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 all group members can submit the same code and output; indicate in your writeup who you worked with. You must submit a unique writeup that answers the problems below. You can discuss answers with your fellow group members, but your writeup must be in your own words. You can use any "canned" routine (e.g., `lm()`, `glm()`, and `mlogit()`) for this problem set.

Data

Download the file `travel_datasets.zip` from the course website (`github.com/woerman/ResEcon703`). This zipped file contains two datasets, `travel_binary.csv` and `travel_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 commuting to campus. The `travel_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). The `travel_multinomial.csv` dataset corresponds to commuting in spring when riding a bike and walking are feasible alternatives. See the file `travel_descriptions.txt` for descriptions of the variables in each dataset.

```
### Load packages for problem set
library(tidyverse)

## -- Attaching packages ------------------------------------ tidyverse 1.2.1 --
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
## -- Conflicts --------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(lmtest)
```

```
## Loading required package:  zoo
##
## Attaching package:  'zoo'
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

library(sandwich)
library(car)

## Loading required package:  carData
##
## Attaching package:  'car'
## The following object is masked from 'package:dplyr':
##
##     recode
## The following object is masked from 'package:purrr':
##
##     some

library(mlogit)

## Loading required package:  Formula
```

Problem 1: Linear Probability Model

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

a. Model the choice to drive to campus during winter as a linear probability model. Include only alternative-specific data (and not demographic data) in this model. Report the estimated coefficients and heteroskedastic-robust standard errors from this model. Interpret these results, including how a marginal change in each variable would affect the probability of driving.

```
### Part a
## Load dataset
data_binary <- read_csv('travel_binary.csv')

## Parsed with column specification:
## cols(
##   mode = col_character(),
##   time_car = col_double(),
##   cost_car = col_double(),
##   time_bus = col_double(),
##   cost_bus = col_double(),
##   age = col_double(),
##   income = col_double(),
##   marital_status = col_character()
## )
## Clean choice variable
data_binary <- data_binary %>%
  mutate(car = (mode == 'car'))
## Model choice as a linear probability model
reg_lpm <- data_binary %>%
  lm(formula = car ~ time_car + time_bus + cost_car + cost_bus, data = .)
## Calculate heteroskedastic-robust standard errors
reg_lpm %>%
  coeftest(vcov = vcovHC(reg_lpm))

##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  0.8963884  0.2156157  4.1573 3.497e-05 ***
## time_car    -0.0706660  0.0104990 -6.7307 2.847e-11 ***
## time_bus     0.0314546  0.0023972 13.1215 < 2.2e-16 ***
## cost_car    -0.4538255  0.1121724 -4.0458 5.619e-05 ***
## cost_bus     0.5214847  0.0355117 14.6849 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

All four variables have statistically significant and economically meaningful effects on the choice to drive or take a bus to campus. An additional minute of driving reduces the probability of driving by 7.1%, while an additional minute riding the bus increases the probability of driving by 3.1%. An additional 10 cents of driving cost reduces the probability of driving by 4.5%, while an additional 10 cents of bus cost increases the probability of driving by 5.2%. Because there are only two choices, the marginal effects on the choice to ride the bus are the negatives of the driving marginal effects.

b. Do any of your coefficients look like they might be equal (in absolute value)? Test that the two time coefficients are equal (in absolute value) and then test that the two cost coefficients are equal (in absolute value). Interpret the results of these tests and explain why they make intuitive sense. (Hint: There are many ways to conduct a Wald test in R. I like the `linearHypothesis()` function from the `car` (companion to applied regression) package.)

```
### Part b
## Conduct a Wald test on time coefficients
reg_lpm %>%
  linearHypothesis('time_car = -time_bus', vcov = vcovHC(reg_lpm))

## Linear hypothesis test
##
## Hypothesis:
## time_car  + time_bus = 0
##
## Model 1: restricted model
## Model 2: car ~ time_car + time_bus + cost_car + cost_bus
```

```
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F    Pr(>F)
## 1    996
## 2    995  1 20.971 5.253e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Conduct a Wald test on cost coefficients
reg_lpm %>%
  linearHypothesis('cost_car = -cost_bus', vcov = vcovHC(reg_lpm))

## Linear hypothesis test
##
## Hypothesis:
## cost_car  + cost_bus = 0
##
## Model 1: restricted model
## Model 2: car ~ time_car + time_bus + cost_car + cost_bus
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F Pr(>F)
## 1    996
## 2    995  1 0.3799 0.5378
```

The time coefficients appear to be statistically different from own another, but the cost coefficients (in absolute value) are close enough that they are probably not statistically different from each other. The Wald tests confirm these results; the time coefficients are statistically different from one another, but the cost coefficients are not. These results are intuitive. The experience of driving and riding a bus are different, so a minute of each mode might provide a different amount of utility (or, as we find, disutility). But a dollar spent on driving and a dollar spent on riding the bus are identical and there is no reason why the marginal utility of those dollars should differ.

c. Again model the choice to drive to campus during winter as a linear probability model, but now force the two cost variables to have the same coefficient (in absolute value). Report the estimated coefficients and heteroskedastic-robust standard errors from this model. Interpret these results, including how a marginal change in each variable would affect the probability of driving. (Hint: An easy way to force coefficients to be equal is to construct a new variable from the existing data.)

