Homework 3

Submit the solution in the form of R Markdown report, knitted into either of the available formats (HTML, pdf or Word). Provide all relevant code and output. Goal of this homework is to have you 1) familiarized X^2-test of independence for contingency tables; 2) familiarized with permutation test for contingency tables; 3) interpretation of linear regression; 4) practice your R coding.

Problem #1

1. Code up your own `my.permutation.test()` function to conduct permutations tests on contingency tables.

As inputs, it should take

```
- data frame with two categorical variables as columns (first one - explanatory, second one - respons
- # of randomly generated permutations to be executed.
```

As outputs, it should provide

```
- contingency table for the data frame
- permutation p-value
- plot the histogram of permutation distribution for $X^2$ statistic
```

```
my.permutation.test <- function(data, n){
  data <- data %>% as.data.frame()
  contingency.table <- data %>% table

  values <- sapply(1:n, function(x){
    table <- data.frame(sample(data[,1]), data[,2]) %>% table()
    chisq.test(table)$statistic
  })

  x.2 <- chisq.test(contingency.table)$statistic
  p.value <- sum( values >= x.2 )/n

  plot <- values %>% as.data.frame() %>%
    ggplot(aes(.)) +
      geom_histogram(aes(y=..count../n), binwidth = 0.2, fill = "lightblue") +
      stat_function(fun = dchisq,
                    args = list(df = (nrow(contingency.table)-1)*(ncol(contingency.table)-1)),
                    color = "darkred", size = 1) +
      geom_vline(xintercept=x.2) +
      ggtitle("Chi-Squared Distribution") +
```

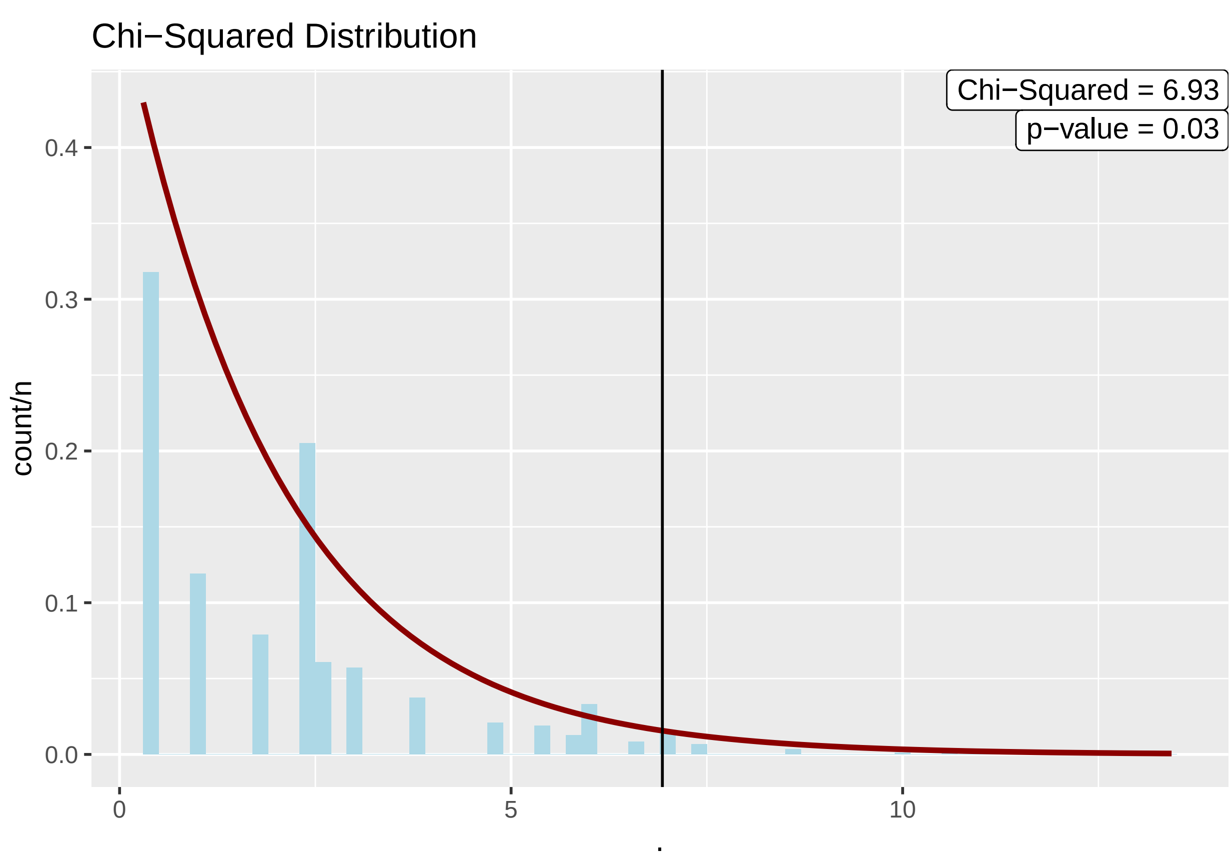
```
      annotate(x=+Inf,y=+Inf,hjust=1,vjust=1,
               label=sprintf("Chi-Squared = %0.2f",x.2), geom="label") +
      annotate(x=+Inf,y=+Inf,hjust=1,vjust=2,
               label=sprintf("p-value = %0.2f",p.value), geom="label")

  list( table = knitr::kable(stats::addmargins(contingency.table), caption = "Contingency Table"),
     plot = plot,
     p.value = p.value
     )
}
```

