Introduction to Econometrics

ECON 7010 Lecture Notes

Ryan T. Godwin[1]

University of Manitoba

[1]email: ryan.godwin@umanitoba.ca

Contents

1

Introduction

1.1 What is Econometrics?

Econometrics is the study of statistical methods applied to economics data. It is a subset of statistics, but Econometrics places more emphasis on methods most suited to economics data, and (Microeconometrics) specializes in dealing with *observational data*, as opposed to *experimental data*. In an experiment, there is some element of *control* - a variable can be changed by the researcher, and the effect of the change on another variable can be more easily measured. In observational data the causal variable is changing on its own, and this can be very problematic. Typically there are important *omitted variables* in observational data, and it is very important to consider the relationship of these omitted variables with the ones included in the model.

Econometrics can be used to estimate *causal effects*, though it should not be used to find them. That is, the theoretical model (e.g. from Micro or Macro) should specify which variable causes which. Typically the goal is to estimate *how much* of an effect a variable has on another, and to test hypothesis concerning the magnitudes of this effect. Econometrics may also be used to forecast or predict economic variables, although this is a topic for *time series* which is only touched on in this course.

1.2 Limitations of Econometrics

It is important to be aware of the limitations of Econometrics. It cannot be used to determine *causation*. Causation must be theorized. If two variables are correlated, Econometrics alone cannot tell which variable causes which, or if there is any causation at all. That is, *correlation does not imply causation*. If, however, we find that two variables are statistically independent from each other, one variable can not cause the other.

1.3 Objective

The main objective of the course is to give you the tools necessary to assess the merits of an econometric model.

2

Basic Multiple Regression

The population model is:

$$\boldsymbol{y} = f\left(x_1, x_2, \ldots, x_k; \boldsymbol{\theta}\right) + \boldsymbol{\varepsilon} \tag{2.1}$$

- $\boldsymbol{y}$ is the dependent variable or "regressand"
- $x_1, x_2, \ldots, x_k$ are the explanatory variables or "regressors"
- $\boldsymbol{\theta}$ is a parameter vector
- $\boldsymbol{\varepsilon}$ is the disturbance term or the random "error"

We'll focus on population models where f is parametric and (usually) linear in the parameters. The first estimation strategy that we'll consider, *Ordinary Least Squares*, requires that the model is linear in the parameters. In general, however, f may be:

- linear or non-linear in the variables
- linear or non-linear in the parameters
- parametric or non-parametric

Questions:

1. What is the role of the error term?
2. What is random, and what is deterministic?
3. What is observable, and what is unobservable?

2.1 Examples of some old population models in economics

You do not have to learn about these economic models. These models are provided in order to relate them to the general population model in equation 2.1. For each model, try to determine the components of the model, and whether or not it is linear/non-linear in the regressors/parameters.

Keynes' consumption function

$$C = \beta_1 + \beta_2 Y + \varepsilon$$

Cobb-Douglas production function

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

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 $\beta_1 = \log A$.

CES production function

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

Taking logs, the CES production function is written as:

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

2.2 Sample information

Suppose that we have a *sample* of n observations:

$$\{y_i; x_{i1}, x_{i2}, \ldots, x_{ik}\}; \quad i = 1, 2, \ldots, n$$

Assuming that the observed values are generated by the population model, and taking the case where the model is *linear in the parameters*, we have:

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i \quad ; \quad i = 1, \ldots, n \tag{2.2}$$

Recall that the βs and ε are unobservable. So, y_i is generated by two components:

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

So, the y_i must be "realized values" of a random variable!

Some typical objectives are to:

(i) Estimate unknown parameters

(ii) Test hypotheses about parameters

(iii) Predict values of y outside sample

2.3 Interpreting the parameters in a model

Once we estimate $\boldsymbol{\theta}$ (i.e. all the βs), how do we interpret them? A major advantage of the linear model is the ease in which the parameters may be interpreted. That is, the βs in equation 2.2 have an important economics interpretation. For example:

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

The parameters are the marginal effects of the x on y, with other factors held constant (ceteris paribus). For example, from Keynes' consumption function:

$$\frac{\partial C}{\partial Y} = \beta_2 = \text{Marginal Propensity to Consume}$$

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 might not be interpreted as marginal effects. For example, after taking logs of the Cobb-Douglas production function in, 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 in all but the most simple of cases.

Question: 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).

2.4 Assumptions of the Classical Linear Regression Model

In this section, we are going to state six "classical" assumptions, and refer back to them frequently throughout the course. These simplifying assumptions are a starting point, and are likely not satisfied in real data. One of the main objectives of this course is to re-consider these assumptions - are they realistic; can they be tested; what if they are wrong; can they be "relaxed"? When these assumptions are violated, and we consider how to fix the resulting consequence, we will be led to different *estimation strategies*, such as *instrumental variables* estimation or *generalized least squares*.

