Lab 5

Keara Dreyfuss

3/26/2024

Directions: Workout the problems using R markdown. Hand in both the *.rmd file and the knitted *.pdf file. (3 points for correctly submitting)

1. Job Changes Data with Dummy Variables (2 points each part) Using the `JobChanges.csv` file, answer the following questions.

a. Fit a simple linear regression model to predict the annual salary in thousands of USD (`Salary`) as a function of the number of job changes (`Jobs`). Print a summary of the model.

```
df = read.csv("JobChanges.csv")
reg = lm(Salary ~ Jobs, data = df)
summary(reg)
```

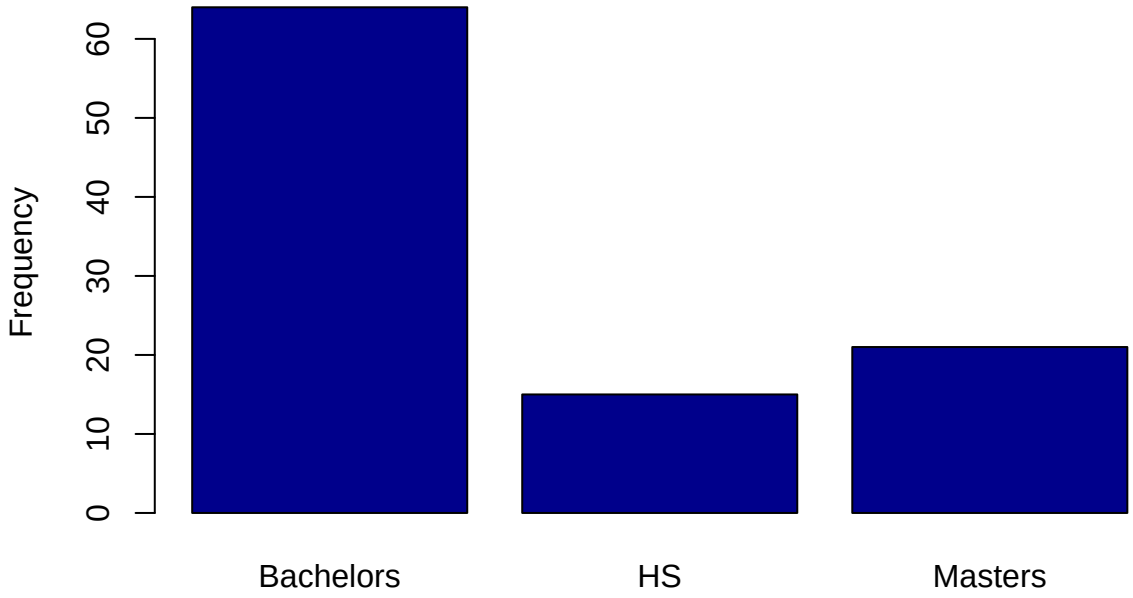
```
##
## Call:
## lm(formula = Salary ~ Jobs, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.395 -13.509  -3.850   7.627  60.151
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   51.124      7.425   6.885 5.54e-10 ***
## Jobs           5.727      1.019   5.621 1.79e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.81 on 98 degrees of freedom
## Multiple R-squared:  0.2438, Adjusted R-squared:  0.2361
## F-statistic: 31.59 on 1 and 98 DF,  p-value: 1.788e-07
```

b. What percent of the variation in `Salary` can be explained by the model in the previous part?

24.38% of Variation in Salary can be explained by the model.

c. Create a barplot of the 3 education levels in the `Education` variable by passing a `summary` of the Education factor into the `barplot` function.

```r
barplot(summary(factor(df$Education)), ylab = "Frequency", col = "darkblue")
```



d. Fit a multiple regression model to predict the `Salary` using `Jobs` and 2 dummy variables representing education levels as predictor variables. Print a summary of the model.

