EC 320: Introduction to Econometrics

Lecture Notes

David M. Hall
University of Oregon

Fall 2025

1 Introduction

In this chapter we will introduce the simple linear regression model (SLRM). In our statistics review, we covered concepts such as **correlation** which can tell us if two variables are associated with each other. Correlation is more concerned with the sign of the relationship rather than the magnitude.

In particular, we are concerned with the relationship between some variable, let's call it Y, that *depends* on one or more other variables. For example, you may recall the following:

Example 1.1 (Known Relationship)**.** Recall that the definition of GDP is:

$$GDP = C + I + G + NX \tag{1}$$

by definition, GDP *depends* on consumption, investment, government spending and net exports.

On the other hand, definitions are not great examples of what we would estimate in a regression context. The regression context is used for when there is a *hypothetical* or *theoretical* relationship between some set of variables. For example:

Example 1.2 (Hypothetical Relationship)**.** Suppose that student test scores are directly and linearly related to the number of hours they study.

- **Correlation** would tell us that the sign between number of hours and test scores is positive.
  - It would give us a magnitude, but there might be other things besides hours of study that contribute to test scores.
- In a linear regression model, we can directly test for what that relationship is and include other determinants of good test scores (e.g. hours of sleep).

2 Simple Linear Regression Model

We begin first with the "simple" linear regression model. This model will deal with only *two* variables. Later, we will explore what happens when we add more to the model.

A quick note on variable notation. Recall that there is a difference between *population* and *sample* parameters. We will denote population parameters with Greek letters ($\alpha, \beta, \gamma, \phi, etc.$) and we will denote sample parameters/data with English (typically X for independent variables, Y for dependent variables, etc.).

Definition 2.1. The **Simple Linear Regression Model** (SLRM) describes a linear relationship between a dependent variable Y and an independent variable X, specified as:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

where u_i is the error term capturing unobserved factors.

Remark 2.1. The goal of regression analysis is to estimate the coefficients/parameters β_0 and β_1 using observed data, and to understand the relationship between X and Y.

Remark 2.2. The disturbance term allows us to measure with some sort of error (*it is also called an error term for that reason*). We do not expect that we can explain *all* of the variation in Y. In fact, if we knew all of the determinants of Y(e.g. if Ywere GDP), it would not be a very good regression.

In the book, Dougherty details five reasons for the existence of the disturbance term. Note that the disturbance term is the aggregation of *all* of these reasons (if relevant):

1. Omission of explanatory variable.
   - e.g. Not including education in a wage regression.
2. Aggregation of many variables.
   - Using average wage rather than individual wages.
3. Model misspecification.
   - e.g. lag structure
4. Functional misspecification.
   - Is this thing even linear?
5. Measurement error
   - Let your five-year-old cousin collect data on NBA statistics then tell me how well you can estimate the relationship between number of 3 pointers made and win probability.

3 Least Squares Algorithm

The first algorithm we will be using to estimate the population parameter β is Ordinary Least Squares (OLS). The least squares algorithm does the following:

- Isolate the residual.
- **Square** it.
- Sum the squared residuals.
- Minimize (least) it with respect to coefficients.

1. Isolate the residual.
   We are regressing:
   $$Y_i = b_0 + b_1 X_i + e_i \tag{2}$$
   to isolate the residual (e_i):
   $$e_i = Y_i - b_0 - b_1 X_i \tag{3}$$
2. Square each residual.
   $$e_i^2 = (Y_i - b_0 - b_1 X_i)^2 \tag{4}$$
3. Sum the squared residuals.
   Suppose that we have 10 different observations, then.
   $$RSS = \sum_{i=1}^{10} e_i^2 = e_1^2 + e_2^2 + \cdots + e_{10}^2 \tag{5}$$
   $$RSS = \sum_i (Y_i - b_0 - b_1 X_i)^2 \tag{6}$$
4. Take the derivative w.r.t. b_0 and b_1.
5. Solve.

Definition 3.1 (Sum of Squared Residuals (SSR))**.**

$$RSS = \sum_{i=1}^{I} e_i^2 = e_1^2 + e_2^2 + \cdots + e_I^2 \tag{7}$$

Remark 3.1. *What are the subscripts i on the regression equation?* Subscripts typically denote a unit of measurement. In this case, we assumed we had *individual* data, so the equation $Y_i = b_0 + b_1 X_i + e_i$ tells us that for each person, denoted i, what we are measuring for them, Y_i depends linearly upon an intercept and some other variable we measure for them X_i. You can also have *time* subscripts (e.g. in each period...) or county/state (e.g. for every county, number of hospitals is a linear function of...).

Remark 3.2. Why do we have to square it? Why not minimize the sum of residuals?

Just let $b_0 = \bar{Y}$ and $b_1 = 0$, for example. That would let the residuals be 0.

