Homework 3

Miles Tweed

2/20/2021

***my.chisq.test()* function from HW2**

```
my.chisq.test <-function(ct=NULL) {
  d <- dim(ct)
  total <- sum(ct)
  ect <- ct
  df <- (d[1]-1) * (d[2]-1)
  for(i in 1:d[1]){
    r.tot <- sum(ct[i,])
    for(j in 1:d[2]) {
      c.tot <- sum(ct[,j])
      ect[i,j] <- r.tot * c.tot / total
    }
  }
  chi <- 0
  for(i in 1:d[1]){
    for(j in 1:d[2]) {
      chi <- chi + (ct[i,j] - ect[i,j])^2 / ect[i,j]
    }
  }
  p = pchisq(chi, df = df, lower.tail = FALSE)

  print(paste("The chi-squared statistic is: ",chi))
  if(p != 0) {
    print(paste("The p-value is: ", p))
  } else {
    print("p-value << 0.0001")
  }
}
```

Problem 1

Part 1.a.

my.chisq.test() confirmed the χ^2 value and the p-value.

```
rel.case.a <- matrix(c(0.51*100, 0.49*100,0.49*100,0.51*100),
                     nrow = 2, ncol = 2,
                     dimnames = list(c('Female','Male'),c('yes','no')))
rel.case.b <- matrix(c(0.51*200, 0.49*200,0.49*200,0.51*200),
                     nrow = 2, ncol = 2,
                     dimnames = list(c('Female','Male'),c('yes','no')))
rel.case.c <- matrix(c(0.51*10000, 0.49*10000,0.49*10000,0.51*10000),
```

```
                    nrow = 2, ncol = 2,
                    dimnames = list(c('Female','Male'),c('yes','no')))

# Chi-squared statistic should be 0.08 and p-value should be 0.78
my.chisq.test(rel.case.a)
```

```
## [1] "The chi-squared statistic is:  0.08"
## [1] "The p-value is:  0.777297410789522"
```

```
# Chi-squared statistic should be 0.16 and p-value should be 0.69
my.chisq.test(rel.case.b)
```

```
## [1] "The chi-squared statistic is:  0.16"
## [1] "The p-value is:  0.689156516779352"
```

```
# Chi-squared statistic should be 8.0 and p-value should be 0.005
my.chisq.test(rel.case.c)
```

```
## [1] "The chi-squared statistic is:  8"
## [1] "The p-value is:  0.00467773498104727"
```

Part 1.b.

Proportion differences between males and females that attend weekly religious services.

```
prop.diff.a <- rel.case.a[1,1]/100-rel.case.a[2,1]/100
prop.diff.a
```

```
## [1] 0.02
```

```
prop.diff.b <- rel.case.b[1,1]/200-rel.case.b[2,1]/200
prop.diff.b
```

```
## [1] 0.02
```

```
prop.diff.c <- rel.case.c[1,1]/10000-rel.case.c[2,1]/10000
prop.diff.c
```

```
## [1] 0.02
```

Risk ratios between males and females that attend weekly religious services.

```
rr.a <- (rel.case.a[1,1]/100)/(rel.case.a[2,1]/100)
rr.a
```

```
## [1] 1.040816
```

```
rr.b <- (rel.case.b[1,1]/200)/(rel.case.b[2,1]/200)
rr.b
```

```
## [1] 1.040816
```

```
rr.c <- (rel.case.c[1,1]/10000)/(rel.case.c[2,1]/10000)
rr.c
```

```
## [1] 1.040816
```

Part 1.c.

Based on the answers, it is apparent that, although the statistical significance increases as as the sample size increases, the practical significance does not change at all.

Part 2. Exercise 11.32

```
GSS <- matrix(c(11,31,20,11,215,231,4,34,337),
              nrow = 3,ncol = 3,
              dimnames = list(Marital_Happiness=c('Not Too Happy', 'Happy','Very Happy'),
                              General_Happiness=c('Not Too Happy', 'Happy','Very Happy')))
GSS
```

```
##                  General_Happiness
## Marital_Happiness Not Too Happy Happy Very Happy
##     Not Too Happy            11    11          4
##     Happy                    31   215         34
##     Very Happy               20   231        337
```

```
chisq.test(GSS)
```

```
##
##  Pearson's Chi-squared test
##
## data:  GSS
## X-squared = 213.76, df = 4, p-value < 2.2e-16
```

Part a. Because this is such a large chi-squared I could conclude that we should reject the null hypothesis, that marital happiness and general happiness are independent, in favor of the alternative hypothesis, that marital happiness and general happiness are not independent. This interpretation of this conclusion is that changes in marital happiness are more likely than not to result in changes in general happiness. Since this value is so large there is strong evidence against the independence of the two measures.

