ECN 6330

Chapter 2 - The Simple Regression Model

Most of econometrics deals with relating two random variables y and x that represent some population.

Usually, we are interested in "explaining y in terms of x."

We confront (at least) 3 issues in "explaining y in terms of x":

1. Since there is never an exact relationship between two variables, how do we allow for other factors to affect y?

2. What is the functional relationship between y and x?

3. How can we be sure we are capturing a ceteris paribus relationship between y and x?

We start by writing down an equation relating y to x as follows:

$$y = \beta_0 + \beta_1 x + u$$

This equation is assumed to hold for the population of interest. It is called the Simple Regression Model.

In this model:

y = explained variable (the variable we are interest in explaining)

x = explanatory variable (the variable we believe affects y)

u = error or disturbance term
(represents other factors affecting y)

The Simple Regression Model addresses the issue of the functional relationship between y and x. If the "other" factors in u are held fixed, so that the change in u is zero, then x has a linear effect on y.

$$\Delta y = \beta_1 \Delta x_1 \quad \text{if } \Delta u = 0$$

The change in y is just the change in x multiplied by β_1. We call β_1 the slope parameter; it is of primary interest in econometrics

The parameter β_0 is the constant term or the intercept parameter.

Ex: 2.1 Wooldridge pp. 24

Suppose soybean yield is determined by the model

$$yield = \beta_0 + \beta_1 \, fertilizer + u$$

then

$y = yield$

$x = fertilizer$

u may contain factors such as land quality, rainfall, etc.

The coefficient β_1 measures the effect of fertilizer on soybean yield, holding all else constant

$$\Delta yield = \beta_1 \Delta fertilizer$$

Ex: 2.2 Wage equation

A model that relates an individual's educations to wage:

$$Wage = \beta_0 + \beta_1 educ + u$$

If wage is measured in dollars per hour and educ is measured in years of education, then β_1 measures the change in hourly wage given another year of education.

The linearity of

$$y = \beta_0 + \beta_1 x + u$$

implies that a one-unit change in x has the same effect on y, regardless of the value of y.

Sometimes, this is unrealistic. For example, in the wage-education example we might believe that there is an increasing relationship, such that an additional year of education increases wages by more than the previous year of education does. We will allow for effects like this later.

Another important question is are we able draw ceteris paribus conclusions about how x affects y?

We have seen that β_1 does measure how x affects y, holding all other factors constant (in u).

Unfortunately this not the end of the <u>causality</u> issue. We will deal with this later. Basically though, correlation is not the same as causality. If for example, there is a third factor in u that causes both y and x then we can have problems.

EX: In the wage equation we might have a factor called "natural ability" in u. The econometrician will never observe this variable, but it clearly relates to both y and x.

For now, we will make an assumption about u. As long as an intercept term β_0 is included in the equation, nothing is lost by assuming

$$E(u) = 0$$

Note: if $E(u) \neq 0$ but we include β_0 the mean of u will get "sucked" into β_0

(technical ~~statistical~~ term JK)

R code: nonzero_mean.r [Simulation to show the above]

One additional (crucial) assumption regarding how u and X are related. A natural measure of the association between two variables is the correlation coefficient.

If ~~X~~ and u are uncorrelated, then they are not linearly related. Assuming that X and u are uncorrelated goes a long way to defining how u and x should be unrelated in

$$Y = \beta_0 + \beta_1 X + u$$

But unfortunately, it is not enough. It is possible for u to be uncorrelated with X but correlated with functions of x (such as X^2).

A stronger (better) assumption involves the conditional expectation of u given X. Namely

$$E(u/X) = E(u)$$

[our old friend the conditional distributions.]

This says that the average value of the unobservables is the same across all slices of the population determined by X, and that the common average is just the average of the marginal distribution of u.

NB: This is called mean independence assumption of u

I.E. we can't learn anything about the average value of u knowing X. Here (in this situation) this is desirable.

Combining this with our earlier assumption about $E(u)$ (namely that $E(u)=0$) we have that

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

This is called the <u>zero conditional mean assumption</u>

Ex: Assume that u is the same as "innate ability". This assumption requires that the average level of ability is the same regardless of the years of education.

Note: Probably too strong an assumption as more "able" people are more likely to be more educated

In the fertilizer example, if fertilizer amounts are chosen independently of other features of plots then the zero conditional mean assumption will hold. If however, more fertilizer is given to higher-quality plots of land then the expected value of u changes with the level of fertilizer. And the zero conditional mean assumption will not hold.

The zero conditional mean assumption gives β_1 another interpretation.

Taking the expected value of (given x)

