; \quad u_t \sim i.i.d.N[0, \sigma_u^2]$$

This is an AR( $p$ ) model for the error process. [e.g.,  $p = 4$  with quarterly data.]

### Moving Average Process

$$\varepsilon_t = u_t + \phi u_{t-1} \quad ; \quad u_t \sim i.i.d.N[0, \sigma_u^2]$$

This is an MA(1) model for the error process.

More generally:

$$\varepsilon_t = u_t + \phi_1\varepsilon_{t-1} + \dots + \phi_q u_{t-q} \quad ; \quad u_t \sim i.i.d.N[0, \sigma_u^2]$$

This is an MA( $q$ ) model for the error process.

We can combine both types of process into an **ARMA( $p, q$ ) model**:

$$\varepsilon_t = \rho_1\varepsilon_{t-1} + \rho_2\varepsilon_{t-2} + \dots + \rho_p\varepsilon_{t-p} + u_t + \phi_1 u_{t-1} + \dots + \phi_q u_{t-q}$$

where:  $u_t \sim i.i.d.N[0, \sigma_u^2]$ .

- Note that in the AR(1) process, we said that  $|\rho| < 1$ .
- This condition is needed to ensure that the process is “stationary”.
- Let's see what this actually means, more generally.
- Note – *all MA processes are stationary*.



## Stationarity

Suppose that the following 3 conditions are satisfied:

1.  $E[\varepsilon_t] = 0$  ; for all  $t$
2.  $var. [\varepsilon_t] = \sigma^2$  ; for all  $t$
3.  $cov. [\varepsilon_t, \varepsilon_s] = \gamma_{|t-s|}$  ; for all  $t, s; t \neq s$

Then we say that the time-series sequence,  $\{\varepsilon_t\}$  is “Covariance Stationary”; or “Weakly Stationary”.

- More generally, this can apply to *any* time-series – not just the error process.
- Unless a time-series is stationary, we can’t identify & estimate the parameters of the process that is generating its values.
- Let’s see how this notion relates to the AR(1) model, introduced above.
- We have:  $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$

$$E[u_t] = 0$$

$$var. [u_t] = E[u_t^2] = \sigma_u^2$$

$$cov. [u_t, u_s] = 0 \quad ; \quad t \neq s$$

- So,

$$\begin{aligned} \varepsilon_t &= \rho[\rho\varepsilon_{t-2} + u_{t-1}] + u_t \\ &= \rho^2\varepsilon_{t-2} + \rho u_{t-1} + u_t \\ &= \rho^2[\rho\varepsilon_{t-3} + u_{t-2}] + \rho u_{t-1} + u_t \\ &= \rho^3\varepsilon_{t-3} + \rho^2 u_{t-2} + \rho u_{t-1} + u_t \\ &\text{etc.} \end{aligned}$$

- Continuing in this way, eventually, we get:

$$\varepsilon_t = u_t + \rho u_{t-1} + \rho^2 u_{t-2} + \dots \quad (1)$$

[This is an infinite-order MA process.]

The value of  $\varepsilon_t$  embodies the entire past history of the  $u_t$  values.

- From (1),  $E(\varepsilon_t) = 0$ , and

$$\begin{aligned} var. (\varepsilon_t) &= var. (u_t) + var. (\rho u_{t-1}) + var. (\rho^2 \varepsilon_{t-2}) + \dots \\ &= \sigma_u^2 + \rho^2 \sigma_u^2 + \rho^4 \sigma_u^2 + \dots \end{aligned}$$

$$= \sigma_u^2 \sum_{s=0}^{\infty} \rho^{2s} = \sigma_u^2 \sum_{s=0}^{\infty} (\rho^2)^s$$

- Now, under what conditions will this series converge?

The series will converge to  $\sigma_u^2(1 - \rho^2)^{-1}$ , as long as  $|\rho^2| < 1$ , and this in turn requires that  $|\rho| < 1$ .

