Problem Set 1

Topics in Advanced Econometrics (ResEcon 703)
University of Massachusetts Amherst

Due: September 29, 8:30 am ET

Rules

Email a single .pdf file of your problem set writeup, code, and output to `mwoerman@umass.edu` by the date and time above. You may work in groups of up to three and submit one writeup for the group, and I strongly encourage you to do so. You can use any "canned" routine (e.g., `lm()`) for this problem set.

Data

Download the file `commute_datasets.zip` from the course website. This zipped file contains two datasets—`commute_binary.csv` and `commute_multinomial.csv`—but you will only use the dataset `commute_binary.csv` for this problem set. Both datasets contain simulated data on the travel mode choice of 1000 UMass graduate students who commute to campus from more than one mile away. The `commute_binary.csv` dataset corresponds to commuting in the middle of winter when only driving a car or taking a bus are feasible options. The `commute_multinomial.csv` dataset corresponds to commuting in the spring when riding a bike and walking are feasible alternatives. See the file `commute_descriptions.txt` for descriptions of the variables in each dataset.

Problem 1: Summary Statistics

We are studying how UMass graduate students choose how to commute to campus during winter when only driving a car or taking a bus are feasible options—assume the weather is too severe for even the heartiest graduate students to ride a bike or walk. This information will help the university to plan for parking and bus needs during the winter months. Use the `commute_binary.csv` dataset for this problem. (Reminder: the `read_csv()` function from the `tidyverse` package reads a .csv file into memory.)

a. The three factors that are mostly likely to influence the choice to drive or take the bus are the cost of driving (`cost.car`), the time to drive (`time.car`), and the time to take the bus (`time.bus`); the bus is free for all students. For each of these three important variables, calculate the mean and median for the full sample. (Reminder: the `summarize()` function is helpful when calculating these kinds of summary statistics.)

    i. Report these means and medians for the full sample.

ii. Do you see any patterns in these summary statistics that might be useful to keep in mind when conducting your analysis? If so, briefly describe these patterns.

b. The students who choose to drive and the students who choose to take the bus might face different choice settings—that is, different times and costs. Calculate the same summary statistics separately for the students who drive and for the students who take the bus. (Reminder: the `group_by()` function allows you to perform the same calculations on different subsamples of the data.)

i. Report these means and medians for each group of students.

ii. Do you see any patterns that might begin to explain why some students drive and some students take the bus? If so, briefly describe these patterns.

Problem 2: Linear Probability Model

The summary statistics can provide some initial suggestive evidence for how UMass graduate students choose how to commute to campus during winter, but estimating a model of decision making can yield more robust conclusions. Again use the `commute_binary.csv` dataset for this problem.

a. Model the choice to drive to campus during winter as a linear probability model. Include the cost of driving and the time of each alternative as independent variables in your model:

$$Y_n = \beta_0 + \beta_1 C_{nc} + \beta_2 T_{nc} + \beta_3 T_{nb} + \varepsilon_n$$

where Y_n is a binary indicator if student n drives, C_{nc} is the cost to student n of driving, T_{nc} is the time for student n to drive, T_{nb} is the time for student n to take the bus, and the β coefficients are to be estimated. (Reminder: the `lm()` function estimates an OLS regression model.)

i. Report the estimated coefficients and heteroskedastic-robust standard errors from this model. Briefly interpret these results. For example, what does each coefficient mean? (Reminder: the `coeftest()` function from the `lmtest` package tests the statistical significance of your coefficient estimates, and the `vcovHC()` function from the `sandwich` package estimates the heteroskedastic-robust covariance matrix of coefficient estimates.)

ii. One potential problem with a linear probability model is that predicted probabilities can fall outside the $[0, 1]$ range. How many students have infeasible choice probabilities? Given these results, are you worried about using a linear probability model in this case? (Reminder: the `predict()` function calculates fitted values of an `lm` regression.)

iii. Test if the two time coefficients are equal in absolute value. Interpret the result of this test and briefly explain why it could make intuitive sense. If a delay were to increase equally the time to drive and the time to take the bus, would you expect the proportion of drivers to increase, decrease, or stay the same? (Hint: There are many ways to conduct this Wald test. I like the `linearHypothesis()` function from the `car` (companion to applied regression) package. You may need to use the help file or a Google search to learn how to use this function.)