Part b. Although the chi-squared statistic gives strong evidence against the null hypothesis of independence it does not suggest that there is a strong association between marital happiness and general happiness. To measure the strength of the association we would have to compare proportions between specific categories.

Part c.

```
GSS.prop <- GSS
GSS.prop[1,] <- GSS[1,]/sum(GSS[1,])
GSS.prop[2,] <- GSS[2,]/sum(GSS[2,])
GSS.prop[3,] <- GSS[3,]/sum(GSS[3,])

GSS.prop
```

```
##                  General_Happiness
## Marital_Happiness Not Too Happy     Happy Very Happy
##     Not Too Happy    0.42307692 0.4230769  0.1538462
##     Happy            0.11071429 0.7678571  0.1214286
##     Very Happy       0.03401361 0.3928571  0.5731293
```

```
prop.dif <- GSS.prop[1,1] - GSS.prop[3,1]

prop.dif
```

```
## [1] 0.3890633
```

This suggests that there is a 38.9 percentage point difference between the proportion of people who are 'Not Too Happy' versus people who are 'Very Happy' in their marriage, in favor of those who are 'Not Too Happy' in their marriage, if those people are also 'Not Too Happy' in general.

Part d.

```
rel.risk <- GSS.prop[1,1] / GSS.prop[3,1]
rel.risk
```

```
## [1] 12.43846
```

The relative risk suggests that someone is 12.4 times more likely to be 'Not Too Happy' than to be 'Very Happy' in their marriage if they are 'Not Too Happy' in general.

Problem 2

Part 1

```
my.permutation.test <- function(df,perm){
  require(tidyverse)
  conTable <- table(df)

  orig.chi <- as.numeric(chisq.test(table(df))$statistic)

  vec.chi <- c()
  for (x in c(1:perm)){
    vec.chi <- append(vec.chi, as.numeric(chisq.test(table(sample(df[,1]),df[,2]))$statistic))
  }

  vec.chi <- as_data_frame(vec.chi)

  num.gt.chi <- vec.chi %>% filter(value >= orig.chi) %>% count()

  p <- as.numeric(num.gt.chi/perm)

  print(conTable)
  print(paste('Chi-Squared value:',orig.chi))
  if(p>0){
    print(paste('Permutation p-value:',p))
  } else {
    print('Pemutation p-value << 0.0001')
  }

  ggplot(vec.chi, aes(x = value)) +
    geom_histogram(binwidth = 1,
                   boundary = 0,
                   closed = 'left',
                   color='lightgrey',
                   fill='slateblue') +
    labs(title = "Permutation Distribution") + xlab('Chi-squared Statistic') +
    theme()
}
```

