STATS506HW3

```
library(haven)
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
library(DBI)
library(stringr)
library(tidyr)
library(ggplot2)
library(microbenchmark)
library(knitr)
```

Problem 1.

a.

```
aux <- read_xpt("AUX_I.xpt")
demo <- read_xpt("DEMO_I.xpt")
joint_df <- inner_join(aux, demo, by = "SEQN")

print(dim(joint_df))
```

```
[1] 4582  119
```

b.

```
unique(joint_df$INDHHIN2)
```

```
 [1] 10  4  7  6  3 15  5  1  8 14  2 77 NA  9 99 12 13
```

```
unique(joint_df$RIAGENDR)
```

```
[1] 1 2
```

```
unique(joint_df$DMDCITZN)
```

```
[1]  1  2  9  7 NA
```

```
unique(joint_df$DMDHHSZA)
```

```
[1] 0 1 3 2
```

```
joint_df$INDHHIN2[joint_df$INDHHIN2 %in% c(77, 99)] <- NA

joint_df$INDHHIN2 <- factor(joint_df$INDHHIN2)
joint_df$RIAGENDR <- factor(joint_df$RIAGENDR,
                            levels = c(2, 1),
                            labels = c("Female", "Male"))
joint_df$DMDCITZN[joint_df$DMDCITZN %in% c(7, 9)] <- NA
joint_df$DMDCITZN <- factor(joint_df$DMDCITZN,
                            levels = c(1, 2),
                            labels = c("Citizen", "NonCitizen"))
```

c.

```
glm_1R <- glm(AUXTWIDR ~ RIAGENDR, family = poisson(), data = joint_df)
glm_2R <- glm(AUXTWIDR ~ RIAGENDR + DMDCITZN + DMDHHSZA + INDHHIN2, family = poisson(), data

glm_1L <- glm(AUXTWIDL ~ RIAGENDR, family = poisson(), data = joint_df)
glm_2L <- glm(AUXTWIDL ~ RIAGENDR + DMDCITZN + DMDHHSZA + INDHHIN2, family = poisson(), data
```

```r
coef_1R <- summary(glm_1R)$coefficients
coef_2R <- summary(glm_2R)$coefficients
coef_1L <- summary(glm_1L)$coefficients
coef_2L <- summary(glm_2L)$coefficients

est_1R <- coef_1R[,'Estimate']
est_2R <- coef_2R[,'Estimate']
est_1L <- coef_1L[,'Estimate']
est_2L <- coef_2L[,'Estimate']

IIR_1R <- exp(est_1R)
IIR_2R <- exp(est_2R)
IIR_1L <- exp(est_1L)
IIR_2L <- exp(est_2L)

df_1R <- data.frame(Term = names(IIR_1R), IRR_1R = as.numeric(IIR_1R))
df_2R <- data.frame(Term = names(IIR_2R), IRR_2R = as.numeric(IIR_2R))
df_1L <- data.frame(Term = names(IIR_1L), IRR_1L = as.numeric(IIR_1L))
df_2L <- data.frame(Term = names(IIR_2L), IRR_2L = as.numeric(IIR_2L))

IRR_table <- full_join(df_1R, df_2R, by = "Term") %>%
  full_join(df_1L, by = "Term") %>%
  full_join(df_2L, by = "Term") %>%
  mutate(across(everything(), ~ replace_na(., 0))) %>%
  mutate(across(-Term, ~ round(., 3))) %>%
  arrange(Term)

R2 <- function(model) {
  1 - (as.numeric(logLik(model)) / as.numeric(logLik(update(model, . ~ 1))))
}

model_stats <- data.frame(
  Model = c("1R", "2R", "1L", "2L"),
  N = c(nobs(glm_1R), nobs(glm_2R), nobs(glm_1L), nobs(glm_2L)),
  R2 = c(R2(glm_1R), R2(glm_2R), R2(glm_1L), R2(glm_2L)),
  AIC = c(AIC(glm_1R), AIC(glm_2R), AIC(glm_1L), AIC(glm_2L))
)

model_stats$R2 = round(model_stats$R2, 2)
model_stats$AIC = round(model_stats$AIC, 2)

kable(IRR_table, caption = "Incidence Rate Ratios (IRR) for All Models")
```

