IST772 Problem Set 8

Abhijith Anil Vamadev

The homework for week 8 is based on exercises 2-8 on pages 181-182 but with changes as noted in this notebook (i.e., follow the problems as given in this document and not the textbook).

Attribution statement: (choose only one) 1. I did this homework by myself, with help from the book and the professor

Chapter 8, Exercise 2

The carData package in R contains a small data set called Prestige that contains n = 102 observations of different occupations in Canada in 1971. Load the carData package and use "?Prestige" to display help about the data set.

As always, you should start by examining the data and identifying any problematic variables. (1 pt) You should also look at the relations and apply any needed transformations. (1 pt) But be conservative!

Create and interpret a bivariate correlation matrix keeping in mind the idea that you will be trying to predict the prestige variable. You will have to select out non-numeric variables (i.e., factors) and should eliminate other variables if they are not meaningful. Which other variable might be the single best predictor? (1 pt)

```r
library(carData)
data <- carData::Prestige #loading the data
str(data) #structure of the data
```

```
## 'data.frame':    102 obs. of  6 variables:
##  $ education: num  13.1 12.3 12.8 11.4 14.6 ...
##  $ income   : int  12351 25879 9271 8865 8403 11030 8258 14163 11377 11023 ...
##  $ women    : num  11.16 4.02 15.7 9.11 11.68 ...
##  $ prestige : num  68.8 69.1 63.4 56.8 73.5 77.6 72.6 78.1 73.1 68.8 ...
##  $ census   : int  1113 1130 1171 1175 2111 2113 2133 2141 2143 2153 ...
##  $ type     : Factor w/ 3 levels "bc","prof","wc": 2 2 2 2 2 2 2 2 2 2 ...
```

```r
table(is.na(data$education)) #checking null values
```

```
##
## FALSE
##   102
```

```r
table(is.na(data$income)) #checking null values
```

```
##
## FALSE
##   102
```

```
table(is.na(data$women)) #checking null values
```

```
##
## FALSE
##   102
```

```
table(is.na(data$prestige)) #checking null values
```

```
##
## FALSE
##   102
```

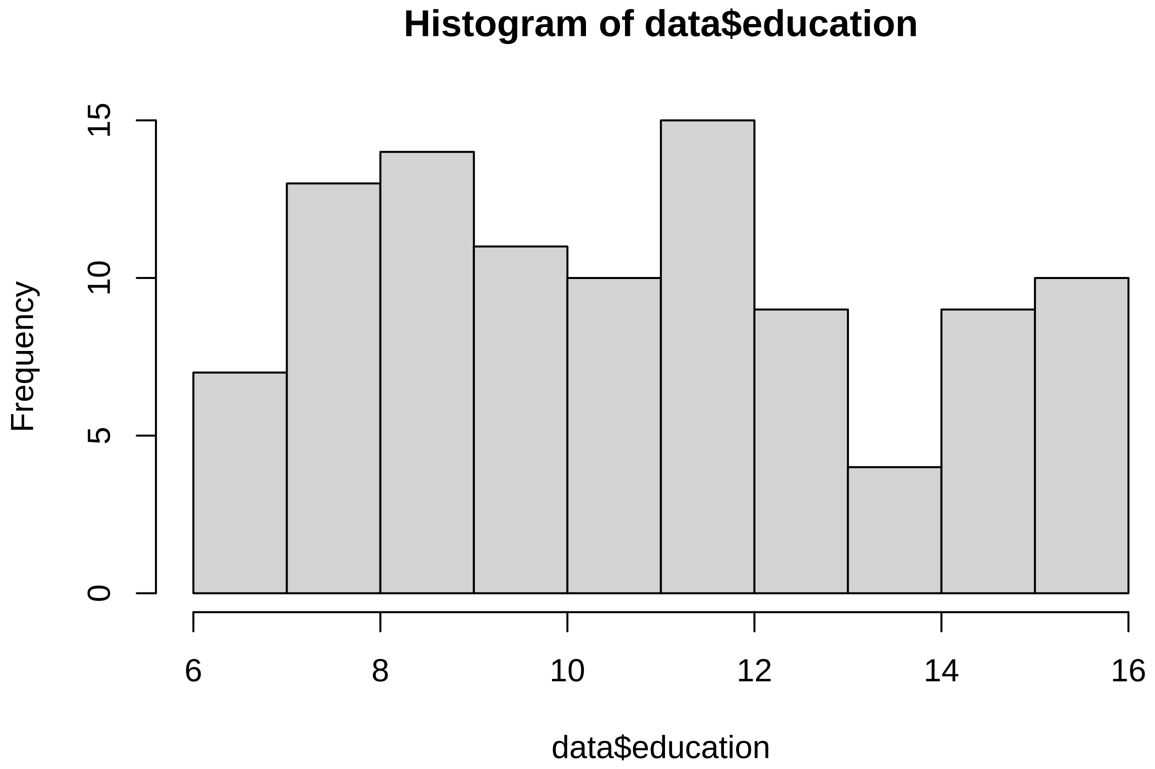
```
table(is.na(data$census)) #checking null values
```

```
##
## FALSE
##   102
```

