Lecture 5: Linear Regression - Variance and Unbiasednes

Pierre Biscaye

Fall 2022

Review of Last Time

Assumptions for causality in SLR

- SLR1: $y_i = \beta_0 + \beta_1 x_i + u_i$
- SLR2: You have a random sample
- SLR3: there is variation in x_i
- SLR4: $E[u|x] = 0$
- If SLR1-SLR 4 hold then $\hat{\beta_1} \approx \beta_1$

These apply to the *population*: cannot test 1 and 4

Aside on functional forms and SLR1

SLR1: $y_i = \beta_0 + \beta_1 x_i + u_i$

- This is still fine if $y_i = \beta_0 + \beta_1 log(x_i) + u_i$ and for similar transformations.
  - After all, just consider $x_2 = log(x_1)$.
  - Then $y_i = \beta_0 + \beta_1 x_2 + u_i$ and SLR1 holds.
- Of course, interpretations change!
- And, you have to specify the *right* model.
  - If the population relationship is given by $y_i = \beta_0 + \beta_1 log(x_i) + u_i$, but you estimate $y_i = \hat{\beta_0} + \hat{\beta_1} x_i + u_i$, you will not recover the true β_1.

Samples and estimators

- For any random variable X, $E[X] = \mu$ and $Var(X) = E[(X - \mu)^2] = \sigma_x^2$ in the population
- In a sample, we don't observe μ or σ_x^2
- We *can* calculate
  - The sample mean $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} x_i$
  - The sample variance $s_x^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2$
- By the Law of Large Numbers, we stated that with a large enough random sample, $\bar{X} \approx \mu$ (and similarly, $s_x^2 \approx \sigma_x^2$)

Unbiased estimators

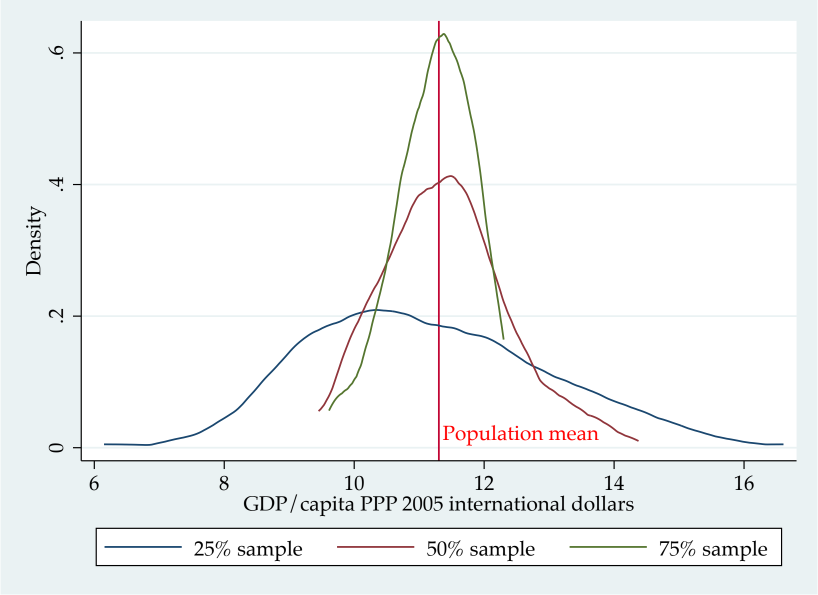
Consider $\bar{X}$. For a random sample,

$$E[\bar{X}] = E[\frac{1}{n}\sum_i x_i] = \frac{1}{n}\sum_i E[x_i] = \frac{1}{n}\sum_i E[x] = \frac{1}{n}n\mu = \mu$$

- Thus, $\bar{X}$ is an *unbiased* estimator of μ.
- However, in any random sample, $\bar{X} \neq \mu$.
- $\bar{X}$ is a random variable, with its own variance.

$$Var(\bar{X}) = Var(\frac{1}{n}\sum_i (x_i)) = \frac{1}{n^2}Var(\sum_i x_i) = \frac{1}{n^2}n * Var(x) = \frac{\sigma_x^2}{n} \quad (1)$$

Example: mean GDP/capita in sample of countries



Variance in the sample mean

$$Var(\bar{X}) = \frac{\sigma_x^2}{n} \tag{2}$$

So that

$$Std.Dev.(\bar{X}) = \sqrt{\frac{\sigma_x^2}{n}} = \frac{\sigma_x}{\sqrt{n}} \tag{3}$$

- More later, but this means that the sample mean will be within about $\frac{2\sigma_x}{\sqrt{n}}$ of μ.