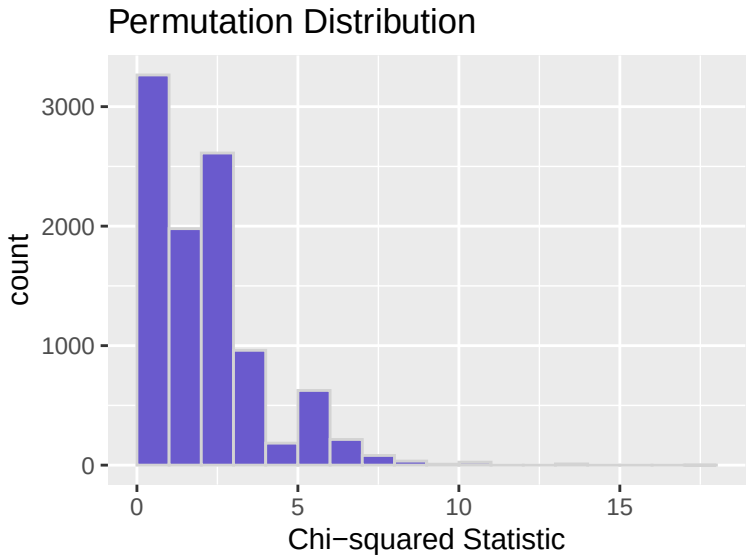
Part 2

a. Using a 0.5 significance value, the results of the permutation test suggests that we should reject the null hypothesis (H_0: Student status and opinion on Edward Snowden are independent) in favor of the alternative hypothesis (H_a: Student status and opinion on Edward Snowden are not independent). The distribution does resemble the one from the book.

```
student.status <- c(rep("US", 12),
                    rep("Intl", 8))
opinion <- c("Hero", rep("Criminal", 9), rep("Neither", 2),
             rep("Hero", 5), rep("Criminal", 2), rep("Neither", 1))
Snowden <- data.frame(student.status, opinion)
```

```
my.permutation.test(Snowden,10000)
```

```
##                opinion
## student.status Criminal Hero Neither
##           Intl        2    5       1
##           US          9    1       2
## [1] "Chi-Squared value: 6.93181818181818"
## [1] "Permutation p-value: 0.0313"
```



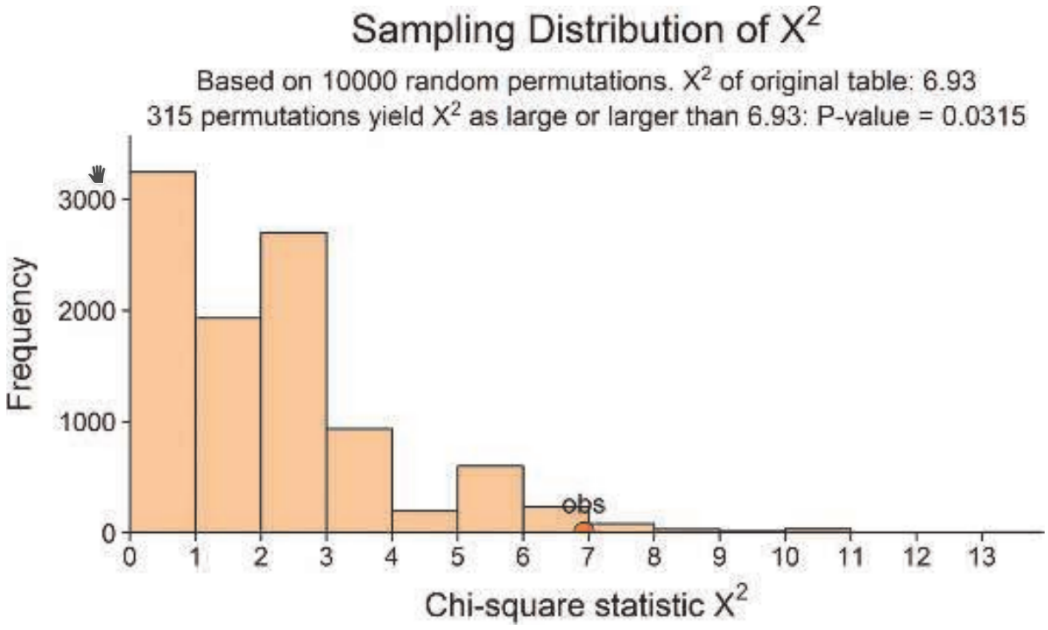


Figure 1: Graph of permutation distribution from the book/slide.

b. The p-value returned by the function was very small ($p \ll 0.0001$) which suggests that we should reject the null hypothesis (H_0: Borough and price range are independent.) in favor of the alternative hypothesis (H_a: Borough and price range are not independent.). The degrees of freedom for this contingency table