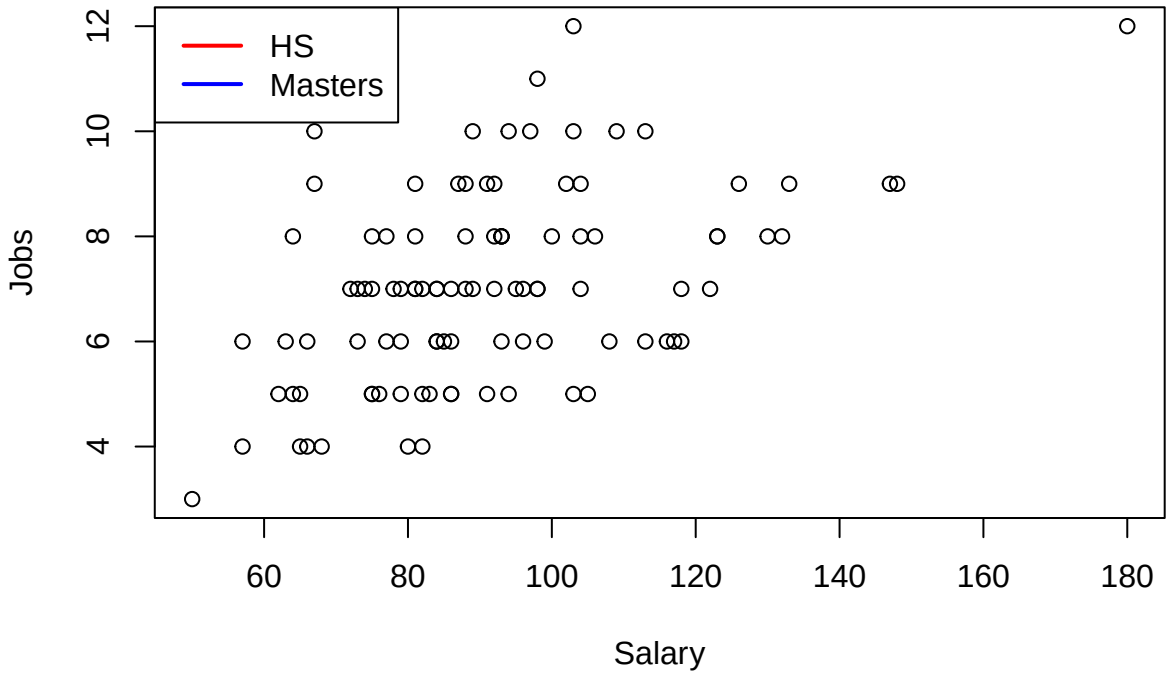
```r
reg2 = lm(Salary ~ Education + Jobs, data = df)
summary(reg2)
```

```
##
## Call:
## lm(formula = Salary ~ Education + Jobs, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.931  -5.423   0.965   6.330  29.638
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        42.6046     4.4375   9.601 1.07e-15 ***
## EducationHS       -11.4896     3.2606  -3.524 0.000654 ***
## EducationMasters   33.6116     2.8001  12.004  < 2e-16 ***
## Jobs                6.1788     0.6158  10.035  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.12 on 96 degrees of freedom
## Multiple R-squared:  0.7409, Adjusted R-squared:  0.7328
## F-statistic: 91.51 on 3 and 96 DF,  p-value: < 2.2e-16
```

e. Create a scatterplot of `Jobs` versus `Salary`. Overlay the scatterplot with the regression line for the `HS` level as given in the previous model. Also, overlay the scatterplot with the regression lines for the other 2 levels as given by the previous model.

```r
plot(df$Salary, df$Jobs, xlab = "Salary", ylab = "Jobs")
abline(a = 42.6046, b = 6.1788, col = "hotpink", lwd = 2) # HS level
abline(a = 76.2162, b = 6.1788, col = "skyblue", lwd = 2) # Master's Degree level
```

```
legend('topleft', c('HS', 'Masters'), col = c("red", "blue"), lwd = 2)
```



f. How should the dummy variables be interpreted?

