506 Problem Set 3

Xiaohan Liu

Table of contents

GitHub repository: https://github.com/EmiiilyLiu/STATS_506

```
setwd("F:/Desktop/STATS 506/STATS_506")
```

Problem 1

(a)

```
. do "K:\PS3 Q1.do"

. * (a) Import and merge data
. * Refer to:
. * chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.stata.com/ma
> nuals13/dimportsasxport.pdf
. * chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.stata.com/ma
> nuals/dmerge.pdf
. import sasxport5 "K:\VIX_D.XPT"

. save "K:\temp_VIX_D.dta", replace
file K:\temp_VIX_D.dta saved

. import sasxport5 "K:\DEMO_D.xpt"

. merge 1:1 seqn using "K:\temp_VIX_D.dta", keep(match) nogenerate

    Result                      Number of obs
    -----------------------------------------
    Not matched                             0
    Matched                             6,980
    -----------------------------------------

. count
  6,980
```

(b)

```
. * (b)
. * Refer to:
. * chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.stata.com/manuals/rtab
. * view the merged data
. * describe
. * age group variable
. gen age_group = "0-9" if ridageyr < 10
(6,980 missing values generated)
```

```
. replace age_group = "10-19" if ridageyr >= 10 & ridageyr < 20
variable age_group was str1 now str5
(2,207 real changes made)

. replace age_group = "20-29" if ridageyr >= 20 & ridageyr < 30
(1,021 real changes made)

. replace age_group = "30-39" if ridageyr >= 30 & ridageyr < 40
(818 real changes made)

. replace age_group = "40-49" if ridageyr >= 40 & ridageyr < 50
(815 real changes made)

. replace age_group = "50-59" if ridageyr >= 50 & ridageyr < 60
(631 real changes made)

. replace age_group = "60-69" if ridageyr >= 60 & ridageyr < 70
(661 real changes made)

. replace age_group = "70-79" if ridageyr >= 70 & ridageyr < 80
(469 real changes made)

. replace age_group = "80-89" if ridageyr >= 80 & ridageyr < 90
(358 real changes made)

. replace age_group = "90-99" if ridageyr >= 90
(0 real changes made)

. gen wear = .
(6,980 missing values generated)

. replace wear = 1 if viq220 == 1
(2,765 real changes made)

. replace wear = 0 if inlist(viq220, 2, 9)
(3,782 real changes made)

. * Use mean of viq220==1 within each age group representing proportion
. table age_group, nototals statistic(mean wear)

---------------------
```

```
          |      Mean
----------+----------
age_group |
  10-19   |   .3208812
  20-29   |   .3258786
  30-39   |   .3586667
  40-49   |   .3699871
  50-59   |   .5500821
  60-69   |   .6222222
  70-79   |   .6689038
  80-89   |   .6688103
---------------------


.
.
```

