Homework 1: Linear Regression

There are six questions (30 total points) in this assignment. The minimum increment is 1 point. Please type in your answers directly in the R Markdown file. After completion, **successfully** knitr it as an html file. Submit **both** the html file and the R Markdown file via Canvas. Please name the R Markdown file in the following format: LastName_FirstName_HW1.Rmd, e.g. Zhao_Zifeng_HW1.Rmd.

Used Car Dataset [9 points]

The used car dataset is the one we analyzed in class. Let's read in the data stored in `UsedCar.csv`.

```
total_data <- read.csv("/Users/TomTheIntern/Desktop/Mendoza/Data Folder/UsedCar.csv", header=T, strings
```

Q1 [2 points]

Since the dependent variable `Price` is severely right-skewed, create a log-transformation of `Price` and store it as a new variable named `Log_price` within the data.frame `total_data`.

```
total_data$Log_price <- log(total_data$Price)
```

Q2 [7 points]

Fit a linear regression model of Log_price w.r.t. two predictors Age and Mileage, name it `lm_short`.

```
lm_short <- lm(Log_price ~ Age + Mileage , data = total_data)
```

Q2(a) [2 points] What is the R^2 of `lm_short`? What are the (in-sample) MAE and RMSE of `lm_short` at the *original scale*?

```
r_2 <- summary(lm_short)$r.squared

library(forecast)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
accuracy(exp(lm_short$fitted.values) - 1, total_data$Price)
```

```
##                ME     RMSE      MAE        MPE     MAPE
## Test set 102.3341 1526.271 1065.202 -0.8676442 9.979393
```

```r
MAE <- 1065.202
RMSE <- 1526.271
```

Answer:

The R^2 value of lm_short is .8054654.

For the original scale:

The Mean Absolute Error is 1065.202

The Root Mean Square Deviation is 1526.271

Q2(b) [**2 point**] What is the estimated coefficient of `lm_short` for Age and Mileage?

```r
lm_short <- lm(Log_price ~ Age + Mileage , data = total_data)

lm_short$coefficients
```

```
##   (Intercept)           Age       Mileage
##  1.003505e+01 -1.217033e-02 -1.757577e-06
```

The coefficient for Age is -1.217033e-02 and is -1.757577e-06 for Mileage

Q2(c) [**3 points**] How should we interpret the estimated coefficient of Age at the log-scale and the original scale of Price?

Answer:

The change in the log price represents a percentage change for the change of price. Meanwhile, the change at the original price scale represents a change in the actual price.

Because both coefficients are negative, we know that as the mileage and age of a vehicle increase, the log of its price should decrease. An older car with more miles should have a lower log price than a newer car with fewer miles.

For one unit increase of Age, the log of price should go down -1.217033e-02. For one unit increase of Mileage, the log of price should go down -1.757577e-06

Car Seat Sales Dataset [21 points]

The car seat sales dataset contains sales of child car seats at 400 different stores and the data is stored in `Carseats.csv`. It contains 9 variables, `Sales`, `CompPrice`, `Income`, `Advertising`, `Population`, `Price`, `ShelveLoc`, `Age` and `Urban`. We would like to build a linear regression model to predict `Sales` at a planned new store. The data description is as follows.

- `Sales`: Unit sales (in thousands) at each location
- `CompPrice`: Price charged by competitor at each location
- `Income`: Community income level (in thousands of dollars)
- `Advertising`: Local advertising budget for company at each location (in thousands of dollars)
- `Population`: Population size in region (in thousands)
- `Price`: Price company charges for car seats at each site
- `ShelveLoc`: A factor with levels Bad, Good and Medium indicating the quality of the shelving location for the car seats at each site
- `Age`: Average age of the local population
- `Urban`: A factor with levels No and Yes to indicate whether the store is in an urban or rural location

Q4 [2 points] Which variable is the dependent variable? Which predictors are categorical variables?

Answer:

The Dependent variable is sales, as we are trying to determine how well the new store will perform.

The categorical predictors are ShelveLoc, which is comprised of three values and Urban, which is comprised of two. The remaining variables are all numeric.

Q5 [9 points] Let's read in the data and perform visualization to get a better sense of the data.

