EEP/IAS 118 - Introductory Applied Econometrics, Lecture 3

Gregory Lane

June 2017

Today

- Simple Regression Framework
- OLS Derivation
- Properties of $\hat{\beta}$
- R^2
- Assignments:
    - Problem Set 1 Due on Monday
    - First Quiz on Tuesday

Economic Model

An economic model is a equation that describes relationships. For example, we can try to describe participation in crime:

$$y = f(x_1, x_2, x_3, x_4, \cdots, x_6)$$

where y =hours spend in criminal activity, x_1 =police enforcement, x_2 =hourly wage in legal employment,..., x_6 =age.

We turn this economic model into a econometric model by assigning a functional form (linear):

$$\underbrace{crime}_{y} = \beta_0 + \underbrace{\beta_1}_{parameter} \; enforcement + \beta_2 \underbrace{wage}_{x} + \cdots + \beta_6 age + u$$

- Note that u here contains all the unobserved variables (e.g. family background, earnings from crime) that we cannot include in the model.

Population Regression Function

Consider a version of this model where crime (y) is only a function of wage in legal activity (x):

$$y = f(x, u) = \beta_0 + \beta_1 x + u$$

- Let us assume this is the "true data generating process"
- $u = y - \beta_0 - \beta_1 x$ is the error term. We make two important assumptions on u
  1. $E(u) = 0$ - this means that across the entire population the average residual is equal to 0. This is essentially just redefining the unobservable to be distributed around 0
  2. $E(u|x) = 0$ - this assumption says that u is "mean independent" of x

Population Regression Function

Assumption 2: $E(u|x) = 0$ - this assumption says that u is "mean independent" of x

- Implies that $cov(u, x) = 0$

$$\begin{aligned} E(u|x) = 0 \rightarrow Cov(u,x) &= 0 \\ Cov(u,x) &= E(ux) - E(u)E(x) \\ &= E(ux) \\ &= 0 \end{aligned}$$

- We will examine the importance of this assumption in more detail later (and it is critically important)

Population Regression Function

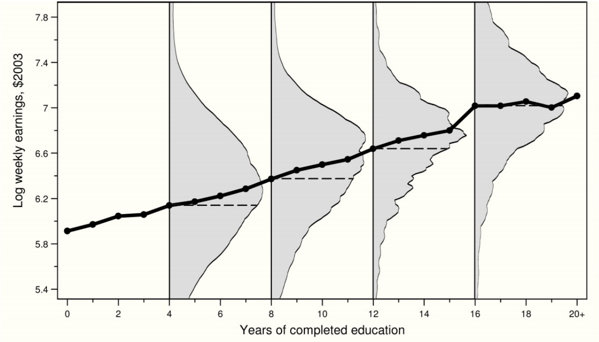
Note that we are making these assumptions on the *true* underlying model. Therefore, we can never directly test these assumptions

- We can summarize these conditions with: $E(u|x) = E(u) = 0$
- This assumption allows us to define a linear **population regression function**:

$$E(y|x) = \beta_0 + \beta_1 x$$

Population Regression Function

$$E(y|x) = \beta_0 + \beta_1 x$$



The PRF describes how the *average* value of y changes with x. Note, the above picture isn't linear, but for this class we will assume it is.

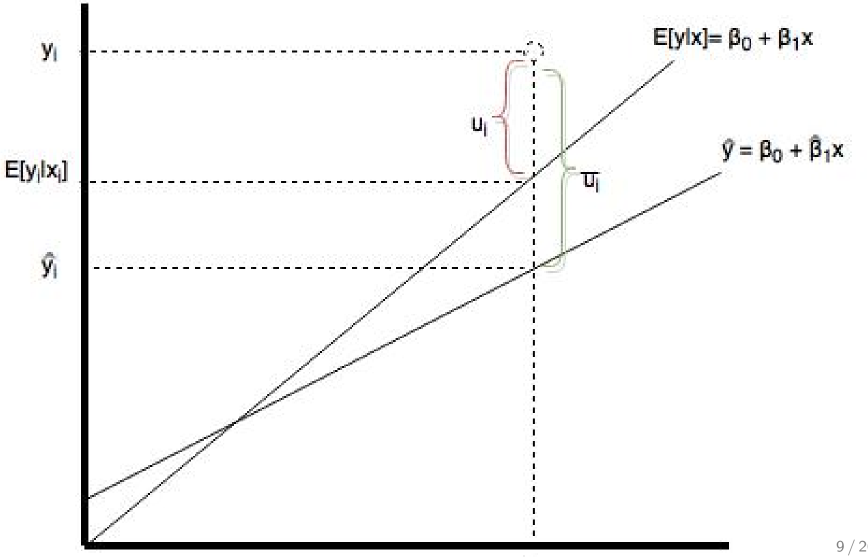
Regression with a Sample

The above example is done with a *population*, which we almost never observe. Instead, we work with *samples*.

