Simple Regression Model: Part 1

October 23, 2019

Ordinary Least Squares

- Last class, we looked at some ways for measuring relationships between two variables (correlation, covariance, scatter plot, bar plot)
- But what if you want to say that the effect of x on y is equal to some number? How do you estimate this effect?
- For the next couple of classes, we will study the **simple linear regression** (also known as two-variable regression model):

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

- This model will allow us to make statements about the effect of one variable on another

Ordinary Least Squares

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

- Terminology: **We regress y on x**
- y is the **dependent variable**, **explained variable**, or **the regressand**
- x is the **independent variable**, **explanatory variable**, or **the regressor**
- u (u for "unobservable") is the **error term.** The error term captures everything other than x that affects y

Ordinary Least Squares

Our model is:

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

Assume x changes by an amount Δx, while u stays the same. What is the change in y?

$$y_{new} = \beta_0 + \beta_1 (x + \Delta x) + u$$

So,

$$\Delta y = y_{new} - y = \beta_1 \Delta x$$

- $\beta_1 = \frac{\Delta y}{\Delta x}$ is the **slope parameter**, the effect of a change in x on y
- β_0 is the **intercept parameter** or the **constant term**. Let's ignore it for now

Ordinary Least Squares

EXAMPLE 2.2 **A SIMPLE WAGE EQUATION**

A model relating a person's wage to observed education and other unobserved factors is

$$wage = \beta_0 + \beta_1 educ + u. \qquad [2.4]$$

If *wage* is measured in dollars per hour and *educ* is years of education, then β_1 measures the change in hourly wage given another year of education, holding all other factors fixed. Some of those factors include labor force experience, innate ability, tenure with current employer, work ethic, and numerous other things.

Notice we assume that the effect of education does not depend on the level of education

We will look at regression models that relax this assumption later

Assumptions

$$E(u) = 0$$

- Notice this assumption is a very simple assumption to make
- Assume $y = \beta_0 + \beta_1 x + u$ and $E(u) = \mu \neq 0$
- Then $y = \beta_0 + \mu + \beta_1 x + u - \mu$ (we just added and subtracted μ)
- Let $\beta_0' = \beta_0 + \mu$ and $u' = u - \mu$
- Our model can be re-written as:

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

- In this re-written version, $E(u') = 0$!

Assumptions (the important one!!!!!!)

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

- The conditional mean of u must be independent of x
- Let's go back to our wage example:

$$wage = \beta_0 + \beta_1 education + u$$

- In the wage example, u might contain ability
- Our assumption requires that $E(ability|education)$ is independent of *education*
- This is what we need in order to interpret β_1 as the effect of education on wages
- Otherwise, β_1 will also capture the effects of other variables

Exercise

Let *kids* denote the number of children ever born to a woman, and let *educ* denote years of education for the woman. A simple model relating fertility to years of education is

$$kids = \beta_0 + \beta_1 educ + u,$$

where u is the unobserved error.

(i) What kinds of factors are contained in u? Are these likely to be correlated with level of education?

(ii) Will a simple regression analysis uncover the ceteris paribus effect of education on fertility? Explain.

Assumptions

- We already assumed $E(u) = 0$ (this assumption comes for free)
- Combined with $E(u|x) = E(u)$, we get $E(u|x) = 0$
- We will call this assumption the **zero conditional mean assumption**
- Notice that

$$\begin{aligned} E(y|x) &= E(\beta_0 + \beta_1 x + u|x) = \\ &= \beta_0 + \beta_1 x + E(u|x) = \\ &= \beta_0 + \beta_1 x \end{aligned}$$

Ordinary Least Squares

- Under our assumptions, $E(y|x) = \beta_0 + \beta_1 x$
- This is our **population regression function**
- **Interpretation:**
  - Before, we were thinking of β_1 as the effect of x on y holding u fixed
  - But u is never fixed
  - Even if $E(u|x) = 0$, sometimes u will be positive, and sometimes it will be negative: **it's a random error!**
  - The right way to think of β_1 is the effect of x on the expected value of y

Visualizing the population regression function