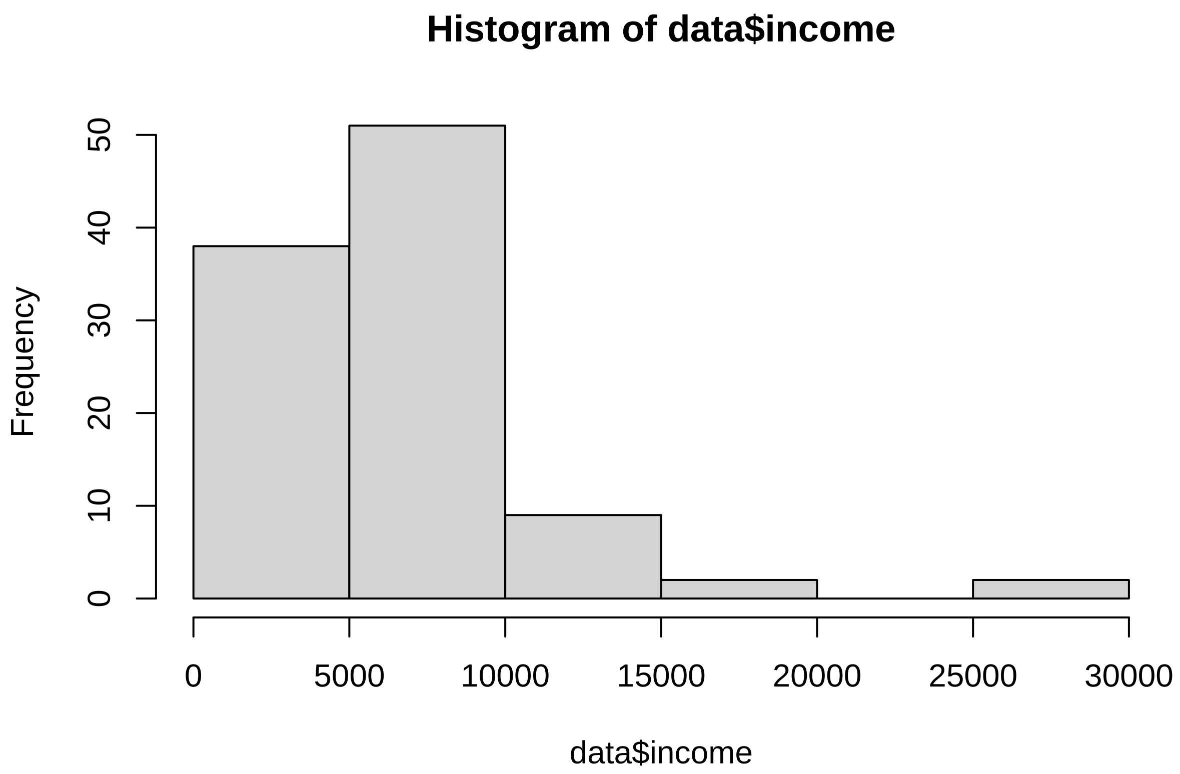
```
table(is.na(data$type)) #checking null values
```

```
##
## FALSE  TRUE
##    98     4
```