are df $= (5-1)\cdot(4-1) = 12$ and, using the shiny app at https://istats.shinyapps.io/ChisqDist/, the permutation distribution is similar to the chi-squared distribution with 12 degrees of freedom. This suggests that the previous results were appropriate since the sample size was large enough and permutation under the assumtion of no association led to a distribution of chi-squared statistics that closely resembles the chi-squared distribution with df=12.

```
airbnb <- read.csv('../../Data/listings.csv')

encode <- function(price){
  result <- vector()
  for(i in 1:length(price)){
    if(price[i] <= 69) {
      result <- append(result, 'Low Priced')
    } else if(price[i] > 69 & price[i] <= 105){
      result <- append(result,'Moderately Low Priced')
    } else if(price[i] > 105 & price[i] <= 175) {
      result <- append(result,'Moderately High Priced')
    } else {
      result <- append(result,'High Priced')
    }
  }
  result
}

airbnb$price_cat <- encode(airbnb$price)

airbnb$price_cat <-  fct_relevel(airbnb$price_cat, c('Low Priced',
                                                     'Moderately Low Priced',
                                                     'Moderately High Priced',
                                                     'High Priced'))

air_price <- airbnb[,c("neighbourhood_group","price_cat")]

my.permutation.test(air_price,1000)
```

```
##                    price_cat
## neighbourhood_group Low Priced Moderately Low Priced Moderately High Priced
##       Bronx                589                   296                    152
##       Brooklyn            6569                  5426                   4858
##       Manhattan           2481                  4582                   6189
##       Queens              2651                  1629                    982
##       Staten Island        158                   117                     64
##                    price_cat
## neighbourhood_group High Priced
##       Bronx                  68
##       Brooklyn             3261
##       Manhattan            8204
##       Queens                549
##       Staten Island          39
## [1] "Chi-Squared value: 6760.98516245231"
## [1] "Pemutation p-value << 0.0001"
```