(c)

```
. * (c)
. * Refer to *chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.st
> ata.com/manuals/rlogistic.pdf
. * https://www.stata.com/support/faqs/statistics/outcome-does-not-vary/
. * chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.stata.com/ma
> nuals13/restatic.pdf
. * http://repec.org/bocode/e/estout/esttab.html
. * http://repec.org/bocode/e/estout/estout.html
. replace viq220 = 0 if viq220 == 2
(3,780 real changes made)

. replace viq220 = . if viq220 == 9
(2 real changes made, 2 to missing)

. logistic viq220 ridageyr

Logistic regression                                     Number of obs =  6,545
                                                        LR chi2(1)    = 443.37
                                                        Prob > chi2   = 0.0000
Log likelihood = -4235.9433                             Pseudo R2     = 0.0497

------------------------------------------------------------------------------
```

```
      viq220 | Odds ratio   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
    ridageyr |    1.02498    .0012356    20.47   0.000     1.022561    1.027405
       _cons |    .283379    .0151461   -23.59   0.000     .2551952    .3146755
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.

. eststo model1

. estat ic

Akaike's information criterion and Bayesian information criterion

-----------------------------------------------------------------------------
       Model |          N   ll(null)  ll(model)      df         AIC        BIC
-------------+---------------------------------------------------------------
      model1 |      6,545  -4457.627  -4235.943       2    8475.887    8489.46
-----------------------------------------------------------------------------
Note: BIC uses N = number of observations. See [R] IC note.

. logistic viq220 ridageyr i.riagendr i.ridreth1

Logistic regression                                     Number of obs =  6,545
                                                        LR chi2(6)    = 641.49
                                                        Prob > chi2   = 0.0000
Log likelihood = -4136.8805                             Pseudo R2     = 0.0720

------------------------------------------------------------------------------
      viq220 | Odds ratio   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
    ridageyr |   1.022831    .0012912    17.88   0.000     1.020303    1.025365
  2.riagendr |    1.65217    .0875831     9.47   0.000     1.489127    1.833064
             |
    ridreth1 |
          2  |   1.169203     .192081     0.95   0.341     .8473273    1.613349
          3  |   1.952149    .1366952     9.55   0.000     1.701803    2.239322
          4  |    1.29936    .0995052     3.42   0.001     1.118264    1.509783
          5  |   1.917442    .2596352     4.81   0.000     1.470495    2.500236
             |
       _cons |   .1593479    .0124169   -23.57   0.000     .1367784    .1856414
------------------------------------------------------------------------------
```

```
Note: _cons estimates baseline odds.

. eststo model2

. estat ic

Akaike's information criterion and Bayesian information criterion

-----------------------------------------------------------------------------
       Model |          N   ll(null)  ll(model)      df        AIC        BIC
-------------+---------------------------------------------------------------
      model2 |      6,545  -4457.627   -4136.88       7   8287.761   8335.266
-----------------------------------------------------------------------------
Note: BIC uses N = number of observations. See [R] IC note.

. logistic viq220 ridageyr i.riagendr i.ridreth1 indfmpir

Logistic regression                                     Number of obs =  6,247
                                                        LR chi2(7)    = 625.30
                                                        Prob > chi2   = 0.0000
Log likelihood = -3946.9041                             Pseudo R2     = 0.0734

------------------------------------------------------------------------------
      viq220 | Odds ratio   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
    ridageyr |   1.022436    .001324    17.14   0.000     1.019845    1.025035
  2.riagendr |   1.675767   .0910025     9.51   0.000      1.50657    1.863967
             |
    ridreth1 |
          2  |   1.123021   .1889653     0.69   0.490     .8075333    1.561764
          3  |   1.651244   .1240886     6.67   0.000     1.425098    1.913277
          4  |   1.230456   .0974736     2.62   0.009     1.053503     1.43713
          5  |   1.703572   .2387583     3.80   0.000     1.294384    2.242114
             |
    indfmpir |   1.120301   .0198376     6.42   0.000     1.082087    1.159865
       _cons |   .1331659   .0116903   -22.97   0.000     .1121161    .1581678
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.

. eststo model3
```

```
. esttab model1 model2 model3, eform cells(b(star fmt(3)) se(par fmt(3))) stats(
> N r2_p aic, labels("Sample Size" "Pseudo R^2" "AIC")) mtitle("Model 1" "Model
> 2" "Model 3") label

--------------------------------------------------------------------
                              (1)             (2)             (3)   
                          Model 1         Model 2         Model 3   
                             b/se            b/se            b/se   
--------------------------------------------------------------------
Glasses/contact le~                                                 
Age at Screening A~R        1.025***        1.023***        1.022***
                          (0.001)         (0.001)         (0.001)   
Gender=1                                    1.000           1.000   
                                              (.)             (.)   
Gender=2                                    1.652***        1.676***
                                          (0.088)         (0.091)   
Race/Ethnicity - R~1                        1.000           1.000   
                                              (.)             (.)   
Race/Ethnicity - R~2                        1.169           1.123   
                                          (0.192)         (0.189)   
Race/Ethnicity - R~3                        1.952***        1.651***
                                          (0.137)         (0.124)   
Race/Ethnicity - R~4                        1.299***        1.230** 
                                          (0.100)         (0.097)   
Race/Ethnicity - R~5                        1.917***        1.704***
                                          (0.260)         (0.239)   
Family PIR                                                  1.120***
                                                          (0.020)   
--------------------------------------------------------------------
Sample Size              6545.000        6545.000        6247.000   
Pseudo R^2                  0.050           0.072           0.073   
AIC                      8475.887        8287.761        7909.808   
--------------------------------------------------------------------
Exponentiated coefficients

.
```

