A3

Tian Chen

03/22/2022

Set up environment and import data

```
#clear environment
rm(list = ls())

#import package
require(tidyverse)

## Loading required package: tidyverse

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.
3.1 ──

## ✓ ggplot2 3.3.5     ✓ purrr   0.3.4
## ✓ tibble  3.1.5     ✓ dplyr   1.0.7
## ✓ tidyr   1.1.4     ✓ stringr 1.4.0
## ✓ readr   2.0.1     ✓ forcats 0.5.1

## ── Conflicts ────────────────────────────────────────── tidyverse_conflict
s() ──
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

#import data
datsss <- read.csv("~/Desktop/Duke study/Econ613/A3/datsss.csv")
datjss <- read.csv("~/Desktop/Duke study/Econ613/A3/datjss.csv")
datstu_v2 <- read.csv("~/Desktop/Duke study/Econ613/A3/datstu_v2.csv")
datsss_tbl <- datsss %>% as_tibble() #school
datjss_tbl <- datjss %>% as_tibble() #geolocation of school
datstu_v2_tbl <- datstu_v2 %>% as_tibble() #students
```

Exercise 1

###1. Number of Students, schools, programs

```
#Number of students
length(unique(datstu_v2$V1))

## [1] 340823

#Number of Schools
length(unique(datsss$schoolname))

## [1] 842
```

```
#Number of programs
program <- select(datstu_v2_tbl, choicepgm1:choicepgm6)
program_name <- names(program)
unique_name <- vector()

#Find unique program name
for (i in c(1:length(program_name))) {
  unique_col <- unique(select(program, program_name[i]))
  unique_col <- rename(unique_col, program = program_name[i])
  unique_name <- unique(rbind(unique_name, unique_col))
}
length(unique_name$program)

## [1] 33
```

###2. Number of Choices

```
school_program <- datstu_v2_tbl %>% select(V1, schoolcode1:choicepgm6)
#transfer from wide to long
program<- school_program %>% select(V1, choicepgm1:choicepgm6) %>% pivot_long
er(cols = starts_with("choice"),
                                                    names_to = "choice_progran
",
                                                    values_to = "program_name")

school <- school_program %>% select(V1, schoolcode1:schoolcode6) %>% pivot_lo
nger(cols = starts_with("schoolcode"),
                                                    names_to = "choice_school",
                                                    values_to = "schoolcode")

school_program <- cbind(school, program) %>% select(program_name, schoolcode)
 %>% unique()
length(school_program$schoolcode)

## [1] 3086
```

###3. Number of students applying to at least one senior high school in the same district to home

```
students_school <- datstu_v2_tbl %>% select(V1, schoolcode1:schoolcode6, jssd
istrict) %>% pivot_longer(cols = starts_with("schoolcode"),

                          names_to = "choice_school",

                          values_to = "schoolcode")
school_location <- datsss_tbl %>% select(schoolcode, sssdistrict)
students_school <- left_join(students_school, school_location, by = "schoolco
de")
students_school <- students_school %>% mutate(same_district = ifelse(students
_school$jssdistrict == students_school$sssdistrict, 1, 0))
#count the at least one same location
```

```
students_school <- students_school %>% group_by(V1) %>% summarise(count = sum
(same_district)) %>% filter(count>=1)
length(students_school$V1)

## [1] 254096
```