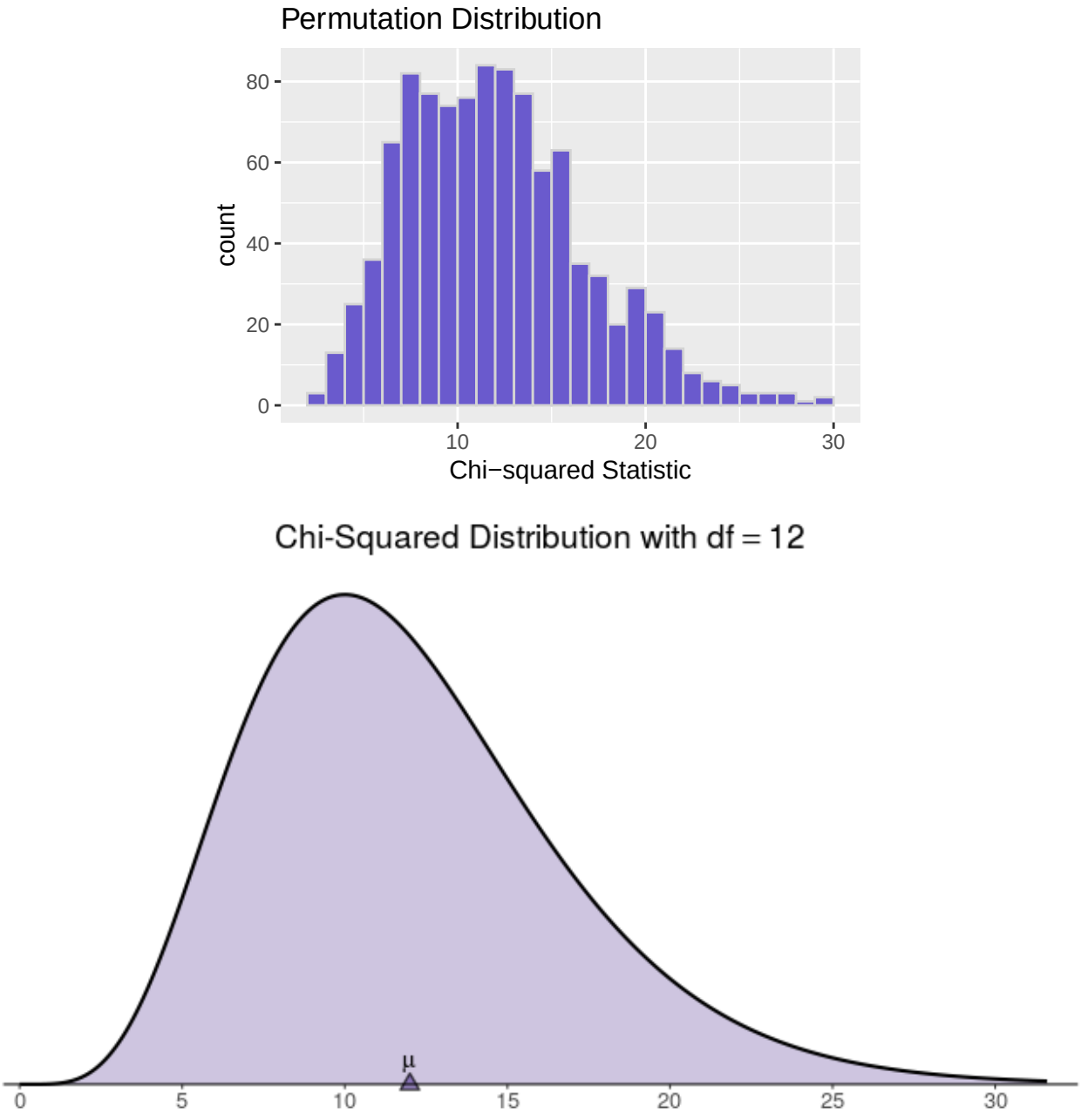


Figure 2: Graph of chi-squared distribution with df = 12 from https://istats.shinyapps.io/ChisqDist/.

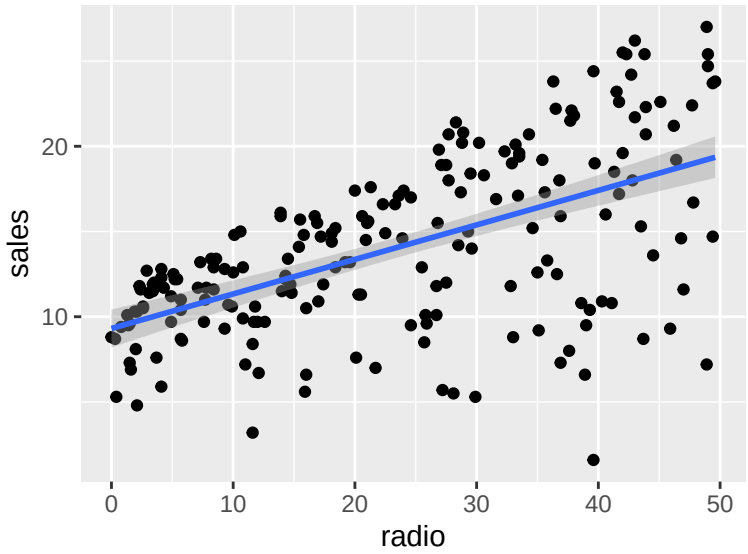
Problem 3

```
ads <- read.csv('../../Data/Advertising.csv')
```

Part a. Based on the plots below, linear regression seems much more appropriate for radio and sales than newspaper and sales. Regardless, linear regression could be used on this data to some effect.

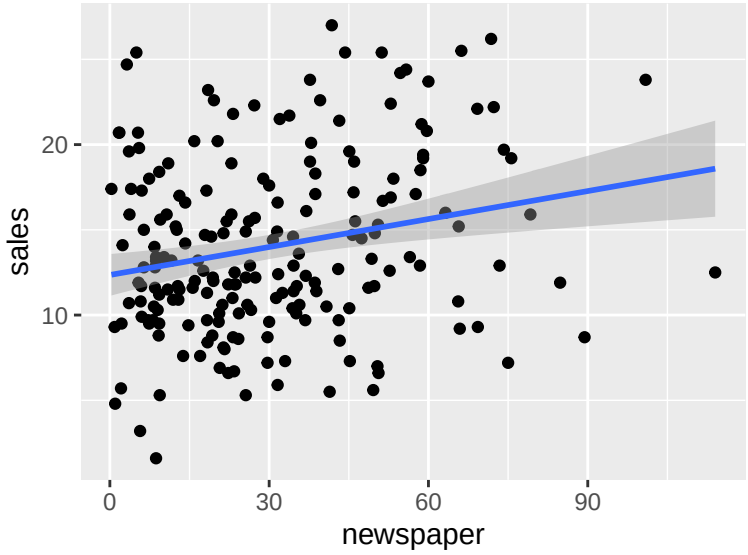
```
ads %>% ggplot(aes(x = radio, y = sales)) +
  geom_point() +
  geom_smooth(method = 'lm')
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
ads %>% ggplot(aes(x = newspaper, y = sales)) +
  geom_point() +
  geom_smooth(method='lm')
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Part b.

