Week 2

Introduction

This week we'll start Chapter 2. We'll talk about the single variable linear regression, the population regression function, the sample regression function, and we'll also look at a technique to estimate the coefficients of linear regression equations. Something worth mentioning, you should use these notes together with the textbook to solidify your understanding of the material.

Single variable linear regression and the population regression line

A linear regression relates one economic variable to another. i.e., how changes in an exogenous variable relate to changes in an endogenous variable.

Let's assume we have information on two variables:

x : daily income

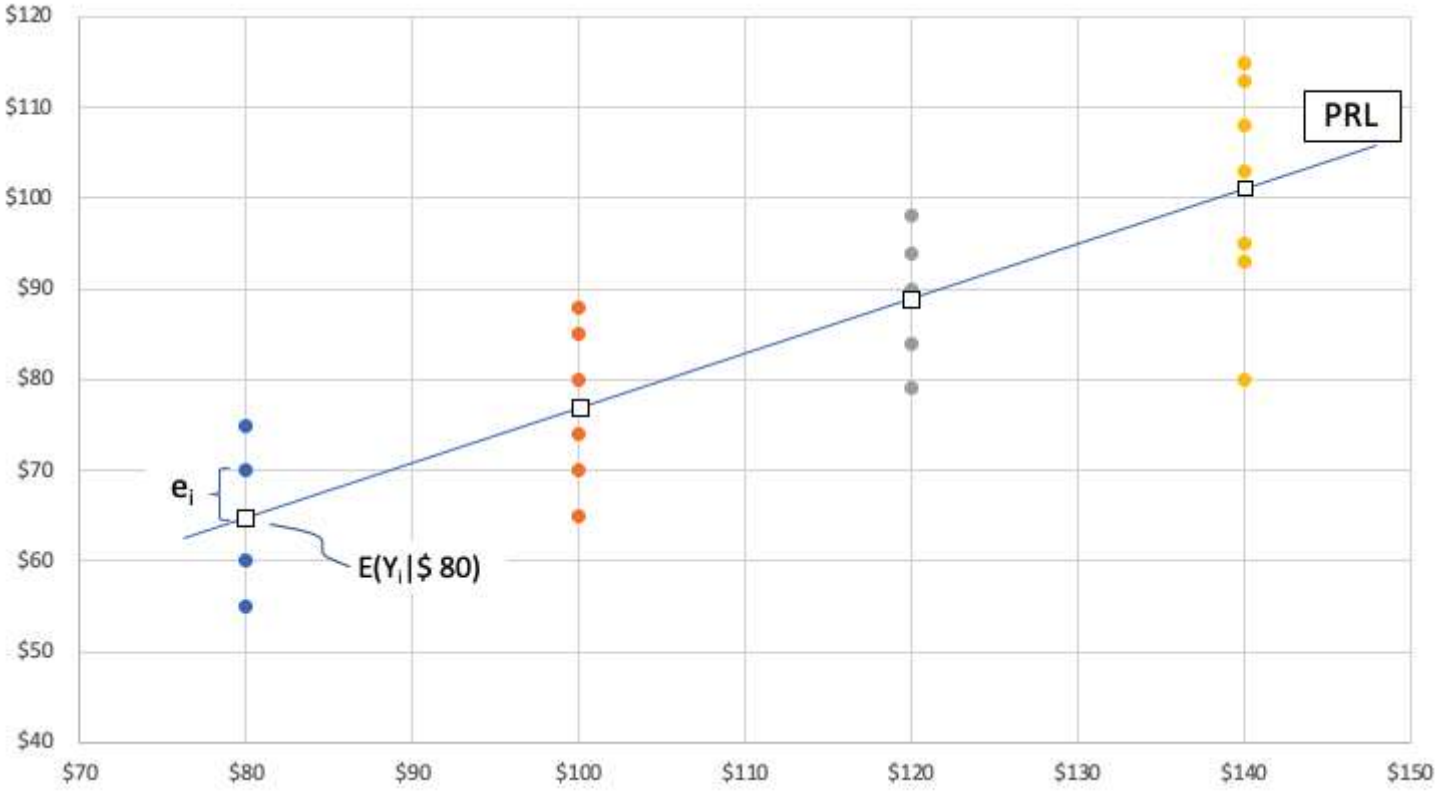
y : daily consumption

	Daily Income			
	$ 80	$ 100	$ 120	$ 140
Daily Consumption	$ 55	$ 65	$ 79	$ 80
	$ 60	$ 70	$ 84	$ 93
	$ 65	$ 74	$ 90	$ 95
	$ 70	$ 80	$ 94	$ 103
	$ 75	$ 85	$ 98	$ 108
		$ 88		$ 113
				$ 115
$E(Y_i \mid X)$	$ 65	$ 77	$ 89	$ 101

> **Recall:** The expected value is a measure of central tendency:
>
> $$E(X) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

We are interested in the conditional expectation $E(Y_i|X)$ from our table above.

Let's take a look at the graph below:



The *population regression line* (PRL) is the line connecting all of the expected values. In the graph above, the expected values are represented with squares. The dots are our data points, and the line connecting the squares is the PRL.

