STATS 506 HW 3-kjaewon

Github Repository: https://github.com/kjaewon-umich/STATS506

Problem 1

```r
library(kableExtra)
library(tidyverse)
library(haven)
```

a.

```r
vix_d <- read_xpt("VIX_D.XPT", col_select = NULL, skip = 0, n_max = Inf,
                  .name_repair = "unique")
demo_d <- read_xpt("DEMO_D.XPT", col_select = NULL, skip = 0, n_max = Inf,
                  .name_repair = "unique")
vision <- inner_join(vix_d, demo_d, by = "SEQN")

nrow(vision)
```

```
[1] 6980
```

We can confirm that the number of rows that match is 6,980.

b.

First, we will simplify the names of relevant columns and manipulate the data for computations.

```
vision <- vision %>%
  rename(
    glasses = VIQ220,
    age = RIDAGEYR,
    race = RIDRETH1,
    gender = RIAGENDR,
    PIR = INDFMPIR
  ) %>%
  drop_na(glasses) %>%
  filter(glasses != 9) %>%
  mutate(
    glasses = glasses - 1,
    race = factor(race, levels = c(1, 2, 3, 4, 5),
                  labels = c("Mexican American", "Other Hispanic",
                             "Non-Hispanic White", "Non-Hispanic Black",
                             "Other Race - Including Multi-Racial")),
    gender = factor(gender, levels = c(1, 2), labels = c("Male", "Female"))
  )
```

Next, we will get the proportion of respondents that wear glasses/contact lenses for distance vision by categorizing them with 10-year age bracket.

```
vision.result <- vision %>%
  mutate(age_group = floor(age / 10)) %>%
  mutate(age_group = factor(age_group, levels = c(1, 2, 3, 4, 5, 6, 7, 8),
                            labels = c("10-19", "20-29", "30-39", "40-49",
                                       "50-59", "60-69", "70-79", "80-89"))) %>%
  group_by(age_group) %>%
  summarize(glasses_proportion = round(mean(glasses == 1, na.rm = TRUE) * 100, 2))

colnames(vision.result) <- c("Age Group", "Proportion")

kable(vision.result, caption = "Proportion of Glasses/Contacts")
```

Table 1: Proportion of Glasses/Contacts

Age Group	Proportion
10-19	67.91
20-29	67.34
30-39	64.13
40-49	63.00
50-59	44.99
60-69	37.78
70-79	33.11
80-89	33.12

c.

Model 1

```r
glm1 <- glm(glasses ~ age, family = binomial(link = logit), data = vision)
summary(glm1)
```

```

Call:
glm(formula = glasses ~ age, family = binomial(link = logit),
    data = vision)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.260970   0.053448   23.59   <2e-16 ***
age         -0.024673   0.001206  -20.47   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8915.3  on 6544  degrees of freedom
Residual deviance: 8471.9  on 6543  degrees of freedom
AIC: 8475.9

Number of Fisher Scoring iterations: 4
```

```r
nobs(glm1)
```

```
[1] 6545
```

```r
pseudo_r2.1 <- 1 - (summary(glm1)$deviance / summary(glm1)$null.deviance)
pseudo_r2.1
```

```
[1] 0.04973123
```

Model 2

```r
glm2 <- glm(glasses ~ age + race + gender, family = binomial(link = logit),
            data = vision)
summary(glm2)
```

```
Call:
glm(formula = glasses ~ age + race + gender, family = binomial(link = logit),
    data = vision)

Coefficients:
                                         Estimate Std. Error z value Pr(>|z|)
(Intercept)                              1.836666   0.077923  23.570  < 2e-16
age                                     -0.022574   0.001262 -17.882  < 2e-16
raceOther Hispanic                      -0.156322   0.164284  -0.952 0.341332
raceNon-Hispanic White                  -0.668931   0.070023  -9.553  < 2e-16
raceNon-Hispanic Black                  -0.261872   0.076580  -3.420 0.000627
raceOther Race - Including Multi-Racial -0.650992   0.135407  -4.808 1.53e-06
genderFemale                            -0.502090   0.053011  -9.471  < 2e-16

(Intercept)                             ***
age                                     ***
raceOther Hispanic
raceNon-Hispanic White                  ***
raceNon-Hispanic Black                  ***
raceOther Race - Including Multi-Racial ***
genderFemale                            ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8915.3  on 6544  degrees of freedom
Residual deviance: 8273.8  on 6538  degrees of freedom
AIC: 8287.8

Number of Fisher Scoring iterations: 4
```

```r
nobs(glm2)
```

```
[1] 6545
```

```r
pseudo_r2.2 <- 1 - (summary(glm2)$deviance / summary(glm2)$null.deviance)
pseudo_r2.2
```

```
[1] 0.07195445
```