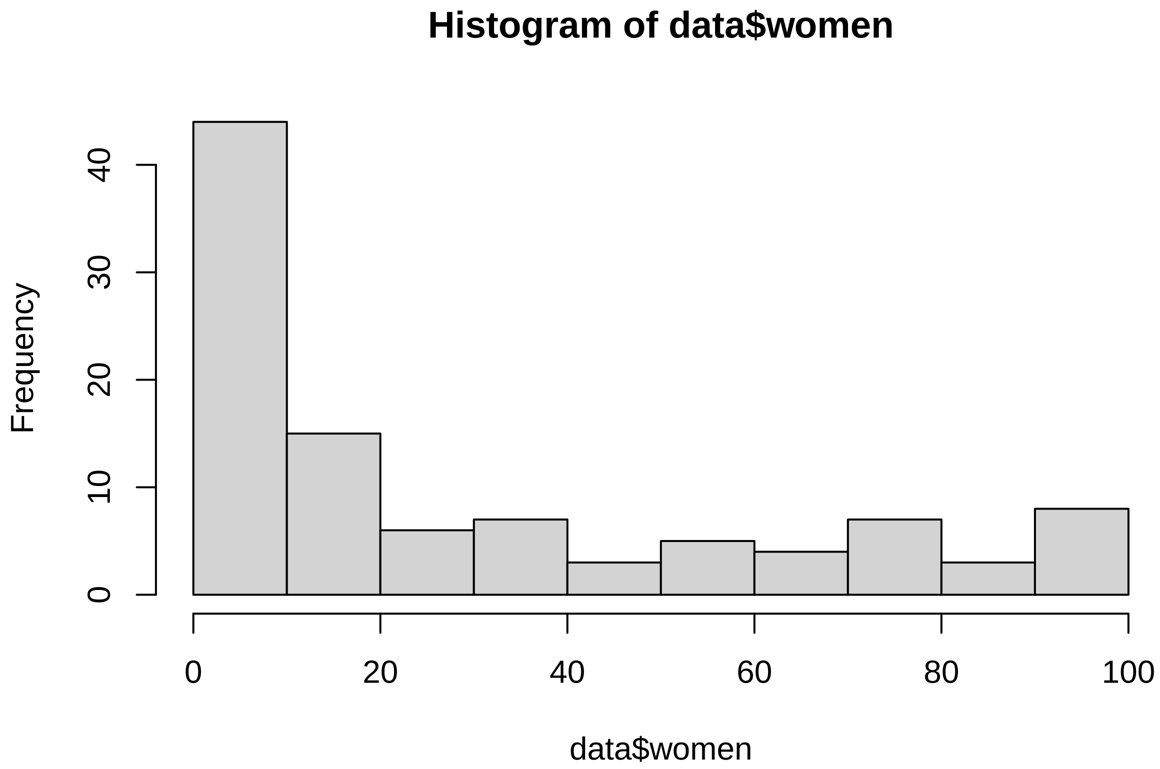
```
hist(data$education)#hist of the data
```