People with nasters degrees will earn $33,611 more than people with only HS degrees

2. Job Changes Data with Interactions (2 points each part) Using the `JobChanges.csv` file, answer the following questions.

a. Fit a multiple regression model on the data set predicting `Salary` as a function of `Jobs` and `Education`. Include an interaction between the predictor variables. Print a summary of the model.

```
reg3 = lm(Salary ~ Jobs + Education + Jobs:Education, data = df)
summary(reg3)
```

```
##
## Call:
## lm(formula = Salary ~ Jobs + Education + Jobs:Education, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.664  -6.178   1.182   7.555  18.329
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            50.7179     5.1213   9.903 2.95e-16 ***
## Jobs                    4.9933     0.7256   6.881 6.55e-10 ***
## EducationHS            -0.6774    11.6826  -0.058    0.954
## EducationMasters       -9.3977     9.8717  -0.952    0.344
## Jobs:EducationHS       -1.2001     1.4738  -0.814    0.418
## Jobs:EducationMasters   6.1370     1.3654   4.495 1.99e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

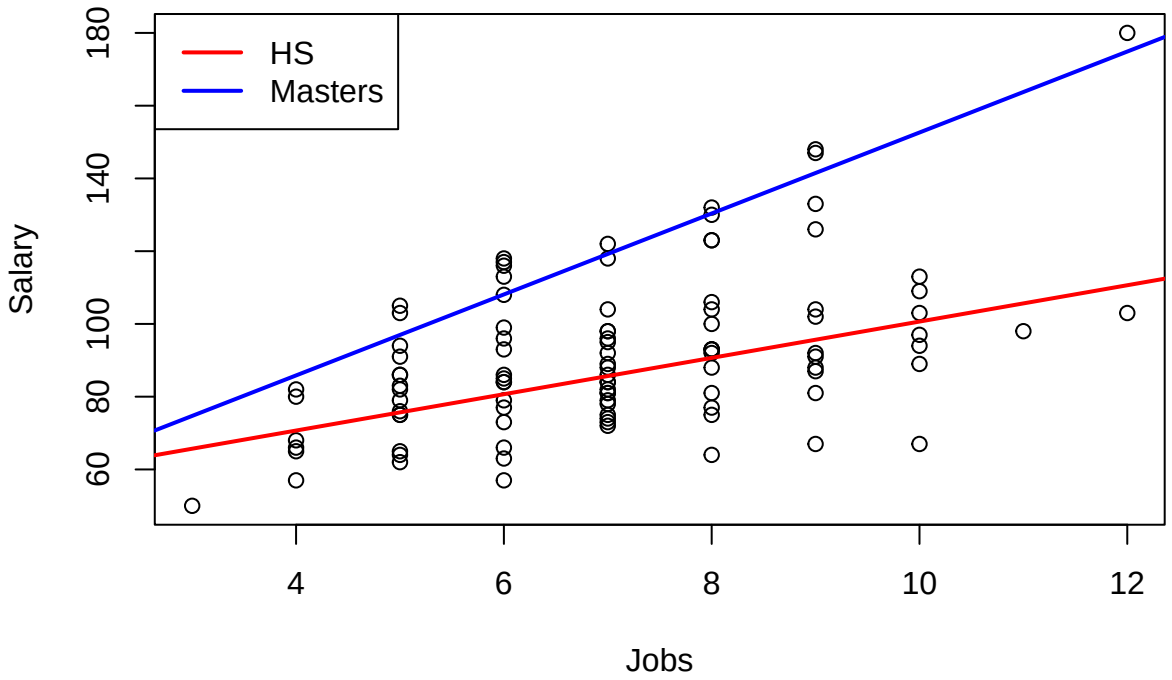
```
## Residual standard error: 10.01 on 94 degrees of freedom
## Multiple R-squared:  0.7944, Adjusted R-squared:  0.7835
## F-statistic: 72.64 on 5 and 94 DF,  p-value: < 2.2e-16
```

b. Interpret the interaction coefficients from the previous model.

For each job, the salary will be 1,200 less if the education is high school, and 6.137 more if the education is Masters.

c. Create a scatterplot of `Jobs` versus `Salary`. Overlay the scatterplot with the regression line for the `HS` level as given in the previous model. Also, overlay the scatterplot with the regression lines for the remaining levels as given by the previous model.