2. Proceed to apply the my.permutation.test() function (and subsequently interpret the results) to:

- Snowden data (from the lecture), with 10, 000 permutations. What's the conclusion? Compare the resulting histogram with the one in the slides (they should be roughly similar).

Table 1: Contingency Table

	Criminal	Hero	Neither	Sum
Intl	2	5	1	8
US	9	1	2	12
Sum	11	6	3	20

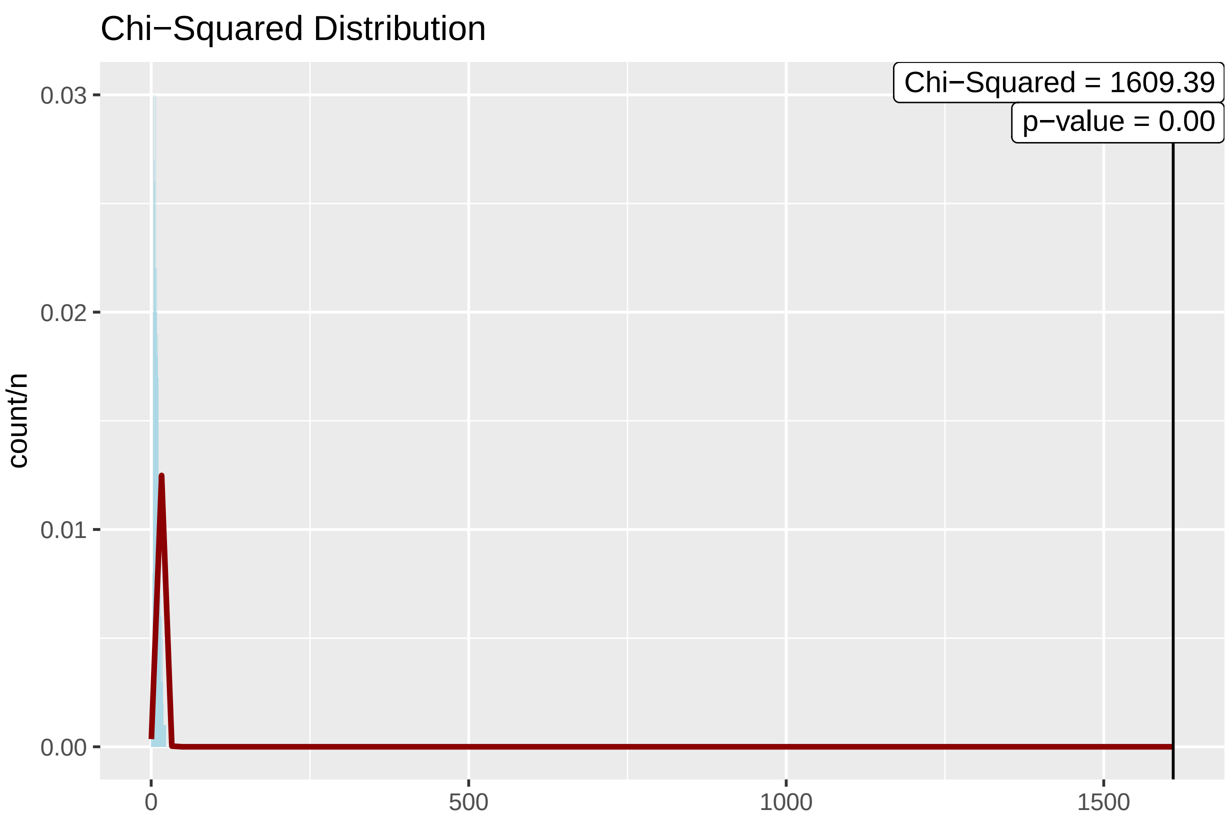


The resulting histogram is similar to the one on the slides.

- Airbnb data (from previous HW), with just 1, 000 permutations. What's the conclusion? Compare the shape of resulting histogram with the density of X^2 distribution with appropriate degrees of freedom. What does it tell us about whether X^2-test results from previous HW were appropriate for Airbnb data?