Table 1: Incidence Rate Ratios (IRR) for All Models

Term	IRR_1R	IRR_2R	IRR_1L	IRR_2L
(Intercept)	85.102	84.041	85.863	87.713
DMDCITZNNonCitizen	0.000	1.067	0.000	1.037
DMDHHSZA	0.000	1.001	0.000	0.983
INDHHIN210	0.000	1.009	0.000	0.958
INDHHIN212	0.000	1.107	0.000	1.067
INDHHIN213	0.000	1.078	0.000	0.901
INDHHIN214	0.000	0.930	0.000	0.950
INDHHIN215	0.000	0.959	0.000	0.954
INDHHIN22	0.000	0.967	0.000	0.980
INDHHIN23	0.000	1.013	0.000	1.031
INDHHIN24	0.000	1.060	0.000	1.048
INDHHIN25	0.000	1.008	0.000	1.004
INDHHIN26	0.000	1.025	0.000	0.942
INDHHIN27	0.000	1.038	0.000	0.997
INDHHIN28	0.000	1.013	0.000	0.950
INDHHIN29	0.000	1.004	0.000	1.007
RIAGENDRMale	0.991	0.987	0.987	0.984

```
kable(model_stats, caption = "Model Statistics for All Models")
```

Table 2: Model Statistics for All Models

Model	N	R2	AIC
1R	4149	0.00	96618.48
2R	3886	0.07	89786.13
1L	4103	0.00	98685.15
2L	3847	0.07	91499.43

d.

```
summary(glm_2L)$coefficients
```

```
                    Estimate  Std. Error     z value     Pr(>|z|)
(Intercept)      4.474071186 0.011679628 383.0662486 0.000000e+00
```

```
RIAGENDRMale       -0.016306018 0.003514585  -4.6395291 3.492039e-06
DMDCITZNNonCitizen  0.036551136 0.004505757   8.1120971 4.975353e-16
DMDHHSZA           -0.017526166 0.002651890  -6.6089336 3.870982e-11
INDHHIN22          -0.020488473 0.015041887  -1.3620946 1.731680e-01
INDHHIN23           0.030513564 0.013675955   2.2311835 2.566898e-02
INDHHIN24           0.046927260 0.013283082   3.5328592 4.110913e-04
INDHHIN25           0.003677600 0.013296090   0.2765926 7.820929e-01
INDHHIN26          -0.059649392 0.012669282  -4.7081904 2.499256e-06
INDHHIN27          -0.002634725 0.012645750  -0.2083486 8.349568e-01
INDHHIN28          -0.051693625 0.012890110  -4.0103323 6.063335e-05
INDHHIN29           0.006953909 0.013169358   0.5280371 5.974736e-01
INDHHIN210         -0.043212618 0.013642945  -3.1673966 1.538104e-03
INDHHIN212          0.064904101 0.014468940   4.4857536 7.265666e-06
INDHHIN213         -0.104489477 0.020196770  -5.1735736 2.296586e-07
INDHHIN214         -0.051091519 0.012643988  -4.0407756 5.327471e-05
INDHHIN215         -0.046963177 0.012158003  -3.8627376 1.121234e-04
```

```
glm_2L_reduced <- glm(AUXTWIDL ~ DMDCITZN + DMDHHSZA + INDHHIN2, family = poisson(), data = j

anova(glm_2L_reduced, glm_2L)
```

```
Analysis of Deviance Table

Model 1: AUXTWIDL ~ DMDCITZN + DMDHHSZA + INDHHIN2
Model 2: AUXTWIDL ~ RIAGENDR + DMDCITZN + DMDHHSZA + INDHHIN2
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      3831      67683
2      3830      67661  1   21.533 3.479e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the summary table, we can see that p-value of RIAGENDR coefficient is less than 0.05, which means that we can reject the null hypothesis(B_gender = 0) in favor of the alternative hypothesis(B_gender != 0). This indicates that gender has significant effect on the incidence rate.

From the anova test, we can also see that the p-value is also less than 0.05, so we reject our null hypothesis (B_gender = 0) in favor of alternative hypothesis(B_gender != 0). This shows that gender influences the incidence rate, which consistent with the result from the summary table.