```
### Part c
## Calculate cost difference to force coefficients to be equal
data_binary <- data_binary %>%
  mutate(cost_difference = cost_car - cost_bus)
## Model choice as LPM with cost difference
reg_lpm_restricted <- data_binary %>%
  lm(formula = car ~ time_car + time_bus + cost_difference, data = .)
## Calculate heteroskedastic-robust standard errors
reg_lpm_restricted %>%
  coeftest(vcov = vcovHC(reg_lpm_restricted))

##
## t test of coefficients:
##
##                    Estimate Std. Error t value  Pr(>|t|)
## (Intercept)       1.0285344  0.0324997  31.648 < 2.2e-16 ***
## time_car         -0.0650474  0.0043183 -15.063 < 2.2e-16 ***
## time_bus          0.0307068  0.0019517  15.733 < 2.2e-16 ***
## cost_difference -0.5206611  0.0354737 -14.677 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

All variables again have statistically significant and economically meaningful effects on the choice to drive or take a bus to campus. An additional minute of driving reduces the probability of driving by 6.5%, while an additional minute riding the bus increases the probability of driving by 3.1%. An additional 10 cents of driving cost reduces the probability of driving by 5.2%, while an additional 10 cents of bus cost has the exact opposite effect, increasing the probability of driving by 5.2%. Because there are only two choices, the marginal effects on the choice to ride the bus are the negatives of the driving marginal effects.

d. One potential problem with a linear probability model is that predicted probabilities can fall outside the $[0, 1]$ range. Using the model from part (c), how many graduate students have infeasible choice probabilities? Given these results, are you worried about using a linear probability model in this case?

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

## [1] 39
```

Only 39 decision makers 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.

Problem 2: Binary Logit Model

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

a. Model the choice to drive to campus during winter as a binary logit model with the same variables as in part (c) of problem 1. Report the estimated coefficients and standard errors from this model. Interpret these results, including how a marginal change in each variable would affect the probability of driving, and compare to the linear probability model you estimated in part (c) of problem 1. Also calculate the dollar value of an hour spent on each travel mode and explain these time-value results.

```r
### Part a
## Model choice as binary logit with equal cost coefficients
reg_logit <- data_binary %>%
  glm(formula = car ~ time_car + time_bus + cost_difference,
      data = ., family = 'binomial')
## Summarize model results
reg_logit %>%
  summary()
```

```
##
## Call:
## glm(formula = car ~ time_car + time_bus + cost_difference, family = "binomial",
##     data = .)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6825  -1.0226   0.3388   0.9753   2.5646
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      2.95555    0.29054  10.173   <2e-16 ***
## time_car        -0.48681    0.04845 -10.049   <2e-16 ***
## time_bus         0.24239    0.02518   9.626   <2e-16 ***
## cost_difference -2.85617    0.26941 -10.602   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1362.5  on 999  degrees of freedom
## Residual deviance: 1074.3  on 996  degrees of freedom
## AIC: 1082.3
##
## Number of Fisher Scoring iterations: 5
```

```r
## Calculate estimated utility and probability of car
data_binary <- data_binary %>%
  mutate(utility_logit = predict(reg_logit),
         car_probability_logit = 1 / (1 + exp(-utility_logit)))
## Calculate mean marginal effects
coef(reg_logit)[2:4] *
```

```
  mean(data_binary$car_probability_logit *
         (1 - data_binary$car_probability_logit))

##        time_car        time_bus cost_difference
##     -0.08924869      0.04443785     -0.52363435

## Calculate hourly time-value for each mode
coef(reg_logit)[2:3] / coef(reg_logit)[4] * 60

##  time_car  time_bus
## 10.226452 -5.091857
```

All variables again have statistically significant and economically meaningful coefficients. Those coefficients, however, are now interpreted as marginal utilities rather than marginal effects. We can calculate the marginal effects, but they now vary across decision makers. Taking the mean across all individuals, we find that an additional minute of driving reduces the probability of driving by 8.9%, an additional minute riding the bus increases the probability of driving by 4.4%, and an additional 10 cents of cost changes the probability of driving by 5.2% . The marginal effects of time are larger than we estimated using the linear probability model, but not substantially so, and the marginal effect of cost is the same. These results imply that an hour of driving has a cost equal to $10.23 and an hour riding the bus has a cost equal to $5.09.

b. Demographic information might also affect a travel mode decision. For example, individuals 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 divide cost by income to allow for this kind of heterogeneity. Report the estimated coefficients and standard errors from this model. Interpret these results, including how a marginal change in each time or cost variable would affect the probability of driving. Also calculate the dollar value of an hour spent on each travel mode, which now varies by income; calculate these values for incomes of $15,000, $25,000, and $35,000.

```
### Part b
## Model choice as binary logit with cost divided by income
reg_logit_income <- data_binary %>%
  glm(formula = car ~ time_car + time_bus + I(cost_difference / income),
      data = ., family = 'binomial')
## Summarize model results
reg_logit_income %>%
  summary()

##
## Call:
## glm(formula = car ~ time_car + time_bus + I(cost_difference/income),
##     family = "binomial", data = .)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7222  -0.9340   0.3180   0.8922   2.5454
```

```
## 
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)    
## (Intercept)                   3.09631    0.27343  11.324   <2e-16 ***
## time_car                     -0.49147    0.04804 -10.231   <2e-16 ***
## time_bus                      0.24239    0.02501   9.692   <2e-16 ***
## I(cost_difference/income)   -73.05742    6.17061 -11.840   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 1362.5  on 999  degrees of freedom
## Residual deviance: 1030.8  on 996  degrees of freedom
## AIC: 1038.8
## 
## Number of Fisher Scoring iterations: 5
```

