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()`, `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.

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

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.)

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

## Parsed with column specification:
## cols(
##   id = col_double(),
##   mode = col_character(),
##   time.car = col_double(),
##   cost.car = col_double(),
##   time.bus = col_double(),
##   cost.bus = col_double(),
##   price_gas = col_double(),
##   snowfall = col_double(),
##   construction = col_double(),
##   bus_detour = col_double(),
##   age = col_double(),
##   income = col_double(),
##   marital_status = col_character()
## )
```

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.)

```
## Clean choice variable
data_binary <- data_binary %>%
  mutate(car = (mode == 'car'))
## Model choice as a linear probability model
reg_1a <- 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.)

```
## Calculate heteroskedastic-robust standard errors
coeftest(reg_1a, vcov = vcovHC(reg_1a))

##
## 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.)

```
## Calculate estimated probability of car for each individual
data_binary <- data_binary %>%
  mutate(prob_car_1a = predict(reg_1a))
## Count number of individuals with probabilities outside [0, 1]
data_binary %>%
  filter(prob_car_1a < 0 | prob_car_1a > 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_1a, 'time.car = -time.bus', vcov = vcovHC(reg_1a))

## 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.

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.)

```
## Model choice as binary logit
model_2a <- glm(formula = car ~ cost.car + time.car + time.bus,
                family = 'binomial',
                data = data_binary)
```

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.)

```
## Summarize model results
summary(model_2a)

##
## Call:
## glm(formula = car ~ cost.car + time.car + time.bus, family = "binomial",
##     data = data_binary)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7722  -0.9983  -0.5338   1.0524   3.1361
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.23327    0.34662   6.443 1.17e-10 ***
## cost.car    -2.07716    0.73245  -2.836  0.00457 **
## time.car    -0.33222    0.03534  -9.400  < 2e-16 ***
## time.bus     0.13257    0.03240   4.092 4.28e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1365.5  on 999  degrees of freedom
## Residual deviance: 1200.9  on 996  degrees of freedom
## AIC: 1208.9
##
## Number of Fisher Scoring iterations: 4
```

All three independent variables again have statistically significant and economically meaningful coefficients. Those coefficients, however, are now interpreted as marginal utilities rather than marginal effects. The cost of driving and the time spent driving both decrease the utility of driving, and the time spent riding the bus increases the utility of driving relative to riding the bus.

ii. Calculate the marginal effect of each independent variable for each student; that is, 3 variables $\times$ 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.)

```
## Calculate estimated utility and probability of car
data_binary <- data_binary %>%
  mutate(utility_2a = predict(model_2a),
         prob_car_2a = 1 / (1 + exp(-utility_2a)))
## Calculate marginal effects
data_binary <- data_binary %>%
  mutate(prob_prod_2a = prob_car_2a * (1 - prob_car_2a),
         mfx_cost_car = coef(model_2a)[2] * prob_prod_2a,
         mfx_time_car = coef(model_2a)[3] * prob_prod_2a,
         mfx_time_bus = coef(model_2a)[4] * prob_prod_2a)
## Summarize marginal effects
data_binary %>%
  select(starts_with('mfx')) %>%
  summary()
```

```
##   mfx_cost_car       mfx_time_car        mfx_time_bus
##  Min.   :-0.51929   Min.   :-0.083054   Min.   :0.0009629
##  1st Qu.:-0.51007   1st Qu.:-0.081579   1st Qu.:0.0248228
##  Median :-0.47723   Median :-0.076326   Median :0.0304589
##  Mean   :-0.43143   Mean   :-0.069001   Mean   :0.0275357
##  3rd Qu.:-0.38892   3rd Qu.:-0.062203   3rd Qu.:0.0325551
##  Max.   :-0.01509   Max.   :-0.002413   Max.   :0.0331436
```

The means of these marginal effects—reported above—are comparable to the estimated coefficients in problem 1, but there is heterogeneity around these means. For each marginal effect, there is a long tail that approaches zero, corresponding to students that have a probability of driving close to 0 or 1.

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.)

```
## Calculate hourly time-value for each commute mode at different incomes
abs(coef(model_2a)[3:4] / coef(model_2a)[2]) * 60

## time.car time.bus
## 9.596257 3.829494
```

Each hour of driving has a dollar value of $9.60 and each hour of bus riding has a dollar value of $3.83. In other words, a student would be willing to pay $9.60 to spend one less hour commute by car but only $3.83 to spend one less hour commuting by bus.