```r
female_test <- data.frame(
  RIAGENDR = factor("Female", levels = c("Female", "Male")),
  DMDCITZN = factor("Citizen", levels = c("Citizen", "NonCitizen")),
  DMDHHSZA = mean(joint_df$DMDHHSZA, na.rm = TRUE),
  INDHHIN2 = factor(levels(joint_df$INDHHIN2)[1], levels = levels(joint_df$INDHHIN2))
)

male_test <- female_test
male_test$RIAGENDR <- factor("Male", levels = c("Female", "Male"))

female_pred <- predict(glm_2L, newdata = female_test, type = "response")
male_pred <- predict(glm_2L, newdata = male_test, type = "response")

female_pred
```

```
       1
87.18927
```

```r
male_pred
```

```
       1
85.77908
```

In model 2L, the coefficient for gender is -0.0163, and $\exp(-0.0163) = 0.984$, indicating that males have an expected tympanometric width about 1.6% lower than females. The predicted values for males and females are consistent with the previous testing results.

Problem 2.

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

a.

```
sql_q1 <- "
SELECT store_id, Count(customer_id) as total_customer, SUM(CASE WHEN active = 1 THEN 1 ELSE 0
  FROM customer
  Group By store_id
"

microbenchmark(
  r = {
    customer_df <- dbGetQuery(sakila, "SELECT * FROM customer")
    customer_df %>%
      group_by(store_id) %>%
      summarise(
        total_customers = n(),
        active_customers = sum(active == 1),
        pct_active = 100 * mean(active == 1)
      )
  },
  sql = {
    dbGetQuery(sakila, sql_q1)
  },
  times = 50
)
```

```
Unit: microseconds
 expr    min     lq     mean  median     uq    max neval
    r 1846.4 1899.7 2091.740 1947.15 2036.1 5227.5    50
  sql  194.1  220.0  277.052  271.10  288.2 1058.1    50
```

b.

```
sql_q2 <-  "
SELECT s.staff_id, s.first_name || ' ' || s.last_name AS full_name, co.country
  FROM staff s
  Left Join address a on s.address_id = a.address_id
  Left Join city c on a.city_id = c.city_id
  Left Join country co on c.country_id = co.country_id
"

microbenchmark(
  r = {
```

```
    staff_df <- dbGetQuery(sakila, "SELECT * FROM staff")
    address_df <- dbGetQuery(sakila, "SELECT * FROM address")
    city_df <- dbGetQuery(sakila, "SELECT * FROM city")
    country_df <- dbGetQuery(sakila, "SELECT * FROM country")
    staff_df %>%
      left_join(address_df, by = "address_id") %>%
      left_join(city_df, by = "city_id") %>%
      left_join(country_df, by = "country_id") %>%
      transmute(
        staff_id,
        full_name = paste(first_name, last_name),
        country
      )
  },
  sql = {
    dbGetQuery(sakila, sql_q2)
  },
  times = 50
)
```

```
Unit: microseconds
 expr    min     lq     mean  median     uq    max neval
    r 3785.9 3939.0 4282.620 4069.45 4413.9 6949.0    50
  sql  129.8  156.7  242.572  219.20  231.0 2314.7    50
```

c.

```
sql_q3 <- "
SELECT f.title, SUM(p.amount) AS total_amount
FROM film f
JOIN inventory i ON f.film_id = i.film_id
JOIN rental r ON i.inventory_id = r.inventory_id
JOIN payment p ON r.rental_id = p.rental_id
GROUP BY f.title
HAVING total_amount = (
  SELECT MAX(total_rev)
  FROM (
    SELECT SUM(p2.amount) AS total_rev
    FROM film f2
    JOIN inventory i2 ON f2.film_id = i2.film_id
    JOIN rental r2 ON i2.inventory_id = r2.inventory_id
```