On the following equation:

$$E(Y_i|X) = \beta_0 \pm \beta_1 X_i$$

...we will be tasked with finding β_0 and β_1.

On the graph above we can also see that e_i is the difference between the expected value and a given data point: this is the random error for the specific x_i value.

Knowing this, we can see that:

$$e_i = Y_i - E\left(Y_i \mid X\right)$$

$$Y_i = E\left(Y_i \mid X\right) + e_i$$

$$Y_i = \beta_0 + \beta_1 X_i + e_i \tag{1}$$

Equation (1) is our Population Regression Function (PRF).

$E(Y_i|X) = \beta_0 \pm \beta_1 X_i$ $\rightarrow$ the deterministic or systematic portion of the PRF

$e_i \rightarrow$ the non-deterministic or non-systematic portion of the PRF

We will now take a look at how to derive the following:

$$E\left(e_i \mid x\right) = 0$$

We start by taking expectations of our PRF:

$$E(y \mid x) = E(E(y \mid x) \mid x) + E(e \mid x)$$

We know that the expected Value of the Expected Value is the expected value:

$$E(E(y|x)) = E(y|x)$$

So we can simplify the prior step and rewrite it as:

$$E(y \mid x) = E(y \mid x) + E(e \mid x)$$

We keep manipulating the expression to get:

$$E(e \mid x) = E(y \mid x) - E(y \mid x)$$

$$E(e \mid x) = 0$$

So this is *very* important. We just derived an expression that tells us that the expected error is zero! What does this mean? The error term is basically capturing the variation in the dependent variable that the independent variable does not explain. It's a stochastic error term, meaning it's determined by random chance. For our model to be unbiased, we need the average of the error term to be zero.

Ok I know, there is a lot to unpack there. Let's look at it from another angle. If the average error is, say, +10, we would mean that the model is systematically *UNDERPREDICTING* the observed value. This is what we call *BIAS*, and it tells us that our model is not adequate because, on average, it is not correct.

Let's try this on our example above:

$$E(e|80) = \frac{1}{5}\sum_{i=1}^{n}(e_i|80) = (55-65) + (60-65) + (65-65) + (70-65) - (75-65) =$$

$$= \frac{-10-5+0+5+10}{5} = 0$$

$$E(e|x) = 0$$

Sample regression line

The sample regression line is an estimate of the population regression line.

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Where:

$\hat{y}_i$ is the estimate of the truve value of y, or $E(y_i|x)$

$\hat{\beta}_0$ is the estimate of the β_0 found in the PRL

$\hat{\beta}_1$ is the estimate of the β_1 found in the PRL

Our **sample regression function (SRF)** is then:

$$y_i = \hat{y}_i + \hat{e}_i$$

Or:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{e}_i$$

Now that we have our SRF, we can move on to the next task. We know that we want to run a line through our sample, minimizing the distance between each data point and the line. We are interested in having the line go through the data in such a way that our predicted error term is as small as possible. In other words, we want to *minimize* the predicted error term.

Ordinary Least Squares (OLS)

OLS tries to minimize the sum of the squared residuals (SSR). On the graph above, you can see that you have data points above and below the line, so we square the distances to the line as a clever way to deal with positive numbers only.

So here we go:

$$\min_{\hat{\beta}_0, \hat{\beta}_1} SSR : \sum_{i=1}^{n} \hat{e_i}^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 \, x_i)^2$$

First-order Condition (FOC):

$$\frac{\partial SSR}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0 \tag{I}$$

$$\frac{\partial SSR}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^{n} x_i \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0 \tag{II}$$

> Note: you should try and derive the second-order condition (SOC) for equations (I) and (II) to verify that they are a minimum.

From (I):

$$-2\sum_{i=1}^{n} y_i + 2\sum_{i=1}^{n} \hat{\beta}_0 + 2\,\beta_1 \sum_{i=1}^{n} x_i = 0$$

$$2\sum_{i=1}^{n} \hat{\beta}_0 + 2\,\beta_1 \sum_{i=1}^{n} x_i = 2\sum_{i=1}^{n} y_i$$

$$\sum_{i=1}^{n} \hat{\beta}_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$n \cdot \hat{\beta}_0 + \hat{\beta}_1 \sum x_i = \sum_{i=1}^{n} y_i$$

$$\frac{1}{n}\left(\sum_{i=1}^{n} y_1\right) = \frac{1}{n}\left(n\,\hat{\beta}_0 + \hat{\beta}_1 \sum x_i\right)$$

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

$$\boxed{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\,\bar{x}} \tag{a}$$

We plug (a) in (II):

$$-2\sum x_i \left(y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i\right) = 0$$

$$\boxed{\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y - \bar{y})}{\sum (x_i - \bar{x})^2}} \tag{b}$$

What's another way of writing equation (b) ?

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum (x_i - \bar{x})\,(y_i - \bar{y})}{\frac{1}{n}\sum (x_i - \bar{x})^2} = \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)}$$

This is a good stopping point. If you are not familiar with the variance and covariance, please take some time to review what you learned in your statistics class as we'll be diving a bit deeper into this next week.
