Econ 322: Econometrics

Data Analysis Project: Estimation and Inference

Due Date: 23:00, 11/13/2023, EST

Read the following instructions carefully.

- You can submit your reports unlimited times before the deadline. Do wait until the last minute and tell me that you have an internet problem.
- Report the required results in a word or PDF document and submit it **together with your code**. In the report, make your answers concise and to the point. My TA has a lot of submissions to grade. So make sure that your answers are easy to spot, in case my TA may miss them.
- Name your submitted files in the following format: initial first_full last name_Project (so that we know this is your answer to the big assignment).
- You can refer to the sample codes I gave in my lecture notes and the uploaded Python codes.
- Report the numbers as how they appear in the regression output unless instructed otherwise. Report the predicated value to the same precision as the coefficients.

Series in Data Set

Name	Desrciption
ed	Years of Education Completed (See below)
female	1 = Female/0 = Male
black	1 = Black/0 = Not-Black
Hispanic	1 = Hispanic/0 = Not-Hispanic
bytest	Base Year Composite Test Score. (These are achievement tests given to high school seniors in the sample)
dadcoll	1 = Father is a College Graduate/ 0 = Father is not a College Graduate
momcoll	1 = Mother is a College Graduate/ 0 = Mother is not a College Graduate
incomehi	1 = Family Income > $25,000 per year/ 0 = Income ≤ $25,000 per year.
ownhome	1= Family Owns Home / 0 = Family Does not Own Home
urban	1 = School in Urban Area / = School not in Urban Area
cue80	County Unempolyment rate in 1980
stwmfg80	State Hourly Wage in Manufacturing in 1980
dist	Distance from 4yr College in 10's of miles
tuition	Avg. State 4yr College Tuition in $1000's

The data file **"CollegeDistance.csv"** contains data from a random sample of high school seniors interviewed in 1980 and re-interviewed in 1986. In this exercise, you will

use these data to investigate the relationship between the number of completed years of education for young adults and the distance from each student's high school to the nearest four-year college. (Proximity to college lowers the cost of education, so that students who live closer to a four-year college should, on average, complete more years of higher education.) A detailed description is given below:

You can continue to work on the .py file in Spyder or the .ipynb file in Juptyer Notebook (or Jupyter Lab).

Q1 **(50 points)**. Run a regression of years of completed education (*ED*) on distance to the nearest college (*Dist*), where Dist is measured in tens of miles. (For example, Dist = 2 means that the distance is 20 miles.)

(a) What is the estimated intercept? What is the estimated slope? Use the estimated regression to answer this question: How does the average value of years of completed schooling change when colleges are built close to where students go to high school?

(b) Bob's high school was 20 miles from the nearest college. Predict Bob's years of completed education using the estimated regression. How would the prediction change if Bob lived 10 miles from the nearest college?

(c) Does distance to college explain a large fraction of the variance in educational attainment across individuals? Explain.

(d) Is the estimated regression slope coefficient statistically significant? That is, can you reject the null hypothesis $H_0 : \beta_1 = 0$ versus a two-sided alternative at the 10%, 5%, or 1% significance level? What is the p-value associated with coefficient's t-statistic?

(e) Construct a 95% confidence interval for the slope coefficient.

(f) An education advocacy group argues that, on average, a person's educational attainment would increase by approximately 0.15 year if distance to the nearest college is decreased by 20 miles. Is the advocacy groups' claim consistent with the estimated regression? Explain.

Q2 **(50 points)**. 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 *Bytest*, *Female*, *Black*, *Hispanic*, *Incomehi*, *Ownhome*, *DadColl*, *Cue80*, and *Stwmfg80*. What is the estimated effect of *Dist* on *ED*?

(a) Is the estimated effect of *Dist* on *ED* in the regression in Q2 substantively different from the regression in Q1? Based on this, does the regression in Q1 seem to suffer from important omitted variable bias?

(b) The value of the coefficient on *DadColl* is positive. What does this coefficient measure?

(c) Explain why *Cue80* and *Swmfg80* appear in the regression. Are the signs of their estimated coefficients (+ or -) what you would have believed? Interpret the magnitudes of these coefficients.

(d) Bob is a black male. His high school was 20 miles from the nearest college. His baseyear composite test score (*Bytest*) was 58. His family income in 1980 was $26,000, and his family owned a home. His mother attended college, but his father did 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 Q2.

(e) 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 Q2.

(f) It has been argued that, controlling for other factors, blacks and Hispanics complete more college than whites. Is this result consistent with the regressions in Q2?