(d)

$$\frac{odds(female)}{odds(male)} = \frac{\frac{Pr(viq220=1|female)}{1-Pr(viq220=1|female)}}{\frac{Pr(viq220=1|male)}{1-Pr(viq220=1|male)}} = 1.676$$

The odds ratio is 1.676, with $p-value << 0.05$, meaning it is statistically significant. We can conclude that the odds of men and women being wears of glasess/contact lenses for distance vision differs. The odd of women being wears of glasses/contact lenses for distance vision is larger than that of men.

```
. *(d)
. * Refer to:
. * https://stats.oarc.ucla.edu/stata/webbooks/logistic/chapter1/logistic-regression-with-
> apter-1-introduction-to-logistic-regression-with-stata/#:~:text=Stata%20has%20two%20comm
> for,command%20with%20the%20or%20option.
. logit viq220 ridageyr i.riagendr i.ridreth1 indfmpir

Iteration 0:  Log likelihood = -4259.5533
Iteration 1:  Log likelihood = -3948.3256
Iteration 2:  Log likelihood = -3946.9043
Iteration 3:  Log likelihood = -3946.9041

Logistic regression                                     Number of obs =  6,247
                                                        LR chi2(7)    = 625.30
                                                        Prob > chi2   = 0.0000
Log likelihood = -3946.9041                             Pseudo R2     = 0.0734

------------------------------------------------------------------------------
      viq220 | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
    ridageyr |   .0221883   .0012949    17.14   0.000     .0196504    .0247263
  2.riagendr |   .5162712    .054305     9.51   0.000     .4098355     .622707
             |
    ridreth1 |
          2  |   .1160225   .1682651     0.69   0.490    -.213771     .4458161
          3  |   .5015289   .0751486     6.67   0.000     .3542404    .6488174
          4  |   .2073846   .0792175     2.62   0.009     .0521211     .362648
          5  |   .5327271   .1401516     3.80   0.000     .2580349    .8074192
             |
    indfmpir |   .1135978   .0177073     6.42   0.000      .078892    .1483035
       _cons |   -2.01616   .0877879   -22.97   0.000    -2.188221   -1.844099
------------------------------------------------------------------------------

.
.
.
.
```

```
.
end of do-file

.
```

Since $p - value << 0.05$, meaning the coefficient is statistically significant at 0.05 level, that is the proportion of wearers of glasses/contact lenses for distance vision differs between men and women.