###4. Number of students each senior high school admitted

```
school_number <- datstu_v2_tbl %>% select(V1, schoolcode1:schoolcode6, rankpl
ace) %>% drop_na() %>% pivot_longer(cols = starts_with("schoolcode"),

                            names_to = "choice_school",

                            values_to = "schoolcode")

school_number <- school_number %>% mutate(choice_school_number = 0,
                                          choice_school_number = replace
(choice_school_number, choice_school == "schoolcode1", 1),
                                          choice_school_number = replace
(choice_school_number, choice_school == "schoolcode2", 2),
                                          choice_school_number = replace
(choice_school_number, choice_school == "schoolcode3", 3),
                                          choice_school_number = replace
(choice_school_number, choice_school == "schoolcode4", 4),
                                          choice_school_number = replace
(choice_school_number, choice_school == "schoolcode5", 5),
                                          choice_school_number = replace
(choice_school_number, choice_school == "schoolcode6", 6),
                                          )
school_number <- school_number %>% mutate(admitted = ifelse(school_number$cho
ice_school_number == school_number$rankplace, 1, 0))
school_number <- school_number %>% select(schoolcode, admitted) %>% group_by
(schoolcode) %>% summarise(count = sum(admitted))
```

###5. The cutoff of senior high schools

```
school_score <- datstu_v2_tbl %>% select(V1, score, schoolcode1:schoolcode6,
rankplace) %>% drop_na() %>% pivot_longer(cols = starts_with("schoolcode"),

                            names_to = "choice_school",

                            values_to = "schoolcode")

school_score <- school_score %>% mutate(choice_school_number = 0,
                                          choice_school_number = replace
(choice_school_number, choice_school == "schoolcode1", 1),
                                          choice_school_number = replace
(choice_school_number, choice_school == "schoolcode2", 2),
                                          choice_school_number = replace
(choice_school_number, choice_school == "schoolcode3", 3),
                                          choice_school_number = replace
```

```
(choice_school_number, choice_school == "schoolcode4", 4),
                                                  choice_school_number = replace
(choice_school_number, choice_school == "schoolcode5", 5),
                                                  choice_school_number = replace
(choice_school_number, choice_school == "schoolcode6", 6),
                                                  )
school_score <- school_score %>% mutate(admitted = ifelse(school_score$choice
_school_number == school_score$rankplace, 1, 0)) %>% filter(admitted == 1)

school_cutoff <- school_score %>% select(score, schoolcode) %>% group_by(scho
olcode) %>% summarise(cutoff = min(score))
head(school_cutoff)
```

```
## # A tibble: 6 × 2
##   schoolcode cutoff
##        <int>  <int>
## 1      10101    284
## 2      10102    343
## 3      10103    316
## 4      10104    245
## 5      10105    260
## 6      10106    293
```

###6. The quality of senior high schools

```
school_quality <- school_score %>% select(score, schoolcode) %>% group_by(sch
oolcode) %>% summarise(quality = mean(score))
head(school_quality)
```

```
## # A tibble: 6 × 2
##   schoolcode quality
##        <int>   <dbl>
## 1      10101    320.
## 2      10102    394.
## 3      10103    354.
## 4      10104    297.
## 5      10105    351.
## 6      10106    340.
```