Table 2: Contingency Table

	Bronx	Brooklyn	Manhattan	Queens	Staten Island	Sum
Entire home/apt	378	9565	13054	2118	181	25296
Private room	659	10131	7931	3489	187	22397
Shared room	68	418	471	204	10	1171
Sum	1105	20114	21456	5811	378	48864



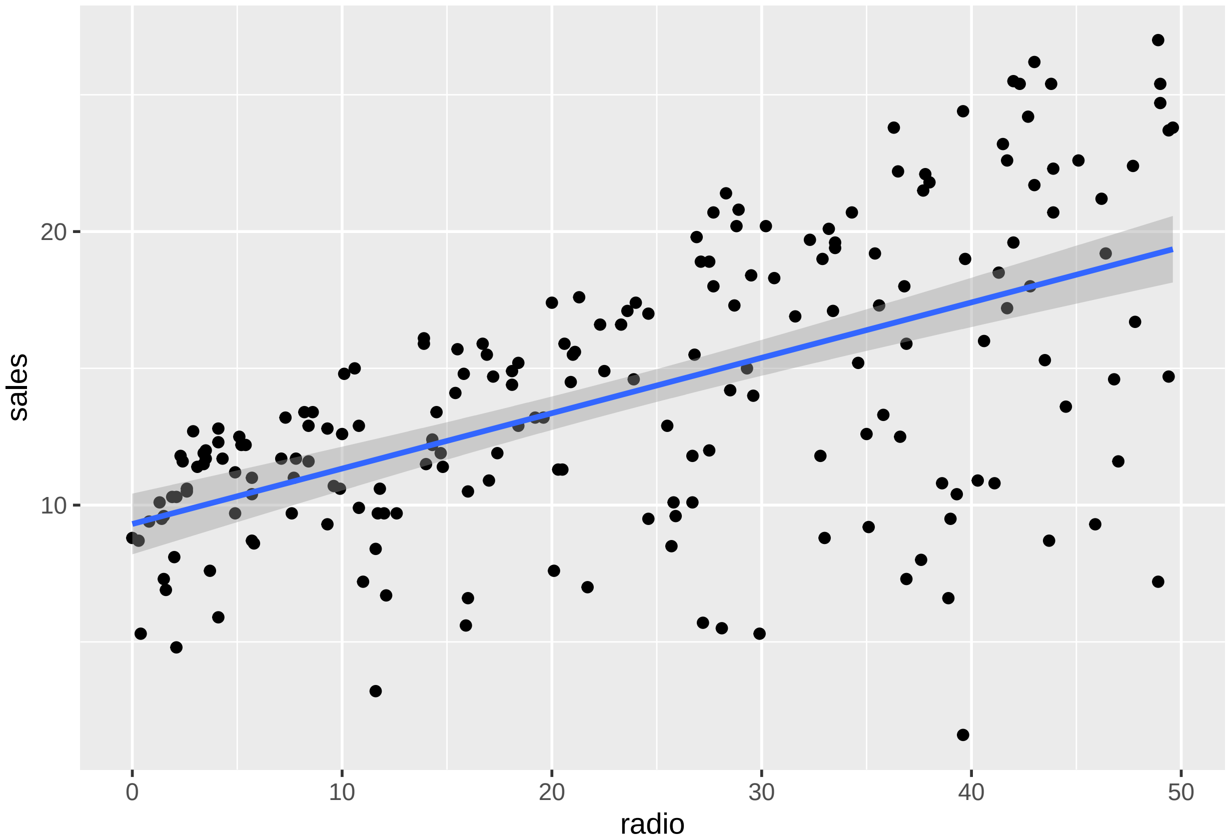
With axes so large it is hard to say definitively, but the histogram appears to roughly follow the density X^2 distribution with the appropriate degrees of freedom, besides the higher than expected volume of values closer to 0. This indicates that the X^2-test results from the previous HW are appropriate.