```
    JOIN payment p2 ON r2.rental_id = p2.rental_id
    GROUP BY f2.title
  )
);
"

microbenchmark(
  r = {
    film_df <- dbGetQuery(sakila, "SELECT * FROM film")
    inventory_df <- dbGetQuery(sakila, "SELECT * FROM inventory")
    rental_df <- dbGetQuery(sakila, "SELECT * FROM rental")
    payment_df <- dbGetQuery(sakila, "SELECT * FROM payment")

    film_df %>%
      left_join(inventory_df, by = "film_id") %>%
      left_join(rental_df, by = "inventory_id") %>%
      left_join(payment_df, by = "rental_id") %>%
      group_by(title) %>%
      summarise(total_revenue = sum(amount, na.rm = TRUE)) %>%
      filter(total_revenue == max(total_revenue))
  },
  sql = {
    dbGetQuery(sakila, sql_q3)
  },
  times = 50
)
```

```
Unit: milliseconds
 expr     min      lq     mean   median      uq      max neval
    r 85.6835 87.7682 96.25786 90.48360 95.9312 179.4811    50
  sql 54.2516 55.2332 56.98420 56.51165 58.0477  77.7295    50
```

Problem 3.

a.

```
aus <- read.csv("au-500.csv", stringsAsFactors = FALSE)
head(aus)
```

```
  first_name last_name                   company_name               address
1   Rebbecca     Didio         Brandt, Jonathan F Esq         171 E 24th St
```

```
2     Stevie     Hallo     Landrum Temporary Services        22222 Acoma St
3     Mariko    Stayer            Inabinet, Macre Esq 534 Schoenborn St #51
4    Gerardo    Woodka       Morris Downing & Sherred     69206 Jackson Ave
5      Mayra      Bena             Buelt, David L Esq     808 Glen Cove Ave
6     Idella  Scotland Artesian Ice & Cold Storage Co      373 Lafayette St
         city state post       phone1       phone2                         email
1       Leith   TAS 7315 03-8174-9123 0458-665-290 rebbecca.didio@didio.com.au
2     Proston   QLD 4613 07-9997-3366 0497-622-620    stevie.hallo@hotmail.com
3       Hamel    WA 6215 08-5558-9019 0427-885-282   mariko_stayer@hotmail.com
4    Talmalmo   NSW 2640 02-6044-4682 0443-795-912  gerardo_woodka@hotmail.com
5   Lane Cove   NSW 1595 02-1455-6085 0453-666-885        mayra.bena@gmail.com
6 Cartmeticup    WA 6316 08-7868-1355 0451-966-921          idella@hotmail.com
                                          web
1       http://www.brandtjonathanfesq.com.au
2 http://www.landrumtemporaryservices.com.au
3         http://www.inabinetmacreesq.com.au
4     http://www.morrisdowningsherred.com.au
5           http://www.bueltdavidlesq.com.au
6 http://www.artesianicecoldstorageco.com.au
```

```r
aus <- aus %>%
  mutate(
    web_reduced = str_remove(web, "^https?://(www\\.)?"),
    domain_ending = str_extract(web_reduced, "[^.]+\\.[^.]+$")
  )

pct_com <- mean(str_detect(aus$domain_ending, "\\.com$"), na.rm = TRUE) * 100
pct_com
```

```
[1] 0
```

b.

```r
aus <- aus %>%
  mutate(email_domain = sub(".*@", "", email))

most_common_domain <- aus %>%
  count(email_domain, sort = TRUE) %>%
  slice(1)

most_common_domain
```

```
  email_domain   n
1  hotmail.com 114
```

c.

```
aus <- aus %>%
  mutate(company_clean = str_remove_all(company_name, "[ ,&]"),
         non_alpha2 = str_detect(company_clean, "[^A-Za-z]"))

prop_non_alpha <- mean(aus$non_alpha2, na.rm = TRUE)
prop_non_alpha
```

```
[1] 0.008
```

d.