Exercise2 Data

```
#(school, program)
#transfer from wide to long
program<- datstu_v2_tbl %>% select(V1, choicepgm1:choicepgm6) %>% pivot_longe
r(cols = starts_with("choice"),
                                                  names_to = "choice_progran
",
                                                  values_to = "program_name")
 %>% select(!V1)
school <- datstu_v2_tbl %>% select(!(choicepgm1:choicepgm6)) %>% pivot_longer
(cols = starts_with("schoolcode"),
```

```
                                                    names_to = "choice_school",
                                                    values_to = "schoolcode")
#find cutoff, quality, and size
school_program <- cbind(program, school) %>% mutate(choice_school_number = 0,

                                                    choice_school_number = replace(ch
oice_school_number, choice_school == "schoolcode1", 1),
                                                    choice_school_number = replace(ch
oice_school_number, choice_school == "schoolcode2", 2),
                                                    choice_school_number = replace(ch
oice_school_number, choice_school == "schoolcode3", 3),
                                                    choice_school_number = replace(ch
oice_school_number, choice_school == "schoolcode4", 4),
                                                    choice_school_number = replace(ch
oice_school_number, choice_school == "schoolcode5", 5),
                                                    choice_school_number = replace(ch
oice_school_number, choice_school == "schoolcode6", 6),
                                                      )
school_program <- school_program %>% mutate(admitted = ifelse(school_program
$choice_school_number == school_program$rankplace, 1, 0))%>% filter(admitted
== 1)
school_program <- school_program %>% group_by(schoolcode, program_name) %>% s
ummarize(cutoff = min(score, na.rm = TRUE),
                                                                          quali
ty = mean(score, na.rm = TRUE),
                                                                          size
= n())

## `summarise()` has grouped output by 'schoolcode'. You can override using t
he `.groups` argument.

#merge (school, program) with datsss
school_info <- datsss_tbl %>% select(-V1) %>% drop_na() %>% unique() %>% muta
te(length = str_count(schoolname))
#because of the data duplication, we use the longest name in each schoolcode
as the school name
school_info_unique <- school_info %>% group_by(schoolcode) %>% slice_max(leng
th, n = 1) %>% select(-length)
school_level_data <- left_join(school_program, school_info_unique, by = "scho
olcode")
head(school_level_data)

## # A tibble: 6 × 9
## # Groups:   schoolcode [1]
##   schoolcode program_name    cutoff quality  size schoolname sssdistrict s
sslong
##        <int> <chr>            <int>   <dbl> <int> <chr>      <chr>
 <dbl>
## 1      10101 Agriculture        288    310.    49 EBENEZER … Accra Metr…
-0.197
```

```
## 2      10101 Business           305    325.   100 EBENEZER … Accra Metr…
-0.197
## 3      10101 General Arts       316    330.   100 EBENEZER … Accra Metr…
-0.197
## 4      10101 General Science    299    329.    50 EBENEZER … Accra Metr…
-0.197
## 5      10101 Home Economics     284    301.    49 EBENEZER … Accra Metr…
-0.197
## 6      10101 Visual Arts        296    312.    50 EBENEZER … Accra Metr…
-0.197
## # … with 1 more variable: ssslat <dbl>
```

Exercise3 Distance

```
#individual school_program data
ind_data <- cbind(school, program)
#merge with school-level data
ind_data <- left_join(ind_data, school_level_data, by = c("program_name", "sc
hoolcode"))
#merge with datjss (home location)
location <- datjss_tbl %>% drop_na() %>% filter(jssdistrict != "")
ind_withdistance <- left_join(ind_data,location, by = "jssdistrict") %>% rena
me(jsslong = point_x, jsslat = point_y)
#calculate the distance
ind_data_withdistance <- ind_withdistance %>% mutate(distance = sqrt((69.172*
(ssslong-jsslong)*cos(jsslat/53))^2 + (69.172*(ssslat-jsslat))^2))
head(ind_data_withdistance)
```

```
##   V1 score agey male                          jssdistrict rankplace
## 1  1    NA   16    0 Bosomtwe/Atwima/Kwanwoma (Kuntanase)        NA
## 2  1    NA   16    0 Bosomtwe/Atwima/Kwanwoma (Kuntanase)        NA
## 3  1    NA   16    0 Bosomtwe/Atwima/Kwanwoma (Kuntanase)        NA
## 4  1    NA   16    0 Bosomtwe/Atwima/Kwanwoma (Kuntanase)        NA
## 5  1    NA   16    0 Bosomtwe/Atwima/Kwanwoma (Kuntanase)        NA
## 6  1    NA   16    0 Bosomtwe/Atwima/Kwanwoma (Kuntanase)        NA
##   choice_school schoolcode choice_progran   program_name cutoff  quality s
ize
## 1   schoolcode1      50112     choicepgm1 Home Economics    293 312.3200
 50
## 2   schoolcode2      50107     choicepgm2   General Arts    375 386.1778
135
## 3   schoolcode3      50202     choicepgm3    Visual Arts    321 333.7000
 50
## 4   schoolcode4      50202     choicepgm4    Visual Arts    321 333.7000
 50
## 5   schoolcode5      50702     choicepgm5 Home Economics    272 289.2833
 60
## 6   schoolcode6      50901     choicepgm6   General Arts    217 254.4417
120
##                                   schoolname                sssdistrict
## 1     KUMASI SENIOR HIGH./TECH. SCHOOL, KUMASI               Kumasi Metro
```

```
## 2     ANGLICAN SENIOR HIGH SCHOOL, ASEM-KUMASI                Kumasi Metro
## 3              TOASE SENIOR HIGH SCHOOL, TOASE Atwima / Nwabiagya (Nkawie)
## 4              TOASE SENIOR HIGH SCHOOL, TOASE Atwima / Nwabiagya (Nkawie)
## 5     SIMMS SENIOR HIGH. COMM. SCHOOL, FAWOADE         Kwabre (Mamponteng)
## 6 EJURAMAN ANGLICAN SENIOR HIGH. SCHOOL, EJURA   Ejura/Sekyedumase (Ejura)
##     ssslong   ssslat  X   jsslong   jsslat  distance
## 1 -1.597187 6.682060 23 -1.562752 6.559323  8.812873
## 2 -1.597187 6.682060 23 -1.562752 6.559323  8.812873
## 3 -1.808757 6.681337 23 -1.562752 6.559323 18.878252
## 4 -1.808757 6.681337 23 -1.562752 6.559323 18.878252
## 5 -1.541420 6.806778 23 -1.562752 6.559323 17.179514
## 6 -1.367965 7.462874 23 -1.562752 6.559323 63.914633
```