- This is a necessary condition needed to ensure that the process,  $\{\varepsilon_t\}$  is stationary, because if this condition isn't satisfied, then  $var. [\varepsilon_t]$  is *infinite*.
- So, for the AR(1) process, as long as  $|\rho| < 1$ , then  $var. [\varepsilon_t] = \sigma_u^2(1 - \rho^2)^{-1}$ .
- In addition, stationarity implies that  $var. [\varepsilon_t] = var. [\varepsilon_{t-s}]$ , for all 's'.
- So, now consider the covariances of terms in the process:

$$\begin{aligned} cov. [\varepsilon_t, \varepsilon_{t-1}] &= E[(\varepsilon_t - E(\varepsilon_t))(\varepsilon_{t-1} - E(\varepsilon_{t-1}))] \\ &= E[\varepsilon_t \varepsilon_{t-1}] \\ &= E[\varepsilon_{t-1}(\rho \varepsilon_{t-1} + u_t)] \\ &= \rho E[\varepsilon_{t-1}^2] + 0 \\ &= \rho var. [\varepsilon_{t-1}] = \rho \sigma_u^2 / (1 - \rho^2) \end{aligned}$$

- Similarly,

$$\begin{aligned} cov. [\varepsilon_t, \varepsilon_{t-2}] &= E[(\varepsilon_t - E(\varepsilon_t))(\varepsilon_{t-2} - E(\varepsilon_{t-2}))] \\ &= E[\varepsilon_{t-2}(\rho \varepsilon_{t-1} + u_t)] \\ &= E[\varepsilon_{t-2}(\rho(\rho \varepsilon_{t-2} + u_{t-1}) + u_t)] \\ &= \rho^2 E[\varepsilon_{t-2}^2] + 0 \\ &= \rho^2 var. [\varepsilon_{t-2}] = \rho^2 \sigma_u^2 / (1 - \rho^2) \end{aligned}$$

- In general, then, for the AR(1) process:

$cov. [\varepsilon_t, \varepsilon_s] = \rho^{(t-s)} \sigma_u^2 / (1 - \rho^2)$  ; depends on  $(t - s)$ , not values of  $t, s$  ; and we can reverse  $t$  and  $s$ , so it actually depends on  $|t - s|$  .

- Also, recall that

$$var. [\varepsilon_t] = \sigma_u^2 / (1 - \rho^2)$$

- So, the full covariance matrix for  $\varepsilon$  is:

$$V(\boldsymbol{\varepsilon}) = \sigma_u^2 \Omega = \frac{\sigma_u^2}{(1 - \rho^2)} \begin{bmatrix} 1 & \rho & \dots & \rho^{n-1} \\ \rho & 1 & \ddots & \rho^{n-2} \\ \vdots & \ddots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \dots & 1 \end{bmatrix}$$

If we can find a matrix,  $P$ , such that  $\Omega^{-1} = P'P$ , and if the value of  $\rho$  were *known*, then we could apply GLS estimation.

- More likely, in practice, find  $P$ , which will depend on  $\rho$ , and then estimate  $\rho$  consistently, and we can implement *feasible* GLS estimation.
- Before we consider GLS estimation any further, let's first see what implications autocorrelation of the errors has for the OLS estimator of  $\boldsymbol{\beta}$ .

### OLS Estimation

- Given that the error term in our model now has a non-scalar covariance matrix, we know that the OLS estimator,  $\mathbf{b}$ , is still linear and unbiased, but it is *inefficient*.
- In general,  $\mathbf{b}$  will still be a consistent estimator. However, there is one important situation where it will be *inconsistent*.
- This will be the case if the errors are autocorrelated, *and* one or more lagged values of the dependent variable enter the model as regressors.  
[The GLS estimator will also be inconsistent in this case.]
- A quick way to observe that inconsistent estimation will result in this case is as follows:

- Suppose that

$$\begin{aligned} y_t &= \beta y_{t-1} + \varepsilon_t & ; & \quad |\beta| < 1 & \quad (2) \\ \varepsilon_t &= \rho \varepsilon_{t-1} + u_t & ; & \quad u_t \sim i.i.d. [0, \sigma_u^2] & \quad ; |\rho| < 1 \end{aligned}$$