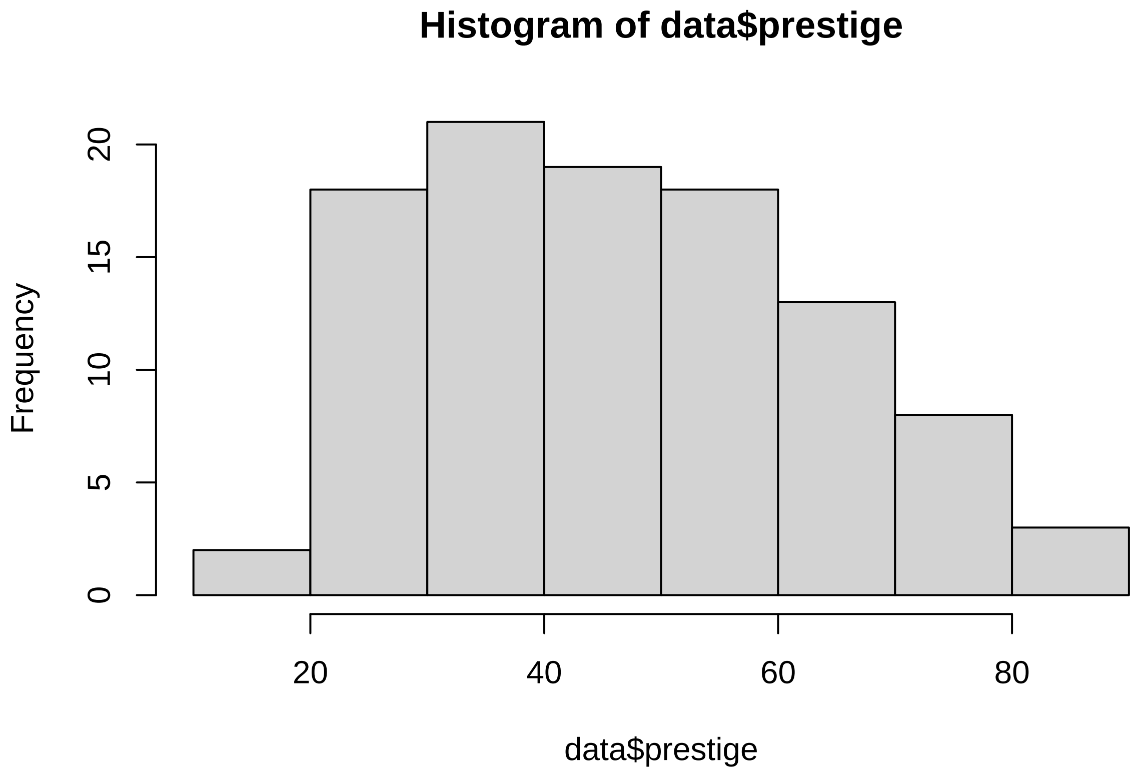
```
hist(data$income)#hist of the data
```



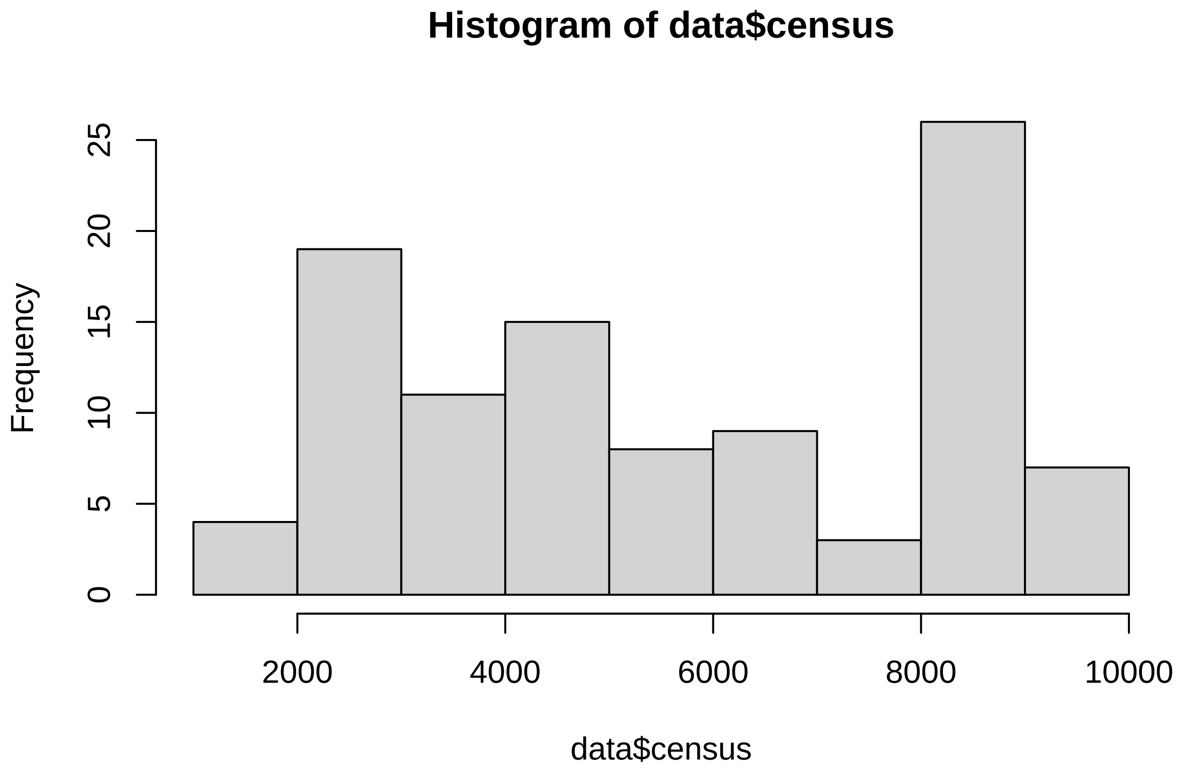
```
hist(data$women)#hist of the data
```



```
hist(data$prestige)#hist of the data
```

