Lecture 4: Simple Regressions and Causality

Pierre Biscaye

Fall 2022

Conditional expectations and regression

- Last time we presented conditional expectations $E[Y|X]$ as a key way of describing a relationship between two variables
- If you know the conditional expectation, then for any $x \in X$, you can predict the average value of Y
- Unfortunately, we never know the true conditional expectation
- Linear regression gives us a way to estimate it

Estimating simple regression models

Last time we proposed a simple linear model for conditional expectations

$$E[Y|X = x] = \beta_0 + \beta_1 x \tag{1}$$

This implies that the relationship in the data is

$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{2}$$

We derived equations (using OLS) to estimate the β parameters for a particular sample

$$\hat{\beta_1} = \frac{\sum_i x_i y_i - N\bar{x}\bar{y}}{\sum_i x_i^2 - N\bar{x}^2} \tag{3}$$

$$\hat{\beta_0} = \bar{y} - \hat{\beta_1}\bar{x} \tag{4}$$

Excel example

Use these expressions

$$\hat{\beta_1} = \frac{\sum_i x_i y_i - N\bar{x}\bar{y}}{\sum_i x_i^2 - N\bar{x}^2} \tag{5}$$

$$\hat{\beta_0} = \bar{y} - \hat{\beta_1}\bar{x} \tag{6}$$

to recover $\hat{\beta}$ estimates for

$$\frac{CO_{2i}}{Pop_i} = \hat{\beta_0} + \hat{\beta_1}\frac{GDP_i}{Pop_i} + \hat{u}_i$$

using real data

Tying back to conditional expectations

We can use our estimates to generate *predicted values*

$$\widehat{E[y_i|X = x_i]} = \hat{y}_i = \hat{\beta_0} + \hat{\beta_1} x_i \tag{7}$$

And, we can predict residuals $\hat{u}$ (also sometimes labeled $\hat{\epsilon}$)

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta_0} - \hat{\beta_1} x_i \tag{8}$$

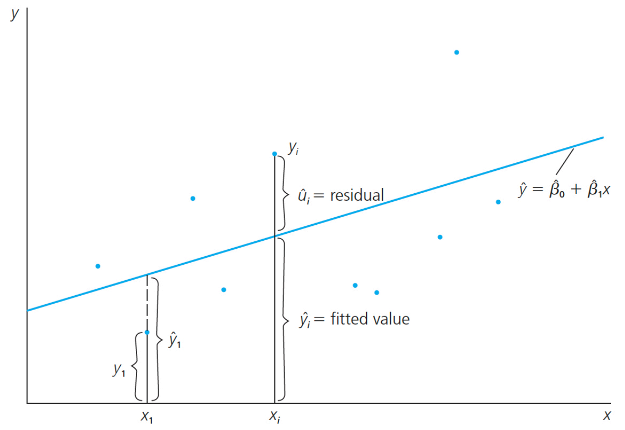
$\Rightarrow$ In Excel

We use the *residuals* to identify how well our "best fit" line fits

Predicted residuals, graphically

FIGURE 2.4

Fitted values and residuals.



How good is our "line of best fit"?

- Goal of ordinary least squares was to minimize the sum of squared residuals
- We use the *residuals* to identify how well our "best fit" line fits
- By construction, we know that $E[u] = 0$
- Because of this, the variance of the residuals is given by

$$s_u^2 = \frac{1}{n-2} \sum_i \hat{u}_i{}^2 \tag{9}$$

- How much total variation is there in the residuals? $SSR = \sum_i \hat{u}_i{}^2$
  - This is referred to as *SSR*, the Sum of Squared Residuals

Evaluating regression fit

- Our OLS objective was to minimize the Sum of Squared Residuals (SSR)
$$SSR = \sum_i (y_i - \hat{y}_i)^2 = \sum_i \hat{u}_i{}^2 = s_u^2 * (n-2) \tag{10}$$

- We also know the total variation in y, or Sum of Squares Total (SST)
$$SST = \sum_i (y_i - \bar{y})^2 = s_y^2 * (n-1) \tag{11}$$

- The Sum of Squares Explained (SSE) is
$$SSE = \sum_i (\hat{y}_i - \bar{y})^2 \tag{12}$$

- We can relate these three expressions through $SST = SSE + SSR$

How much of the variation in Y is explained by the model?