- Goal is to approximate the PRF using a **sample regression function (SRF)**: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
- $y = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{u} = \hat{y} + \hat{u}$ is our estimated model. The "hats" indicate that these are estimates of some true value or parameter.
    - $\hat{y}$ is our best guess at the true $E(y|x)$
    - $\hat{\beta}$ is our best guess at the true relationship between x and y
    - $\hat{u}_i$ is the residual and is the deviation between the real observed y_i and $\hat{y}_i$. That is : $\hat{u}_i = y_i - \hat{y}_i$

Regression with sample

The PRF and SRF will (almost) never be the same!



Deriving OLS

How do we actually go about estimating the model, choosing the best line to fit the data (i.e. finding our best guess for the β_j)?

- Approach should seek to minimize the *errors* ($\hat{u}_i$) between our model prediction and the actual data
- A couple of options for how to do this:
  1. **Least squares:** Minimize sum of square distance between our data (x_i, y_i) and the line $(\hat{y}_i)$ - i.e. minimize the errors

$$Min \sum_i (y_i - \hat{y}_i)^2$$

  2. **Least absolute error:** Minimize the sum of the absolute value of the errors

$$Min \sum_i |y_i - \hat{y}_i|$$

Deriving OLS

For various reasons, Least squares is generally the preferred method. So the goal is to choose the β_j that accomplishes:

$$Min \sum_i (y_i - \hat{y}_i)^2$$

How do we do this: take the derivative of the function above w.r.t. $\hat{\beta}_j$ and set equal to zero. Then solve for β_j (remember calculus and critical values)

Deriving OLS

Let's define W, plugging in our model for $\hat{y}_i$:

$$W = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

We'd like to choose $\hat{\beta}_0$ and $\hat{\beta}_1$ so that W is as small as possible:

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

Taking the first order conditions (partial derivatives):

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

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

Deriving OLS

These equations can be solved for $\hat{\beta}_0$ and $\hat{\beta}_1$.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2} = \frac{cov(x,y)}{var(x)}$$

Our best choice of $\hat{\beta}_1$ is given by the covariance of x and y divided by the variance of x. Any intuition for why this might make sense?

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

The best choice of β_0 - the intercept - is the average of y minus our previously estimated β_1 times the average x

- You can (will) calculate these by hand in excel. However, in the future programs like Stata will calculate them for you

Deriving OLS

Data on CO2 (y) and GDP (x):

	GDP/cap (PPP$ 1000) x	CO2/cap (tons) y	(x-xbar)	(y-ybar)	(x-xbar)^2	(x-xbar)(y-ybar)	yhat	uhat^2	(y-ybar)^2
United States	43.7	19.3	18.5	10.0	342.9	184.5	14.84	19.9	99.3
Ireland	41.0	10.2	15.8	0.9	250.2	13.7	14.04	14.7	0.8
Belgium	33.5	9.7	8.3	0.4	69.2	3.0	11.81	4.4	0.1
Korea, Rep.	25.0	10.4	-0.2	1.1	0.0	-0.2	9.28	1.3	1.1
India	2.7	1.4	-22.5	-7.9	505.5	178.4	2.65	1.6	62.9
China	5.2	5.0	-20.0	-4.3	399.3	86.6	3.39	2.6	18.8
Sum	151.1	56.0	0.0	0.0	1567.1	466.1	56.0	44.5	183.1
Sample mean	25.2	9.3	0.0	0.0					
Sample variance/covariance					313.4	93.2			