```
## Calculate estimated utility and probability of car
data_binary <- data_binary %>%
  mutate(utility_logit_income = predict(reg_logit_income),
         car_probability_logit_income = 1 / (1 + exp(-utility_logit_income)))
## Calculate mean marginal effects
coef(reg_logit_income)[2:4] *
  c(rep(mean(data_binary$car_probability_logit_income *
               (1 - data_binary$car_probability_logit_income)), 2),
    mean(data_binary$car_probability_logit_income *
           (1 - data_binary$car_probability_logit_income) /
           data_binary$income))
```

```
##                  time_car                  time_bus 
##               -0.08539312                0.04211521 
## I(cost_difference/income) 
##               -0.51708462
```

```
## Calculate hourly time-values at different income levels
rep(-abs(coef(reg_logit_income)[2:3]), 3) / coef(reg_logit_income)[4] * 60 *
  c(15, 15, 25, 25, 35, 35)
```

```
##  time_car  time_bus  time_car  time_bus  time_car  time_bus 
##  6.054465  2.986014 10.090775  4.976690 14.127085  6.967367
```

All variables again have statistically significant and economically meaningful coefficients. As in the previous logit model, those coefficients are interpreted as marginal utilities rather than marginal effects. We can calculate the marginal effects, which vary across decision makers. Taking the mean across all individuals, we find that an additional minute of driving reduces the probability of driving by 8.5%, an additional minute riding the bus increases the probability of driving by 4.2%, and an additional 10 cents of cost changes the probability of driving by 5.2% . These marginal effects are nearly identical to those calculated in part (a), which provides support that they are correct. These results imply that an hour of driving and an hour riding the bus have costs equal to $6.05 and $2.99, respectively, at an income of $15,000; $10.09 and $4.98, respectively, at an income of $25,000; and $14.13 and $6.97, respectively, at an income of $35,000. Note that these values are linear functions of income, which is a consequence of how we have modeled income.

c. The family status of a graduate student might also affect their decision making. Estimate the binary logit model of part (b) separately for single students and for married students. Report the estimated coefficients and standard errors from each model. Interpret these results and compare your estimated coefficients across the two models. Also calculate and compare the time-value of each travel mode for single students and for married students; again calculate these values for incomes of $15,000, $25,000, and $35,000. How can you explain that one set of results is very different across these two models, but the other set of results are roughly similar?

```
### Part c
## Model choice as binary logit for single individuals
reg_logit_single <- data_binary %>%
  filter(marital_status == 'single') %>%
  glm(formula = car ~ time_car + time_bus + I(cost_difference / income),
      data = ., family = 'binomial')
## Summarize model results
reg_logit_single %>%
  summary()

##
## Call:
## glm(formula = car ~ time_car + time_bus + I(cost_difference/income),
##     family = "binomial", data = .)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2144  -1.0697   0.5111   0.9988   1.9765
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  2.03295    0.32420   6.271 3.60e-10 ***
## time_car                    -0.29383    0.05173  -5.680 1.34e-08 ***
## time_bus                     0.14776    0.02658   5.559 2.71e-08 ***
## I(cost_difference/income) -49.16771    7.50831  -6.548 5.81e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 647.88  on 473  degrees of freedom
## Residual deviance: 556.87  on 470  degrees of freedom
## AIC: 564.87
```

```
##
## Number of Fisher Scoring iterations: 4

## Model choice as binary logit for married individuals
reg_logit_married <- data_binary %>%
  filter(marital_status == 'married') %>%
  glm(formula = car ~ time_car + time_bus + I(cost_difference / income),
      data = ., family = 'binomial')
## Summarize model results
reg_logit_married %>%
  summary()

##
## Call:
## glm(formula = car ~ time_car + time_bus + I(cost_difference/income),
##     family = "binomial", data = .)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6858  -0.7042   0.1110   0.6953   2.2294
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   5.07681    0.54826   9.260  < 2e-16 ***
## time_car                     -0.93178    0.11103  -8.392  < 2e-16 ***
## time_bus                      0.46096    0.05843   7.889 3.04e-15 ***
## I(cost_difference/income) -116.27520   12.06695  -9.636  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 714.40  on 525  degrees of freedom
## Residual deviance: 431.69  on 522  degrees of freedom
## AIC: 439.69
##
## Number of Fisher Scoring iterations: 6

## Calculate time-values for singles at different incomes
rep(-abs(coef(reg_logit_single)[2:3]), 3) / coef(reg_logit_single)[4] * 60 *
  c(15, 15, 25, 25, 35, 35)

##  time_car  time_bus  time_car  time_bus  time_car  time_bus
##  5.378539  2.704693  8.964232  4.507821 12.549925  6.310949

## Calculate time-values for marrieds at different incomes
rep(-abs(coef(reg_logit_married)[2:3]), 3) / coef(reg_logit_married)[4] * 60 *
  c(15, 15, 25, 25, 35, 35)
```

