Problem Set 1

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

Solutions

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.

```r
### Load packages for problem set
library(tidyverse)
library(lmtest)
library(sandwich)
library(car)
```

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.)

```
## Load dataset
data_binary <- read_csv('commute_binary.csv')

## Rows:  1000 Columns:  13
## -- Column specification --------------------------------------------------------
## Delimiter:  ","
## chr  (2):  mode, marital_status
## dbl (11):  id, time.car, cost.car, time.bus, cost.bus, price_gas, sno...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

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.)

```
## Calculate means and medians of time and cost variables for full sample
summ_stats_1a <- data_binary %>%
  summarize(mn_cost_c = mean(cost.car),
            md_cost_c = median(cost.car),
            mn_time_c = mean(time.car),
            md_time_c = median(time.car),
            mn_time_b = mean(time.bus),
            md_time_b = median(time.bus))
```

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

```
    ## Report summary stats for full sample
    summ_stats_1a

## # A tibble: 1 x 6
##   mn_cost_c md_cost_c mn_time_c md_time_c mn_time_b md_time_b
##       <dbl>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>
## 1     0.411      0.35      10.9        10      14.2        13
```

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.

For the average student, the time to drive is shorter than the time to take the bus. If students value their time, then these shorter commute times would cause them to prefer driving. Of course, driving also incurs a monetary cost that does not occur when taking the bus, so students will trade off this shorter commute time against the monetary cost when choosing how to commute. Additionally, for all three variables, the mean is greater than the median, suggesting there is a long right tail of students who commute from far away and, hence, have especially long costs and times for commuting.

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.)

```
## Calculate means and medians of time and cost variable by commute mode
summ_stats_1b <- data_binary %>%
  group_by(mode) %>%
  summarize(mn_cost_c = mean(cost.car),
            md_cost_c = median(cost.car),
            mn_time_c = mean(time.car),
            md_time_c = median(time.car),
            mn_time_b = mean(time.bus),
            md_time_b = median(time.bus))
```

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

```
    ## Report summary stats for subsamples
    summ_stats_1b

## # A tibble: 2 x 7
##   mode  mn_cost_c md_cost_c mn_time_c md_time_c mn_time_b md_time_b
##   <chr>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>
## 1 bus       0.449      0.38     11.9         11      14.7        13
## 2 car       0.361      0.32      9.70         9      13.5        13
```

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.

The students who drive, on average, have lower costs to drive and shorter driving commute times, compared to students who take the bus. This result is intuitive because, *ceteris paribus*, having a cheaper and shorter drive would make a student more likely to drive. Interestingly, the median time to take the bus is identical in both groups, and the mean time to take the bus is greater for that group that does take the bus. This result is less intuitive, but a more rigorous analysis might reveal an explanation.

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.)

```r
## Clean choice variable
data_binary <- data_binary %>%
  mutate(car = (mode == 'car'))
## Model choice as a linear probability model
reg_2a <- lm(formula = car ~ cost.car + time.car + time.bus,
             data = data_binary)
```

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.)

```r
## Calculate heteroskedastic-robust standard errors
coeftest(reg_2a, vcov = vcovHC(reg_2a))

##
## t test of coefficients:
##
##               Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)  0.9205109  0.0737943  12.4740 < 2.2e-16 ***
## cost.car    -0.4375642  0.1392888  -3.1414  0.001731 **
## time.car    -0.0652505  0.0058747 -11.1070 < 2.2e-16 ***
## time.bus     0.0283670  0.0065722   4.3162 1.746e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

All three independent variables have statistically significant and economically meaningful effects on the choice to drive or take a bus to campus. An additional 10 cents of driving cost reduces the probability of driving by 4.4%, an additional minute of driving reduces the probability of driving by 6.5%, and an additional minute riding the bus increases the probability of driving by 2.8%. Because there are only two alternatives, the marginal effects on the choice to ride the bus are the negatives of the driving marginal effects.

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.)

```r
## Calculate estimated probability of car for each individual
data_binary <- data_binary %>%
  mutate(prob_car_2a = predict(reg_2a))
## Count number of individuals with probabilities outside [0, 1]
data_binary %>%
  filter(prob_car_2a < 0 | prob_car_2a > 1) %>%
  nrow()

## [1] 22
```

Only 22 students, or 2.2% of the sample, have estimated probabilities outside the $[0, 1]$ range. This result suggests that our estimated marginal effects are not likely to be inconsistent and our interpretation of the results is sound.

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.)

```
## Conduct a Wald test on time coefficients
linearHypothesis(reg_2a, 'time.car = -time.bus', vcov = vcovHC(reg_2a))

## Linear hypothesis test
##
## Hypothesis:
## time.car  + time.bus = 0
##
## Model 1: restricted model
## Model 2: car ~ cost.car + time.car + time.bus
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F    Pr(>F)
## 1    997
## 2    996  1 17.203 3.646e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The result of Wald test indicates that the time coefficients are statistically different from one another. This result is intuitive because the experience of driving and riding a bus are different; for example, a student can read or catch up on email while riding the bus, but not while driving. So a minute of each mode might differently affect the decision to drive or ride the bus. The marginal effect of driving time is larger in absolute, so an equal increase in the time of both modes would decrease the utility of driving and cause some drivers to substitute to the bus.