- Exercise4 Dimensionality Reduction

###Recode the schoolcode into its three digits(substr). Call this new variable scode_rev

```
data <- ind_data_withdistance
data <- data %>% mutate(scode_rev = substr(schoolcode, 1, 3))
```

###Recode the program variable into 4 categories. Call this new variable pgm_rev

```
#Recode the program variable into 4 categories
data <- data %>% mutate(pgm_rev = ifelse(data$program_name == "General Arts"|
data$program_name == "Visual Arts", "arts",
                                         ifelse(data$program_name == "Busines
s"|data$program_name == "Home Economics", "economics",
                                                ifelse(data$program_name == "
General Science", "science", "others"))))
```

###Create a new choice variable choice_rev

```
data <- data %>% mutate(choice_rev = paste(data$scode_rev, data$pgm_rev))
```

###Recalculate the cutoff and the quality for each recoded choice

```
cutoff_quality<- data %>% mutate(choice_school_number = 0,
                                         choice_school_number = replace(ch
oice_school_number, choice_school == "schoolcode1", 1),
                                         choice_school_number = replace(ch
oice_school_number, choice_school == "schoolcode2", 2),
                                         choice_school_number = replace(ch
oice_school_number, choice_school == "schoolcode3", 3),
                                         choice_school_number = replace(ch
oice_school_number, choice_school == "schoolcode4", 4),
                                         choice_school_number = replace(ch
oice_school_number, choice_school == "schoolcode5", 5),
                                         choice_school_number = replace(ch
oice_school_number, choice_school == "schoolcode6", 6),
                                           )
cutoff_quality <- cutoff_quality %>% mutate(admitted = ifelse(cutoff_quality
$choice_school_number == cutoff_quality $rankplace, 1, 0))%>% filter(admitted
```

```
 == 1)
cutoff_quality <- cutoff_quality %>% group_by(choice_rev) %>% summarize(cutof
f = min(score,na.rm = TRUE),
                                                                        quali
ty = min(score, na.rm = TRUE))
data <- data %>% select(-cutoff, -quality) %>% left_join(cutoff_quality, by =
 "choice_rev")
```

