Problem Set 1

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

Due: September 24, 11:30 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()`, `glm()`, and `mlogit()`) 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`—that you will use 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—assume the weather is too severe for even the heartiest graduate students to ride a bike or walk this distance. 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: Linear Probability Model

Use the `commute_binary.csv` dataset for this question. (Reminder: the `read_csv()` function from the `tidyverse` package reads a .csv file into memory.)

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.)

Problem 2: Binary Logit Model

Use the `commute_binary.csv` dataset for this question.

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

$$\ln\left(\frac{P_n}{1-P_n}\right) = \beta_0 + \beta_1 C_{nc} + \beta_2 T_{nc} + \beta_3 T_{nb}$$

where P_n is the probability that 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 `glm()` function with argument `family = 'binomial'` estimates a binary logit model.)

i. Report the estimated coefficients and standard errors from this model. Briefly interpret these results. For example, what does each coefficient mean? (Reminder: the `summary()` function summarize the results of a `glm` model.)

ii. Calculate the marginal effect of each independent variable for each student; that is, 3 variables × 1000 students = 3000 marginal effects. For each of these three variables, report the mean, minimum, maximum, and quartiles of its marginal effects. Compare these marginal effects to your estimates in problem 1. (Reminder: the `predict()` function calculates fitted values of a `glm` model, and the `summary()` function reports these summary statistics for a vector or data frame.)

iii. Use your coefficient estimates to calculate the dollar value that a student places on each hour spent driving and on each hour spent on the bus. (Hint: think about how to use your coefficient estimates to convert a student's time to money.)

b. Demographic information might affect a student's commute decision or underlying preferences. For example, students with different incomes might have different sensitivities to cost. Again model the choice to drive to campus during winter as a binary logit model, but now allow the parameter on cost to vary inversely with income:

$$\ln\left(\frac{P_n}{1-P_n}\right) = \beta_0 + \frac{\beta_1}{I_n}C_{nc} + \beta_2 T_{nc} + \beta_3 T_{nb}$$

where I_n is the income of student n. (Reminder: the `I()` function allows you to include math inside a `formula` object.)

i. Report the estimated coefficients and standard errors from this model. Briefly interpret these results. For example, what does each coefficient mean?

ii. Use your coefficient estimates to calculate the marginal utility of income for a student at three different income levels: \$15,000, \$25,000, and \$35,000. For each of these three income levels, also calculate the dollar value that a student places on each hour spent driving and on each hour spent on the bus.

Problem 3: Multinomial Logit Model

Use the `commute_multinomial.csv` dataset for this question.

a. Model the commute choice during spring as a multinomial logit model. Express the representative utility of each alternative as a linear function of its cost and time. Include an alternative-specific intercept, allow the common parameter on cost to vary inversely with income, and allow the parameter on time to be alternative-specific. That is, the representative utility to student n from alternative j is

$$V_{nj} = \alpha_j + \frac{\beta_1}{I_n}C_{nj} + \beta_j T_{nj}$$

where V_{nj} is the representative utility to student n from alternative j, I_n is the income of student n, C_{nj} is the cost to student n of alternative j, T_{nj} is the time for student n of alternative j, and the α and β parameters are to be estimated. (Reminder: the `mlogit()` function from the `mlogit` package estimates a multinomial logit model, but the data must first be converted to an indexed data frame using the `dfidx()` function from the `dfidx` package. See the Week 4 slides or the `mlogit` vignettes at `cran.r-project.org/web/packages/mlogit/index.html` for information on specifying a `formula` for the `mlogit()` function.)

i. Report the estimated parameter and standard errors from this model. Briefly interpret these results. For example, what does each parameter mean?

ii. Calculate the elasticity of each commute alternative with respect to the cost of driving for each student; that is, 4 alternatives × 1000 students = 4000 elasticities. For each alternative, report the mean, minimum, maximum, and quartiles of its elasticity with respect to the cost of driving. Describe how these elasticities and substitution patterns relate to an important property of the logit model. (Reminder: the `fitted()` function with argument `type = 'probabilities'` calculates the choice probabilities of each alternative for each decision maker.)

b. A student's family status might also affect their commute decision or underlying preferences. Estimate the model from part (a) on two subsets of the data based on student marital status; that is, estimate one model using only single students, and estimate a second model using only married students.

i. Report the estimated parameters and standard errors from both models. Briefly interpret these results. For example, what does each parameter mean?

ii. For each marital status, use the corresponding parameter estimates to calculate the marginal utility of income for a student with $25,000 income. For each marital status, also calculate the dollar value that a student with $25,000 income places on one hour of commute time for each of the four commute alternatives.

iii. You should have found—when comparing these two models—that parameter estimates and marginal utilities differ by 50% or more in most cases, but that the dollar values of commute times tend be much more similar. Give a potential econometric explanation for why these models yield such different parameters but also yields similar values of commute time. (Hint: think about what component of the random utility model could cause parameter estimates to differ, even for the same underlying preferences.)

c. The university has a strong commitment to environmental sustainability and would like to convince graduate students to take the bus rather than drive to campus. One proposal is to introduce more buses on the existing bus routes, which would reduce bus commute time by 20%. Use your parameter estimates from part (a) to simulate this counterfactual.

i. How many additional students—of the 1000 students in this dataset—do you expect will commute by bus because of this reduction in bus commute time? How many fewer students do you expect will choose each of the three other commute alternatives?

ii. How much additional economic surplus do you expect this reduction in bus commute time will generate for the 1000 students in this dataset?