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.)

```
## Model choice as binary logit with cost divided by income
model_2b <- glm(formula = car ~ I(cost.car / income) + time.car + time.bus,
                family = 'binomial',
                data = data_binary)
```

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

```
## Summarize model results
summary(model_2b)
```

```
##
## Call:
## glm(formula = car ~ I(cost.car/income) + time.car + time.bus,
##     family = "binomial", data = data_binary)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7489  -0.9967  -0.5234   1.0442   3.1429
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)           2.26541    0.33110   6.842 7.81e-12 ***
## I(cost.car/income) -53.63314   14.54884  -3.686 0.000227 ***
## time.car            -0.33521    0.03484  -9.622  < 2e-16 ***
## time.bus             0.13589    0.02880   4.719 2.37e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1365.5  on 999  degrees of freedom
## Residual deviance: 1194.9  on 996  degrees of freedom
## AIC: 1202.9
##
## Number of Fisher Scoring iterations: 4
```

All three independent variables again have statistically significant and economically meaningful coefficients. The time coefficients are comparable to those estimated in part (a). The cost coefficient now varies with income; a higher level of income yields a lower marginal utility of income.

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.

```
## Calculate marginal utility of income at different incomes
-coef(model_2b)[2] / c(15, 25, 35)

## [1] 3.575543 2.145326 1.532375

## Calculate hourly time-value for each commute mode at different incomes
rep(abs(coef(model_2b)[3:4] / coef(model_2b)[2]), 3) *
  c(rep(15, 2), rep(25, 2), rep(35, 2)) * 60

##  time.car  time.bus  time.car  time.bus  time.car  time.bus
##  5.625096  2.280260  9.375160  3.800433 13.125224  5.320607
```

The marginal utility of income at each of these progressively higher incomes is 3.58, 2.15, and 1.53, respectively. At each of these incomes, each hour of driving has a dollar value of $5.63,

$9.38, and $13.13, respectively; and each hour of bus riding has a dollar value of $2.28, $3.80, and $5.32, respectively.

Problem 3: Multinomial Logit Model

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

```
## Load dataset
data_multi <- read_csv('commute_multinomial.csv')

## Parsed with column specification:
## cols(
##  id = col_double(),
##  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()
## )
```

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. The `dfidx()` function sometimes does not work on a `tibble`, so you may need to use the `as.data.frame()` function to ensure your data are in a `data.frame` format. 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.)

```
## Convert dataset to data frame format
data_df <- as.data.frame(data_multi)
## Convert dataset to mlogit format
```

```r
data_dfidx <- dfidx(data_df, shape = 'wide', choice = 'mode', varying = 3:10)
## Model choice as multinomial logit with common cost/income coefficient,
## alternative intercepts, and alternative-specific time coefficients
model_3a <- mlogit(formula = mode ~ I(cost / income) | 1 | time,
                   data = data_dfidx)
```

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

```r
## Summarize model results
summary(model_3a)

##
## Call:
## mlogit(formula = mode ~ I(cost/income) | 1 | time, data = data_dfidx,
##     method = "nr")
##
## Frequencies of alternatives:choice
##  bike   bus   car  walk
## 0.113 0.453 0.375 0.059
##
## nr method
## 8 iterations, 0h:0m:0s
## g'(-H)^-1g = 7.44E-06
## successive function values within tolerance limits
##
## Coefficients :
##                     Estimate Std. Error z-value  Pr(>|z|)
## (Intercept):bus   -0.256237   0.382073 -0.6706 0.5024438
## (Intercept):car    2.816387   0.438281  6.4260 1.310e-10 ***
## (Intercept):walk   3.000947   0.784617  3.8247 0.0001309 ***
## I(cost/income)   -57.797383  15.464584 -3.7374 0.0001859 ***
## time:bike         -0.283030   0.036182 -7.8224 5.107e-15 ***
## time:bus          -0.134744   0.031030 -4.3424 1.410e-05 ***
## time:car          -0.413471   0.044998 -9.1887 < 2.2e-16 ***
## time:walk         -0.295081   0.038231 -7.7184 1.177e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Log-Likelihood: -980.22
## McFadden R^2:  0.14008
## Likelihood ratio test : chisq = 319.34 (p.value = < 2.22e-16)
```