Problem 2

```r
library(DBI)

## Import the SQLite database of "sakila" data
sakila <- dbConnect(RSQLite::SQLite(), "sakila_master.db")

#' @param x a string as input SQL query
gg <- function(x){
  dbGetQuery(sakila, x)
}

## Get the list of tables in "sakila" database
dbListTables(sakila)
```

```
 [1] "actor"                  "address"                "category"
 [4] "city"                   "country"                "customer"
 [7] "customer_list"          "film"                   "film_actor"
[10] "film_category"          "film_list"              "film_text"
[13] "inventory"              "language"               "payment"
[16] "rental"                 "sales_by_film_category" "sales_by_store"
[19] "staff"                  "staff_list"             "store"
```

```r
## Get lists of columns of the tables
for (i in dbListTables(sakila)){
  cat("Table:", i, "\n")
  print(dbListFields(sakila, i))
  cat("\n")
}
```

```
Table: actor
```

```
[1] "actor_id"    "first_name"  "last_name"   "last_update"

Table: address
[1] "address_id"  "address"     "address2"    "district"    "city_id"
[6] "postal_code" "phone"       "last_update"

Table: category
[1] "category_id" "name"        "last_update"

Table: city
[1] "city_id"     "city"        "country_id"  "last_update"

Table: country
[1] "country_id"  "country"     "last_update"

Table: customer
[1] "customer_id" "store_id"    "first_name"  "last_name"   "email"
[6] "address_id"  "active"      "create_date" "last_update"

Table: customer_list
[1] "ID"       "name"     "address"  "zip_code" "phone"    "city"     "country"
[8] "notes"    "SID"

Table: film
 [1] "film_id"              "title"                "description"
 [4] "release_year"         "language_id"          "original_language_id"
 [7] "rental_duration"      "rental_rate"          "length"
[10] "replacement_cost"     "rating"               "special_features"
[13] "last_update"

Table: film_actor
[1] "actor_id"    "film_id"     "last_update"

Table: film_category
[1] "film_id"     "category_id" "last_update"

Table: film_list
[1] "FID"         "title"       "description" "category"    "price"
[6] "length"      "rating"      "actors"

Table: film_text
[1] "film_id"     "title"       "description"
```

```
Table: inventory
[1] "inventory_id" "film_id"      "store_id"     "last_update"

Table: language
[1] "language_id" "name"        "last_update"

Table: payment
[1] "payment_id"   "customer_id"  "staff_id"     "rental_id"    "amount"
[6] "payment_date" "last_update"

Table: rental
[1] "rental_id"    "rental_date"  "inventory_id" "customer_id"  "return_date"
[6] "staff_id"     "last_update"

Table: sales_by_film_category
[1] "category"    "total_sales"

Table: sales_by_store
[1] "store_id"    "store"       "manager"     "total_sales"

Table: staff
 [1] "staff_id"    "first_name"  "last_name"   "address_id"  "picture"
 [6] "email"       "store_id"    "active"      "username"    "password"
[11] "last_update"

Table: staff_list
[1] "ID"       "name"     "address"  "zip_code" "phone"    "city"     "country"
[8] "SID"

Table: store
[1] "store_id"         "manager_staff_id" "address_id"       "last_update"
```

(a)

```
gg("SELECT l.name as language, COUNT(f.film_id) AS frenquncy
   FROM language l
   LEFT JOIN film f
   ON f.language_id = l.language_id
   GROUP BY l.language_id
   ORDER BY COUNT(f.film_id) DESC")
```

```
  language frenquncy
1  English      1000
2  Italian         0
3 Japanese         0
4 Mandarin         0
5   French         0
6   German         0
```

All the films for which we have relevant language information are in English in this database. Therefore, we cannot determine which language, aside from English, is most common for films.

(b)

```r
## R
## Extract appropriate tables
category <- gg("SELECT * FROM category")
film <- gg("SELECT * FROM film_category")

genre_count <- table(film$category_id)
## Get the most common genre id
most_common_genreID <- names(which.max(genre_count))

## Get the corresponding genre name
most_common_genre <- category$name[category$category_id
                                   == most_common_genreID]

most_common_genre
```

```
[1] "Sports"
```

```r
## SQL answer
gg("SELECT c.name AS genre, COUNT(fc.film_id) AS frequency
   FROM film_category fc
   LEFT JOIN category c ON fc.category_id = c.category_id
   GROUP BY c.name
   ORDER BY frequency DESC
   LIMIT 1")
```

```
   genre frequency
1 Sports        74
```

Both two methods generate the same result: *Sports* is the most common movie genre.

(c)

```r
## R
## Get the appropriate table
customer <- gg("SELECT * FROM customer_list")

country_count <- table(customer$country)

## Countries have exact 9 customers
country_with9customers <- names(country_count[country_count == 9])

country_with9customers
```

```
[1] "United Kingdom"
```

```r
## SQL answer
gg("SELECT country, COUNT(country) AS frequency
   FROM customer_list
   GROUP BY country
   HAVING frequency=9")
```

```
         country frequency
1 United Kingdom         9
```

Both two methods generate the same result: *United Kingdom* is the country with exact 9 customers.