$$y = \beta_0 + \beta_1 x + u$$

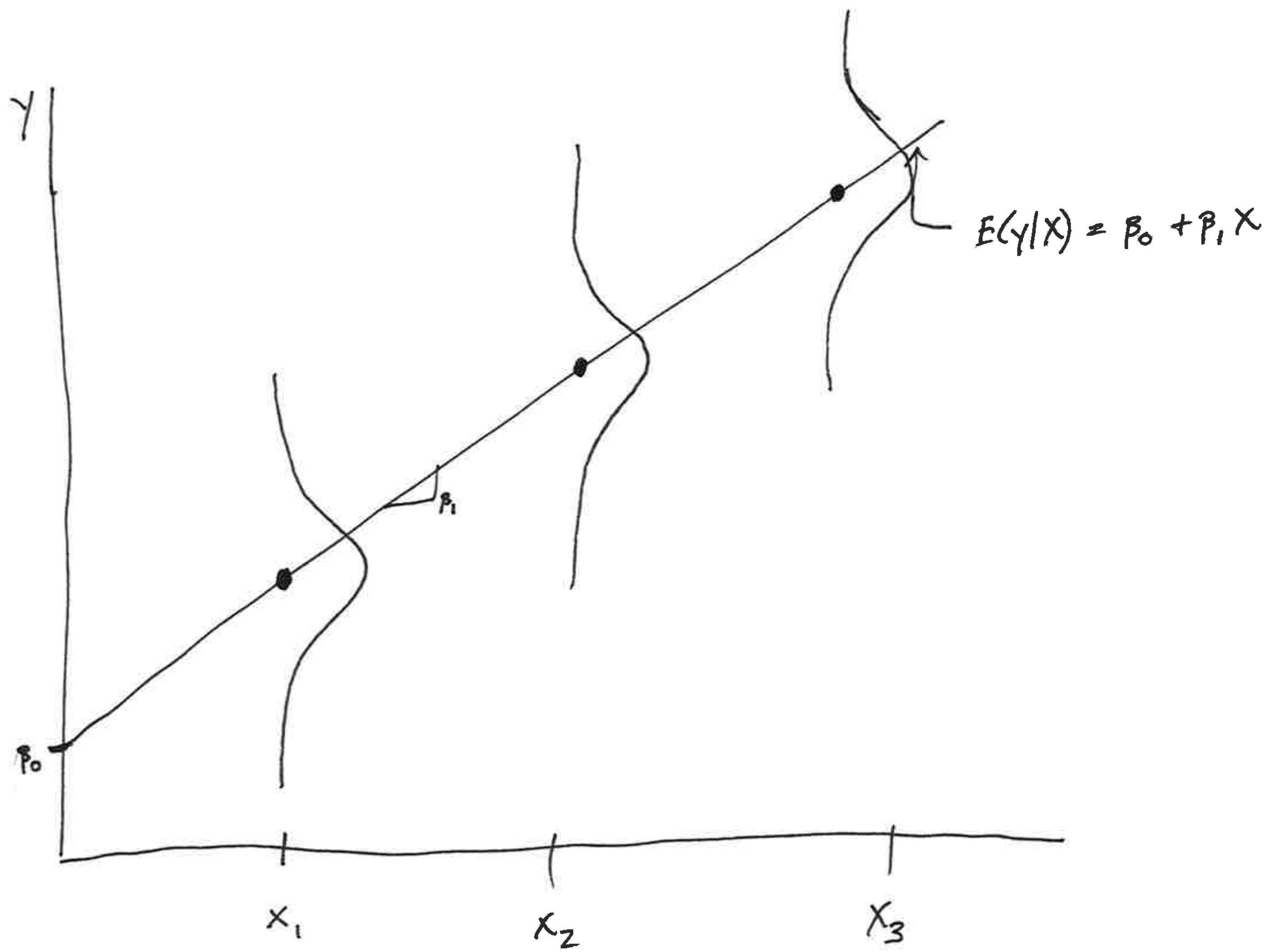
we get

$$E(y|x) = \beta_0 + \beta_1 x + \overset{0}{E(u|x)}$$

$$= \beta_0 + \beta_1 x$$

This says that a one-unit change in x changes the expected value of y by β_1.

Graphically (see p. 26 in Wooldridge)



NB: $E(y|x) = \beta_0 + \beta_1 X$ tells how the average value of y changes with X.

Ordinary Least Squares Estimates

Q: How to get estimates for population parameters β_0, β_1?

A: Define an estimator and use a sample of data!

Let $\{(x_i, y_i) : i = 1, \ldots, n\}$ denote a random sample of size n from the population. These are drawn from the population

So for every $i = 1, \ldots, n$ we can write

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

Here u_i is the error term for the ith observation because it contains all factors affecting Y_i other than X_i.

Ex:

Y_i = annual savings

X_i = annual income

u_i = all other factors affecting household i's annual savings other than X_i

We assume that u is uncorrelated with X

$$E(u) = 0$$

$$\mathrm{Cov}(X, u) = E(Xu) = 0$$

treat X as fixed
Ex: fertilizer

which can be written as

$$E(y - \beta_0 - \beta_1 X) = 0$$

and

$$E(X(y - \beta_0 - \beta_1 X)) = 0$$

Given a sample of data we will choose estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ to solve the sample counter parts

$$\frac{1}{n}\sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \qquad *$$

$$\frac{1}{n}\sum x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \qquad **$$

Equation (*) can be re written as

$$\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}$$

Where $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ and $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Then we can write $\hat{\beta}_0$ in terms of $\hat{\beta}_1$, $\bar{Y}$, and $\bar{X}$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Therefore, once we have the slope estimate $\hat{\beta}_1$ it is simple to get $\hat{\beta}_0$ given $\bar{Y}$ and $\bar{X}$

Dropping the $\frac{1}{n}$ (does not affect the solution) and plugging

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ into equation (**) yields

$$\sum_{i=1}^{n} x_i \left[y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 x_i \right] = 0$$

which we can re arrange as

$$\sum_{i=1}^{n} x_i (y_i - \bar{y}) = \hat{\beta}_1 \sum_{i=1}^{n} x_i (x_i - \bar{x})$$

Using properties of summation operators (see Appendix A A.7 & A.8)

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

and

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

Then as long as $\sum_{i=1}^{n} (x_i - \bar{x})^2 > 0$

the slope estimator is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Note: This is simply the sample covariance of X and Y

divided by the sample variance of X