ECON 308: Econometrics
Assignment 2

Complete each problem to the best of your ability and turn them in on Thursday, October 7. You are encouraged to collaborate with other students, but you should turn in the problem solutions individually. Write legibly, denote your final answers clearly, and show all of your work. Your complete assignment should include written/typed responses to the questions along with a printout of the code you ran to generate them (your do-file) and any graphs produced.[1]

1. Suppose the population regression function is $y = \beta_0 + \beta_1 x + \epsilon$, where y is an indicator for whether or not a child eventually attends college and x is the child's score on a 5th grade academic assessment (which ranges from 0 to 100, with mean μ and variance σ^2). Assume the Gauss-Markov assumptions hold.

   (a) Write the expression for $\mathrm{E}\left[\hat{\beta}_1\right]$ in the estimated regression of y on x.

   (b) What is the interpretation for $\hat{\beta}_1$?

   (c) Suppose that, instead of using x in the regression, we use z, defined as $z = \frac{x-\mu}{\sigma}$. What is the expression for $\widetilde{\beta}_1$ now? How do we interpret its magnitude?

   (d) Show that $\mathrm{E}\left[\widetilde{\beta}_1\right] = \sigma\,\mathrm{E}\left[\hat{\beta}_1\right]$.

2. Omitted variable bias: Suppose you are given data on the variables y, x_1, and x_2. Suppose that $y = 1 + 2x_1 - 4x_2 + \epsilon$, where $\mathrm{Cov}\left(x_1, \epsilon\right) = \mathrm{Cov}\left(x_2, \epsilon\right) = 0$ and $\mathrm{Var}\left(\epsilon\right) = \sigma^2$.

   (a) Find $\mathrm{Cov}\left(y, \epsilon\right)$. (Hint: Plug in the equation for y and use the rules for covariances, in particular the fact that $\mathrm{Cov}\left(U + W, Z\right) = \mathrm{Cov}\left(U, Z\right) + \mathrm{Cov}\left(W, Z\right)$ for any random variables U, W, Z.)

   (b) Suppose we attempt to estimate $y = \beta_0 + \beta_1 x_1 + \tilde{\epsilon}$. Write an expression for $\tilde{\epsilon}$ in terms of ϵ and x_2.

   (c) What is $\mathrm{Cov}\left(x_1, \tilde{\epsilon}\right)$? Under what condition will x_1 *not* be correlated with the error term $\tilde{\epsilon}$ in this case?

3. Suppose you wish to estimate a linear regression of Y on X_1 and X_2. Write a model describing each of the following scenarios, and fill in actual numbers which match the patterns described.

   (a) Write a model where Y increases at an increasing rate with X_1 and a constant rate with X_2.

   (b) Write a model where Y increases with X_2, but the increase is larger at smaller values of X_1.

   (c) Write a model where Y increases at a constant rate with X_1 and X_2, but X_2 only enters the model if X_1 is greater than 5.

   (d) Write a model where the coefficients on both X_1 and X_2 can be interpreted as elasticities.

[1] Example code from our lectures is available on Blackboard. If you need help understanding the syntax for any particular command, the `help` command in Stata can give you information and examples for each. There are also a great deal of online resources for Stata, and most questions can be resolved with a simple online search of the form "how do I do [x] in Stata." Here is an additional resource that may be helpful: `https://www.ssc.wisc.edu/sscc/pubs/sfs/home.htm`.

4. This exercise explores data from a sex discrimination investigation involving the University of California, Berkeley. The data contains information on the number of applicants to different Berkeley graduate programs in 1973, broken down by gender and acceptance/rejection.

   (a) from the Data folder on Blackboard.

   (b) Load the dataset `berkeley.dta`. How many departments are represented in the dataset? You can use the `tabulate` or `distinct` commands to determine this.

   (c) Generate an indicator variable which equals 1 if the applicant is female and zero otherwise.

   (d) What share of applicants were women?

   (e) Generate an indicator variable which equals 1 if the applicant was admitted.

   (f) Find out what share of women are admitted by computing the sample mean.

   (g) Regress the admission status indicator on the female indicator and interpret the estimated coefficient; are women accepted at a lower or higher rate than men?

   (h) Generate indicator variables for each department and re-estimate the above regression separately for each individual department. How does the admissions rate for women vs. men compare by department?

   (i) Estimate the overall acceptance rate for each department.

   (j) Estimate the share of applicants that are female for each department. How can this help explain the previous results?

5. In this exercise, you will create and analyze data on U.S. cities pertaining to education, income, and income inequality. We will rank cities from most to least educated, examine the correlation between city size and income, and see how these relate to inequality.

   (a) To do this, you will need estimates (at the city level) of the share of individuals with different levels of education, as well as measures of income and inequality. Since you only need the aggregate data (rather than the microdata), you will use NHGIS. Select the most recent 5-year binned ACS data (2015-2019).

   (b) Now, decide on the level of geography you would like to use. Since we're interested in "cities," tract- or block-level data are unnecessary. Choose either Census place-level or CBSA-level data.

   (c) Select the variable best representing the education data you need - it's generally sensible to focus on those above a certain age (say, 25).

   (d) Select the measure of income you prefer - median household or per capita would probably be best. Also, select the Gini index of income inequality. Download the data in fixed-width format.

   (e) The income inequality measure will be in a separate dataset from the rest of your data. You can use the `merge` command to join these two data sets.

   (f) Convert the education data into a usable variable for measuring "most educated." You must decide on how you will measure "most educated" - fraction with a BA or higher? Graduate degree? Something else?

   (g) You can additionally decide if you want to restrict your analysis to larger cities. Why might you want (or not want) to impose such a restriction?

   (h) Once you've generated a measure of education level for the city, use the `sort` command on this variable to put the list in order. Comment on the pattern you observe; what types of cities seem to be the most/least educated? (You do NOT need to print out the whole list for me.)

   (i) Now, you'll examine the relationship between city size and your measure of residents' incomes. Regress income on city population. What relationship do you observe?

   (j) Repeat the previous regression, but replace income with the Gini index. Do larger cities exhibit more or less inequality?

   (k) Now, regress the Gini index on income. Are richer cities associated with more or less inequality?

   (l) Include a table in your writeup that reports the results of the previous three regressions (using the `estout` package).