Model 3

```r
glm3 <- glm(glasses ~ age + race + gender + PIR,
            family = binomial(link = logit), data = vision)
summary(glm3)
```

```
Call:
glm(formula = glasses ~ age + race + gender + PIR, family = binomial(link = logit),
    data = vision)

Coefficients:
                                         Estimate Std. Error z value Pr(>|z|)
(Intercept)                              2.016160   0.087788  22.966  < 2e-16
age                                     -0.022188   0.001295 -17.135  < 2e-16
raceOther Hispanic                      -0.116023   0.168265  -0.690 0.490495
raceNon-Hispanic White                  -0.501529   0.075149  -6.674 2.49e-11
raceNon-Hispanic Black                  -0.207385   0.079217  -2.618 0.008847
raceOther Race - Including Multi-Racial -0.532727   0.140152  -3.801 0.000144
genderFemale                            -0.516271   0.054305  -9.507  < 2e-16
```

```
PIR                                     -0.113598   0.017707  -6.415 1.41e-10

(Intercept)                             ***
age                                     ***
raceOther Hispanic
raceNon-Hispanic White                  ***
raceNon-Hispanic Black                  **
raceOther Race - Including Multi-Racial ***
genderFemale                            ***
PIR                                     ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8519.1  on 6246  degrees of freedom
Residual deviance: 7893.8  on 6239  degrees of freedom
  (298 observations deleted due to missingness)
AIC: 7909.8

Number of Fisher Scoring iterations: 4
```

```r
nobs(glm3)
```

```
[1] 6247
```

```r
pseudo_r2.3 <- 1 - (summary(glm3)$deviance / summary(glm3)$null.deviance)
pseudo_r2.3
```

```
[1] 0.07339952
```

Table

Now we summarize the results in a table. This table will include the estimated odd ratios for the coefficients, the sample size, the pseudo-R^2, and AIC value for each model.

```r
library(broom)

# Extract coefficients for each model
coef_glm1 <- tidy(glm1)
```

```r
coef_glm2 <- tidy(glm2)
coef_glm3 <- tidy(glm3)

# Combine coefficients with pseudo-R², AIC, and number of observations
c.results <- data.frame(
  Variable = c("age", "Other Hispanic", "Non-Hispanic Black", "Non-Hispanic White",
               "Multi-racial", "gender", "PIR", "Constant", "N", "p_r2", "AIC"),

  Model_1 = c(round(exp(coef_glm1$estimate[2]), 4), NA, NA, NA, NA, NA, NA,
             round(exp(coef_glm1$estimate[1]), 4), nobs(glm1),
             round(pseudo_r2.1, 4), round(AIC(glm1), 4)),

  Model_2 = c(round(exp(coef_glm2$estimate[2]), 4),
             round(exp(coef_glm2$estimate[3]), 4),
             round(exp(coef_glm2$estimate[4]), 4),
             round(exp(coef_glm2$estimate[5]), 4),
             round(exp(coef_glm2$estimate[6]), 4), NA, NA,
             round(exp(coef_glm2$estimate[1]), 4), nobs(glm2),
             round(pseudo_r2.2, 4), round(AIC(glm2), 4)),

  Model_3 = c(round(exp(coef_glm3$estimate[2]), 4),
             round(exp(coef_glm3$estimate[3]), 4),
             round(exp(coef_glm3$estimate[4]), 4),
             round(exp(coef_glm3$estimate[5]), 4),
             round(exp(coef_glm3$estimate[6]), 4),
             round(exp(coef_glm3$estimate[7]), 4),
             round(exp(coef_glm3$estimate[8]), 4),
             round(exp(coef_glm3$estimate[1]), 4),
             nobs(glm3), round(pseudo_r2.3, 4), round(AIC(glm3), 4))
)

# Create the formatted table
kable(c.results, col.names = c("Variable", "Model 1", "Model 2", "Model 3"),
      caption = "Logistic Regression Results")
```

Table 2: Logistic Regression Results

Variable	Model 1	Model 2	Model 3
age	0.9756	0.9777	0.9781
Other Hispanic	NA	0.8553	0.8905
Non-Hispanic Black	NA	0.5123	0.6056