```
plot(Salary ~ Jobs, data = df)
abline(a = 50.7179, b = 4.9933, col = "red", lwd = 2)
abline(a = 50.7179 - 9.3977, b = 4.9933 + 6.1370, col = "blue", lwd = 2)
legend('topleft', c('HS', 'Masters'), col = c("red", "blue"), lwd = 2)
```



d. Using the multiple regression model with the interaction from part a, specify the simple linear regression equation that predicts `Salary` using `Jobs` as a predictor for high school graduates (`HS`). Also, specify the same simple linear regressions for college graduates (`bachelors`) and people with a graduate degree (`masters`).

HS graduates: Salary = (50.7179 + 4.9933 Jobs)

Bachelors: Salary = (50.7179 + 4.9933 * Jobs+ (-1.2001) *Jobs)

Masters: Salary = (50.7179 + 4.9933 * Jobs + 6.1370 * Jobs)*

e. Create and add a variable called `JobsM` which mean-centers the Jobs variable.

```
df$Jobs_MC = df$Jobs - mean(df$Jobs)
summary(df$Jobs_MC)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -4.05   -1.05   -0.05    0.00    0.95    4.95
```

f. Fit a multiple regression model which uses the following predictors:

- a dummy variable for observations at the `bachelors` level,
- a dummy variable for observations at the `masters` level,
- the mean-centered Jobs, and
- interactions between each dummy variable and the mean-centered Jobs.

```
df$Education = factor(df$Education, levels = c("Bachelors", "Masters", "HS"))
reg4 = lm(Salary ~  Jobs_MC + Education + Jobs_MC:Education, data = df)
summary(reg4)
```

```
##
## Call:
## lm(formula = Salary ~ Jobs_MC + Education + Jobs_MC:Education,
##     data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.664  -6.178   1.182   7.555  18.329
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 85.9205     1.2606  68.157  < 2e-16 ***
## Jobs_MC                      4.9933     0.7256   6.881 6.55e-10 ***
## EducationMasters            33.8679     2.5227  13.425  < 2e-16 ***
## EducationHS                 -9.1378     3.0916  -2.956  0.00394 **
## Jobs_MC:EducationMasters     6.1370     1.3654   4.495 1.99e-05 ***
## Jobs_MC:EducationHS         -1.2001     1.4738  -0.814  0.41756
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.01 on 94 degrees of freedom
## Multiple R-squared:  0.7944, Adjusted R-squared:  0.7835
## F-statistic: 72.64 on 5 and 94 DF,  p-value: < 2.2e-16
```

3. Carseat Sales with Categorical Variables (2 points each part)

Here we will use the car seats data to solve the questions below. From the `ISLR` package, load in the `Carseats` data by running the code chunk below.

```
library(ISLR)
data(Carseats)
```

a. Get a summary of the data using the `summary` function.

```
summary(Carseats)
```

```
##      Sales          CompPrice       Income        Advertising
##  Min.   : 0.000   Min.   : 77   Min.   : 21.00   Min.   : 0.000
##  1st Qu.: 5.390   1st Qu.:115   1st Qu.: 42.75   1st Qu.: 0.000
##  Median : 7.490   Median :125   Median : 69.00   Median : 5.000
##  Mean   : 7.496   Mean   :125   Mean   : 68.66   Mean   : 6.635
##  3rd Qu.: 9.320   3rd Qu.:135   3rd Qu.: 91.00   3rd Qu.:12.000
##  Max.   :16.270   Max.   :175   Max.   :120.00   Max.   :29.000
##    Population        Price        ShelveLoc        Age          Education
##  Min.   : 10.0   Min.   : 24.0   Bad   : 96   Min.   :25.00   Min.   :10.0
##  1st Qu.:139.0   1st Qu.:100.0   Good  : 85   1st Qu.:39.75   1st Qu.:12.0
##  Median :272.0   Median :117.0   Medium:219   Median :54.50   Median :14.0
```