```
model.one <- with(ads,lm(sales~radio))
summary(model.one)
```

```
##
## Call:
## lm(formula = sales ~ radio)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.7305  -2.1324   0.7707   2.7775   8.1810
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.31164    0.56290  16.542   <2e-16 ***
## radio        0.20250    0.02041   9.921   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.275 on 198 degrees of freedom
## Multiple R-squared:  0.332,  Adjusted R-squared:  0.3287
## F-statistic: 98.42 on 1 and 198 DF,  p-value: < 2.2e-16
```

```
model.two <- with(ads,lm(sales~newspaper))
summary(model.two)
```

```
##
## Call:
## lm(formula = sales ~ newspaper)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.2272  -3.3873  -0.8392   3.5059  12.7751
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.35141    0.62142   19.88  < 2e-16 ***
## newspaper    0.05469    0.01658    3.30  0.00115 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.092 on 198 degrees of freedom
## Multiple R-squared:  0.05212,    Adjusted R-squared:  0.04733
## F-statistic: 10.89 on 1 and 198 DF,  p-value: 0.001148
```

The two fitted model equations are:

$$\text{sales} = 9.312 + 0.203 \cdot \text{radio}$$

$$\text{sales} = 12.351 + 0.055 \cdot \text{newspaper}$$

Part c.

The units for sales is in 1000 items while the units of radio and newspaper is in $1000.

For sales ~ radio:

The slope (0.203) implies that increasing the radio advertisement expenditure by $1000 should increase sales by 203 items on average and the intercept (9.312) implies that if $0k is spent on radio advertisement the company should sell 9,312 items on average. The average of the slope is over all situation where two radio advertisement expenditures differ by $1000 and the intercept is averaged across all markets where $0k is spent on radio advertisement.

For sales ~ newspaper:

The slope (0.055) implies that increasing the newspaper advertisement expenditure by \$1000 should increase sales by 55 items on average and the intercept (12.351) implies that if \$0k is spent on newspaper advertisement the company should sell 12,351 items on average. The averages of the slopes are across all situation where two newspaper advertisement expenditures differ by \$1000 and the intercepts are averaged across all markets where \$0k is spent on newspaper advertisement.

Part d.

```
# Prediction of 50,000 spent on radio advertisement
predict(model.one, newdata = data.frame(radio=50))
```

```
##        1
## 19.43643
```

```
# Prediction of 50,000 spent on newspaper advertisement
predict(model.two, newdata = data.frame(newspaper=50))
```

```
##        1
## 15.08606
```

These predictions imply that a company spending $\$50,000$ on radio advertisement would sell $19.43643 \cdot 1000 = 19,436.43$ items on average across all companies that spent $\$50,000$ on radio advertisement and, similarly, a company spending $\$50,000$ on newspaper advertisement would sell $15.08606 \cdot 1000 = 15,086.06$ items on average across all companies that spent $\$50,000$ on newspaper advertisement.

Part e. The RSE for the fit of radio and sales was 4.275 while the RSE for the fit of newspaper and sales was 5.092. These values represent the average deviation of the actual data points from the fitted trend line. In this case the estimates for sales using values of radio advertisement expenditure (in \$1000's) will differ from the true values by 4,275 units on average with model one, while the estimates for sales using values of newspaper advertisement expenditure (in \$1000's) will differ from the true values by 5,092 units on average with model two. These both seem like very high margins of error, however, because the R^2 is fairly low ($R^2 < 0.5$) we would not expect a great fit and the RSE can also be thought of as a measure of the lack of fit.

Part f. The R^2 statistic for the interaction of radio and sales was 0.33 while the R^2 statistic for the interaction of newspaper and sales was 0.05. This indicates that radio rather than newspaper described more of the variance in sales on average. In other words, a greater proportion of the variance in sales is explained by expenditure in radio advertisement alone than by expenditure in newspaper advertisement alone.