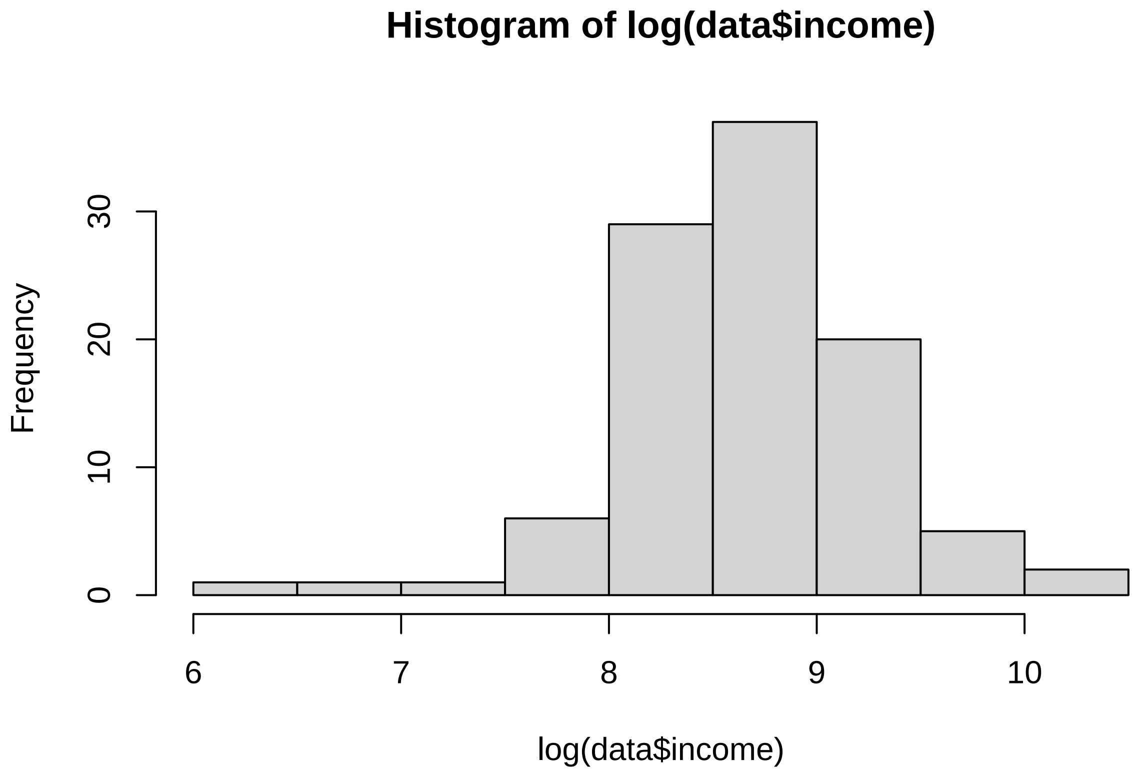


```
hist(data$census)#hist of the data
```