```
##  time_car  time_bus  time_car  time_bus  time_car  time_bus
##  7.212247  3.567917 12.020412  5.946528 16.828576  8.325139
```

Both models have coefficients that are statistically significant and economically meaningful. Importantly, though, the coefficients for married students are much larger than the coefficients for single students. The time values of these two subsets are much closer, though. The results for single students imply that an hour of driving and an hour riding the bus have costs equal to $5.38 and $2.70, respectively, at an income of $15,000; $8.96 and $4.51, respectively, at an income of $25,000; and $12.55 and $6.31, respectively, at an income of $35,000. Similarly, the results for married students imply that an hour of driving and an hour riding the bus have costs equal to $7.21 and $3.57, respectively, at an income of $15,000; $12.02 and $5.95, respectively, at an income of $25,000; and $16.83 and $8.33, respectively, at an income of $35,000. Married students have a higher value of their time than do single students, but these values are much more in line than are the model coefficients. These results suggest that single students and married students have very different scale parameters, so estimated coefficients are very different. But the scale parameter does not affect the ratio of parameters within a model, so the time valuations are not affected by the different scale parameters.

d. 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 express bus lines that would speed up the longest bus commutes. The express buses would reduce bus travel time by 10 minutes for the long-distance bus routes; long-distance bus routes are those that cost $3 to ride. But new bus routes are costly, so the university will only implement this plan if an extra 10% of graduate students start taking the bus. Using your results from part (b), how many students will switch from driving to taking the bus because of these express routes?

```
### Part d
## Create new dataset with express bus times
data_binary_express <- data_binary %>%
  mutate(time_bus = ifelse(cost_bus == 3, time_bus - 10, time_bus))
## Calculate estimated utility and probability of car with express buses
data_binary_express <- data_binary_express %>%
  mutate(utility_logit_express = predict(reg_logit_income,
                                         newdata = data_binary_express),
         car_probability_logit_express = 1 / (1 + exp(-utility_logit_express)))
## Count number of bus riders in original model
bus_logit_income <- data_binary %>%
  filter(car_probability_logit_income < 0.5) %>%
  nrow()
## Count number of bus riders with express buses
bus_logit_express <- data_binary_express %>%
  filter(car_probability_logit_express < 0.5) %>%
  nrow()
## Calculate difference in bus ridership due to express buses
bus_logit_express - bus_logit_income

## [1] 95
```

The express bus routes would cause 95 of these 1,000 graduate students to switch from driving to riding the bus. This is only 9.5% of our sample, not the 10% required by the university, so the university will not implement these express bus routes.

e. An alternative proposal is to reduce the cost of these long-distance bus routes by $0.50. Using only the results from part (b) and no new calculations, you should be able to form a good idea of how this proposal compares to the express bus routes. Do you think think this $0.50 subsidy will yield more or fewer new bus riders than the express bus routes, and how did you come to this conclusion? Now calculate the effect of this subsidy. Using your results from part (b), how many students will switch from driving to taking the bus because of this subsidy?

```
### Part e
## Create new dataset with bus subsidies
data_binary_subsidy <- data_binary %>%
  mutate(cost_difference = ifelse(cost_bus == 3,
                                  cost_difference + 0.5,
                                  cost_difference))
## Calculate estimated utility and probability of car with bus subsidies
data_binary_subsidy <- data_binary_subsidy %>%
  mutate(utility_logit_subsidy = predict(reg_logit_income,
                                         newdata = data_binary_subsidy),
         car_probability_logit_subsidy = 1 / (1 + exp(-utility_logit_subsidy)))
## Count number of bus riders with express buses
bus_logit_subsidy <- data_binary_subsidy %>%
  filter(car_probability_logit_subsidy < 0.5) %>%
  nrow()
## Calculate difference in bus ridership due to bus subsidies
bus_logit_subsidy - bus_logit_income

## [1] 69
```

In part (b), we found that the cost of an hour riding the bus is in the range of $3–7 for most of the graduate students we observe, which is equivalent to a cost of roughly $0.50–1.15 for ten minutes riding the bus. Thus, for a subsidy to yield an effect as large as the express routes, which reduced bus times by ten minutes, the subsidy would have to be larger than $0.50. The subsidy would cause only 69 drivers to switch to riding the bus, which is smaller than the effect of the express routes routes, as expected.

Problem 3: Multinomial Logit Model

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

a. Model the travel mode choice to commute to campus during spring as a multinomial logit model with the same variables as in part (b) of problem 2. Report the estimated coefficients and standard errors from this model. Interpret these results, including the elasticity of each alternative with respect to the cost to drive, the cost to take the bus, and the time to take the bus (i.e., you should discuss 12 elasticities, 4 alternatives × 3 elasticities each). You should notice a pattern in these elasticities.

Describe this pattern and how it relates to the logit model. Also calculate the time-value of each travel mode; again calculate these values for incomes of $15,000, $25,000, and $35,000.

```
### Part a
## Load dataset
data_multi <- read_csv('travel_multinomial.csv')

## Parsed with column specification:
## cols(
##   mode = col_character(),
##   time_car = col_double(),
##   cost_car = col_double(),
##   time_bus = col_double(),
##   cost_bus = col_double(),
##   time_bike = col_double(),
##   cost_bike = col_double(),
##   time_walk = col_double(),
##   cost_walk = col_double(),
##   age = col_double(),
##   income = col_double(),
##   marital_status = col_character()
## )

## Convert dataset to mlogit format
data_mlogit <- data_multi %>%
  mlogit.data(shape = 'wide', choice = 'mode', varying = 2:9, sep = '_')

## Warning:  Setting row names on a tibble is deprecated.

## Warning:  Setting row names on a tibble is deprecated.

## Warning:  Setting row names on a tibble is deprecated.

## Warning:  Setting row names on a tibble is deprecated.

## Model choice as multinomial logit with common cost/income coefficient,
## alternative intercepts, and alternative-specific time coefficients
model_mlogit <- data_mlogit %>%
  mlogit(mode ~ I(cost / income) | 1 | time, data = .)
## Summarize model results
model_mlogit %>%
  summary()