Problem 3

```r
data <- read.csv("us-500.csv")
```

(a)

```r
## This data set has NO missing value
## so I use the number of the row as the total number of email
sum(is.na(data))
```

```
[1] 0
```

```r
## email ending with ".net"
sum(grepl("\\.net$", data$email))/nrow(data)
```

```
[1] 0.14
```

(b)

```r
## Since an email address must have a "@" and a ".",
## so I delete a "@" and  "." firstly
#' @param x a string
#' @return string deleted an "at" and a "dot"
delete_at_dot <- function(x){
  delete_at <- sub("@", "", x, fixed = TRUE)
  modified_x <- sub(".", "", delete_at, fixed = TRUE)

  return(modified_x)
}

modified_emails <- sapply(data$email, delete_at_dot)

## Email with non alphanumeric character
## except for an "@" and a ".'
sum(grepl("[^a-zA-Z0-9]", modified_emails))/nrow(data)
```

```
[1] 0.506
```

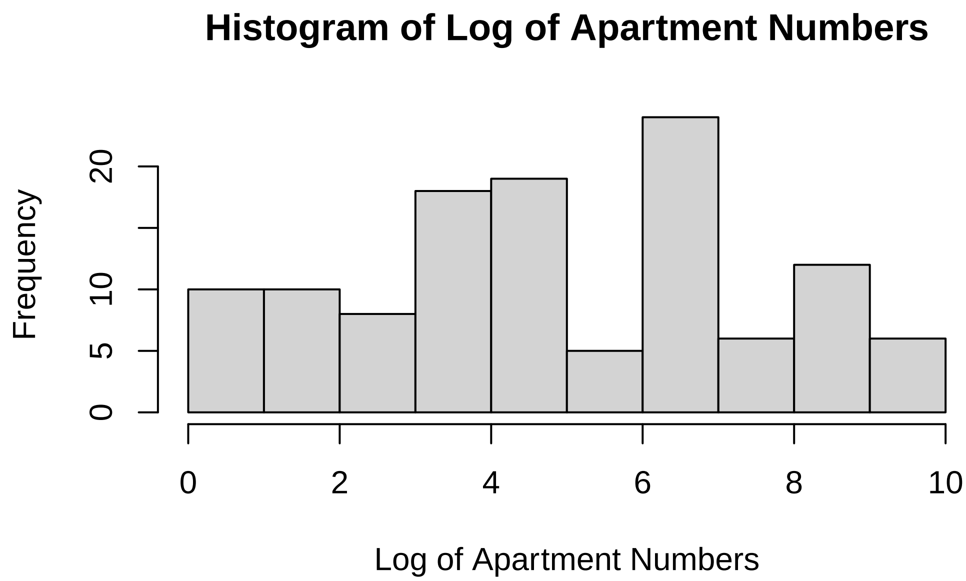
(c)

```r
## Get area code
AreaCode_phone1 <- substr(data$phone1,1,3)
AreaCode_phone2 <- substr(data$phone2,1,3)
AreaCode_all <- c(AreaCode_phone1, AreaCode_phone2)

## Count the occurrence of each area code
AreaCode_count <- table(AreaCode_all)

## Get the area code with the highest frequency
names(AreaCode_count[which.max(AreaCode_count)])
```

```
[1] "973"
```

(d)

```r
library(stringr)

## Get the apartment number
apartment_num <- str_extract(data$address, "(?<=\\D)(\\d+)$")
numeric_apt_num <- as.numeric(apartment_num)

log_apt_num <- log(numeric_apt_num)

## histogram of the log of the apartment numbers
hist(log_apt_num, main="Histogram of Log of Apartment Numbers",
     xlab="Log of Apartment Numbers")
```