Other estimators

- We don't know σ_x^2, but s_x^2 is an *unbiased* estimator of σ_x^2

$$E[s_x^2] = E[\frac{1}{n-1}\sum_i (x_i - \bar{X})^2] = \sigma_x^2 \tag{4}$$

- SLR1-SLR4: $E[\bar{y} - \hat{\beta_1}\bar{x}] = E[\hat{\beta_0}] = \beta_0$: $\hat{\beta_0}$ is an *unbiased* estimator of β_0
- SLR1-SLR4: $E[\frac{\widehat{cov(x,y)}}{\widehat{var(x)}}] = E[\hat{\beta_1}] = \beta_1$: $\hat{\beta_1}$ is an *unbiased* estimator of β_1

What about the variance of $\hat{\beta}_1$?

- Critical for knowing range of true β_1

To Jupyter!

Understanding variability of coefficient estimates

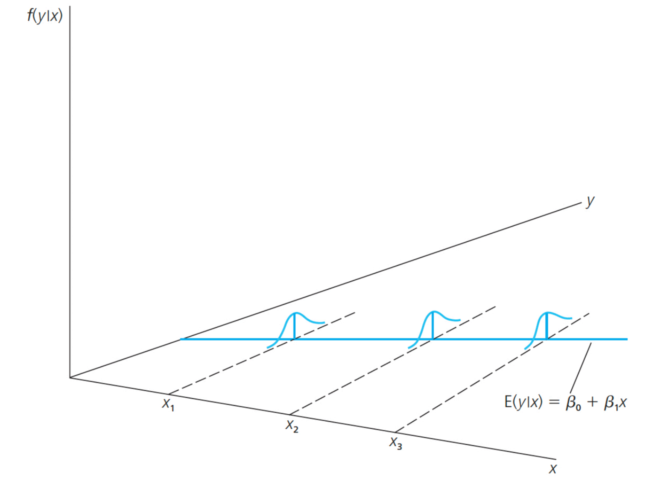
- We need one more assumption
- SLR5 (*homoskedasticity*): the error u has the same variance given any value of the explanatory variable

$$var(u|x) = \sigma_u^2 \tag{5}$$

Homoskedastic errors

FIGURE 2.8

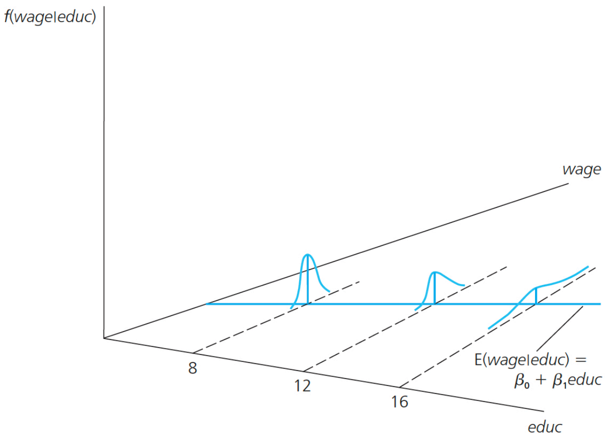
The simple regression model under homoskedasticity.