- How much of the variation in y (the SST) are we explaining with X?
- The R^2 calculates how much variation we explained

$$R^2 = 1 - \frac{\sum_i {\hat{u}_i}^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{SSR}{SST} \tag{13}$$

Or equivalently,

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = \frac{SSE}{SST} \tag{14}$$

R^2 example

- What is our R^2 for the example with CO_2 and GDP? $\Rightarrow$ *Excel*

- What does it mean for R^2 to be closer to 1?
- What does it mean for R^2 to be farther from 1 (closer to 0)?

What about *Causality*?

- We know how to estimate $\hat{\beta_1}$ for a given sample after assuming that $E[Y|X = x] = \beta_0 + \beta_1 x$
  - This is a Simple Linear Regression (SLR)
- When is $\hat{\beta_1}$ an estimate of the causal effect of X on Y? Five assumptions about the *population* are needed
- We need an *economic model*
  - Suppose the true, causal effect of a one unit increase in x on y is β_1 units
  - Then, in the population $y_i = \beta_0 + \beta_1 x_i + u_i$: *SLR1*

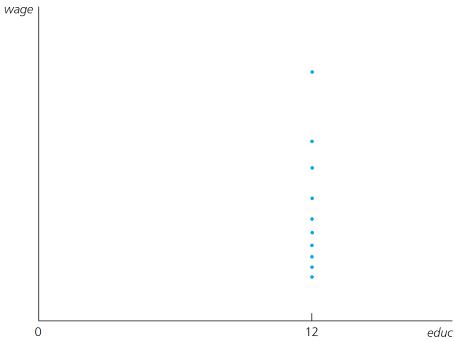
Is our Sample adequate?

- We need to make sure our sample reflects population characteristics
- Consider estimating the effect of education on wages: what issues might arise?
- *SLR2*: we have a *random* sample from the population of size n of variables x, y such that for each observation i we observe x_i, y_i.
- Under SLR1 and SLR2, in our sample we have: $y_i = \beta_0 + \beta_1 x_i + u_i$

No Variance in x_i

FIGURE 2.3

A scatterplot of wage against education when $educ_i = 12$ for all i.



- We want to estimate a slope: We need variance in x_i
- *SLR3*: The sample outcomes of x_i are not all the same value ($Var(x) \neq 0$)

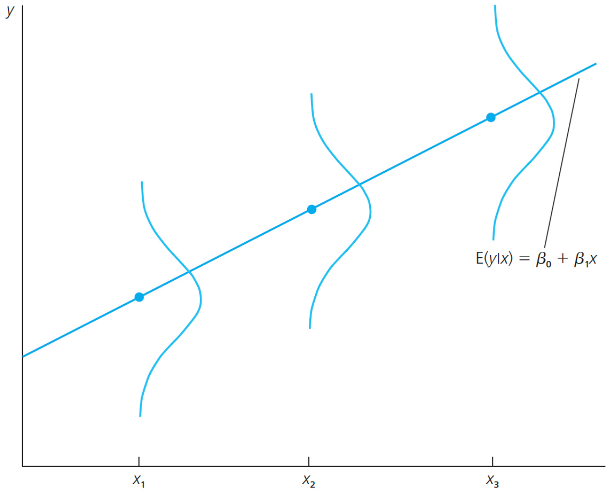
Independence of error term

- *SLR4*: The error u has an expected value of 0 given any value of the explanatory variable x: $E[u|x] = 0$
- Toughest assumption: nothing in u (that helps to explain Y) can be associated with X
- Remember, SLR 1: $y_i = \beta_0 + \beta_1 x_i + u_i$.
  - So, whenever x_i goes up by one unit, y_i goes up by β_1 units

$E[y|x]$ as a linear model, graphically

FIGURE 2.1

E(*y*|*x*) as a linear function of *x*.



Example: Section attendance and course grades

- Let's build intuition for why $E[u|x] = 0$ is a tough assumption
- Suppose we wanted to evaluate the effectiveness of course sections
- Suppose we had data on attendance (or video streaming)
- We want to estimate $Grade_i = \beta_0 + \beta_1 Section_i + u_i$
- What would it mean for $E[u_i|section_i] = 0$?

What if $E[u_i|x_i] \neq 0$?

- It *does not* mean that our population model is wrong.
- It *does* mean our estimates will be biased (will discuss this *Omitted Variables Bias* more later in the class)
- Recall, we got our estimating equations by minimizing

$$\sum_i (y_i - \beta_0 - \beta_1 x_i)^2 \tag{15}$$

which generated the FOCs

$$\sum_i (y_i - \hat{\beta_0} - \hat{\beta_1} x_i) = 0 \tag{16}$$

$$\sum_i x_i (y_i - \hat{\beta_0} - \hat{\beta_1} x_i) = 0 \tag{17}$$

What do the OLS FOCs imply?

