Week 1 - Introduction to Regression and Simple Linear Regression

Written by /u/econpanda

Problems with a * are not necessary but may provide additional insight. The readings for this problem set are

- Chapter 1
- 2.1, 2.2, 2.4, 2.6

Pay attention to the following key topics

- Meaning of **ceteris paribus**
- Examples 1.3, 1.4, 1.5, 1.6
- Problems with nonrandom assignment (pages 14-15)
- What the regression error term u captures
- Assumptions about relation between x and u
- Estimation of regression coefficients
- Interpretation of regression coefficients
- How $\log(\cdot)$ changes interpretation of regression coefficients (Table 2.3)
- Regression assumptions (section 2.3) and properties of OLS estimators (section 2.5) will be covered next week

1. (Wooldridge 1.1) Suppose that you are asked to conduct a study to determine whether smaller class sizes lead to improved performance of fourth graders.[1]
   (a) If you could conduct any experiment you want, what would you do?
   (b) More realistically, suppose you can collect observational data on several thousand fourth graders in a given state. You can obtain the size of their fourth-grade class and a standardized test score taken at the end of fourth grade. Why might you expect a negative correlation between class size and test score?
   (c) Would a negative correlation necessarily show that smaller class sizes cause better performance?

[1] For a good answer to this question see the Tennessee STAR experiment and Krueger (1999)

2. (Wooldridge 2.1) Let *kids* denote the number of children born to a woman, and let *educ* denote years of education for the woman. A simple model relating education to fertility to years of education is

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

   (a) List 5 specific variables that are in u

   (b) Are any of these things likely to be correlated with *educ*?

   (c) Would this simple regression uncover the ceteris paribus effect of education on fertility (Is $E(u|x)$ likely to hold)?

3. You are interested in finding the relation between time allowed for college students to take an exam and their performance on the exam. You notice that at your university a class is offered on MWF (for 50 minutes) and on TTH (for 75 minutes) and it is taught by the same professor that uses the same exam.

   (a) Is a simple regression of time on test score likely to uncover a ceteris paribus effect of time on test score (Hint: Are students randomly assigned between classes? Is $E(u|x)$ likely to hold)?

   (b) Alternatively you can convince the professor to pool the sections for exams and flip a coin for each student to determine their time allotment, heads means they get 75 minutes and tails means they get 50 minutes. Would this approach uncover a ceteris paribus effect?

4. * Wooldridge derives OLS through the method of moments estimator, an alternative way to estimate $\hat{\beta}_0$ and $\hat{\beta}_1$ is through minimizing the sum of squared residuals. Define the residuals as $\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i,1}$, the objective function is:

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

   (a) Show that the derivatives of this function with respect to β_0 and β_1 are[2]

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

   (b) Set these equations equal to 0 and solve for $\hat{\beta}_0$ and $\hat{\beta}_1$ and show that they are equivalent to equations 2.17 and 2.19 from Wooldridge. You will need the following properties

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

[2] If you need a review of calculus and summation operators see Appendix A in Wooldridge

5. (Wooldridge 2.2) In the simple linear regression model $y = \beta_0 + \beta_1 x_i$, suppose that $E(u) = \alpha_o \neq 0$. Show that the model can always be written with the same slope, but a new intercept and new error, where the new error has zero mean.

6. For each of the following regressions on the relation between a persons high school GPA and the ACT score provide a general interpretation of β_1

   (a) $GPA = \beta_0 + \beta_1 ACT + u$

   (b) $\log(GPA) = \beta_0 + \beta_1 ACT + u$

   (c) $GPA = \beta_0 + \beta_1 \log(ACT) + u$

   (d) $\log(GPA) = \beta_0 + \beta_1 \log(ACT) + u$

References

Krueger, A. B. (1999). Experimental estimates of education production functions*. *The Quarterly journal of Economics 114*(2), 497–532.