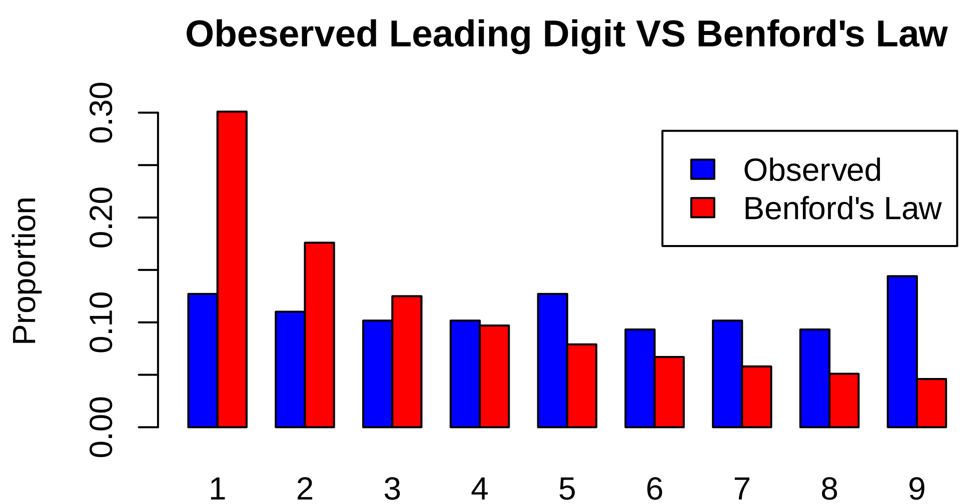
(e)

```
## the leading digit of each apartment number
leading_digit <- substr(apartment_num,1,1)

## distribution of the leading digit
dist_leading_digit <- table(leading_digit)/sum(!is.na(leading_digit))

## Refer to https://en.wikipedia.org/wiki/Benford's_law
benford_prob <- c(0.301, 0.176, 0.125, 0.097, 0.079, 0.067,
                  0.058, 0.051, 0.046)

comparison <- data.frame(
  Digit = 1:9,
  Observed = as.numeric(dist_leading_digit),
  Benford = benford_prob
)

comparison
```

```
  Digit   Observed Benford
```

```
1       1 0.12711864    0.301
2       2 0.11016949    0.176
3       3 0.10169492    0.125
4       4 0.10169492    0.097
5       5 0.12711864    0.079
6       6 0.09322034    0.067
7       7 0.10169492    0.058
8       8 0.09322034    0.051
9       9 0.14406780    0.046
```

```r
## visualize the comparison
barplot(rbind(comparison$Observed, comparison$Benford),
        beside = TRUE, col = c("blue", "red"), names.arg = 1:9,
        legend.text = c("Observed", "Benford's Law"),
        ylab = "Proportion",
        main = "Obeserved Leading Digit VS Benford's Law")
```



The distribution of observed apartment numbers shows significant difference from that of Benford's Law. We can conclude that the apartment numbers do **not** appear to follow Benford's law and they would **not** pass as real data.

(f)

```r
## get the street number of each address
all <- str_extract_all(data$address, "\\d+")
street_num <- sapply(all, function(x) x[1])

## last digit of each street number
last_digit <- str_sub(street_num, start= -1)

## distribution of the last digit
dist_last_digit <- table(last_digit)/sum(!is.na(last_digit))

## assume they are uniform
expected_dist <- rep(1/10, 10)

comparison_last_digit <- data.frame(
  Digit = 0:9,
  Observed = as.numeric(dist_last_digit),
  Expected = expected_dist
)

comparison_last_digit
```

```
   Digit Observed Expected
1      0    0.078      0.1
2      1    0.082      0.1
3      2    0.108      0.1
4      3    0.102      0.1
5      4    0.106      0.1
6      5    0.104      0.1
7      6    0.086      0.1
8      7    0.112      0.1
9      8    0.090      0.1
10     9    0.132      0.1
```

```r
## visualize the comparison
barplot(rbind(comparison_last_digit$Observed,
              comparison_last_digit$Expected),
        beside = TRUE, col = c("blue", "red"), names.arg = 0:9,
        ylab = "Proportion",
        main = "Observed Last Digit VS Expected",
```

```
         legend.text  =  c("Observed", "Expected"))
```