All independent variables again have statistically significant and economically meaningful parameters. The cost parameter again varies with income, yielding a greater marginal utility of income at lower incomes. The time parameter varies by alternative and is negative for all four alternatives, indicating that the marginal utility of commute time is consistently negative

regardless of commute mode. The parameter values differ, however, providing evidence that time driving creates the greatest disutility and time riding the bus creates the least disutility.

ii. Calculate the elasticity of each commute alternative with respect to the cost of driving for each student; that is, 4 alternatives $\times$ 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.)

```
## Calculate the choice probabilities for car
data_multi <- data_multi %>%
  mutate(prob_car_3a = fitted(model_3a, type = 'probabilities')[, 3])
## Calculate the own elasticity of car cost
data_multi <- data_multi %>%
  mutate(elas_own_car_cost_3a =
           coef(model_3a)[4] * (cost.car / income) * (1 - prob_car_3a))
## Calculate the cross-elasticity of car cost
data_multi <- data_multi %>%
  mutate(elas_cross_car_cost_3a =
           -coef(model_3a)[4] * (cost.car / income) * prob_car_3a)
## Summarize elasticities
data_multi %>%
  select(starts_with('elas')) %>%
  summary()
```

```
##  elas_own_car_cost_3a elas_cross_car_cost_3a
##  Min.   :-4.4549      Min.   :0.01524
##  1st Qu.:-0.8268      1st Qu.:0.22315
##  Median :-0.4802      Median :0.31333
##  Mean   :-0.6645      Mean   :0.31519
##  3rd Qu.:-0.3070      3rd Qu.:0.40649
##  Max.   :-0.1000      Max.   :0.83561
```

The summary statistics for own-elasticity and cross-elasticity of driving cost are reported above. Note that all three other alternatives—biking, riding the bus, and walking—have the same elasticity with respect to the cost of driving. This common cross-elasticity is an example of the independence of irrelevant alternatives (IIA), which implies proportional substitution to or from all other alternatives.

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.

```
## Create a separate dataset of single students
data_dfidx_single <- data_dfidx %>%
  filter(marital_status == 'single')
## Create a separate datasets of married students
data_dfidx_married <- data_dfidx %>%
```

```
  filter(marital_status == 'married')
## Model choice for single students
model_3b_single <- mlogit(formula = mode ~ I(cost / income) | 1 | time,
                          data = data_dfidx_single)
## Model choice for single students
model_3b_married <- mlogit(formula = mode ~ I(cost / income) | 1 | time,
                           data = data_dfidx_married)
```

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

```
## Summarize model results for single students
summary(model_3b_single)

##
## Call:
## mlogit(formula = mode ~ I(cost/income) | 1 | time, data = data_dfidx_single,
##     method = "nr")
##
## Frequencies of alternatives:choice
##     bike      bus      car     walk
## 0.136508 0.412698 0.373016 0.077778
##
## nr method
## 7 iterations, 0h:0m:0s
## g'(-H)^-1g = 2.57E-07
## gradient close to zero
##
## Coefficients :
##                     Estimate Std. Error z-value  Pr(>|z|)
## (Intercept):bus   -0.617806   0.446835 -1.3826 0.1667792
## (Intercept):car    2.027036   0.494410  4.0999 4.133e-05 ***
## (Intercept):walk   2.719800   0.815557  3.3349 0.0008533 ***
## I(cost/income)   -51.619079  18.633104 -2.7703 0.0056007 **
## time:bike         -0.257675   0.042466 -6.0677 1.297e-09 ***
## time:bus          -0.109750   0.038796 -2.8289 0.0046705 **
## time:car          -0.330645   0.052043 -6.3534 2.107e-10 ***
## time:walk         -0.264483   0.039102 -6.7639 1.343e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Log-Likelihood: -666.28
## McFadden R^2:  0.12129
## Likelihood ratio test : chisq = 183.94 (p.value = < 2.22e-16)

## Summarize model results for married students
summary(model_3b_married)
```

```
##
## Call:
## mlogit(formula = mode ~ I(cost/income) | 1 | time, data = data_dfidx_married,
##     method = "nr")
##
## Frequencies of alternatives:choice
##     bike      bus      car     walk
## 0.072973 0.521622 0.378378 0.027027
##
## nr method
## 9 iterations, 0h:0m:0s
## g'(-H)^-1g = 4.67E-06
## successive function values within tolerance limits
##
## Coefficients :
##                     Estimate Std. Error z-value  Pr(>|z|)
## (Intercept):bus     0.155878   0.796138  0.1958 0.8447722
## (Intercept):car     4.811120   0.978671  4.9160 8.834e-07 ***
## (Intercept):walk    4.666696   2.463856  1.8941 0.0582168 .
## I(cost/income)    -79.357983  29.275838 -2.7107 0.0067142 **
## time:bike          -0.369977   0.075932 -4.8725 1.102e-06 ***
## time:bus           -0.189161   0.053586 -3.5301 0.0004154 ***
## time:car           -0.649090   0.096271 -6.7423 1.559e-11 ***
## time:walk          -0.445265   0.130962 -3.4000 0.0006740 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Log-Likelihood: -289.87
## McFadden R^2:  0.21327
## Likelihood ratio test : chisq = 157.16 (p.value = < 2.22e-16)
```