* Looking at the data, only type has 4 values missing every other variable has no null values. * Only income and women looks a bit skewed to the right, maybe giving it a log transformation will make it less skewed. * Income is way more skewed than women, so we will do a transformation to income.

```
library(base)
hist(log(data$income)) #histogram of the log of  income
```



```
data$income <- log(data$income) #converting all values to log values
cor(x = data$prestige, y = data[,1:5]) #correlation of the data
```

```
##      education    income      women prestige     census
## [1,] 0.8501769 0.7410561 -0.1183342        1 -0.6345103
```

```
new_data <- subset(data, select = - c(women, type)) #subsetting a new dataframe
str(new_data) #structure of the new data frame
```

```
## 'data.frame':    102 obs. of  4 variables:
##  $ education: num  13.1 12.3 12.8 11.4 14.6 ...
##  $ income   : num  9.42 10.16 9.13 9.09 9.04 ...
##  $ prestige : num  68.8 69.1 63.4 56.8 73.5 77.6 72.6 78.1 73.1 68.8 ...
##  $ census   : int  1113 1130 1171 1175 2111 2113 2133 2141 2143 2153 ...
```

```
cor(new_data) #correlation of the 2 by 2 matrix
```

```
##            education     income   prestige     census
## education  1.0000000  0.5481051  0.8501769 -0.8230882
## income     0.5481051  1.0000000  0.7410561 -0.3048716
## prestige   0.8501769  0.7410561  1.0000000 -0.6345103
## census    -0.8230882 -0.3048716 -0.6345103  1.0000000
```

- From the correlation of variables between prestige and other variables it looks like education, income and census all correlate well with prestige. Women can be dropped because the variable dosen't correlate well with prestige.
- Education might be the best predictor.

Chapter 8, Exercise 3

Run a multiple regression analysis on the Prestige data with lm(), using prestige as the dependent variable and education and women as the predictors. (1 pt) Check and interpret the diagnostics. (1 pt) Say whether or not the overall R-squared was significant. If it was significant, report the value and say in your own words whether it seems like a strong result or not. (1 pt) Review the significance tests on the coefficients (B-weights). For each one that was significant, report its value and say in your own words what it means and whether it seems like a strong result or not. (1 pt)

```
library(lm.beta)
result <- lm(prestige~education + women, data = data) #linear regreesion
summary(result) #summary of the results
```

```
##
## Call:
## lm(formula = prestige ~ education + women, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.010  -4.069   1.050   5.027  18.942
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.75416    3.54213  -2.471 0.015164 *
## education    5.42780    0.31612  17.170  < 2e-16 ***
## women       -0.09305    0.02719  -3.422 0.000904 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.652 on 99 degrees of freedom
## Multiple R-squared:  0.7521, Adjusted R-squared:  0.7471
## F-statistic: 150.2 on 2 and 99 DF,  p-value: < 2.2e-16
```

```
lm.beta(result) #b-weights
```

```
##
## Call:
## lm(formula = prestige ~ education + women, data = data)
##
## Standardized Coefficients::
## (Intercept)   education       women
##   0.0000000   0.8607894  -0.1715765
```

- The Median is around 1 and the residulas appear to be distributed without any skewness. If the median was more closer to 0, it suggests there is no skewneess in the data.

- The y-intercept is about -8.75 and both the ceofficents for education and women are 5.427 and -0.093 respectively. Each of the values including the intercept all have values below 0.05 showing that they are all significatnt, and that we should reject the null hypothesis.
- The t-value shows the Student's t-test of the null hypothesis that each estimated coefficient is equal to zero.
- THe r-squared is 0.7521 and the adjusted r-squared being 0.74 which is a very high indication that the variable prestige can be well predicted by women and education.
- The f-statistic is 150.2 with df of 2 and 99. THe p-value associated with it is 2.2e-16 which is less than the alpha so we reject that the null hypothesis that R-squared is equal to 0.
- Looking at the b weights, education has the highest with 5.427 indicating it is the most impact predictor in predicting the variable prestige compared to women.