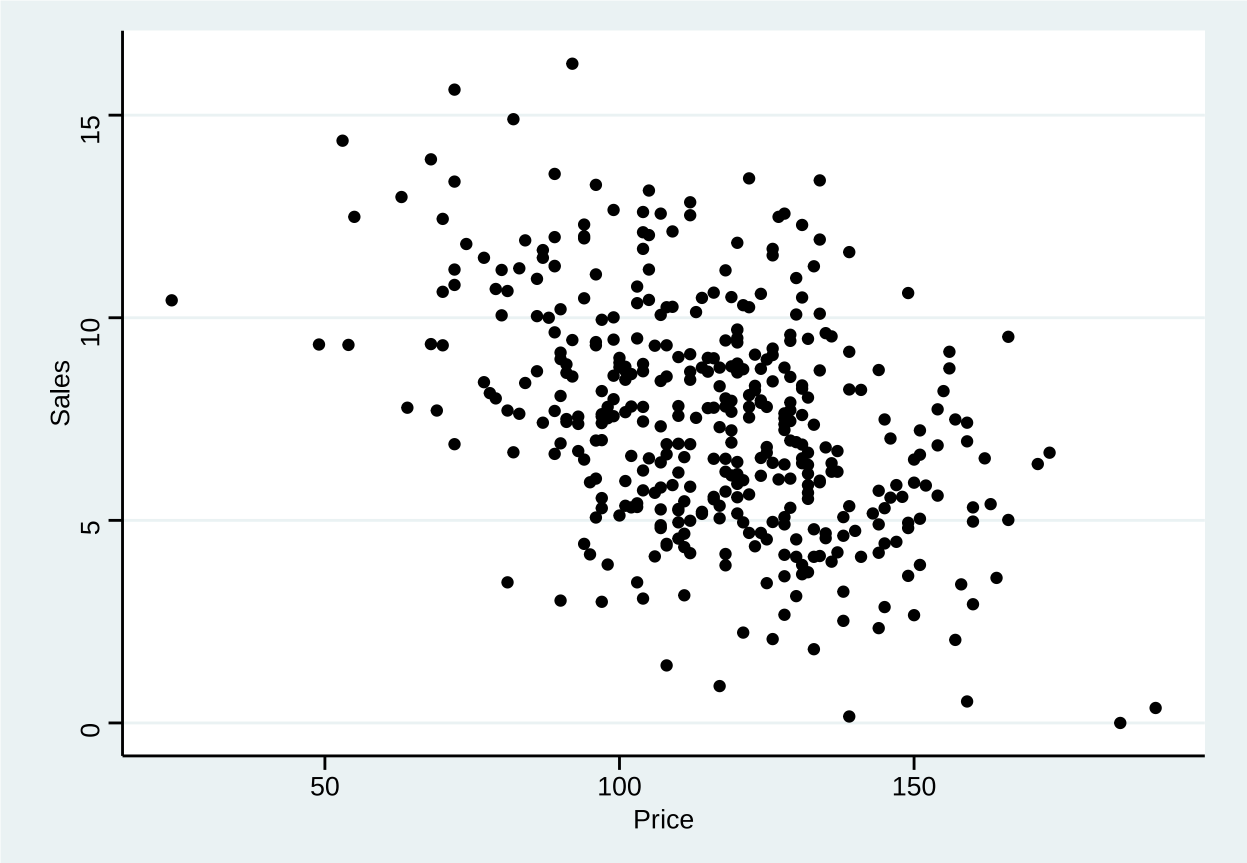
Q5(a) [2 points] Correctly read in the data stored at `Carseats.csv`.

```
Carseats <- read.csv("/Users/TomTheIntern/Desktop/Mendoza/Data Folder/Carseats.csv", header=T, stringsA

str(Carseats)
```

```
## 'data.frame':    400 obs. of  9 variables:
##  $ Sales      : num  9.5 11.22 10.06 7.4 4.15 ...
##  $ CompPrice  : int  138 111 113 117 141 124 115 136 132 132 ...
##  $ Income     : int  73 48 35 100 64 113 105 81 110 113 ...
##  $ Advertising: int  11 16 10 4 3 13 0 15 0 0 ...
##  $ Population : int  276 260 269 466 340 501 45 425 108 131 ...
##  $ Price      : int  120 83 80 97 128 72 108 120 124 124 ...
##  $ ShelveLoc  : Factor w/ 3 levels "Bad","Good","Medium": 1 2 3 3 1 1 3 2 3 3 ...
##  $ Age        : int  42 65 59 55 38 78 71 67 76 76 ...
##  $ Urban      : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 1 2 2 1 1 ...
```

####**Q5(b) [2 points]** Produce a scatterplot between `Sales` and `Price`. What is the general pattern from the scatterplot?