Heteroskedastic errors: wages and education

FIGURE 2.9

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



Theorem

Theorem: suppose SLR1-SLR5 hold. Then

$$E[\hat{\beta_1}] = \beta_1 \tag{6}$$

$$var(\hat{\beta_1}) = \frac{\sigma_u^2}{\sum_i (x_i - \bar{x})^2} = \frac{\sigma_u^2}{SST_x} \tag{7}$$

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

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

- s_u^2 is an *unbiased* estimator for σ_u^2
- If SLR1-SLR5 hold, we therefore have

$$\widehat{var(\hat{\beta}_1)} = \frac{s_u^2}{SST_x} = \frac{SSR}{(n-2)\sum_i (x_i - \bar{X})^2} \tag{8}$$

Estimator variance and standard errors

- Variances are not always intuitive to interpret or as useful for inference.
- Instead will typically look at *standard errors* (SE): what we call the standard deviation for estimated values

$$SE(\hat{\beta_1}) = \widehat{Std.Dev.}(\hat{\beta_1}) = \sqrt{\widehat{var}(\hat{\beta_1})} = \frac{s_u}{\sqrt{SST_x}} \tag{9}$$

- Aside: SLR5 is the least important of the assumptions
  - Can easily estimate $\widehat{var(\hat{\beta_1})}$ in R even if it fails
- Back to Jupyter!

Pulling all of this together

- We have an estimator $\hat{\beta_1}$
- if SLR1-SLR4 hold $E[\hat{\beta_1}] = \beta_1$
  - If we drew many random samples, and calculated $\hat{\beta_1}$ in each of them, on average $\bar{\hat{\beta_1}} \approx \beta_1$
- In any individual random sample, $\hat{\beta_1} \neq \beta_1$
- How close it is will depend on (if SLR1-SLR5 hold)

$$\widehat{var(\hat{\beta_1})} = \frac{SSR}{(n-2)\sum_i (x_i - \bar{X})^2} \tag{10}$$

To Jupyter!

Assessing the variance of $\hat{\beta}_1$

$$\widehat{var(\hat{\beta_1})} = \frac{SSR}{(n-2)\sum_i (x_i - \bar{X})^2} \tag{11}$$

- Variance will be small when:
  - n is large
  - $\sigma_x^2 \approx s_x^2 = \frac{1}{n-1}\sum_i (x_i - \bar{X})^2$ is large
  - SSR is small
- How to reduce SSR?

Multiple Linear Regression

- The goal of reducing SSR, and concerns about SLR4, motivate *multiple linear regression* (MLR)
- Example: Education and wages. Suppose we started from the simple linear regression model

$$ln(wage_i) = \beta_0 + \beta_1 Educ_i + u_i \tag{12}$$

- What do we need to assume for $E[\hat{\beta_1}] = \beta_1$?

Omitting a variable

- Suppose the true model is

$$ln(wage_i) = \beta_0 + \beta_1 Educ_i + \beta_2 Exper_i + \epsilon_i \tag{13}$$

- What happens if we instead estimate the simple regression model?

$$ln(wage_i) = \beta_0 + \beta_1 Educ_i + u_i \tag{14}$$

Conditional expectations interpretation

If we estimate

$$ln(wage_i) = \beta_0 + \beta_1 Educ_i + u_i$$

We are estimating

$$\begin{aligned} E[ln(wage_i)|Educ_i] &= \beta_0 + \beta_1 Educ_i + E[u_i|Educ_i] \\ &= \beta_0 + \beta_1 Educ_i + \beta_2 E[exper_i|Educ_i] + E[\epsilon_i|Educ_i] \end{aligned}$$

Quantifying bias

Suppose

$$E[Exper|Educ_i] = \delta_0 + \delta_1 Educ_i \tag{15}$$

Then

$$E[ln(wage_i)|Educ_i] = \beta_0 + \beta_2\delta_0 + (\beta_1 + \beta_2\delta_1)Educ_i + E[\epsilon_i|Educ_i] \tag{16}$$

- Our line of best fit will find $\hat{\beta_1} \approx \beta_1 + \beta_2\delta_1$!
- This is what is called *omitted variables bias*: $E[\hat{\beta_1}] - \beta_1 = \beta_2\delta_1$ (in this case - more on this in future lecture)

Multiple regression

Suppose instead we estimate

$$ln(wage_i) = \beta_0 + \beta_1 Ed_i + \beta_2 Exper_i + u_i \tag{17}$$

- Experience is no longer in u
- Interpretations change: How does Education relate to wages *holding experience constant* (also referred to as "*ceteris paribus*")
- Or, compare two people with the same amount of experience. If one has one more year of education, how much more do they earn? β_1