1. What is $\hat{\beta}_1$?
2. What is $\hat{\beta}_0$?

Deriving OLS

	GDP/cap (PPP$ 1000) x	CO2/cap (tons) y	(x-xbar)	(y-ybar)	(x-xbar)^2	(x-xbar)(y-ybar)	yhat	uhat^2	(y-ybar)^2
United States	43.7	19.3	18.5	10.0	342.9	184.5	14.84	19.9	99.3
Ireland	41.0	10.2	15.8	0.9	250.2	13.7	14.04	14.7	0.8
Belgium	33.5	9.7	8.3	0.4	69.2	3.0	11.81	4.4	0.1
Korea, Rep.	25.0	10.4	-0.2	1.1	0.0	-0.2	9.28	1.3	1.1
India	2.7	1.4	-22.5	-7.9	505.5	178.4	2.65	1.6	62.9
China	5.2	5.0	-20.0	-4.3	399.3	86.6	3.39	2.6	18.8
Sum	151.1	56.0	0.0	0.0	1567.1	466.1	56.0	44.5	183.1
Sample mean	25.2	9.3	0.0	0.0					
Sample variance/covariance					313.4	93.2			

1. $\hat{\beta}_1 = \frac{93.2}{313.4} = 0.297$ tons/\$1000
2. $\hat{\beta}_0 = 9.3 - 0.297 * (25.2) = 1.843$ tons

Don't forget units!

Summary: β_0, β_1

The equations of β_0 and β_1:

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

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

As we saw, these are derived by minimizing the sum of squared errors ($\hat{u}$). This process is called Ordinary Least Squares (OLS). OLS has some nice estimation properties (e.g. it is the least variance estimator under certain conditions), which is why we use it.

- Don't worry about doing the full derivation

Interpreting $\hat{\beta}$ - Sign, Significance, Size

When asked to "interpret your results" you should check 3 things:

1. **Sign:**
   - What sign did you expect the estimated parameter to have? Why? Does your estimate have this sign (i.e. are you surprised or reassured by your results)?
2. **Significance:**
   - Is the estimate statistically different from zero? What is the t-statistic of this hypothesis?
   - Don't worry about this for now, we will deal with this in more detail later in the course.
3. **Size:**
   - How do changes in this variable affect the dependent variable according to your estimation? Is this an economically meaningful effect size?

Interpreting $\hat{\beta}$ vs. β

- We interpret β_1 as the marginal effect (change of one unit) of our x variable (e.g. education) on the EXPECTED value of the y (denoted $E(y)$, $E(wage)$) in the population. Note that without the entire population of data, we never actually see β_1.
- We interpret $\hat{\beta}_1$ as the marginal effect of our x variable (e.g. education) on PREDICTED y (denoted $\hat{y}$, $w\hat{a}ge$). This is an estimated parameter, for which we get a value from our data.
    - You will need to use the word *predicted* when interpreting $\hat{\beta}$

Example Interpretation

Example: Exercise 2.4 Woolridge: Let's examine a regression of baby birthweight on number of daily cigarettes smoked by the mother:

$$\widehat{bwght} = 119.77 - 0.514cigs$$

1. Interpret the coefficient on cigs.
2. What is the predicted birthweight when $cigs = 0$?
3. To predict a birthweight of 125, what would $cigs$ have to be?

Example Interpretation

$$\widehat{bwght} = 119.77 - 0.514cigs$$

1. Interpret the coefficient on cigs.: *Sign: The coefficient on* $cigs$ *is negative, as we would expect. Significance: Leave for now, but assume it is. Size: Smoking an additional cigarette per day is associated with a 0.514 ounce decrease in predicted birth weight - this seems important*
2. What is the predicted birthweight when $cigs = 0$? *Predicted birth weight is 119.77 ounces (the intercept)*
3. To predict a birthweight of 125, what would $cigs$ have to be? *Solve for cigs and plug in 125 for birthweight:*

$$cigs = (125 - 119.77)/(-0.524) \approx -10$$

Example Interpretation: $\hat{\beta}_0$

How do we interpret $\hat{\beta}_0$?

- This is the best prediction of $\hat{y}$ when all x are equal to zero

$$\widehat{bwght} = 119.77 - 0.514cigs$$

- Here, 119.77 is the predicted birthweight for a mother who does not smoke any cigarettes
- In many cases, $\hat{\beta}_0$ does not mean very much. For instance if the regression were $income = \beta_0 + \beta_1 educ$, then $\hat{\beta}_0$ would be predicted income for someone without *any* education - something that we probably don't see in the data
- To see if β_0 is meaningful, think about what it means when all $x = 0$. If this makes sense in the context, $\hat{\beta}_0$ may be useful

Properties of $\hat{\beta}$

There are several useful facts about our estimated model

1. We have that the regression line goes through the mean of the data:

$$\frac{1}{n}\sum_i (y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i) = 0$$