Now subtract  $\rho y_{t-1}$  from the expression for  $y_t$  in equation (2):

$$(y_t - \rho y_{t-1}) = (\beta y_{t-1} + \varepsilon_t) - \rho(\beta y_{t-2} + \varepsilon_{t-1})$$

or,

$$\begin{aligned}
 y_t &= (\beta + \rho)y_{t-1} - \beta\rho y_{t-2} + (\varepsilon_t - \rho\varepsilon_{t-1}) \\
 &= (\beta + \rho)y_{t-1} - \beta\rho y_{t-2} + u_t
 \end{aligned}$$

- So, if we estimate the model with just  $y_{t-1}$  as the only regressor, then we are effectively omitting a relevant regressor,  $y_{t-2}$ , from the model.
- This amounts to imposing a false (zero) restriction on the coefficient vector, and we know that this causes OLS to be not only **biased**, but also **inconsistent**.
- As was noted when we were discussing the general situation involving a regression model whose error vector has a non-scalar covariance matrix (in Topic 6), the estimated  $V(\mathbf{b})$  will be **inconsistent**, regardless of the form of the regressors.
- So, to get consistent standard errors for the elements of  $\mathbf{b}$ , we can use the Newey-West correction when estimating  $V(\mathbf{b})$ .

### Testing for Serial Independence

- Let's consider the problem of testing the hypothesis,  $H_0$ : "The errors in our regression model are serially independent".
- We'll need to formulate both the null, and an alternative hypothesis, expressing them in terms of the underlying parameters of the model.
- First, consider the possibility that the errors follow an AR(1) process, if they are not serially independent.
- That is:

$$y_t = \mathbf{x}'_t \boldsymbol{\beta} + \varepsilon_t \quad ; \quad t = 1, 2, \dots, n \quad (3)$$

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t \quad ; \quad u_t \sim i.i.d. [0, \sigma_u^2] \quad ; \quad |\rho| < 1$$

Then, we have  $H_0: \rho = 0$  vs.  $H_A: \rho \neq 0$  ( $> 0$  ;  $< 0$  )

- Notice that, as usual, we can learn something about the behaviour of the **errors** in our regression model by looking at the **residuals** obtained when we estimate the model.
- So, estimate (3) by OLS (ignoring any possibility of serial correlation), and get the residuals,  $\{e_t\}$ .

- Then, fit the following “auxiliary regression”:

$$e_t = r e_{t-1} + v_t \quad ; \quad t = 2, 3, \dots, n$$

- The OLS estimator of the coefficient, “ $r$ ”, is:

$$\hat{r} = \left[ \sum_{t=2}^n e_t e_{t-1} \right] / \left[ \sum_{t=2}^n e_{t-1}^2 \right]$$

- We could think of using a “z-test” to test if  $r = 0$ . This test will be valid, *asymptotically*:

$$z = \frac{(\hat{r} - 0)}{s.e.(\hat{r})} \xrightarrow{d} N[0, 1]$$

- Now, testing for serial independence, against the alternative hypothesis that the process is AR(1) is very interesting.
- Anderson (1948) proved that **there does not exist any UMP test** for this problem!
- So, historically, there were lots of attempts to construct tests that were “approximately” most powerful.
- These days we generally use tests from the so-called “**Lagrange Multiplier Test**” family. Also called the family of “**Score Tests**”.
- Tests of this type can be used for all sorts of testing problems – not just for testing for serial independence.
- They are especially useful when it is relatively easy to estimate the model under the assumption that the null hypothesis is true.
- Here, such estimation involves just OLS.
- LM tests have only *asymptotic validity*. Asymptotically, the distribution of the test statistic is Chi-Square, with d.o.f. equal to the number of restrictions being tested, if the null hypothesis is true.
- The pay-off is that the test can be applied under *very general conditions*.
- We don’t need to have normally distributed errors in our regression model.
- The regressors can be random; *etc.*