3.1 Examples

Example 3.1 (OLS Derivation: Dougherty's Three-Point Example)**.** Suppose we observe the following three data points:

i	Y_i	X_i
1	3	1
2	5	2
3	6	3

We assume a simple linear model:

$$Y_i = b_1 + b_2 X_i + e_i$$

The predicted values from the regression line are:

$$\begin{aligned}\hat{Y}_1 &= b_1 + b_2 \cdot 1 \\ \hat{Y}_2 &= b_1 + b_2 \cdot 2 \\ \hat{Y}_3 &= b_1 + b_2 \cdot 3\end{aligned}$$

Thus, the residuals are:

$$\begin{aligned}e_1 &= 3 - b_1 - b_2 \\ e_2 &= 5 - b_1 - 2b_2 \\ e_3 &= 6 - b_1 - 3b_2\end{aligned}$$

We now construct the Residual Sum of Squares (RSS):

$$\text{RSS} = (3 - b_1 - b_2)^2 + (5 - b_1 - 2b_2)^2 + (6 - b_1 - 3b_2)^2$$

Expanding each term:

$$\begin{aligned}(3 - b_1 - b_2)^2 &= 9 + b_1^2 + b_2^2 - 6b_1 - 6b_2 + 2b_1b_2 \\ (5 - b_1 - 2b_2)^2 &= 25 + b_1^2 + 4b_2^2 - 10b_1 - 20b_2 + 4b_1b_2 \\ (6 - b_1 - 3b_2)^2 &= 36 + b_1^2 + 9b_2^2 - 12b_1 - 36b_2 + 6b_1b_2\end{aligned}$$

Adding all terms:

$$\text{RSS} = 70 + 3b_1^2 + 14b_2^2 - 28b_1 - 62b_2 + 12b_1b_2$$

To minimize RSS, take first-order conditions:

$$\frac{\partial \text{RSS}}{\partial b_1} = 6b_1 + 12b_2 - 28 = 0 \tag{1}$$
$$\frac{\partial \text{RSS}}{\partial b_2} = 28b_2 + 12b_1 - 62 = 0 \tag{2}$$

Solving the system:
From (1): $b_1 = \frac{28-12b_2}{6} = \frac{14-6b_2}{3}$
Substitute into (2):

$$\begin{aligned} 28b_2 + 12\left(\frac{14-6b_2}{3}\right) &= 62 \\ 28b_2 + 4(14-6b_2) &= 62 \\ 28b_2 + 56 - 24b_2 &= 62 \\ 4b_2 = 6 \Rightarrow b_2 &= 1.5 \end{aligned}$$

Then:

$$b_1 = \frac{14-6(1.5)}{3} = \frac{14-9}{3} = \frac{5}{3} \approx 1.67$$

Final regression line:

$$\boxed{Y_i = 1.67 + 1.5X_i}$$

Example 3.2 (OLS Derivation: General Case with One Explanatory Variable)**.** We consider the simple linear regression model with one explanatory variable and n observations:

$$Y_i = b_1 + b_2X_i + e_i$$

Our goal is to choose b_1 and b_2 to minimize the residual sum of squares:

$$\text{RSS} = \sum_{i=1}^{n}(Y_i - b_1 - b_2X_i)^2$$

We first expand the square in the residual for a single observation:

$$e_i^2 = (Y_i - b_1 - b_2X_i)^2 = Y_i^2 + b_1^2 + b_2^2X_i^2 - 2b_1Y_i - 2b_2X_iY_i + 2b_1b_2X_i$$

Summing over all observations gives:

$$\text{RSS} = \sum_{i=1}^{n}Y_i^2 + nb_1^2 + b_2^2\sum_{i=1}^{n}X_i^2 - 2b_1\sum_{i=1}^{n}Y_i - 2b_2\sum_{i=1}^{n}X_iY_i + 2b_1b_2\sum_{i=1}^{n}X_i$$

Taking partial derivatives with respect to b_1 and b_2 and setting them to zero:

$$\frac{\partial \text{RSS}}{\partial b_1} = 2nb_1 + 2b_2 \sum X_i - 2 \sum Y_i = 0 \tag{1}$$
$$\frac{\partial \text{RSS}}{\partial b_2} = 2b_2 \sum X_i^2 + 2b_1 \sum X_i - 2 \sum X_i Y_i = 0 \tag{2}$$

Dividing both equations by 2, we obtain the normal equations:

$$nb_1 + b_2 \sum X_i = \sum Y_i \tag{3}$$
$$b_1 \sum X_i + b_2 \sum X_i^2 = \sum X_i Y_i \tag{4}$$

Let $\bar{X} = \frac{1}{n} \sum X_i$ and $\bar{Y} = \frac{1}{n} \sum Y_i$. Then equation (3) becomes:

$$b_1 = \bar{Y} - b_2 \bar{X}$$

Substituting into equation (4):

$$(\bar{Y} - b_2 \bar{X}) \sum X_i + b_2 \sum X_i^2 = \sum X_i Y_i$$

$$n\bar{X}\bar{Y} - b_2 n\bar{X}^2 + b_2 \sum X_i^2 = \sum X_i Y_i$$

$$b_2 \left(\sum X_i^2 - n\bar{X}^2 \right) = \sum X_i Y_i - n\bar{X}\bar{Y}$$

This leads to the formula for the slope:

$$\boxed{b_2 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}}$$

Then, the intercept is:

$$\boxed{b_1 = \bar{Y} - b_2 \bar{X}}$$

These are the ordinary least squares estimators for the intercept and slope in the simple linear regression model.

4 Be Able To

- Write the simple regression model.
- Define/explain the disturbance term.
- Explain the least squares algorithm in three steps.
- Carry out OLS for simple linear model.

5 Key Words

- Independent Variable/Regressor
- Dependent variable
- Disturbance Term
- Residual