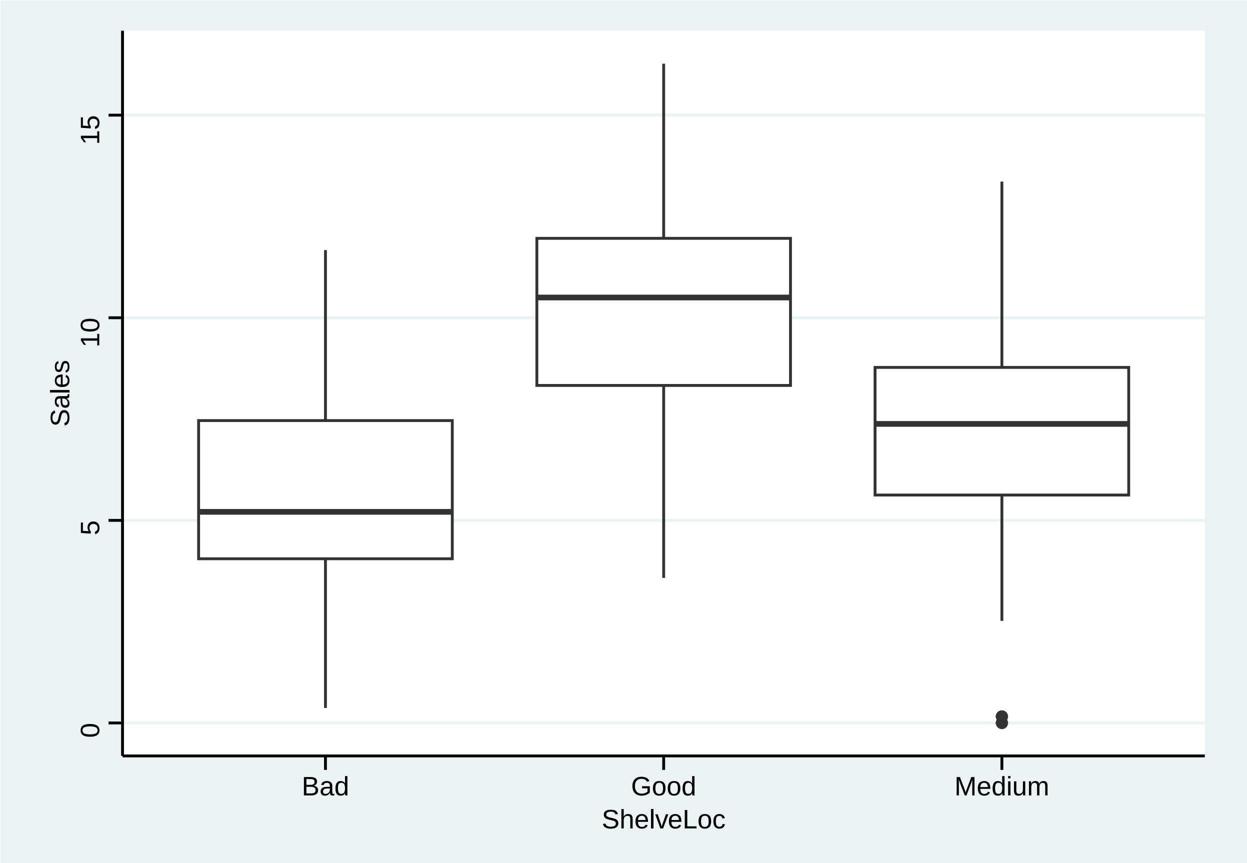
```
library(ggplot2)
library(ggthemes)

ggplot(data = Carseats, aes(Price, Sales)) +
  geom_point()+
  theme_stata()
```



Answer: The general trend for the scatterplot is that as prices increases sales decrease. Conversely, as prices decrease, the sales increase.

Q5(c) [2 points] Produce a boxplot between `Sales` and `ShelveLoc`. What is the general pattern from the boxplot?