Variable	Model 1	Model 2	Model 3
Non-Hispanic White	NA	0.7696	0.8127
Multi-racial	NA	0.5215	0.5870
gender	NA	NA	0.5967
PIR	NA	NA	0.8926
Constant	3.5288	6.2756	7.5094
N	6545.0000	6545.0000	6247.0000
p_r2	0.0497	0.0720	0.0734
AIC	8475.8866	8287.7609	7909.8082

d.

```r
summary(glm3)
```

```

Call:
glm(formula = glasses ~ age + race + gender + PIR, family = binomial(link = logit),
    data = vision)

Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                             2.016160   0.087788  22.966  < 2e-16
age                                    -0.022188   0.001295 -17.135  < 2e-16
raceOther Hispanic                     -0.116023   0.168265  -0.690 0.490495
raceNon-Hispanic White                 -0.501529   0.075149  -6.674 2.49e-11
raceNon-Hispanic Black                 -0.207385   0.079217  -2.618 0.008847
raceOther Race - Including Multi-Racial -0.532727   0.140152  -3.801 0.000144
genderFemale                           -0.516271   0.054305  -9.507  < 2e-16
PIR                                    -0.113598   0.017707  -6.415 1.41e-10

(Intercept)                             ***
age                                     ***
raceOther Hispanic
raceNon-Hispanic White                  ***
raceNon-Hispanic Black                  **
raceOther Race - Including Multi-Racial ***
genderFemale                            ***
PIR                                     ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8519.1  on 6246  degrees of freedom
Residual deviance: 7893.8  on 6239  degrees of freedom
  (298 observations deleted due to missingness)
AIC: 7909.8

Number of Fisher Scoring iterations: 4
```

We can observe that the estimated odds ratio for females is approximately 0.5967 amd statistically significant from the previous part. This implies that the odds of females wearing glasses/contacts for distance vision is statistically significantly lower than the odds for males.

```r
coef_gender <- summary(glm3)$coefficients[7, 1]
se_gender <- summary(glm3)$coefficients[7, 2]

# Wald test statistic (z-value)
z_gender <- coef_gender / se_gender
p_value_gender <- 2 * pnorm(-abs(z_gender))  # two-tailed test

# Output p-value to determine statistical significance
p_value_gender
```

```
[1] 1.96446e-21
```

We also see evidence that females have a statistically significantly lower probability of wearing glasses/contact lenses for distance vision than males.

Problem 2.

```r
library(DBI)
sakila <- dbConnect(RSQLite::SQLite(), "sakila_master.db")
```

a.

```r
dbGetQuery(sakila, "
           SELECT MIN(release_year) AS oldest_year,
           count(release_year) AS count
           FROM film AS f
           ")
```

```
  oldest_year count
1        2006  1000
```

We can observe that the oldest movies were released in 2006 and their quantity is 1000.

b.

```r
dbGetQuery(sakila, "
SELECT c.name, count(c.category_id) AS count
  FROM category as c
RIGHT JOIN film_category AS fc ON fc.category_id = c.category_id
GROUP BY c.category_id
ORDER by count
LIMIT 1
")
```

```
   name count
1 Music    51
```

We can observe that the least common genre is Music and there are 51 movies in this dataset.

c.

```r
customer <- dbGetQuery(sakila, "SELECT * FROM customer")
address <- dbGetQuery(sakila, "SELECT * FROM address")
city <- dbGetQuery(sakila, "SELECT * FROM city")
country <- dbGetQuery(sakila, "SELECT * FROM country")
```

```r
cities <- address$city_id[match(customer$address_id, address$address_id)]
countries <- city$country_id[match(cities, city$city_id)]
countries.table <- table(country$country[match(countries, country$country_id)])
countries.table[countries.table == 13]
```

```

Argentina   Nigeria 
       13        13 
```

We can observe that Argentina and Nigeria have exactly 13 customers.