All "models" are simplifications of reality. Presumably we want our econometric model to be simple but "realistic" – at least in the sense that we can *identify* our objective.

Traditionally the objective of most econometric models was to describe an economic process. Assumptions, such as the ones to follow, were to ensure the "quality" of the estimated economic model (we will soon measure quality in terms of *unbiasedness*, *efficiency*, and *consistency*). More recently, applied econometrics has been focused on obtaining *causal inference* from *observational data*. The emphasis is usually on estimating the marginal effect of just one of the x variables on y. From this perspective, the following assumptions can in part be viewed as necessary for estimating a causal relationship between x and y.

A.1: Linearity

The model is linear in the parameters:

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

Linearity in the parameters allows the model to be written in matrix nota-

tion. Let,

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

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

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.3}$$

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

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

Notational points

- Vectors are in bold.
- The dimensions of vectors/matrices are written (rows × columns).
- The first subscript denotes the row, the second subscript the column.
- 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 $\text{rank}(X) = k$. 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, $\text{rank}(X) = k - 1 = 1$. 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*, $\mathrm{E}(\varepsilon_i) = 0 \quad ; \quad i = 1, 2, \ldots, n$. 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:

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

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

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

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

but...

$$E(\varepsilon_i\varepsilon_i) = E\left(\varepsilon_i^2\right) = E\left[(\varepsilon_i - 0)^2\right] = \mathrm{var}\,(\varepsilon_i) = \sigma^2$$

and

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

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 X to be random, but restrict the form of their randomness – the process that generates X is unrelated to the process that generates $\boldsymbol{\varepsilon}$ in the population.

This is likely the most important assumption. We will soon see that it is imperative that X and $\boldsymbol{\varepsilon}$ are statistically independent.

So, if X is random, we need to assume *strict exogeneity*:

$$\mathrm{E}(\boldsymbol{\varepsilon}|X) = \mathbf{0} \tag{2.4}$$

or make the weaker assumption that X and $\boldsymbol{\varepsilon}$ are uncorrelated:

$$\mathrm{cov}(\boldsymbol{x_j}, \boldsymbol{\varepsilon}) = \mathbf{0} \quad ; \quad \text{for } j = 1, \ldots, k$$

Note that independence implies zero correlation, but not necessarily the other way around due to correlation measuring only *linear* dependencies between variables. (**Question:** Does the implication go both ways in the present case?)

Prove that the zero correlation assumption is equivalent to each column of X being *orthogonal* to $\boldsymbol{\varepsilon}$, that is, prove that A.5 at least implies that:

$$\mathrm{E}(X'\boldsymbol{\varepsilon}) = \mathbf{0} \tag{2.5}$$

Finally, note that the strict exogeneity assumption implies that: $\mathrm{E}(f(X)'\boldsymbol{\varepsilon}) = \mathbf{0}$.

A.6: Normality of errors

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

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:

- $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
- $(\boldsymbol{\varepsilon}|X) \sim N\left[\mathbf{0}, \sigma^2 I_n\right]$
- $\text{rank}(X) = k$
- The data generating process (DGP) of X and $\boldsymbol{\varepsilon}$ are unrelated

Implications for 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:

$$\boldsymbol{y} \sim N\left[X\boldsymbol{\beta}, \sigma^2 I_n\right]$$

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?

2.5 Least Squares Estimator

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

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

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

- There are 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:

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

and the population regression is:

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

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

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

In the population, the true (unobserved) disturbance is

$$\varepsilon_i = y_i - \boldsymbol{x}_i'\boldsymbol{\beta}.$$

When we use $\boldsymbol{b}$ to estimate $\boldsymbol{\beta}$, there will be some "estimation error", and the value,

$$e_i = y_i - \boldsymbol{x}_i'\boldsymbol{b}$$

will be called the i^{th} "residual". So,

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

Question: Which terms are unobserved (from the population) and which are observed (determined by the sample)?

2.5.1 The Least Squares criterion

We will "choose $\boldsymbol{b}$ so as to minimize the sum of squared residuals".
Questions:

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

2.5.2 Minimizing the sum of squared residuals: an optimization problem

We will solve a minimization problem using the least-squares criterion in order to derive the "Least Squares" estimator. The problem that we are trying to solve can be stated as:

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

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:

$$\begin{gathered}\boldsymbol{b}'X'\boldsymbol{y} = \boldsymbol{y}'X\boldsymbol{b} \\ (1 \times k)(k \times n)(n \times 1) = (1 \times 1)\end{gathered}$$

So,

$$S = \boldsymbol{y}'\boldsymbol{y} - 2\left(\boldsymbol{y}'X\right)\boldsymbol{b} + \boldsymbol{b}'\left(X'X\right)\boldsymbol{b}$$

Two rules involving the differentiation of vectors and matrices that we need are:

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

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

Applying these two results:

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

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

$$\begin{gathered}X'X\boldsymbol{b} = X'\boldsymbol{y} \\ (k \times n)(n \times k)(k \times 1) = (k \times n)(n \times 1)\end{gathered}$$

This gives us k equations in k unknowns, sometimes called the "normal equations". Finally, provided that $(X'X)^{-1}$ exists:

$$\boldsymbol{b} = \left(X'X\right)^{-1}X'\boldsymbol{y} \tag{2.6}$$

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

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

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

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

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:

(i) The Least Squares estimator, $\boldsymbol{b}$, *exists*.