FOC 1

$$\sum_i (y_i - \hat{\beta_0} - \hat{\beta_1} x_i) = 0$$

$$\frac{1}{n} \sum_i (y_i - \hat{y_i}) = 0$$

$$E[\hat{u}_i] = 0$$

FOC 2

$$\frac{1}{n} \sum_i x_i (y_i - \hat{\beta_0} - \hat{\beta_1} x_i) = 0$$

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

$$E[x \hat{u}] = cov(x, \hat{u}) = 0$$

- We found a $\hat{\beta_1}$ where we made sure that x_i is uncorrelated with $\hat{u}_i$

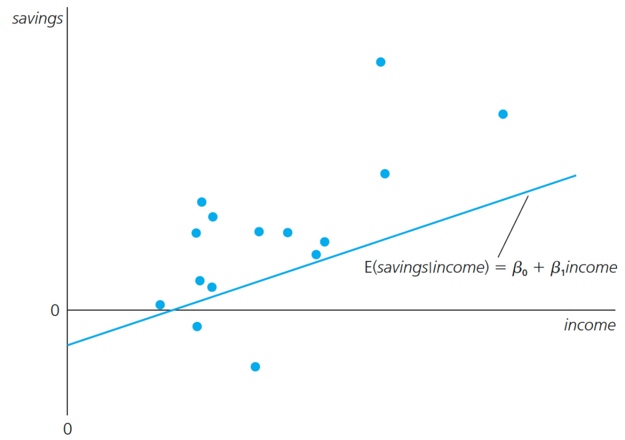
What does OLS ensure about $\hat{u}_i$?

- *Every time* you use OLS, you will find that:
  1. $E[\hat{u}_i] = 0$
  2. $cov(\hat{u}_i, x_i) = 0$
- But, nothing guarantees that these relationships hold in the population! These are just assumptions we made to generate our β estimates

$E[u_i|x_i] \neq 0$ in the population

FIGURE 2.2

Scatterplot of savings and income for 15 families, and the population regression $E(savings|income) = \beta_0 + \beta_1 income$.



SLR 4 and causality

- With OLS, we know that $\hat{u}$ will be uncorrelated with x
- When SLR 4 fails, u *is* correlated with x
- This means the one thing we are sure if is that we've estimated the wrong values for β_0 and β_1.
- This is the issue that leads to *correlation* and not *causation*
- OLS *always* estimates the best linear predictor of $y|x$. It only finds the *causal* estimate when $E[u|x] = 0$ (and the other 3 assumptions also hold).

$E[u_i|x_i] \neq 0$, Wooldridge example 2.12

- Suppose we want to estimate the effect of a federal free school lunch program (targeting low-income students) on student performance
- Have data on
  - *math*10 is share of 10th graders in a school that pass a standardized math exam
  - *lnchprg* is share of 10th graders in a school eligible for free school lunch
- Estimate SLR $math10 = \beta_0 + \beta_1 lnchprg + u$
- What sign do you expect for β_1?

$E[u_i|x_i] \neq 0$, Wooldridge example 2.12

- Have data on
  - *math*10 is percentage of 10th graders in a school that pass a standardized math exam
  - *lnchprg* is percentage of 10th graders in a school eligible for free school lunch
- Using data from Michigan high schools in 1992-1993 ($n = 408$), obtain

$$\widehat{math10} = 32.14 - 0.319 lnchprg$$

- How do we interpret $\widehat{\beta_1}$?
- What can explain this?

Example: CO2 and GDP

- Consider $CO_{2i} = \beta_0 + \beta_1 GDP_i + u_i$
- Do we think $E[u|GDP] = 0$?

