STATS 506 Problem Set #3

Haiming Li

Vision

a. Read data and merge data

```
library(haven)
vix <- read_xpt('./VIX_D.XPT')
demo <- read_xpt('./DEMO_D.XPT')
df <- merge(vix, demo, by='SEQN')
cat('Sample size:', nrow(df))
```

```
Sample size: 6980
```

b. The max age is 85, so there will only be 9 age brackets.

```
library(dplyr)
df <- subset(df, (VIQ220 == 1) | (VIQ220 == 2))
df$VIQ220 <- ifelse(is.na(df$VIQ220), 0, df$VIQ220)
age_groups <- c('10-19' ,'20-29', '30-39', '40-49',
                '50-59', '60-69', '70-79', '80-89')
df$age_cat <- age_groups[floor(df$RIDAGEYR / 10)]
res <- df %>% group_by(age_cat) %>%
          summarise(proportion = round(100 * mean(VIQ220 == 1, na.rm=TRUE), 2))
knitr::kable(res, 'simple', col.names = c('Age Group', 'Proportion'))
```

Age Group	Proportion
10-19	32.09
20-29	32.66
30-39	35.87
40-49	37.00
50-59	55.01

Age Group	Proportion
60-69	62.22
70-79	66.89
80-89	66.88

c. Here are fitted models and their summary.

```
#' Create Summary Table for Logistic Regression
#'
#' @param model a fitted logistic regression model
#' @return a table with required stats
summary_table <- function(model) {
  odds_ratios <- as.data.frame(t(exp(coef(model))))
  res <- data.frame(
  'Sample Size' = nobs(model),
  'Pseudo R2' = 1 - model$deviance / model$null.deviance,
  'AIC' = AIC(model)
  )
  res <- cbind(odds_ratios, res)
  return(t(res))
}

# data cleaning
df_mod <- subset(df, select=c(VIQ220, RIAGENDR, RIDAGEYR, RIDRETH1, INDFMPIR))
df_mod <- na.omit(df_mod)
df_mod$VIQ220 <- as.factor(df_mod$VIQ220)
df_mod$RIAGENDR <- as.factor(df_mod$RIAGENDR)
df_mod$RIDRETH1 <- as.factor(df_mod$RIDRETH1)

# model fitting
mod1 <- glm(VIQ220 ~ RIDAGEYR, data = df_mod,
            family = binomial(link = 'logit'))
knitr::kable(summary_table(mod1))
```

(Intercept)	3.4170427
RIDAGEYR	0.9760682
Sample.Size	6247.0000000
Pseudo.R2	0.0473330
AIC	8119.8714676

```
mod2 <- glm(VIQ220 ~ RIDAGEYR + RIDRETH1 + RIAGENDR, data = df_mod,
            family = binomial(link = 'logit'))
knitr::kable(summary_table(mod2))
```

(Intercept)	6.0521094
RIDAGEYR	0.9779121
RIDRETH12	0.8560069
RIDRETH13	0.5277604
RIDRETH14	0.7737545
RIDRETH15	0.5310998
RIAGENDR2	0.6053170
Sample.Size	6247.0000000
Pseudo.R2	0.0685555
AIC	7949.0753329

```
mod3 <- glm(VIQ220 ~ RIDAGEYR + RIDRETH1 + RIAGENDR + INDFMPIR, data = df_mod,
            family = binomial(link = 'logit'))
knitr::kable(summary_table(mod3))
```

(Intercept)	7.5094311
RIDAGEYR	0.9780560
RIDRETH12	0.8904552
RIDRETH13	0.6056040
RIDRETH14	0.8127071
RIDRETH15	0.5870020
RIAGENDR2	0.5967415
INDFMPIR	0.8926169
Sample.Size	6247.0000000
Pseudo.R2	0.0733995
AIC	7909.8082208

