Topic 1: Basic Multiple Regression

Population "model" –

$$\boldsymbol{y} = f(x_1, x_2, \ldots., x_k\,;\, \boldsymbol{\theta}) + \boldsymbol{\varepsilon}$$

Dependent variable ("regressand")

Explanatory variables ("regressors")

Parameter vector

Disturbance term (random "error")

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:

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

2) Cobb-Douglas production function:

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

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

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

3) CES production function

$Y = \varphi(aK^r + (1-a)L^r)^{1/r}e^{\varepsilon}$ (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: $\{y_i ; x_{i1}, x_{i2}, \ldots, x_{ik}\}$; $i = 1, 2, \ldots, n$
- 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*:

$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i \;\; ; \;\; i = 1, \ldots, n$ (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:

$\frac{\partial y_i}{\partial x_{1i}} = \beta_1$; 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:

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

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

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

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

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

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

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

Notational points

i. Vectors are in bold.
ii. The dimensions of vectors/matrices are written (*rows* × *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 $y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i$.

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 $Rank(X) = k$. So we are also implicitly assuming that $n > k$, since $Rank(A) \leq min.\{\#rows, \; \#cols\}$.

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*, $E(\varepsilon_i) = 0$; $i = 1, 2, \ldots, n$. 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:

$var(\varepsilon_i) = \sigma^2 \ ; \ i = 1,2, \ldots, n$ (Homoskedasticity)

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

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

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

but...

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

and

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

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

A.6: Normality of Errors

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

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:

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

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

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

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

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

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,

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

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:

$$y_i = \boldsymbol{x_i'\beta} + \varepsilon_i \quad ,$$

and the population regression is:

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

We'll estimate $E(y_i \mid \boldsymbol{x'}_i)$ by

$$\hat{y}_i = \boldsymbol{x_i'b}.$$

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

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

So,

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

unobserved — observed

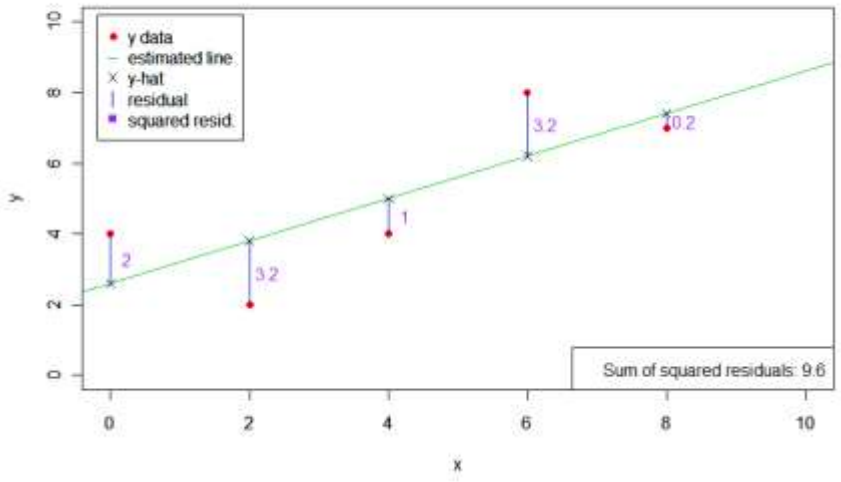
[*Population*] — [*Sample*]

The Least Squares Criterion:

"Choose $\boldsymbol{b}$ 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

$$Min._{(\boldsymbol{b})} \sum_{i=1}^{n} e_i^2 \quad \Leftrightarrow \quad Min._{(b)} \ (\boldsymbol{e}'\boldsymbol{e})$$

$$\Leftrightarrow \quad Min._{(\boldsymbol{b})} \ [(\boldsymbol{y} - X\boldsymbol{b})'(\boldsymbol{y} - X\boldsymbol{b})].$$

Now, let:

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

Note that,

$$\underset{(1\times k)(k\times n)(n\times 1)}{\boldsymbol{b}'X'\boldsymbol{y}} \quad = \quad \underset{(1\times 1)}{\boldsymbol{y}'X\boldsymbol{b}}.$$

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

Note:

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

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

Applying these 2 results –

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

Set this to zero (*for a turning point*):

$$\underset{(k\times n)(n\times k)(k\times 1)}{X'X\boldsymbol{b}} \quad = \quad \underset{(k\times n)(n\times 1)}{X'\boldsymbol{y}} \ , \qquad (k \text{ equations in } k \text{ unknowns})$$

(the "normal equations")

so:

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

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

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

We need the "full rank" assumption for the Least Squares estimator, $\boldsymbol{b}$, to *exist*.