```
format_cell <- function(x) {
  digits <- str_replace_all(x, "\\D", "")
  if (nchar(digits) == 10) {
    paste0(substr(digits, 1, 4), "-", substr(digits, 5, 7), "-", substr(digits, 8, 10))
  } else {
    x
  }
}

aus$phone1 <- sapply(aus$phone1, format_cell)
aus$phone2 <- sapply(aus$phone2, format_cell)

head(aus$phone1, 10)
```

```
 [1] "0381-749-123" "0799-973-366" "0855-589-019" "0260-444-682" "0214-556-085"
 [6] "0878-681-355" "0865-228-931" "0252-269-402" "0731-849-989" "0868-904-661"
```

```
head(aus$phone2, 10)
```

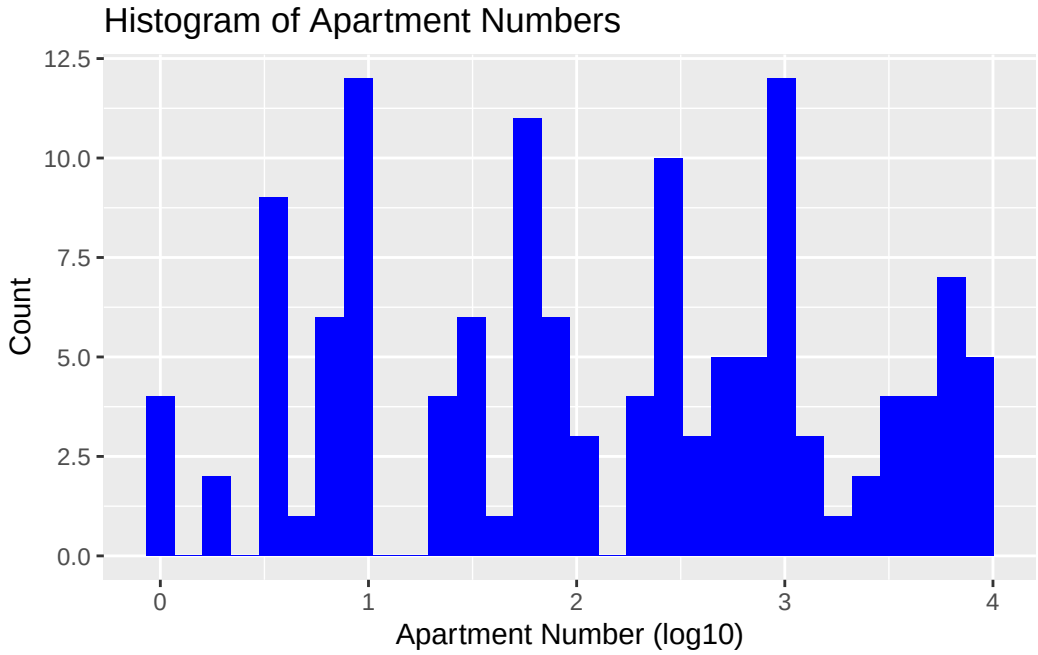
```
 [1] "0458-665-290" "0497-622-620" "0427-885-282" "0443-795-912" "0453-666-885"
 [6] "0451-966-921" "0427-991-688" "0415-961-606" "0411-732-965" "0461-862-457"
```

e.

```r
aus <- aus %>%
  mutate(apt_num = as.numeric(str_extract(address, "\\d+$")))

apt_nums <- aus$apt_num[!is.na(aus$apt_num)]

ggplot(data.frame(apt_nums), aes(x = log10(apt_nums))) +
  geom_histogram(bins = 30, fill = "blue") +
  labs(title = "Histogram of Apartment Numbers",
       x = "Apartment Number (log10)",
       y = "Count")
```



f.

```r
leading_digit <- as.numeric(str_sub(apt_nums, 1, 1))

obs <- table(leading_digit) / length(leading_digit)

benford <- log10(1 + 1 / (1:9))
chisq.test(x = obs, p = benford)
```

```
    Chi-squared test for given probabilities
```

```
data:  obs
X-squared = 0.47898, df = 8, p-value = 0.9999
```

The chi-square test shows that the distribution of the leading digits of apartment numbers closely follows Benford's law. This suggests that, based on this test, the apartment numbers would likely pass as real data.