##
## Call:
## mlogit(formula = mode ~ I(cost/income) | 1 | time, data = .,
##     method = "nr")
##
## Frequencies of alternatives:
##  bike   bus   car  walk
## 0.103 0.290 0.465 0.142 
## 
## nr method
## 11 iterations, 0h:0m:1s 
## g'(-H)^-1g = 7.46E-06 
## successive function values within tolerance limits 
## 
## Coefficients :
##                     Estimate Std. Error  z-value  Pr(>|z|)    
## bus:(intercept)    3.392381   0.279235  12.1488 < 2.2e-16 ***
## car:(intercept)    6.650311   0.464905  14.3047 < 2.2e-16 ***
## walk:(intercept)   3.711605   0.338860  10.9532 < 2.2e-16 ***
## I(cost/income)   -76.745551   5.590969 -13.7267 < 2.2e-16 ***
## bike:time         -0.362973   0.026469 -13.7129 < 2.2e-16 ***
## bus:time          -0.232457   0.023900  -9.7263 < 2.2e-16 ***
## car:time          -0.475974   0.045422 -10.4790 < 2.2e-16 ***
## walk:time         -0.401968   0.034372 -11.6946 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Log-Likelihood: -760.7
## McFadden R^2:  0.3797 
## Likelihood ratio test : chisq = 931.28 (p.value = < 2.22e-16)
```

```
## Calculate estimated probabilities for each mode
fitted_mlogit <- fitted(model_mlogit, type = 'probabilities')
## Assign probabilities as variables in dataset
data_multi <- data_multi %>%
  mutate(bike_probability_mlogit = fitted_mlogit[, 1],
         bus_probability_mlogit = fitted_mlogit[, 2],
         car_probability_mlogit = fitted_mlogit[, 3],
         walk_probability_mlogit = fitted_mlogit[, 4])
## Calculate own- and cross-price elasticities with respect to car cost
data_multi <- data_multi %>%
  mutate(elasticity_cost_car_own_mlogit = coef(model_mlogit)[4] *
           cost_car / income * (1 - car_probability_mlogit),
         elasticity_cost_car_other_mlogit = -coef(model_mlogit)[4] *
           cost_car / income * car_probability_mlogit)
## Report means of own- and cross-price elasticities with respect to car cost
data_multi %>%
  select(elasticity_cost_car_own_mlogit, elasticity_cost_car_other_mlogit) %>%
  summarize_all(mean) %>%
  unlist()
```

```
##   elasticity_cost_car_own_mlogit elasticity_cost_car_other_mlogit 
##                        -4.261091                         3.728308
```

```
## Calculate own- and cross-price elasticities with respect to bus cost
```

```r
data_multi <- data_multi %>%
  mutate(elasticity_cost_bus_own_mlogit = coef(model_mlogit)[4] *
           cost_bus / income * (1 - bus_probability_mlogit),
         elasticity_cost_bus_other_mlogit = -coef(model_mlogit)[4] *
           cost_bus / income * bus_probability_mlogit)
## Report means of own- and cross-price elasticities with respect to bus cost
data_multi %>%
  select(elasticity_cost_bus_own_mlogit, elasticity_cost_bus_other_mlogit) %>%
  summarize_all(mean) %>%
  unlist()

##   elasticity_cost_bus_own_mlogit elasticity_cost_bus_other_mlogit
##                        -3.969960                         1.700537

## Calculate own- and cross- elasticities with respect to bus time
data_multi <- data_multi %>%
  mutate(elasticity_time_bus_own_mlogit = coef(model_mlogit)[6] *
           time_bus * (1 - bus_probability_mlogit),
         elasticity_time_bus_other_mlogit = -coef(model_mlogit)[6] *
           time_bus * bus_probability_mlogit)
## Report means of own- and cross- elasticities with respect to bus time
data_multi %>%
  select(elasticity_time_bus_own_mlogit, elasticity_time_bus_other_mlogit) %>%
  summarize_all(mean) %>%
  unlist()

##   elasticity_time_bus_own_mlogit elasticity_time_bus_other_mlogit
##                        -2.826442                         1.193433

## Calculate hourly time-value for each mode at different incomes
rep(coef(model_mlogit)[5:8], 3) / coef(model_mlogit)[4] * 60 *
  c(rep(15, 4), rep(25, 4), rep(35, 4))