Chapter 8, Exercise 4

Using the results of the analysis from Exercise 2, construct a prediction equation for prestige using all three of the coefficients from the analysis (the intercept along with the two B-weights). (1 pt) If you observed an occupation with scores for education of 8 and women of 35, what would you predict the prestige of the occupation to be (you can use your equation or the predict function)? Show your calculation and the resulting value of prestige (1 pt)

```
pred <- data.frame(education = 8, women = 35) #convertnig it into a dataframe
predict(result, pred) #using prediction function
```

```
##        1
## 31.41159
```

```
#Prediction equation : y = education*5.42 + women* -0.093 -8.75
```

- The prediction equation is y = education5.42 - women0.093 -8.75
- We can calculate the prestige value by inputing in all the numbers or by using the predict function
- We get a value of 31.411 Prestige.

Chapter 8, Exercise 5

Run a multiple regression analysis on the Prestige data with lmBF(), using prestige as the dependent variable and education and women as the predictors. (1 pt) Interpret the resulting Bayes factor in terms of the odds in favor of the alternative hypothesis (be sure to note the hypotheses being compared). Do these results strengthen or weaken your conclusion to exercise 2? (1 pt)

```
library(BayesFactor)
```

```
## Loading required package: coda

## Loading required package: Matrix

## ************
## Welcome to BayesFactor 0.9.12-4.2. If you have questions, please contact Richard Morey (richarddmorey
##
## Type BFManual() to open the manual.
## ************
```

```
bf <- lmBF(prestige~education + women, data = data) #linear regresion model
summary(bf) #summary of the model
```

```
## Bayes factor analysis
## --------------
## [1] education + women : 2.564849e+27 ±0%
##
## Against denominator:
##   Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

- This shows that the odds are overwhelmingly in favor of the alternative hypothesis, in the sense that the model containing education and women as predictors is hugely favored over a model means that only contains the y-intercept.
- Calculating overall mean-rsquared value we get 0.739 which indicates the model is a good predictor.
- This strengthens the conclusion to exercise 2.

Chapter 8, Exercise 6

Run lmBF() with the same model as for Exercise 4, but with the options posterior=TRUE and iterations=10000. (1 pt) Interpret the resulting information about the coefficients. (1 pt)

```
bfp <- lmBF(prestige~education + women, data = data, posterior = TRUE, iterations = 10000)
#lmBF function
summary(bfp) #summary of the results
```

```
##
## Iterations = 1:10000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##                 Mean      SD Naive SE Time-series SE
## mu        46.83983  0.8660 0.008660       0.008345
## education  5.37568  0.3254 0.003254       0.003198
## women     -0.09193  0.0273 0.000273       0.000273
## sig2      77.23615 11.4144 0.114144       0.118061
## g          2.74403  8.4152 0.084152       0.084152
##
## 2. Quantiles for each variable:
##
##              2.5%     25%      50%      75%     97.5%
## mu        45.1561 46.2624 46.84434 47.41711  48.53836
## education  4.7283  5.1598  5.37560  5.59019   6.01631
## women     -0.1453 -0.1105 -0.09222 -0.07375  -0.03805
## sig2      58.4596 69.1889 76.17867 83.99225 102.19500
## g          0.2955  0.7043  1.25042  2.51381  13.68491
```

- Calculating overall mean-rsquared value we get 0.739 which indicates the model is a good predictor.
- Sig2 is a indication of model precision, the smaller the sig2 the better the quality of prediction.
- From the mean of education adn women, it seems that education is a better indicator of prestige than women.
- All the HDIS for the variables education and women, are 4.73 and 5.994 and -0.14 and -0.03 respectively neither of these HDIS include 0, indicating they are statistically significant

Chapter 8, Exercise 7

Run install.packages() and library() for the "car" package. The car package is "companion to applied regression" rather than more data about automobiles. Read the help file for the vif() procedure and then look up more information online about how to interpret the results. Then write down in your own words a "rule of thumb" for interpreting vif. (1 pt)