- The Breusch-Godfrey Test for serial independence of the errors can be implemented as follows:
  1. Estimate the model,  $y_t = \mathbf{x}'_t \boldsymbol{\beta} + \varepsilon_t$  ;  $t = 1, 2, \dots, n$  by OLS, and get the residuals  $\{e_t\}$ .
  2. If the Alternative Hypothesis is that the errors follow *either* an AR( $p$ ) process, *or* an MA( $p$ ) process, then estimate the following auxiliary regression:
 
$$e_t = \mathbf{x}'_t \boldsymbol{\gamma} + \delta_1 e_{t-1} + \dots + \delta_p e_{t-p} + v_t \quad (4)$$
  3. The test statistic is  $LM = nR^2$ , where  $R^2$  is the “uncentered” coefficient of determination from (4).
  4. Reject  $H_0 : \varepsilon_t$  *serially independent*; if  $LM > \chi^2_{(p)}$  critical value.
- If we reject  $H_0$ , we're left with *incomplete information* about the particular form of the autocorrelation.

### Estimation Allowing for Autocorrelation

- Suppose we have a regression model with AR(1) errors:

$$y_t = \mathbf{x}'_t \boldsymbol{\beta} + \varepsilon_t \quad ; \quad t = 1, 2, \dots, n$$

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t \quad ; \quad u_t \sim i.i.d. [0, \sigma_u^2] \quad ; \quad |\rho| < 1$$

- So, the full covariance matrix for  $\boldsymbol{\varepsilon}$  is:

$$V(\boldsymbol{\varepsilon}) = \sigma_u^2 \Omega = \frac{\sigma_u^2}{(1 - \rho^2)} \begin{bmatrix} 1 & \rho & \dots & \rho^{n-1} \\ \rho & 1 & \ddots & \rho^{n-2} \\ \vdots & \ddots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \dots & 1 \end{bmatrix}$$

- We need to find a matrix,  $P$ , such that  $\Omega^{-1} = P'P$ , and then we can apply GLS estimation.
- In the AR(1) case, we can show that:

$$P = \begin{bmatrix} \sqrt{1-\rho^2} & 0 & 0 & 0 & \cdots & 0 \\ -\rho & 1 & 0 & 0 & \cdots & 0 \\ 0 & -\rho & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & \cdots & & & -\rho & 1 \end{bmatrix}$$

- GLS is simply OLS, using the data  $\mathbf{y}^*$  and  $X^*$ , where:

$$\mathbf{y}^* = \begin{bmatrix} y_1\sqrt{1-\rho^2} \\ y_2 - \rho y_1 \\ \vdots \\ y_n - \rho y_{n-1} \end{bmatrix} ; \quad \mathbf{x}_j^* = \begin{bmatrix} x_{1j}\sqrt{1-\rho^2} \\ x_{2j} - \rho x_{1j} \\ \vdots \\ x_{nj} - \rho x_{n-1,j} \end{bmatrix} ; \quad j = 1, 2, \dots, k$$

- What if  $\rho$  is unknown, as is likely to be the case?
- We can apply feasible GLS – this is essentially what Cochrane & Orcutt (1949) did, except that they “dropped” the first observation as they didn’t know the leading (1, 1) element of the  $P$  matrix.

- The steps are:

1. Estimate the model,  $y_t = \mathbf{x}'_t \boldsymbol{\beta} + \varepsilon_t$ , by OLS and get the residuals,  $\{e_t\}$ .
2. Estimate  $\rho$ , using

$$\hat{\rho} = \left[ \sum_{t=2}^n e_t e_{t-1} \right] / \left[ \sum_{t=2}^n e_{t-1}^2 \right]$$