## bike:time  bus:time  car:time walk:time bike:time  bus:time  car:time
##  4.256609  2.726035  5.581777  4.713906  7.094348  4.543392  9.302962
## walk:time bike:time  bus:time  car:time walk:time
##  7.856511  9.932087  6.360749 13.024146 10.999115
```

All variables again have statistically significant and economically meaningful coefficients. The marginal utility of cost is negative, and the marginal utility of time for each mode is also negative, which is intuitive since people generally like having both money and time. The own-price elasticity of driving is -4.3, and the cross-price elasticity of all other modes with respect to the cost of driving is 3.7. The own-price elasticity of riding the bus is -4.0, and the cross-price elasticity of all other modes with respect to the cost of the bus is 1.7. The own elasticity of riding the bus with respect to time is -2.8, and the cross elasticity of all other modes with respect to time on the bus is 1.2. Because of the independence of irrelevant alternatives, which is a property of the logit model, each cross elasticity is the same for all alternative travel modes; for example, in response to a 1% increase in the cost of driving, the number of graduate students choosing each other mode increases by 3.7%. The results imply that an hour of biking, riding the bus, driving, and walking have costs equal to $4.25, $2.73, $5.58, and $4.71, respectively, at an income of $15,000; $7.09, $4.54, $9.30, and $7.86, respectively, at an income of $25,000; and $9.93, $6.36, $13.02, and $11.00, respectively, at an income of $35,000.

b. Age is another demographic variable that might affect a graduate student's commute decision, and it might have differential effects on the travel modes. For example, aging might have no direct effect on the utility of driving to campus, but biking through the rolling hills of the Pioneer Valley might become more difficult as a student ages. Again model the travel mode choice to commute to campus during spring as a multinomial logit model, but now include age with alternative-specific coefficients. Report the estimated coefficients and standard errors from this model. Interpret these results, including the elasticity of each alternative with respect to the cost to drive, the cost to take the bus, and the time to take the bus. Also calculate the time-value of each travel mode; again calculate these values for incomes of $15,000, $25,000, and $35,000.

```
### Part b
## Model choice as multinomial logit with common cost/income coefficient and
## alternative-specific age and time coefficients
model_mlogit_age <- data_mlogit %>%
  mlogit(mode ~ I(cost / income) | age | time, data = .)
## Summarize model results
model_mlogit_age %>%
  summary()

##
## Call:
## mlogit(formula = mode ~ I(cost/income) | age | time, data = .,
##     method = "nr")
##
## Frequencies of alternatives:
##  bike   bus   car  walk
## 0.103 0.290 0.465 0.142
##
## nr method
## 11 iterations, 0h:0m:1s
## g'(-H)^-1g = 7.12E-06
## successive function values within tolerance limits
##
## Coefficients :
##                     Estimate Std. Error  z-value  Pr(>|z|)
## bus:(intercept)    -4.062897   1.932908  -2.1020 0.0355567 *
## car:(intercept)    -0.258814   1.933707  -0.1338 0.8935264
## walk:(intercept)   -1.165225   2.326471  -0.5009 0.6164732
## I(cost/income)    -77.134644   5.639246 -13.6782 < 2.2e-16 ***
## bus:age             0.276657   0.071783   3.8541 0.0001162 ***
## car:age             0.257121   0.070904   3.6264 0.0002875 ***
## walk:age           0.182304   0.085568   2.1305 0.0331293 *  
## bike:time         -0.365816   0.026576 -13.7650 < 2.2e-16 ***
## bus:time          -0.232275   0.023935  -9.7046 < 2.2e-16 ***
## car:time          -0.475904   0.045509 -10.4573 < 2.2e-16 ***
## walk:time         -0.404021   0.034472 -11.7201 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Log-Likelihood: -751.88
## McFadden R^2:  0.38689 
## Likelihood ratio test : chisq = 948.92 (p.value = < 2.22e-16)
```

```
## Calculate estimated probabilities for each mode
fitted_mlogit_age <- fitted(model_mlogit_age, type = 'probabilities')
## Assign probabilities as variables in dataset
data_multi <- data_multi %>%
  mutate(bike_probability_mlogit_age = fitted_mlogit_age[, 1],
         bus_probability_mlogit_age = fitted_mlogit_age[, 2],
         car_probability_mlogit_age = fitted_mlogit_age[, 3],
         walk_probability_mlogit_age = fitted_mlogit_age[, 4])
## Calculate own- and cross-price elasticities with respect to car cost
data_multi <- data_multi %>%
  mutate(elasticity_cost_car_own_mlogit_age = coef(model_mlogit_age)[4] *
           cost_car / income * (1 - car_probability_mlogit_age),
         elasticity_cost_car_other_mlogit_age = -coef(model_mlogit_age)[4] *
           cost_car / income * car_probability_mlogit_age)
## Report means of own- and cross-price elasticities with respect to car cost
data_multi %>%
  select(elasticity_cost_car_own_mlogit_age,
         elasticity_cost_car_other_mlogit_age) %>%
  summarize_all(mean) %>%
  unlist()
```

```
##   elasticity_cost_car_own_mlogit_age elasticity_cost_car_other_mlogit_age 
##                            -4.282841                             3.747065
```

```
## Calculate own- and cross-price elasticities with respect to bus cost
data_multi <- data_multi %>%
  mutate(elasticity_cost_bus_own_mlogit_age = coef(model_mlogit_age)[4] *
           cost_bus / income * (1 - bus_probability_mlogit),
         elasticity_cost_bus_other_mlogit_age = -coef(model_mlogit_age)[4] *
           cost_bus / income * bus_probability_mlogit_age)
## Report means of own- and cross-price elasticities with respect to bus cost
data_multi %>%
  select(elasticity_cost_bus_own_mlogit_age,
         elasticity_cost_bus_other_mlogit_age) %>%
  summarize_all(mean) %>%
  unlist()
```

```
##   elasticity_cost_bus_own_mlogit_age elasticity_cost_bus_other_mlogit_age
##                            -3.990087                             1.709305
```

```r
## Calculate own- and cross- elasticities with respect to bus time
data_multi <- data_multi %>%
  mutate(elasticity_time_bus_own_mlogit_age = coef(model_mlogit_age)[9] *
           time_bus * (1 - bus_probability_mlogit_age),
         elasticity_time_bus_other_mlogit_age = -coef(model_mlogit_age)[9] *
           time_bus * bus_probability_mlogit_age)
## Report means of own- and cross- elasticities with respect to bus time
data_multi %>%
  select(elasticity_time_bus_own_mlogit_age,
         elasticity_time_bus_other_mlogit_age) %>%
  summarize_all(mean) %>%
  unlist()
```