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

2. The sum of our estimated errors $\hat{u}$ is zero:

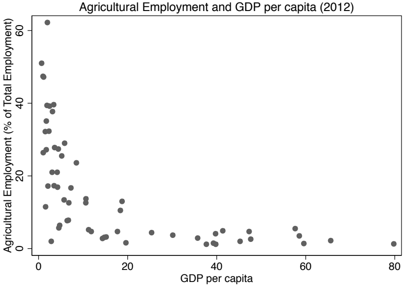
$$\frac{1}{n}\sum_i \hat{u}_i = 0$$

3. Every real data point will be equal to the predicted $\hat{y}$ plus the estimated residual $\hat{u}$

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

Goodness of fit: R^2

The R^2 is a useful measure of how well our model "fits" or explains the data. This can be informative about whether our specified model is close to the true relationship between two variables. For example:



If we fit a simple linear model to this line, it would be poor fit. The low R^2 would help indicate this fact.

Goodness of fit: R^2

Three main terms to define to understand the R^2 and how to calculate it:

1. Sum of Square Total (SST) $= \sum_i^n (y_i - \bar{y})^2$
   - Measure of the total variability of y in our sample data
2. Sum of Squares Explained (SSE) $=\sum_i^n (\hat{y}_i - \bar{y})^2$
   - This is a measure of the total variability of the predicted $\hat{y}$
3. Sum of Squared Residuals (SSR) $= \sum_i^n (y_i - \hat{y})^2$
   - This is a measure of the total variability of our error term ($\hat{u}_i$)

- Note that $SST = SSE + SSR$ (remember property 3)

Goodness of fit: R^2

- The R^2 is defined: $R^2 = \frac{SSE}{SST} = \frac{SSE}{SSE+SSR} = 1 - \frac{SSR}{SST}$
- You can think of the R^2 as how much of the *total* sample variation in y is explained by our model
- R^2 is always less than 1. Being closer to one indicates a better model fit
- Having a model that fits the data better does not necessarily mean it is a "good" model - what makes the model good or bad depends on what you want to use it for:
  - If you are interested in *prediction*, maximizing R^2 is somewhat sensible (but you don't want to "over-fit" the model)
  - If you are interested in *causality* or even just uncovering a relevant correlation, then R^2 should not be your primary concern

Goodness of fit: R^2

	GDP/cap (PPP$ 1000)	CO2/cap (tons)							
	x	y	(x-xbar)	(y-ybar)	(x-xbar)^2	(x-xbar)(y-ybar)	yhat	uhat^2	(y-ybar)^2
United States	43.7	19.3	18.5	10.0	342.9	184.5	14.84	19.9	99.3
Ireland	41.0	10.2	15.8	0.9	250.2	13.7	14.04	14.7	0.8
Belgium	33.5	9.7	8.3	0.4	69.2	3.0	11.81	4.4	0.1
Korea, Rep.	25.0	10.4	-0.2	1.1	0.0	-0.2	9.28	1.3	1.1
India	2.7	1.4	-22.5	-7.9	505.5	178.4	2.65	1.6	62.9
China	5.2	5.0	-20.0	-4.3	399.3	86.6	3.39	2.6	18.8
Sum	151.1	56.0	0.0	0.0	1567.1	466.1	56.0	44.5	183.1
Sample mean	25.2	9.3	0.0	0.0					
Sample variance/covariance					313.4	93.2			

1. What is the R^2?

Goodness of fit: R^2

	GDP/cap (PPP$ 1000) x	CO2/cap (tons) y	(x-xbar)	(y-ybar)	(x-xbar)^2	(x-xbar)(y-ybar)	yhat	uhat^2	(y-ybar)^2
United States	43.7	19.3	18.5	10.0	342.9	184.5	14.84	19.9	99.3
Ireland	41.0	10.2	15.8	0.9	250.2	13.7	14.04	14.7	0.8
Belgium	33.5	9.7	8.3	0.4	69.2	3.0	11.81	4.4	0.1
Korea, Rep.	25.0	10.4	-0.2	1.1	0.0	-0.2	9.28	1.3	1.1
India	2.7	1.4	-22.5	-7.9	505.5	178.4	2.65	1.6	62.9
China	5.2	5.0	-20.0	-4.3	399.3	86.6	3.39	2.6	18.8
Sum	151.1	56.0	0.0	0.0	1567.1	466.1	56.0	44.5	183.1
Sample mean	25.2	9.3	0.0	0.0					
Sample variance/covariance					313.4	93.2			

1. $R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{44.5}{183.1} = 0.757$