Problem 3.

```r
US.data <- read.csv("us-500.csv", header = TRUE)
head(US.data)
```

```
  first_name last_name            company_name               address        city
1      James      Butt       Benton, John B Jr    6649 N Blue Gum St New Orleans
2  Josephine   Darakjy  Chanay, Jeffrey A Esq  4 B Blue Ridge Blvd    Brighton
3        Art    Venere    Chemel, James L Cpa 8 W Cerritos Ave #54  Bridgeport
4      Lenna  Paprocki Feltz Printing Service           639 Main St   Anchorage
5    Donette    Foller    Printing Dimensions           34 Center St    Hamilton
6     Simona   Morasca    Chapman, Ross E Esq          3 Mcauley Dr     Ashland
      county state   zip       phone1       phone2
1    Orleans    LA 70116 504-621-8927 504-845-1427
2 Livingston    MI 48116 810-292-9388 810-374-9840
3 Gloucester    NJ  8014 856-636-8749 856-264-4130
4  Anchorage    AK 99501 907-385-4412 907-921-2010
5     Butler    OH 45011 513-570-1893 513-549-4561
6    Ashland    OH 44805 419-503-2484 419-800-6759
                          email                                web
1               jbutt@gmail.com        http://www.bentonjohnbjr.com
2 josephine_darakjy@darakjy.org    http://www.chanayjeffreyaesq.com
3                art@venere.org      http://www.chemeljameslcpa.com
4         lpaprocki@hotmail.com http://www.feltzprintingservice.com
5        donette.foller@cox.net   http://www.printingdimensions.com
6             simona@morasca.com      http://www.chapmanrosseesq.com
```

a.

```r
length(US.data$email[grep("com$", US.data$email)]) / nrow(US.data)
```

```
[1] 0.732
```

b.

We will first extract the usernames and detect non-alphanumeric characters other than "@".

```r
emails <- strsplit(US.data$email, "@")
id <- sapply(emails, "[[", 1)
id.non_alphanumeric <- grepl("[^a-zA-Z0-9]", id)
```

Then, we will repeat the same process for the domains by stripping off the TLD.

```r
domains <- sapply(emails, "[[", 2)
domains <- gsub("\\.[a-z]{3}", "", domains)
domains.non_alphanumeric <- grepl("[^a-zA-Z0-9]", domains)
```

We can get the proportion by getting the union of non-alphanumeric IDs and non-alphanumeric domains.

```r
mean(id.non_alphanumeric | domains.non_alphanumeric)
```

```
[1] 0.506
```

c.

First, we will check whether there is any missing row for each column.

```r
nrow(US.data)
```

```
[1] 500
```

```r
table(sapply(US.data$phone1, nchar))
```

```
 12
500
```

```r
table(sapply(US.data$phone2, nchar))
```

```
 12
500
```

Since they both have no missing rows, we proceed. Based on the format of US phone numbers, we can assume that the first three characters will represent the area code. Since we want the top 5 the most common numbers, we will sort the table by decreasing order and display the 5 most common occurrences.

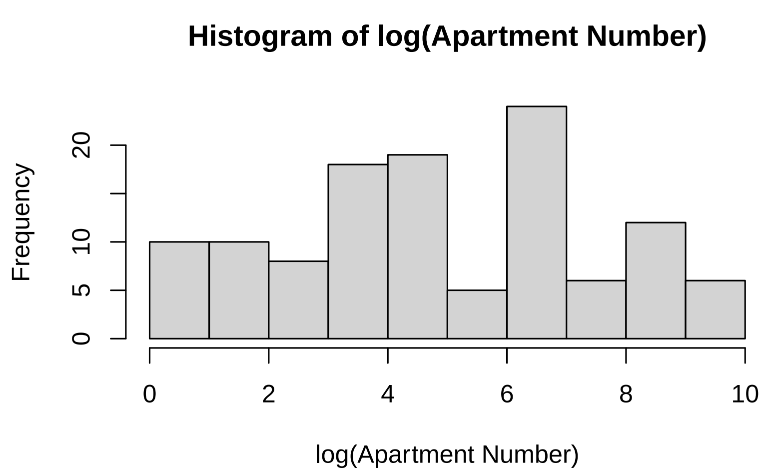
```
phone1area <- substr(US.data$phone1, 1, 3)
phone2area <- substr(US.data$phone2, 1, 3)
sort(table(c(phone1area, phone2area)), decreasing = TRUE)[1:5]
```

```
973 212 215 410 201
 36  28  28  28  24
```

d.

We will assume that any address that ends in a number represents apartments. First, we identify such cases, then split the string on spaces and store the last entry.

```
apts <- US.data$address[grep("[0-9]+$", US.data$address)]
num <- sapply(strsplit(apts, " "), function(x) x[length(x)])
num <- as.numeric(gsub("#", "", num))
hist(log(num), xlab = "log(Apartment Number)",
     main = "Histogram of log(Apartment Number)")
```



e.

```
table(substr(num, 1, 1))
```

```
 1  2  3  4  5  6  7  8  9 
15 13 12 12 15 11 12 11 17 
```

We can observe that this is approximately a uniform distribution, rather than the decreasing distribution as expected by Benford's law. This data does not seem to be real.