```
##   elasticity_time_bus_own_mlogit_age elasticity_time_bus_other_mlogit_age
##                            -2.824227                             1.192498
```

```r
## Calculate hourly time-value for each mode at different incomes
rep(coef(model_mlogit_age)[8:11], 3) / coef(model_mlogit_age)[4] * 60 *
  c(rep(15, 4), rep(25, 4), rep(35, 4))
```

```
## bike:time  bus:time  car:time walk:time bike:time  bus:time  car:time
##  4.268310  2.710159  5.552808  4.714083  7.113850  4.516932  9.254679
## walk:time bike:time  bus:time  car:time walk:time
##  7.856804  9.959390  6.323704 12.956551 10.999526
```

All variables again have statistically significant and economically meaningful coefficients.The marginal utilities of cost and time for each travel mode are again negative, as expected. The alternative-specific coefficients for age are all positive. Note that biking is the reference mode, so these coefficients imply that older graduate students are more likely to ride the bus, drive, or walk, as compared to biking. The own-price elasticity of driving is -4.3, and the cross-price elasticity of all other modes with respect to the cost of driving is 3.7. The own-price elasticity of riding the bus is -4.0, and the cross-price elasticity of all other modes with respect to the cost of the bus is 1.7. The own elasticity of riding the bus with respect to time is -2.8, and the cross elasticity of all other modes with respect to time on the bus is 1.2. The results imply that an hour of biking, riding the bus, driving, and walking have costs equal to $4.27, $2.71, $5.55, and $4.71, respectively, at an income of $15,000; $7.11, $4.52, $9.25, and $7.86, respectively, at an income of $25,000; and $9.96, $6.32, $12.96, and $11.00, respectively, at an income of $35,000. These elasticities and time-values are approximately equal to those in part (a), which suggests that age is an important factor in the choice of travel mode, but its effect is orthogonal to the effects of cost and time.

c. The university is again considering the express bus route proposal from part (d) of problem 2. Using your results from part (b), how many students will begin taking the bus because of these express routes in the spring when biking and walking are also options? Also calculate how many students will no longer use each alternative because of the bus (i.e., how many current bikers, how many current drivers, and how many current walkers will start taking the bus). Compare these results to your results from part (d) of problem 2 and explain why they are similar or different.

```
### Part c
## Create index of individuals
data_multi <- data_multi %>%
  mutate(id = 1:n())
## Find estimated choice for each individual and join to dataset
data_multi <- data_multi %>%
  select(id, contains('probability_mlogit_age')) %>%
  gather(mode_mlogit_age, probability_mlogit_age, -id) %>%
  group_by(id) %>%
  arrange(id, desc(probability_mlogit_age)) %>%
  slice(1) %>%
  ungroup() %>%
  select(id, mode_mlogit_age) %>%
  mutate(mode_mlogit_age = str_remove(mode_mlogit_age,
                                      '_probability_mlogit_age')) %>%
  inner_join(data_multi)

## Joining, by = "id"

## Create new dataset with express bus times
data_multi_express <- data_multi %>%
  mutate(time_bus = if_else(cost_bus == 3, time_bus - 10, time_bus))
## Convert new dataset of relevant variables to mlogit format
data_mlogit_express <- data_multi_express %>%
  select(mode, time_car, cost_car, time_bus, cost_bus, time_bike, cost_bike,
         time_walk, cost_walk, age, income) %>%
  mlogit.data(shape = 'wide', choice = 'mode', varying = 2:9, sep = '_')

## Warning:  Setting row names on a tibble is deprecated.

## Warning:  Setting row names on a tibble is deprecated.

## Warning:  Setting row names on a tibble is deprecated.

## Warning:  Setting row names on a tibble is deprecated.

## Calculate estimated probabilities for each mode with express buses
predict_mlogit_express <- predict(model_mlogit_age,
                                  newdata = data_mlogit_express)
## Assign probabilities as variables in new dataset
data_multi_express <- data_multi_express %>%
  mutate(bike_probability_mlogit_express = predict_mlogit_express[, 1],
         bus_probability_mlogit_express = predict_mlogit_express[, 2],
         car_probability_mlogit_express = predict_mlogit_express[, 3],
         walk_probability_mlogit_express = predict_mlogit_express[, 4])
## Create index of individuals in new dataset
data_multi_express <- data_multi_express %>%
  mutate(id = 1:n())
## Find estimated choice for each individual with express buses
data_multi_express <- data_multi_express %>%
  select(id, contains('probability_mlogit_express')) %>%
  gather(mode_mlogit_express, probability_mlogit_express, -id) %>%
  group_by(id) %>%
  arrange(id, desc(probability_mlogit_express)) %>%
  slice(1) %>%
  ungroup() %>%
  select(id, mode_mlogit_express) %>%
  mutate(mode_mlogit_express = str_remove(mode_mlogit_express,
                                          '_probability_mlogit_express')) %>%
  inner_join(data_multi_express)

## Joining, by = "id"

## Count number of bus riders in original model
bus_mlogit_age <- data_multi_express %>%
  filter(mode_mlogit_age == 'bus') %>%
  nrow()
## Count number of bus riders with express buses
bus_mlogit_express <- data_multi_express %>%
  filter(mode_mlogit_express == 'bus') %>%
  nrow()
## Calculate difference in bus ridership due to express buses
bus_mlogit_express - bus_mlogit_age

