Problem Set 2

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., `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. 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(mlogit)
```

Problem 1: Binary Logit Model

We are again 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—as in problem set 1—but we will use a different model of student decision making. The model in problem 2 of problem set 1 assumed the probability of driving is a linear function of the data. In reality, however, a different functional form may provide a better fit for the data. Use the `commute_binary.csv` dataset for this problem.

```r
## 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.

## Clean choice variable
data_binary <- data_binary %>%
  mutate(car = (mode == 'car'))
```

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_1a <- 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_1a)

##
## 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 have statistically significant and economically meaningful coefficients, which are interpreted as marginal utilities. 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 × 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 2 of problem set 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_1a = predict(model_1a),
         prob_car_1a = 1 / (1 + exp(-utility_1a)))
## Calculate marginal effects
data_binary <- data_binary %>%
  mutate(prob_prod_1a = prob_car_1a * (1 - prob_car_1a),
         mfx_cost_car = coef(model_1a)[2] * prob_prod_1a,
         mfx_time_car = coef(model_1a)[3] * prob_prod_1a,
         mfx_time_bus = coef(model_1a)[4] * prob_prod_1a)
## 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 2 of problem set 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
abs(coef(model_1a)[3:4] / coef(model_1a)[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 commuting 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_1b <- 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_1b)

##
## 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 car cost at different incomes
-coef(model_1b)[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_1b)[3:4] / coef(model_1b)[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 2: Multinomial Logit Model

We are again studying how UMass graduate students choose how to commute to campus, but we are now interested in this choice in the spring when riding a bike and walking are feasible alternatives. This information will help the university to plan for car parking, bike racks, and bus needs during this time of year. Additionally, the university is considering a change to bus routes, and they want to know how this change will affect commute choices. Use the `commute_multinomial.csv` dataset for this problem.

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

## Rows:  1000 Columns:  13
## -- Column specification --------------------------------------------------------
## Delimiter:  ","
## chr  (2):  mode, marital_status
## dbl (11):  id, time.car, cost.car, time.bus, cost.bus, time.bike, cos...
##
## 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. 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, assume cost has a common parameter that does not vary with income, and allow the parameter on time to be alternative-specific. That is, the representative utility to student n from alternative j is

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

where V_{nj} is the representative utility to student n from alternative j, 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
data_dfidx <- dfidx(data_df, shape = 'wide', choice = 'mode', varying = 3:10)
## Model choice as multinomial logit with common cost coefficient,
## alternative intercepts, and alternative-specific time coefficients
model_2a <- mlogit(formula = mode ~ cost | 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?

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

```
##
## Call:
## mlogit(formula = mode ~ cost | 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 = 6.84E-06
## successive function values within tolerance limits
##
## Coefficients :
##                   Estimate Std. Error z-value  Pr(>|z|)
## (Intercept):bus  -0.219000   0.385544 -0.5680 0.5700161
## (Intercept):car   2.745681   0.442595  6.2036 5.518e-10 ***
## (Intercept):walk  2.975472   0.783182  3.7992 0.0001452 ***
## cost             -2.604415   0.823533 -3.1625 0.0015643 **
## time:bike        -0.289389   0.038564 -7.5041 6.195e-14 ***
## time:bus         -0.143175   0.035108 -4.0781 4.540e-05 ***
## time:car         -0.404666   0.046377 -8.7255 < 2.2e-16 ***
## time:walk        -0.296615   0.038420 -7.7204 1.155e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Log-Likelihood: -982.36
## McFadden R^2:  0.1382
## Likelihood ratio test : chisq = 315.07 (p.value = < 2.22e-16)
```

All independent variables again have statistically significant and economically meaningful parameters. The cost parameter is negative, indicating that the marginal utility of cost is negative and, hence, the marginal utility of income is positive. 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 × 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_2a = fitted(model_2a, type = 'probabilities')[, 3])
## Calculate the own elasticity of car cost
```

```
data_multi <- data_multi %>%
  mutate(elas_own_car_cost_2a =
           coef(model_2a)[4] * cost.car * (1 - prob_car_2a))
## Calculate the cross-elasticity of car cost
data_multi <- data_multi %>%
  mutate(elas_cross_car_cost_2a =
           -coef(model_2a)[4] * cost.car * prob_car_2a)
## Summarize elasticities
data_multi %>%
  select(starts_with('elas')) %>%
  summary()
```

```
##  elas_own_car_cost_2a elas_cross_car_cost_2a
##  Min.   :-4.6129      Min.   :0.02001
##  1st Qu.:-0.8759      1st Qu.:0.25086
##  Median :-0.5256      Median :0.34872
##  Mean   :-0.7237      Mean   :0.34964
##  3rd Qu.:-0.3550      3rd Qu.:0.44536
##  Max.   :-0.1088      Max.   :0.91753
```

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_2b_single <- mlogit(formula = mode ~ cost | 1 | time,
                          data = data_dfidx_single)
## Model choice for single students
model_2b_married <- mlogit(formula = mode ~ cost | 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_2b_single)
```