###Consider the 20,000 highest score students

```
top_20000_students <- data %>% select(V1, score) %>% unique() %>% arrange(des
c(score))
data_top20000 <- data %>% filter(V1 %in% top_20000_students$V1[1:20000])
head(data_top20000)
```

```
##       V1 score agey male       jssdistrict rankplace choice_school schoolc
ode
## 1 179982   375   17    0 Ga East (Abokobi)         1   schoolcode1      21
001
## 2 179982   375   17    0 Ga East (Abokobi)         1   schoolcode2      21
002
## 3 179982   375   17    0 Ga East (Abokobi)         1   schoolcode3      21
006
## 4 179982   375   17    0 Ga East (Abokobi)         1   schoolcode4      21
009
## 5 179982   375   17    0 Ga East (Abokobi)         1   schoolcode5      21
401
## 6 179982   375   17    0 Ga East (Abokobi)         1   schoolcode6      21
201
##   choice_progran   program_name size
## 1     choicepgm1       Business   80
## 2     choicepgm2       Business   80
## 3     choicepgm3       Business  230
## 4     choicepgm4       Business   13
## 5     choicepgm5 Home Economics   49
## 6     choicepgm6 Home Economics   30
##                                    schoolname           sssdistrict
## 1   ABETIFI PRESBY SENIOR HIGH. SCHOOL, ABETIFI Kwahu South (Mpraeso)
## 2          MPRAESO SENIOR HIGH. SCHOOL, MPRAESO Kwahu South (Mpraeso)
## 3 NKWATIA PRESBY SENIOR HIGH./COMM SCH, NKWATIA Kwahu South (Mpraeso)
## 4              BEPONG SENIOR HIGH/COMM., BEPONG Kwahu South (Mpraeso)
## 5     YILO KROBO SENIOR HIGH./COMM SCH, SOMANYA  Yilo Krobo (Somanya)
## 6              PRESBY SENIOR HIGH. SCH., BEGORO    Fanteakwa (Begoro)
##      ssslong   ssslat  X    jsslong   jsslat distance scode_rev   pgm_rev
## 1 -0.6355287 6.619226 81 -0.2411459 5.721143 67.78456       210 economics
## 2 -0.6355287 6.619226 81 -0.2411459 5.721143 67.78456       210 economics
## 3 -0.6355287 6.619226 81 -0.2411459 5.721143 67.78456       210 economics
## 4 -0.6355287 6.619226 81 -0.2411459 5.721143 67.78456       210 economics
## 5 -0.1815932 6.186355 81 -0.2411459 5.721143 32.43923       214 economics
## 6 -0.3560941 6.436071 81 -0.2411459 5.721143 50.08081       212 economics
```

```
##      choice_rev cutoff quality
## 1 210 economics    203     203
## 2 210 economics    203     203
## 3 210 economics    203     203
## 4 210 economics    203     203
## 5 214 economics    207     207
## 6 212 economics    213     213
```

- Exercise 5 First Model

###Using the new data with recoded choices, we want to understand the effect of the student test score on his first choice. ###Propose a model specification. Write the likelihood function.

###likelihood function

```
## create a data set
test_data <- data_top20000 %>% select(score, choice_rev, choice_school)
test_data <- test_data %>% filter(choice_school == "schoolcode1") %>%mutate(c
hoice=as.numeric(factor(choice_rev,ordered=TRUE))) %>% drop_na()
data <- test_data
##likelihood function
like_fun = function(guess)
{
  score =  data$score
  choice = data$choice

  row = nrow(data)
  column = length(unique(choice))
  ut = mat.or.vec(row,column)

  for (j in 1:column)
  {
    ut[,j] = guess[1] + guess[2]*score[j]
  }
  prob   = exp(ut)
  prob   = sweep(prob,MARGIN=1,FUN="/",STATS=rowSums(prob))
  prob2 = prob[,1]
  prob2[prob2>0.999999] = 0.999999
  prob2[prob2<0.000001] = 0.000001
  like = sum(log(prob2))
  return(-like)
}


#we guess the parameter, set the second one as negative
guess <- runif(2)
guess[2] <- -guess[2]
like_fun(guess = guess)
```

```
## [1] 154190.3

#simulation
optim(par = guess,fn=like_fun)

## $par
## [1]  0.71658723 -0.02138275
##
## $value
## [1] 107985.6
##
## $counts
## function gradient
##       45       NA
##
## $convergence
## [1] 0
##
## $message
## NULL
```