None of our other assumptions have been used so far.

Check – have we *minimized* S ?

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

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

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

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, $\boldsymbol{b}$, *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:

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

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

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 –

$X'X\boldsymbol{b} - X'\boldsymbol{y} = \boldsymbol{0}$ ("normal equations"; (k×1))

So,

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

or,

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

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

$$X'\boldsymbol{e} = \begin{pmatrix} 1 & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ 1 & \cdots & x_{nk} \end{pmatrix}' \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} = \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ x_{1k} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$$

$$= \begin{pmatrix} \sum_i e_i \\ ? \\ ? \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

From the first element: $\sum_{i=1}^{n} e_i = 0$

2. Fitted regression passes through sample mean

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

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

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

From the first row of this vector equation –

$$\sum_i y_i = nb_1 + b_2 \sum_i x_{i2} + \cdots + b_k \sum_i x_{ik}$$

or, $$\bar{y} = b_1 + b_2\overline{x_2} + \cdots + b_k\overline{x_k}$$

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

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

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 –

$$\underset{(n\times 1)}{\boldsymbol{y}} = \underset{(n\times k_1)(k_1\times 1)}{X_1\boldsymbol{\beta}_1} + \underset{(n\times k_2)(k_2\times 1)}{X_2\boldsymbol{\beta}_2} + \underset{(n\times 1)}{\boldsymbol{\varepsilon}}$$

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:

$$\underset{(n\times 1)}{\boldsymbol{y}} = \underset{(n\times (k_1+k_2))}{[X_1 : X_2]} \underset{((k_1+k_2)\times 1)}{\begin{bmatrix}\boldsymbol{\beta_1}\\ \boldsymbol{\beta_2}\end{bmatrix}} + \underset{(n\times 1)}{\boldsymbol{\varepsilon}} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}\ ,$$

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

We can write this LS estimator as:

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

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

So,

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

The "normal equations" underlying this are –

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

or:

$$\begin{bmatrix}X_1'X_1 & X_1'X_2\\ X_2'X_1 & X_2'X_2\end{bmatrix}\begin{pmatrix}\boldsymbol{b_1}\\ \boldsymbol{b_2}\end{pmatrix} = \begin{pmatrix}X_1'\boldsymbol{y}\\ X_2'\boldsymbol{y}\end{pmatrix}\ .$$

Let's solve these "normal equations" for $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$:

$$X_1'X_1\boldsymbol{b_1} + X_1'X_2\boldsymbol{b_2} = X_1'\boldsymbol{y} \qquad [1]$$

$$X_2'X_1\boldsymbol{b_1} + X_2'X_2\boldsymbol{b_2} = X_2'\boldsymbol{y} \qquad [2]$$

From [1]:

$$(X_1'X_1)\boldsymbol{b_1} = X_1'\boldsymbol{y} - X_1'X_2\boldsymbol{b_2} \ ,$$

or, $$\boldsymbol{b_1} = (X_1'X_1)^{-1}X_1'\boldsymbol{y} - (X_1'X_1)^{-1}X_1'X_2\boldsymbol{b_2}$$

$$= (X_1'X_1)^{-1}[X_1'\boldsymbol{y} - X_1'X_2\boldsymbol{b_2}] \qquad [3]$$

Note: If $X_1'X_2 = 0$, then $\boldsymbol{b_1} = (X_1'X_1)^{-1}X_1'\boldsymbol{y}$.