3. Construct  $\mathbf{y}^*$  and  $X^*$ , using  $\hat{\rho}$  in place of  $\rho$ .
  4. Apply OLS using the transformed data. This is **feasible GLS** estimation.
  5. Iterate Steps 1 through 4.
  6. Continue until convergence is achieved.
- Convergence is guaranteed in a *finite number of steps*, unless the model includes lagged values of the dependent variable.
  - The same approach can be used if the errors follow a (“simple”) AR( $p$ ) process:  $\varepsilon_t = \rho \varepsilon_{t-p} + u_t$  ;  $u_t \sim i.i.d. [0, \sigma_u^2]$
  - Things are more complicated if the errors follow an MA( $q$ ) or ARMA( $p, q$ ) process

## Topic 9: Maximum Likelihood Estimation

There are many other estimation methodologies besides OLS. For example: GMM, Bayesian, non-parametric, and maximum likelihood (ML). In some of these methodologies, the OLS estimator is just a special case.

- ML proposed by R. A. Fisher, 1921-1925
- MLE is a parametric method.
- That is, we assume each sample data is generated from a known probability distribution function (p.d.f.),  $p(y_i|\boldsymbol{\theta})$ . i.e.  $y_i$  comes from a “family”.

Consider:

|                  |                                                      |
|------------------|------------------------------------------------------|
| Random data      | $\mathbf{y} = \{y_1, \dots, y_n\}$                   |
| Parameter vector | $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)'$ |

Objective: estimate  $\boldsymbol{\theta}$ .

The probability of jointly observing the data is

$$p(y_1, \dots, y_n|\boldsymbol{\theta}) \quad \text{“joint p.d.f.”}$$

We can view  $p(y_1, \dots, y_n|\boldsymbol{\theta})$  in two different ways:

- i. As a function of  $\{y_1, \dots, y_n\}$ , given  $\boldsymbol{\theta}$ .
- ii. As a function of  $(\theta_1, \dots, \theta_k)$ , given  $\mathbf{y}$ . i.e., the **data** is *given*, the **parameters** *vary*.

The latter is called the **likelihood function**.

$$\text{Note: } L(\boldsymbol{\theta}) = L(\boldsymbol{\theta}|y_1, \dots, y_n) = p(y_1, \dots, y_n|\boldsymbol{\theta})$$

**Definition:** The Maximum Likelihood Estimator (MLE) of  $\boldsymbol{\theta}$  (say,  $\tilde{\boldsymbol{\theta}}$ ) is that value of  $\boldsymbol{\theta}$  such that  $L(\tilde{\boldsymbol{\theta}}) > L(\hat{\boldsymbol{\theta}})$ , for all other  $\hat{\boldsymbol{\theta}}$ .

**Idea:** “given the  $y_i$ ’s, what is the most likely  $\boldsymbol{\theta}$  to have generated such a sample?”

Note:

- i.  $\tilde{\boldsymbol{\theta}}$  need not be unique.



- ii.  $\tilde{\theta}$  should locate the global max. of  $L(\theta)$ .
- iii. If the sample data are independent then  $L(\theta|\mathbf{y}) = p(\mathbf{y}|\theta) = \prod_{i=1}^n p(y_i|\theta)$
- iv. Any monotonic transformation of  $L(\theta)$  leaves location of extremum unchanged (e.g.  $\log L(\theta)$ )

### Some Basic Concepts and Notation:

- i. “Gradient/Score Vector”:  $\left[ \frac{\partial \log L(\theta)}{\partial \theta} \right] \quad (k \times 1)$
- ii. “Hessian Matrix”:  $\left[ \frac{\partial^2 \log L(\theta)}{\partial \theta \partial \theta'} \right] \quad (k \times k)$
- iii. “Likelihood Equations”:  $\frac{\partial \log L(\theta)}{\partial \theta} = 0 \quad (k \times 1)$

The optimization problem is:

$$\max_{\theta} \prod_{i=1}^n L(\theta|y_i).$$

So, to obtain the MLE,  $\tilde{\theta}$ , we solve the likelihood equation(s) and then check the second-order condition(s) to make sure we have maximized (not minimized)  $L(\theta)$ . If the Hessian matrix is at least n.s.d., then  $\log L(\theta)$  is concave, and this is sufficient for a maximum.