```
##
## Call:
## mlogit(formula = mode ~ cost | 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 = 2.36E-07
## gradient close to zero
##
## Coefficients :
##                   Estimate Std. Error z-value  Pr(>|z|)
## (Intercept):bus  -0.552670   0.451603 -1.2238 0.2210298
## (Intercept):car   1.934913   0.498573  3.8809 0.0001041 ***
## (Intercept):walk  2.687992   0.815153  3.2975 0.0009754 ***
## cost             -2.711833   1.018920 -2.6615 0.0077799 **
## time:bike        -0.272877   0.046166 -5.9108 3.404e-09 ***
## time:bus         -0.128379   0.044525 -2.8833 0.0039357 **
## time:car         -0.316254   0.053243 -5.9398 2.854e-09 ***
## time:walk        -0.269539   0.039545 -6.8159 9.366e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Log-Likelihood: -666.61
## McFadden R^2:  0.12086
## Likelihood ratio test : chisq = 183.29 (p.value = < 2.22e-16)
```

```
## Summarize model results for married students
summary(model_2b_married)
```

```
##
## Call:
## mlogit(formula = mode ~ cost | 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 = 4.33E-06
## successive function values within tolerance limits
##
## Coefficients :
##                   Estimate Std. Error z-value  Pr(>|z|)
## (Intercept):bus   0.154527   0.798997  0.1934 0.8466446
## (Intercept):car   4.781125   0.987391  4.8422 1.284e-06 ***
## (Intercept):walk  4.610041   2.452583  1.8797 0.0601533 .
## cost             -2.726736   1.489068 -1.8312 0.0670753 .
## time:bike        -0.362261   0.078168 -4.6344 3.580e-06 ***
## time:bus         -0.182139   0.058976 -3.0884 0.0020125 **
## time:car         -0.656574   0.100850 -6.5104 7.494e-11 ***
## time:walk        -0.439187   0.130516 -3.3650 0.0007654 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Log-Likelihood: -292.07
## McFadden R^2:  0.2073
## Likelihood ratio test : chisq = 152.76 (p.value = < 2.22e-16)
```

In both models, the marginal utility parameters are again statistically significant—at the 10% level in some cases—and economically meaningful. They also have the same signs as in part (a), so the general interpretation is the same.

ii. Can you use your estimated parameters to compare the marginal utility of income for single students to the marginal utility of income for married students? If so, describe the similarity or difference in these values. If not, explain why you cannot make this comparison using your estimated parameters. (Hint: think about what component of the random utility model is assumed to be the same in both models.)

    The marginal utility parameters are not directly comparable because they are estimated in different models. In every logit model, the variance of the random utility component has the same assumed value. In reality, however, this variance may differ for single students and married students. As a result, the parameters of these models are estimated relative to the true variance of the random utility component—greater variance of random utility yields lower parameter estimates. Thus, if single students and married students have different variances of random utility, these models could yield different parameter estimates, even for the same underlying preferences and valuations of commute time. Conversely, these models could yield similar parameter estimates, even for the very different underlying preferences and valuations of commute time, so long as the variances differ similarly.

iii. For each marital status, use the corresponding parameter estimates to calculate the dollar value that a student places on one hour of commute time for each of the four commute alternatives. Can you compare these dollar values for single students to those for married students? If so, describe the similarity or difference in these values. If not, explain why you cannot make this comparison.

```
## Calculate hourly time-value for each commute mode for single students
abs(coef(model_2b_single)[5:8] / coef(model_2b_single)[4]) * 60

## time:bike  time:bus  time:car time:walk
##  6.037481  2.840412  6.997210  5.963612

## Calculate hourly time-value for each commute mode for married students
abs(coef(model_2b_married)[5:8] / coef(model_2b_married)[4]) * 60

## time:bike  time:bus  time:car time:walk
##  7.971306  4.007852 14.447473  9.664010
```

A single student has an hourly dollar value of $6.04 for biking, $2.84 for riding the bus, $7.00 for driving, and $5.96 for walking. A married student has an hourly dollar value of $7.97 for biking, $4.01 for riding the bus, $14.45 for driving, and $9.66 for walking. Yes, these dollar values can be compared. Intuitively, dollars provide a standardized metric for comparison, unlike utility. Mathematically, taking the ratio of two model parameters cancels out the normalization due to the assumed variance of the random utility component. These hourly dollar values indicate that married students place a greater value on their commute time, especially time spent driving.

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_2a, newdata = data_dfidx)
## Calculate aggregate choices using counterfactual data
agg_choices_counter <- predict(model_2a, newdata = data_counter_dfidx)
## Calculate difference between aggregate choices
colSums(agg_choices_counter - agg_choices_obs)

##       bike        bus        car       walk
## -17.646719  78.164883 -54.675570  -5.842595
```

This reduction in bus commute time is expected to yield an additional 78.2 students riding the bus, or an additional 7.82% of the students in the dataset. Of these 78.2 additional bus riders, 17.6 previously biked, 54.7 previously drove, and 5.8 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?

```r
## Calculate log-sum values using observed data
logsum_obs <- logsum(model_2a, data = data_dfidx)
## Calculate log-sum values using counterfactual data
logsum_counter <- logsum(model_2a, data = data_counter_dfidx)
## Calculate change in consumer surplus from subsidy
sum((logsum_counter - logsum_obs) / -coef(model_2a)[4])
```

```
## [1] 82.74048
```

This reduction in bus commute time is expected to generate $82.74 in economic surplus for these 1000 students each day.