SLR 4 summary

- A regression *always* identifies $\hat{\beta_0}$ and $\hat{\beta_1}$ so that

$$y_i = \hat{\beta_0} + \hat{\beta_1} x_i + \hat{u}_i \tag{18}$$

- and $\hat{u}_i$ is uncorrelated (mean-independent of) x_i.
- But, for $\hat{\beta_1} \approx \beta_1$ we need u_i to be uncorrelated with (mean-independent of) x_i in the population.
- We can *never* test this assumption: can only lay out a careful and considered defense of why it is reasonable

Theorem

Suppose SLR1-SLR4 all hold. Then

$$E[\hat{\beta_1}] = \beta_1$$
$$E[\hat{\beta_0}] = \beta_0$$

So we know that $\hat{\beta_1} \approx \beta_1$, and we can interpret the estimate as *causal*

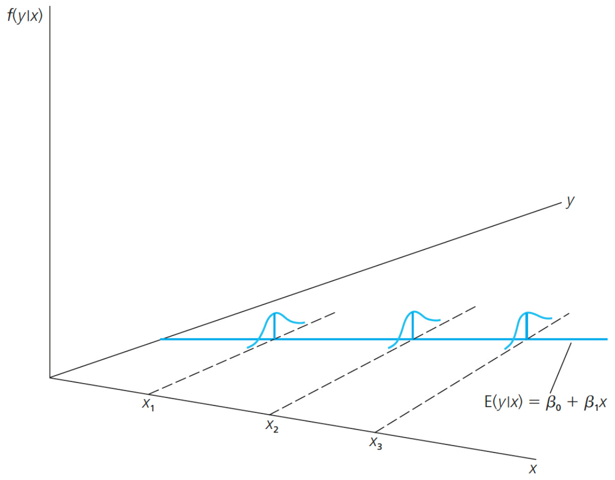
How variable is $\hat{\beta}_1$?

- How approximate is the β_1 estimate?
- Important to know: tells us ultimately how confident be that the true β_1 is within some particular range
- Need *SLR 5*: The error u has the same variance given any value of the explanatory variable.
- $var(u|x) = \sigma_u^2$
- SLR 5 is the *homoskedasticity* assumption

Homoskedastic errors

FIGURE 2.8

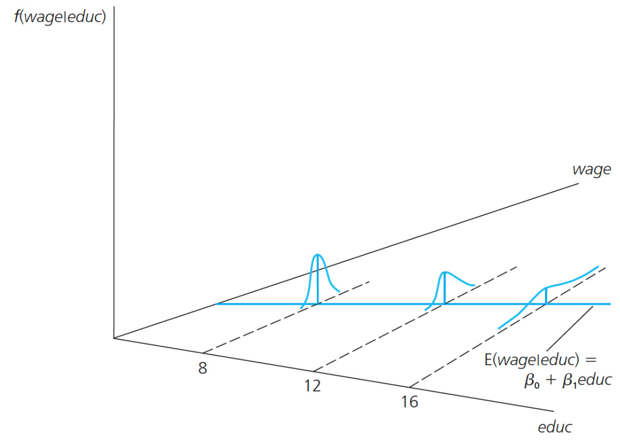
The simple regression model under homoskedasticity.



Heteroskedastic errors: wages and education

FIGURE 2.9

Var(*wage*|*educ*) increasing with *educ*.



Theorem 2

Theorem: suppose SLR1-SLR5 hold. Then

$$E[\hat{\beta_1}] = \beta_1$$

$$var(\hat{\beta_1}) = \frac{\sigma_u^2}{\sum_i (x_i - \bar{x})^2} = \frac{\sigma_u^2}{SST_x}$$

Estimating σ_u^2

- We don't observe u, so we can't observe σ_u^2.
- As with s_x^2, s_y^2, we can calculate the sample variance s_u^2
- Suppose SLR1-SLR5 hold, we can show that

$$E[s_u^2] = E[\frac{1}{n-2}\sum_i \hat{u}_i{}^2] = E[\frac{1}{n-2}SSR] = \sigma_u^2$$

Theorem

Suppose SLR1-SLR5 hold

$$E[\hat{\beta_0}] = \beta_0 \tag{19}$$

$$E[\hat{\beta_1}] = \beta_1 \tag{20}$$

$$\widehat{var(\hat{\beta_1})} = \frac{s_u^2}{\sum_i (x_i - \bar{x})^2} = \frac{SSR}{(n-2)SST_x} \tag{21}$$

Summary: Assumptions for causality in SLR

- SLR1: $y_i = \beta_0 + \beta_1 x_i + u_i$ in the population
- SLR2 : You have a random sample from the population
- SLR3: There is variation in x_i
- SLR4: $E[u|x] = 0$
- If SLR1-SLR 4 hold then $\hat{\beta_1} \approx \beta_1$
- SLR5: $var(u|x) = \sigma_u^2$
- If SLR5 also holds, then also get $\widehat{var(\hat{\beta_1})} = \frac{SSR}{(n-2)SST_x}$
    - But, can correct for this if SLR5 does not hold