```
##  Mean   :264.8   Mean   :115.8                Mean   :53.32   Mean   :13.9
##  3rd Qu.:398.5   3rd Qu.:131.0                3rd Qu.:66.00   3rd Qu.:16.0
##  Max.   :509.0   Max.   :191.0                Max.   :80.00   Max.   :18.0
##  Urban       US
##  No :118   No :142
##  Yes:282   Yes:258
##
##
##
##
```

b. List the variables that are numeric and the variables that are categorical.

Numerical Variables are Sales, Comprice, Income, Adervistiing, Population, Price, Age, and Education. Categorical variables are ShelveLoc, Urban, and US.

c. Generate a linear model to predict the response variable (`Sales`) from all of the other variables in the data set.

```
reg5 = lm(Sales ~ ., data = Carseats)
summary(reg5)
```

```
##
## Call:
## lm(formula = Sales ~ ., data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8692 -0.6908  0.0211  0.6636  3.4115
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.6606231  0.6034487   9.380  < 2e-16 ***
## CompPrice        0.0928153  0.0041477  22.378  < 2e-16 ***
## Income           0.0158028  0.0018451   8.565 2.58e-16 ***
## Advertising      0.1230951  0.0111237  11.066  < 2e-16 ***
## Population       0.0002079  0.0003705   0.561    0.575
## Price           -0.0953579  0.0026711 -35.700  < 2e-16 ***
## ShelveLocGood    4.8501827  0.1531100  31.678  < 2e-16 ***
## ShelveLocMedium  1.9567148  0.1261056  15.516  < 2e-16 ***
## Age             -0.0460452  0.0031817 -14.472  < 2e-16 ***
## Education       -0.0211018  0.0197205  -1.070    0.285
## UrbanYes         0.1228864  0.1129761   1.088    0.277
## USYes           -0.1840928  0.1498423  -1.229    0.220
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.019 on 388 degrees of freedom
## Multiple R-squared:  0.8734, Adjusted R-squared:  0.8698
## F-statistic: 243.4 on 11 and 388 DF,  p-value: < 2.2e-16
```

d. Display a summary of the model given from the previous part.

```
summary(reg5)
```

```
##
## Call:
```

```
## lm(formula = Sales ~ ., data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8692 -0.6908  0.0211  0.6636  3.4115
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.6606231  0.6034487   9.380  < 2e-16 ***
## CompPrice        0.0928153  0.0041477  22.378  < 2e-16 ***
## Income           0.0158028  0.0018451   8.565 2.58e-16 ***
## Advertising      0.1230951  0.0111237  11.066  < 2e-16 ***
## Population       0.0002079  0.0003705   0.561    0.575
## Price           -0.0953579  0.0026711 -35.700  < 2e-16 ***
## ShelveLocGood    4.8501827  0.1531100  31.678  < 2e-16 ***
## ShelveLocMedium  1.9567148  0.1261056  15.516  < 2e-16 ***
## Age             -0.0460452  0.0031817 -14.472  < 2e-16 ***
## Education       -0.0211018  0.0197205  -1.070    0.285
## UrbanYes         0.1228864  0.1129761   1.088    0.277
## USYes           -0.1840928  0.1498423  -1.229    0.220
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.019 on 388 degrees of freedom
## Multiple R-squared:  0.8734, Adjusted R-squared:  0.8698
## F-statistic: 243.4 on 11 and 388 DF,  p-value: < 2.2e-16
```