```
library(regclass)#library to load VIF
```

```
## Loading required package: bestglm

## Loading required package: leaps

## Loading required package: VGAM

## Loading required package: stats4

## Loading required package: splines

##
## Attaching package: 'VGAM'

## The following object is masked from 'package:coda':
##
##     nvar

## Loading required package: rpart

## Loading required package: randomForest

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

## Important regclass change from 1.3:
## All functions that had a . in the name now have an _
## all.correlations -> all_correlations, cor.demo -> cor_demo, etc.
```

```
?VIF()
```

```
## starting httpd help server ...
##  done
```

- Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables.
- A high VIF indicates that the associated independent variable is highly collinear with the other variables in the model.
- Values of VIF that exceed 10 are often regarded as indicating multicollinearity.

Chapter 8, Exercise 8

Run vif() on the results of the model from Exercise 3. (1 pt) Interpret the results. Then run a model that predicts prestige from all of the predictors in Prestige. Run vif() on those results and interpret what you find. (1 pt) Construct and report on a final model that addresses any problems you found. (1 pt)

```
library(regclass)
VIF(result) #VIF results of exercise 3, first model
```

```
## education     women
##   1.00384   1.00384
```

```
print("**************************************")
```

```
## [1] "**************************************"
```

```
final_model <- lm(prestige~education+women+income+census, data = data)
VIF(final_model)#VIF results of all the rpedicotrs
```

```
## education     women    income    census
##  4.613998  1.847239  2.625469  3.502688
```

```
summary(final_model) #summary of the model lwith all predictors
```

```
##
## Call:
## lm(formula = prestige ~ education + women + income + census,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.2039  -4.4726  -0.0698   4.4158  19.6516
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.139e+02  1.573e+01  -7.244 1.04e-10 ***
## education    3.983e+00  5.575e-01   7.144 1.68e-10 ***
## women        4.954e-02  3.034e-02   1.633    0.106
## income       1.328e+01  1.940e+00   6.844 6.99e-10 ***
## census       2.943e-04  5.010e-04   0.587    0.558
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.116 on 97 degrees of freedom
## Multiple R-squared:  0.8357, Adjusted R-squared:  0.8289
## F-statistic: 123.3 on 4 and 97 DF,  p-value: < 2.2e-16
```

- VIF is an in indication of multi-colinearity among variables.
- For the model from excercise 3, all the VIF scores for education and women are 1.003 and 1.003 low, indicating low multicolinearity.
- For the final model with all the variables, the VIF scores for education, women, income and cesus,are as follows: 4.613, 1.84, 2.62 and 3.502 all the values are lower than 10 indicating low multicomlinearity.

```
final_model1 <- lm(prestige~education+income, data = data)#creating final model
summary(final_model1) #summary of final model
```

```
##
## Call:
## lm(formula = prestige ~ education + income, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.0346  -4.5657  -0.1857   4.0577  18.1270
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -95.1940    10.9979  -8.656 9.27e-14 ***
## education     4.0020     0.3115  12.846  < 2e-16 ***
## income       11.4375     1.4371   7.959 2.94e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.145 on 99 degrees of freedom
## Multiple R-squared:  0.831,  Adjusted R-squared:  0.8275
## F-statistic: 243.3 on 2 and 99 DF,  p-value: < 2.2e-16
```

- For the final model, we dropped women and census as they two variables previously had p-values that were 0.106 for women and 0.558 which were above 0.05 indicating that the two variables might not be statistically signicant.
- The p-value of both the intercept and the two variables education and income, were 2e-16 and 2.94e-12 which are very low p-values indicating that we can reject the null hypothesis, that each estimated coefficient is equal to zero.
- So we drop the two variables and run the final regression model.
- The r-squared value for the model is 0.82 showing that it is a very good model that predicts prestige well.
- The f-statistic of 243.3 on df of 2 and 99 show that the p-value is 2.2e-16 which is very low, showing that we can reject the null hypothesis that R-squared is equal to 0.
- This final model is good as it captures all the necesscary variables and predicts prestige well.