d. From previous part, we have the odds ratio for women is 0.5967415. This can be interpreted as the value of female odds divided by male odds. From the summary of model 3, the coefficient is significant, thus implying that female odds differs from male odds for being a glass wearer is statistically significant. Since the positive class of glm is the last level of the factor, 2 in this case, the positive class for the model is actually 'not a glass wearer'. Thus, we need to invert our interpretation. Having an odds ratio less than 1 from the model actually should imply that the odds of females wearing glasses/contacts for distance vision is higher than male.

```r
summary(mod3)
```

```
Call:
glm(formula = VIQ220 ~ RIDAGEYR + RIDRETH1 + RIAGENDR + INDFMPIR,
    family = binomial(link = "logit"), data = df_mod)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.016160   0.087788  22.966  < 2e-16 ***
RIDAGEYR    -0.022188   0.001295 -17.135  < 2e-16 ***
RIDRETH12   -0.116023   0.168265  -0.690 0.490495
RIDRETH13   -0.501529   0.075149  -6.674 2.49e-11 ***
RIDRETH14   -0.207385   0.079217  -2.618 0.008847 **
RIDRETH15   -0.532727   0.140152  -3.801 0.000144 ***
RIAGENDR2   -0.516271   0.054305  -9.507  < 2e-16 ***
INDFMPIR    -0.113598   0.017707  -6.415 1.41e-10 ***
---
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
AIC: 7909.8

Number of Fisher Scoring iterations: 4
```

As shown by the two sample proportion test, the p-value is extremely small. Thus, we can reject the null hypothesis and conclude that the proportion of wearers of glasses/contact lenses for distance vision differs between men and women. We can even say that male proportion is less than female proportion. (according to the fact that the confidence interval is below 0)

```r
tab <- table(df_mod$RIAGENDR, df_mod$VIQ220)
prop.test(tab[,1], rowSums(tab))
```

```
    2-sample test for equality of proportions with continuity correction

data:  tab[, 1] out of rowSums(tab)
X-squared = 69.683, df = 1, p-value < 2.2e-16
```

```
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.12945505 -0.08007986
sample estimates:
   prop 1    prop 2
0.3714379 0.4762054
```

Sakila

a. It appears that the earliest release year is 2006, and there're 1000 movies released that year.

```
library(DBI)
sakila <- dbConnect(RSQLite::SQLite(), './sakila_master.db')
dbGetQuery(sakila,'
  SELECT release_year, COUNT(*) AS count
  FROM film
  WHERE release_year = (SELECT MIN(release_year) FROM film)
  GROUP BY release_year
')
```

```
  release_year count
1         2006  1000
```

b. Here's the R approach, note that the min value is unique.

```
film_cat <- dbGetQuery(sakila, 'SELECT * FROM film_category')
category <-  dbGetQuery(sakila, 'SELECT * FROM category')
cat_count <- table(film_cat$category_id)
min_cat <- which.min(cat_count)
cat(category$name[category$category_id == min_cat], cat_count[min_cat])
```

```
Music 51
```

Here's the SQL approach

```
dbGetQuery(sakila, '
  SELECT c.name, COUNT(*) AS count
  FROM film_category AS fc
  JOIN category AS c
  ON fc.category_id = c.category_id
  GROUP BY c.category_id
  ORDER BY count ASC
  LIMIT 1
')
```

```
   name count
1 Music    51
```

c. Here's the R approach

```
customer <- dbGetQuery(sakila, 'SELECT * FROM customer')
address <- dbGetQuery(sakila, 'SELECT * FROM address')
city <- dbGetQuery(sakila, 'SELECT * FROM city')
country <- dbGetQuery(sakila, 'SELECT * FROM country')
merged_df <- merge(customer, address, by='address_id')
merged_df <- merge(merged_df, city, by='city_id')
merged_df <- merge(merged_df, country, by='country_id')
res <- table(merged_df$country)
res[res == 13]
```

```
Argentina   Nigeria
       13        13
```

Here's the SQL approach.

```
dbGetQuery(sakila, '
  SELECT country.country, COUNT(*) AS count
  FROM customer, address, city, country
  WHERE customer.address_id = address.address_id AND
    address.city_id = city.city_id AND
    city.country_id = country.country_id
  GROUP BY country.country_id
  HAVING count = 13
')
```

```
    country count
1 Argentina    13
2   Nigeria    13
```

US Records

a. Here's the proportion of TLD with ".com'

```
df <- read.csv('./us-500.csv')
cat('Proportion of .com:', mean(grepl('\\.com$', df$email)))
```

```
Proportion of .com: 0.732
```

b. Here's the proportion of email with at least one non alphanumeric character in them. Since it's possible to have "." in the username, we need to seperate the username and domain, then check each part separately.

```
emails <- strsplit(df$email, '@')
usernames <- lapply(emails, '[[', 1)
domains <- lapply(emails, '[[', 2)
domains <- gsub('\\.[a-z]{3}', '', domains)
mean(grepl('[^a-zA-Z0-9]+', usernames) | grepl('[^a-zA-Z0-9]+', domains))
```

```
[1] 0.506
```