Problem #2

In Advertisement.csv data set, proceed to study the relationship between the sales and radio advertising expenses. In particular, proceed to

```
df <- readr::read_csv(sprintf("https://docs.google.com/uc?id=%s&export=download", "1UJIu7Ku3rRWTnFJpK4u
```

1. *Plot their relationship. Does linear regression appear as appropriate model here?*



Although the points are very spread, they do seem to slightly follow a linear trend, thus making the usage of a linear regression model appropriate.

2. *Regardless of the answer to Part 1, proceed to fit the linear regression and write down the fitted model equation.*

```
lm <- lm(sales~radio, data=df)
```

Sales = 9.3116381 + 0.2024958TV

3. *Interpret both the slope and the intercept.*

```
- slope: Per 1000 dollar increase in radio advertisement, the number of items sold increases by 202 ite
- intercept: For companies who spend 0 dollars on radio advertisement, the number of items sold is 9311
```

4. *Provide and interpret the prediction for a 50,000 investment into this advertisement media.* For companies who spend 50,000 dollars on radio advertisement, the number of items sold is 19436, on average.

5. *Report and interpret the Residual Standard Error (RSE)* Our predicted Sales miss the true Sales by 4274.9443549 items, on average

6. *Report and interpret the R^2 statistic* Our linear regression model with radio advertisement as a predictor explains 33.2032455 percent of variety in Sales.

Problem #3

1. Proceed to write your own function which will calculate β_0 and β_1 estimates given vectors X and Y as input.

```
coef.calculation <- function(x, y){
  x.mean <- mean(x)
  y.mean <- mean(y)
  beta.1 <- sum( (x - x.mean)*(y - y.mean) ) / sum( (x - x.mean)^2 )
  beta.0 <- y.mean - beta.1 * x.mean

  c(beta.0 = beta.0,
    beta.1 = beta.1)
}
```

```
coef.calculation(df$TV, df$sales)
```

```
##     beta.0     beta.1
## 7.03259355 0.04753664
```

```
lm(sales~TV, data = df)
```

```
##
## Call:
## lm(formula = sales ~ TV, data = df)
##
## Coefficients:
## (Intercept)           TV
##     7.03259      0.04754
```

2.Write your own function that, for a simple linear regression will calculate Risidual Standard Error (RSE) and R^2 statistic given vectors X and Y as input.

```
my.lm <- function(x, y){
  lm <- lm(y~x)
  rss <- sum(lm$residuals^2)

  c(rse = sqrt( rss / (length(lm$residuals) - 2) ),
    r.squared = (sum( (y - mean(y))^2 ) - rss) / sum( (y - mean(y))^2 )
  )
}
```

```
my.lm(df$TV, df$sales)
```

```
##       rse r.squared
## 3.2586564 0.6118751
```

```
lm(sales~TV, data = df) %>% summary()
```

```
##
## Call:
## lm(formula = sales ~ TV, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3860 -1.9545 -0.1913  2.0671  7.2124
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.032594   0.457843   15.36   <2e-16 ***
## TV          0.047537   0.002691   17.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.259 on 198 degrees of freedom
## Multiple R-squared:  0.6119, Adjusted R-squared:  0.6099
## F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```