(ii) Our optimization problem leads to the *minimization* of S, not its maximization!

2.5.3 Least Squares estimator 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:

$$\begin{aligned} b_1 &= \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2} = \frac{s_{x,y}}{s_x^2} \\ b_0 &= \overline{y} - b_1\overline{x}, \end{aligned} \tag{2.7}$$

where $s_{x,y}$ is the sample covariance between x and y, and s_x^2 is the sample variance of x.

Question: Why do population models typically include an intercept and how is the intercept effected in the population in matrix form (2.3)?

2.6 Method of Moments

The least squares criterion may seem dubious and unmotivated. We have yet to see the benefits of using an estimator that minimizes the sum of squared residuals. Rather than starting from this seemingly arbitrary criterion, we can instead derive the least squared estimator using the *Method of Moments* (MM).

The Method of Moments relies on the principle that the *sample mean* is a good way of estimating a *population mean* (we will see this later in the *law of*

large numbers). The MM is widely used in statistics, and many estimators in econometrics can be motivated using it or the closely related *Generalized Method of Moments*.

Take the simple population model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{2.8}$$

and take assumptions A.3 and A.5:

$$\mathrm{E}[\varepsilon_i] = 0 \quad ; \quad \mathrm{E}[x_i \varepsilon_i] = 0 \tag{2.9}$$

That is, assumptions A.3 and A.5 imply two *moment conditions*, expressed in equation 2.9. By replacing the expectations with sample averages, it can be seen that the MM estimator for the above population model 2.8 is identical to that in equation 2.7.

Note that the population model 2.8 above can be generalized to include k regressors; A.3 will provide one moment condition while A.5 will provide the remaining $(k-1)$ moment conditions necessary to solve for the unknown β. The MM estimator, in matrix form, will be identical to equation 2.6.

2.7 Exercises

1. Let the $\boldsymbol{y}$ and X data be:

$$\boldsymbol{y} = \begin{bmatrix} 1 \\ 4 \\ 5 \\ 4 \end{bmatrix} \; ; \quad X = \begin{bmatrix} 1 & 2 \\ 1 & 4 \\ 1 & 6 \\ 1 & 8 \end{bmatrix}$$

   (a) Calculate the Least Squares estimators for β_0 and β_1 for the population model:

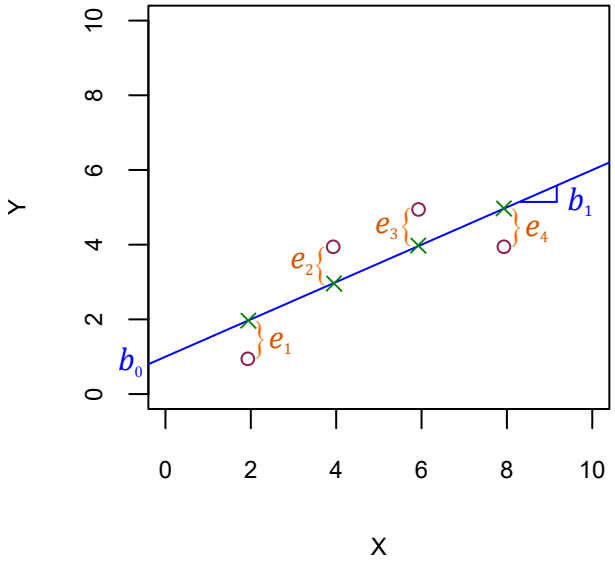
$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

   (b) Calculate the predicted values, and residuals, for the above data and model.

   (c) Verify that equation 2.6 and equation 2.7 are identical for the above situation.

   The data points, Least Squares estimates, predicted values, and residuals, are shown in Figure 2.1.

Figure 2.1: A simple data set with the estimated OLS line in blue. b_0 is the OLS intercept, and b_1 is the OLS slope. The OLS residuals (e_i) are the vertical distances between the actual data points ($\circ$) and the OLS predicted values ($\times$).



2. Prove that the Method of Moments estimator for the population model in equation 2.8 is identical to the least squared estimator in scalar form (equation 2.7). Hint: use the results that

$$\sum_{i=1}^{n}(x_i y_i) - n\bar{x}\bar{y} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

and

$$\sum_{i=1}^{n}\left(x_i^2\right) - n\overline{x}^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2$$