

Empirical Exercises for Chapter 5

In this exercise you will investigate the relationship between the number of completed years of education for young adults and the distance from their high school to the nearest four-year college.

You will examine this question using data on a random sample of high school seniors, contained in the data file **CollegeDistance** (in Excel and Stata formats). These same students were interviewed as high school seniors in 1980, then reinterviewed in 1986 to determine how many years of education they had completed. A detailed description is given in **CollegeDistance_Description**, available on the Web site.

Use these data to answer the following questions.

1. Run a regression of years of completed education (ED) on distance to the nearest college ($Dist$). What is the estimated slope?
2. Run a regression of ED on $Dist$, but include some additional regressors to control for characteristics of the student, the student's family, and the local labor market. In particular include as additional regressors *Female*, *Black*, and *Hispanic*, *Incomehi*, *Ownhome*, *DadColl*, *Cue80* and *Stwmfg80*. What is the estimated effect of $Dist$ on ED ? Construct a 95% confidence interval for the coefficient on $Dist$ in the regression.
3. Is the estimated effect of $Dist$ on ED in the regression in (2) substantively different from the regression in (1)? Based on this, does the regression in (1) seem to suffer from important omitted variable bias?
4.
 - a. Bob is a black male. His high school was 20 miles from the nearest college. His base year composite test score (*Bytest*) was 58. His family income in 1980 was \$26,000, and his family owned a home. His mother had attended college, but his father had not. The unemployment rate in his county was 7.5%, and the state average manufacturing hourly wage was \$9.75. Predict Bob's years of completed schooling using the regression in (2).
 - b. Jim has the same characteristics as Bob except that his high school was 40 miles from the nearest college. Predict Jim's years of completed schooling using the regression in (2).
5. Compare the fit of the regression in (1) and (2) using the regression standard errors, R^2 and \bar{R}^2 . Why are the R^2 and \bar{R}^2 so similar in regression (2)?
6. The value of the coefficient on *DadColl* is positive and statistically significant. What does this coefficient measure?
7. Explain why *Cue80* and *Swmfg80* appear in the regression. Are the signs of their estimated coefficients (+ or -) what you would have guessed? Interpret the magnitudes of these coefficients.

8. The dataset contains two other variables, *Urban* and *tuition*. Explain why these variables might be important omitted variables. Include these variables in the regression. Are their coefficients statistically significant when tested one at a time? Are their coefficients statistically significant when tested jointly?