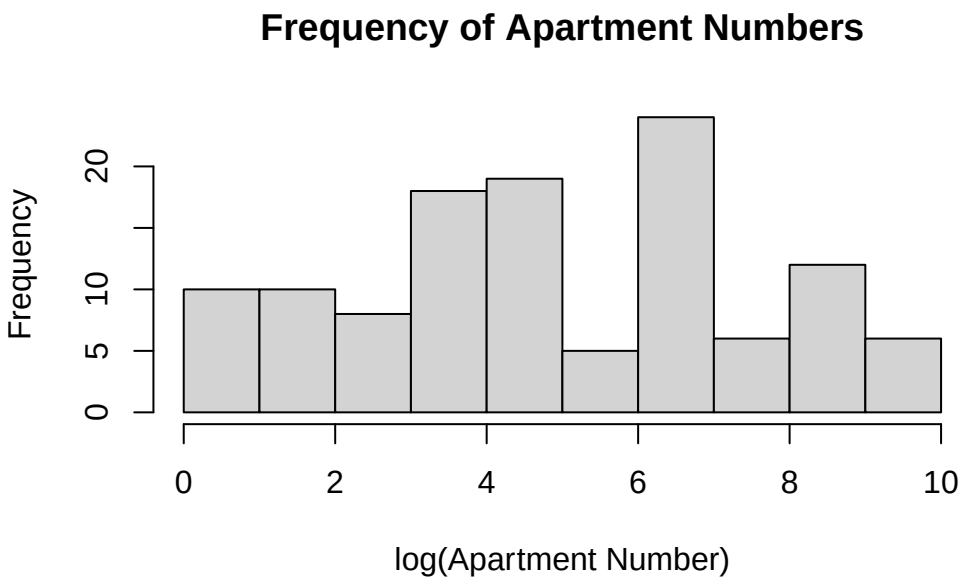
c. Here's the top 5 area code. Notice that there is no ties in top 5, so I can directly use the top 5 element.

```
res <- table(c(substr(df$phone1, 1, 3), substr(df$phone2, 1, 3)))
sort(res, decreasing=TRUE)[1:5]
```

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

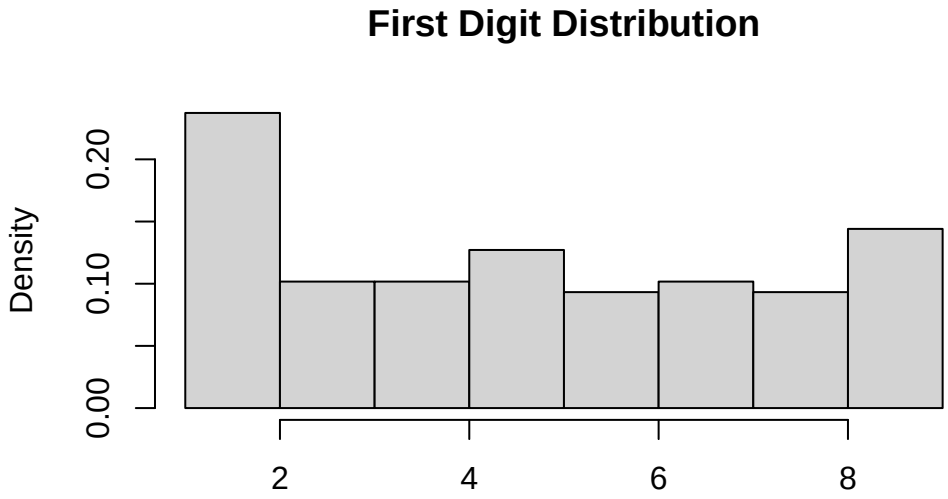
d. Here's the frequency histogram of the apartment numbers.

```
apt_nbrs <- regmatches(df$address, regexpr('[0-9]+$', df$address))
hist(log(as.numeric(apt_nbrs)), main = 'Frequency of Apartment Numbers',
     xlab = 'log(Apartment Number)')
```



e. The data appears to be synthetic, as the first digit does not have a decreasing trend.

```
first_digit <- as.numeric(substring(apt_nbrs, 1, 1))
hist(first_digit, main = 'First Digit Distribution', xlab = '', freq = FALSE)
```



Citaton & Link to GitHub

- Logistic Regression Positive Class in R
- Interpretation of LR coefficients
- GitHub Repo of this Pset