## [1] 94

## Count number of bikers in original model
bike_mlogit_age <- data_multi_express %>%
  filter(mode_mlogit_age == 'bike') %>%
  nrow()
## Count number of bikers with express buses
bike_mlogit_express <- data_multi_express %>%
  filter(mode_mlogit_express == 'bike') %>%
  nrow()
## Calculate difference in bikers due to express buses
bike_mlogit_express - bike_mlogit_age

## [1] 0

## Count number of drivers in original model
car_mlogit_age <- data_multi_express %>%
  filter(mode_mlogit_age == 'car') %>%
  nrow()
## Count number of drivers with express buses
```

```
car_mlogit_express <- data_multi_express %>%
  filter(mode_mlogit_express == 'car') %>%
  nrow()
## Calculate difference in drivers due to express buses
car_mlogit_express - car_mlogit_age

## [1] -94

## Count number of walkers in original model
walk_mlogit_age <- data_multi_express %>%
  filter(mode_mlogit_age == 'walk') %>%
  nrow()
## Count number of walkers with express buses
walk_mlogit_express <- data_multi_express %>%
  filter(mode_mlogit_express == 'walk') %>%
  nrow()
## Calculate difference in walkers due to express buses
walk_mlogit_express - walk_mlogit_age

## [1] 0
```

The express bus routes would cause 94 graduate students to switch to riding the bus. All 94 of these students currently drive; no students would switch from biking or walking to taking a bus because of these express routes. This is intuitive since the express bus routes only reduce time on the long-distance bus routes. Any student who chooses to bike or walk such a long distance likely has a strong preference for that travel mode, so less travel time on the bus is unlikely to affect their decisions. Thus, the result is nearly identical to the result of part (d) in problem 2, when the only alternatives were riding the bus and driving.

d. The university may consider implementing express bus routes even though they do not cause an extra 10% of graduate students start taking the bus. Being an altruistic public institution, the university will begin running express buses if the value they provide graduate students exceeds the cost. The university expects the express routes will cost an extra $500 each day. Using your results from part (b), should the university implement these express bus routes? (Reminder: This is a random sample of 1,000 graduate students, not the entire graduate student population.)

```
### Part d
## Create express bus time variable
data_multi <- data_multi %>%
  mutate(time_bus_express = if_else(cost_bus == 3, time_bus - 10, time_bus))
## Calculate representative utility of each original mode
data_multi <- data_multi %>%
  mutate(utility_bike_mlogit_age = 0 +
           coef(model_mlogit_age)[4] * cost_bike / income +
           0 * age +
           coef(model_mlogit_age)[8] * time_bike,
         utility_bus_mlogit_age = coef(model_mlogit_age)[1] +
           coef(model_mlogit_age)[4] * cost_bus / income +
           coef(model_mlogit_age)[5] * age +
           coef(model_mlogit_age)[9] * time_bus,
         utility_car_mlogit_age = coef(model_mlogit_age)[2] +
           coef(model_mlogit_age)[4] * cost_car / income +
           coef(model_mlogit_age)[6] * age +
           coef(model_mlogit_age)[10] * time_car,
         utility_walk_mlogit_age = coef(model_mlogit_age)[3] +
           coef(model_mlogit_age)[4] * cost_walk / income +
           coef(model_mlogit_age)[7] * age +
           coef(model_mlogit_age)[11] * time_walk)
## Calculate representative utility of express buses
data_multi <- data_multi %>%
  mutate(utility_bus_mlogit_express = coef(model_mlogit_age)[1] +
           coef(model_mlogit_age)[4] * cost_bus / income +
           coef(model_mlogit_age)[5] * age +
           coef(model_mlogit_age)[9] * time_bus_express)
## Calculate sum of exponential of representative utilities for original modes
data_multi <- data_multi %>%
  mutate(sum_exp_mlogit_age = exp(utility_bike_mlogit_age) +
           exp(utility_bus_mlogit_age) + exp(utility_car_mlogit_age) +
           exp(utility_walk_mlogit_age))
## Calculate sum of exponential of representative utilities with express buses
data_multi <- data_multi %>%
  mutate(sum_exp_mlogit_express = exp(utility_bike_mlogit_age) +
           exp(utility_bus_mlogit_express) + exp(utility_car_mlogit_age) +
           exp(utility_walk_mlogit_age))
## Calculate change in consumer surplus from express buses
data_multi <- data_multi %>%
  mutate(surplus_change_express = (log(sum_exp_mlogit_express) -
                                     log(sum_exp_mlogit_age)) /
           abs(coef(model_mlogit_age)[4] / income))
## Scale consumer surplus change for total number of graduate students
sum(data_multi$surplus_change_express) * 7078 / 1000

## [1] 675.7171
```

The university will implement express bus routes if the additional consumer surplus they create is greater than $500. Let us assume that, because the daily cost of commuting is a small proportion of income, the marginal utility of income is approximately equal for a given decision maker, regardless of which alternative is chosen. Then we can calculate the change in consumer surplus as we discussed in lecture and as is presented in the Train textbook. When scaled up to account for all 7,078 graduate students, rather than only the 1,000 in this sample, the express bus routes would yield an additional $676 of consumer surplus each day. Since this is greater than the daily cost, the university should implement express bus routes and increase total social welfare.