- Question 6

###Using the new data with recoded choices, we want to understand the effect of the school_quality on his first choice. ###Propose a model specification. Write the likelihood function.

```
## create a data set
test_data <- data_top20000 %>% select(quality, choice_rev, choice_school)
test_data <- test_data %>% filter(choice_school == "schoolcode1") %>%mutate(c
hoice=as.numeric(factor(choice_rev,ordered=TRUE))) %>% drop_na()
data <- test_data
##likelihood function
like_fun = function(guess)
{
  quality =  data$quality
  choice = data$choice

  row = nrow(data)
  column = length(unique(choice))
  ut = mat.or.vec(row,column)

  for (j in 1:column)
  {
    ut[,j] = guess[1] + guess[2]*quality[j]
  }
  prob   = exp(ut)
  prob   = sweep(prob,MARGIN=1,FUN="/",STATS=rowSums(prob))
  prob2 = prob[,1]
  prob2[prob2>0.999999] = 0.999999
  prob2[prob2<0.000001] = 0.000001
```

```
  like = sum(log(prob2))
  return(-like)
}


#we guess the parameter, set the second one as negative
guess <- -runif(2)
like_fun(guess = guess)

## [1] 95910.58

#simulation
optim(par = guess,fn=like_fun)

## $par
## [1] -0.82267427 -0.05358334
##
## $value
## [1] 95903.39
##
## $counts
## function gradient
##       33       NA
##
## $convergence
## [1] 0
##
## $message
## NULL
```

- Question 7 Counterfactual simulations

##construct the data

```
test_data <- data_top20000 %>% select(quality, choice_rev, choice_school)
test_data <- test_data %>% filter(choice_school == "schoolcode1") %>%mutate(c
hoice=as.numeric(factor(choice_rev,ordered=TRUE))) %>% drop_na()
data <- test_data
```

##Which model is proper?

```
#Schools' quality may affects the types of program they can provide to
students. When excluding the other programs, school's quality should be less
correlative to the choice outcome. Thus, we should use the first model
```

##Simulate the model

```
test_data <- data_top20000 %>% select(quality, choice_rev, choice_school)
test_data <- test_data %>% filter(choice_school ==
"schoolcode1") %>%mutate(choice=as.numeric(factor(choice_rev,ordered=TRUE)))
%>% drop_na()
data <- test_data
```

```r
##likelihood function
like_fun = function(guess)
{
  score =  data$score
  choice = data$choice

  row = nrow(data)
  column = length(unique(choice))
  ut = mat.or.vec(row,column)

  for (j in 1:column)
  {
    ut[,j] = guess[1] + guess[2]*score[j]
  }
  prob   = exp(ut)
  prob   = sweep(prob,MARGIN=1,FUN="/",STATS=rowSums(prob))
  prob2 = prob[,1]
  prob2[prob2>0.999999] = 0.999999
  prob2[prob2<0.000001] = 0.000001
  like = sum(log(prob2))
  return(-like)
}



#we guess the parameter, set the second one as negative
guess <- runif(2)
guess[2] <- -guess[2]
like_fun(guess = guess)

## [1] 212639.5

#simulation
optim(par = guess,fn=like_fun)

## $par
## [1]  0.19878297 -0.02134544
##
## $value
## [1] 107985.6
##
## $counts
## function gradient
##       45       NA
##
## $convergence
## [1] 0
##
```

```
## $message
## NULL
```