(*Why do the "partial" and "full" regression estimators coincide in this case?*)

Now substitute [3] into [2]:

$$(X_2'X_1)[(X_1'X_1)^{-1}X_1'\boldsymbol{y} - (X_1'X_1)^{-1}X_1'X_2\boldsymbol{b_2}] + (X_2'X_2)\boldsymbol{b_2} = X_2'\boldsymbol{y} \ ,$$

or,

$$[(X_2'X_2) - (X_2'X_1)(X_1'X_1)^{-1}(X_1'X_2)]\boldsymbol{b_2} = X_2'\boldsymbol{y} - (X_2'X_1)(X_1'X_1)^{-1}X_1'\boldsymbol{y} \ ,$$

and so:

$$\boldsymbol{b_2} = [(X_2'X_2) - (X_2'X_1)(X_1'X_1)^{-1}(X_1'X_2)]^{-1}[X_2'(I - X_1(X_1'X_1)^{-1}X_1')\boldsymbol{y}].$$

Define:

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

Then, we can write –

$$\boldsymbol{b_2} = (X_2'M_1X_2)^{-1}X_2'M_1\boldsymbol{y}$$

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

$$\boldsymbol{b_1} = (X_1'M_2X_1)^{-1}X_1'M_2\boldsymbol{y}$$

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

- M_1 and M_2 are "*idempotent*" matrices
- $M_iM_i = M_iM_i' = M_i = M_i'M_i \quad ; \qquad i = 1, 2.$

So, finally, we can write:

$$\boldsymbol{b_1} = (X_1^{*\prime}X_1^*)^{-1}X_1^{*\prime}\boldsymbol{y_1^*} \qquad \boldsymbol{b_2} = (X_2^{*\prime}X_2^*)^{-1}X_2^{*\prime}\boldsymbol{y_2^*}$$

where:

$$X_1^* = M_2 X_1 \;;\; X_2^* = M_1 X_2 \;;\; \boldsymbol{y_1^*} = M_2 \boldsymbol{y} \;\;;\;\; \boldsymbol{y_2^*} = M_1 \boldsymbol{y}$$

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 $\boldsymbol{y}$.
- Sample variance of $\boldsymbol{y}$ 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

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

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

- 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:

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

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

Note that:

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

Let's check the third of these results:

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

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

Returning to our "fitted" model:

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

So, we have:

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

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

Then –

$$\boldsymbol{y}'M^0\boldsymbol{y} = \boldsymbol{y}'M^{0\prime}M^0\boldsymbol{y} = (M^0\hat{\boldsymbol{y}} + \boldsymbol{e})'(M^0\hat{\boldsymbol{y}} + \boldsymbol{e})$$

$$= \hat{\boldsymbol{y}}'M^0\hat{\boldsymbol{y}} + \boldsymbol{e}'\boldsymbol{e} + 2\boldsymbol{e}'M^0\hat{\boldsymbol{y}}$$

However,

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

So, we have –

$$\boldsymbol{y}'M^0\boldsymbol{y} = \hat{\boldsymbol{y}}'M^0\hat{\boldsymbol{y}} + \boldsymbol{e}'\boldsymbol{e}$$

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

$$\textbf{SST} = \textbf{SSR} + \textbf{SSE}$$

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?

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

Then:

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

(A) Applying LS to [2]:

$$min.(\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}) \quad ; \quad \hat{\boldsymbol{u}} = \boldsymbol{y} - X_1\boldsymbol{b_1} - X_2\boldsymbol{b_2}$$

(B) Applying LS to [1]:

$$min.(\boldsymbol{e}'\boldsymbol{e}) \quad ; \quad \boldsymbol{e} = \boldsymbol{y} - X_1\hat{\boldsymbol{\beta}}_{\boldsymbol{1}}$$

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, $\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}} \leq \boldsymbol{e}'\boldsymbol{e}$.

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

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

- 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.

- $\boldsymbol{e = y - Xb}$. 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 ≠ 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) $\widehat{C}_i = 0.5 + 0.8Y_i$; $R^2 = 0.90$

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

  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)))
```