So, MLE is accomplished by:

- 1) Specifying the likelihood function.
  - This involves writing down an equation which states the joint likelihood (or joint probability) of observing the sample data, conditional on the unknown parameter values of the probability distribution function.
  - Independence of the  $y$  data is usually assumed (and will be for the purposes of this course).
  - Given independence, the likelihood function is obtained by multiplying together the probability of each  $y_i$  occurring.
- 2) Taking the natural log of the likelihood function. This usually simplifies the next step. The location of the maximum will not change.

- 3) Taking the first derivative of the log-likelihood function with respect to all parameters, setting each derivative equal to zero, and solving for the parameter values. The solution of the FOCs provides the formulas for the MLEs.
- 4) Checking to make sure the estimator in (3) attains a maximum (not a minimum). This involves taking the second derivatives of the log-likelihood function with respect to all parameters, so as to construct the Hessian matrix. If the Hessian is *n.s.d.*, then the MLE achieves a global max.
- 5) Obtaining the variance of the MLEs for use in hypothesis testing. A variance-covariance matrix can be found by inverting the negative of the expected Hessian.

### Properties of MLE

- MLE has very desirable asymptotic properties.
- Namely, MLE is Best Asymptotically Normal.
- That is, under mild assumptions, ML estimators are consistent, asymptotically efficient, and asymptotically Normally distributed.
- These properties are obtained by examining the asymptotic distribution of the MLE (which we will not derive in class):

$$\sqrt{n}(\tilde{\theta} - \theta_0) \xrightarrow{d} N[0, IA^{-1}(\theta)],$$

where

$$IA^{-1}(\theta) = \lim_{n \rightarrow \infty} \left( \frac{1}{n} [-E[H(\theta)]]^{-1} \right)$$

- $IA^{-1}(\theta)$  is the asymptotic information matrix, and  $H(\theta)$  is the Hessian.
- The statement of the asymptotic distribution shows that the MLEs are consistent, asymptotically normal, and asymptotically efficient.
- The efficiency result relies on the Cramer-Rao lower bound. The Cramer-Rao lower bound is a theoretical minimum variance that any estimator can obtain. The MLE attains this minimum, that is,  $IA^{-1}(\theta)$  is equal to the asymptotic Cramer-Rao lower bound.

The asymptotic distribution also allows us to see the variance of the MLEs in finite samples. The variance-covariance of  $\tilde{\theta}$  for finite samples can be solved from the asymptotic variance:

$$\text{var}[\sqrt{n}(\tilde{\theta})] = n \times \text{var}(\tilde{\theta}) = \frac{1}{n} [-E[H(\theta)]]^{-1}, \text{ so}$$

$$\text{var}(\tilde{\theta}) = [-E[H(\theta)]]^{-1}.$$

The matrix  $-E[H]$  is termed the “Information Matrix” and is denoted by  $I(\theta)$ .

A very useful property of MLEs is their “invariance.” That is, the estimator for  $g(\theta)$  is  $g(\tilde{\theta})$ .

Hence, an estimator for the variance-covariance of  $\tilde{\theta}$  is:

$$\widehat{\text{var}}(\tilde{\theta}) = [-E[H(\tilde{\theta})]]^{-1}.$$

Note that if misspecification occurs (if we have selected the wrong probability density function to begin with), we are not assured of any of the asymptotic properties.

### **Finite sample properties of MLEs**

MLEs can be biased in finite samples (and typically are). We can evaluate bias much like we have done in previous parts of the course; by taking  $E(\tilde{\theta})$ . This knowledge can be used to correct for any bias (as in the case of  $\tilde{\sigma}^2$ ). However, in most cases, there is no closed-form solution for the MLE itself, and numerical methods must be used to solve for the estimate. When the estimator does not have a closed form solution, we cannot take  $E(\tilde{\theta})$ , and we will not be able to “see” whether or not the estimator is biased. In this case, approximations or Monte Carlo experiments may be used to evaluate bias.