In both models, the marginal utility parameters are again statistically significant and economically meaningful. They have the same sign as in part (a), although the parameter values differ by approximately 50% in these two models. But we should not directly compare these marginal utility values because they come from different models.

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.

```
## Calculate marginal utility of car cost for single students
-coef(model_3b_single)[4] / 25

## I(cost/income)
##       2.064763

## Calculate marginal utility of car cost for married students
-coef(model_3b_married)[4] / 25
```

```
## I(cost/income)
##       3.174319

## Calculate hourly time-value for each commute mode for single students
abs(coef(model_3b_single)[5:8] / coef(model_3b_single)[4]) * 25 * 60

## time:bike  time:bus  time:car time:walk
##  7.487796  3.189233  9.608231  7.685613

## Calculate hourly time-value for each commute mode for married students
abs(coef(model_3b_married)[5:8] / coef(model_3b_married)[4]) * 25 * 60

## time:bike  time:bus  time:car time:walk
##  6.993187  3.575471 12.268890  8.416260
```

The marginal utility of income for a single student with $25,000 income is 2.06, and the marginal utility of income for a married students with $25,000 income is 3.17, but these values are not directly comparable because they come from different models. A single student with $25,000 income has an hourly dollar value of $7.49 for biking, $3.19 for riding the bus, $9.61 for driving, and $7.69 for walking. A married student with $25,000 income has an hourly dollar value of $6.99 for biking, $3.58 for riding the bus, $12.27 for driving, and $8.42 for walking. These hourly dollar values differ by only 10% in most cases.

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.)

The parameter of a logit model are estimated relative to the variance of the random utility component—greater variance of random utility yields lower parameter estimates. When taking the ratio of two parameters, however, the scale parameter drops out of the calculation. Thus, if single students and married students tend to have a different variance of random utility, these model would yield different parameter estimates, even for the same underlying preferences and valuations of commute time.

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.

```
## Create counterfactual data with more frequent buses
data_df_counter <- data_df %>%
  mutate(time.bus = 0.8 * time.bus)
## Convert counterfactual data to dfidx format
data_counter_dfidx <- dfidx(data_df_counter, shape = 'wide',
                            choice = 'mode', varying = 3:10)
```

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?

```
## Calculate aggregate choices using observed data
agg_choices_obs <- predict(model_3a, newdata = data_dfidx)
## Calculate aggregate choices using counterfactual data
agg_choices_counter <- predict(model_3a, newdata = data_counter_dfidx)
## Calculate difference between aggregate choices
colSums(agg_choices_counter - agg_choices_obs)
```

```
##       bike        bus        car       walk
## -16.699988  73.545557 -51.337560  -5.508009
```

This reduction in bus commute time is expected to yield an additional 73.5 students riding the bus, or an additional 7.35% of the students in the dataset. Of these 73.5 additional bus riders, 16.7 previously biked, 51.3 previously drove, and 5.5 previously walked.

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

```
## Calculate log-sum values using observed data
logsum_obs <- logsum(model_3a, data = data_dfidx)
## Calculate log-sum values using counterfactual data
logsum_counter <- logsum(model_3a, data = data_counter_dfidx)
## Calculate change in consumer surplus from subsidy
sum((logsum_counter - logsum_obs) / (-coef(model_3a)[4] / data_df$income))
```

```
## [1] 86.79554
```

This reduction in bus commute time is expected to generate $86.80 in economic surplus for these 1000 students each day.