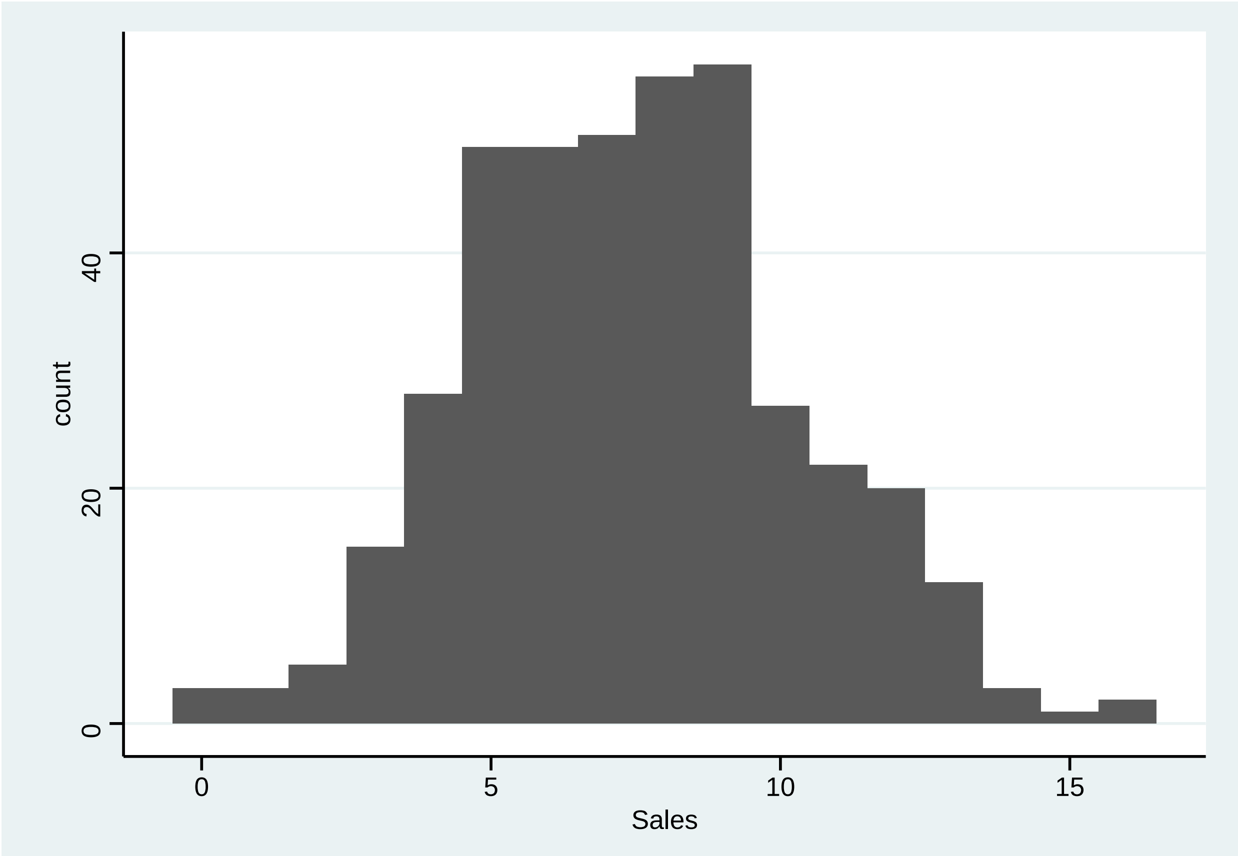
```
ggplot(data = Carseats, aes(ShelveLoc, Sales)) +
  geom_boxplot()+
  theme_stata()
```



Answer: Sales in the Good ShelveLoc category tended to be the highest, while sales in the bad ShelveLoc were the lowest. The sales in the Medium ShelveLoc were split between the two.

Q5(d) [**3 points**] Produce a histogram of `Sales`. Is `Sales` severely right-skewed?

```
ggplot(data = Carseats, aes(Sales)) +
  geom_histogram(binwidth = 1)+
  theme_stata()
```



Answer: Based on the plot I would saw that Sales is somewhat right skewed, but not severely right skewed.

Q6 [10 points]

Q6(a) [4 points] Fit a linear regression model of the original scale `Sales` w.r.t. all the predictors available in the dataset, name it `lm_full`. Take a look at the summary of `lm_full`.

```
lm_full <- lm(Sales ~ CompPrice + Income + Advertising + Population + Price + ShelveLoc + Age + Urban,
              data = Carseats)

summary(lm_full)
```

```
##
## Call:
## lm(formula = Sales ~ CompPrice + Income + Advertising + Population +
##     Price + ShelveLoc + Age + Urban, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7868 -0.6988  0.0160  0.6786  3.2736
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       5.3037420  0.5263440  10.077  < 2e-16 ***
## CompPrice         0.0926753  0.0041490  22.337  < 2e-16 ***
```

```
## Income           0.0157694  0.0018400   8.570 2.44e-16 ***
## Advertising      0.1136057  0.0080241  14.158  < 2e-16 ***
## Population       0.0003255  0.0003627   0.897    0.370
## Price           -0.0954335  0.0026720 -35.716  < 2e-16 ***
## ShelveLocGood    4.8538486  0.1530955  31.705  < 2e-16 ***
## ShelveLocMedium  1.9681961  0.1259228  15.630  < 2e-16 ***
## Age             -0.0461324  0.0031829 -14.494  < 2e-16 ***
## UrbanYes         0.1268438  0.1129471   1.123    0.262
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.02 on 390 degrees of freedom
## Multiple R-squared:  0.8726, Adjusted R-squared:  0.8697
## F-statistic: 296.9 on 9 and 390 DF,  p-value: < 2.2e-16
```

Q6(b) [**2 points**] What is the estimated coefficient of `lm_full` for CompPrice and Price? How should we interpret the estimated coefficients?

```r
lm_full$coefficients
```

```
##     (Intercept)       CompPrice          Income     Advertising      Population
##    5.3037419949    0.0926752906    0.0157693826    0.1136056503    0.0003254507
##           Price   ShelveLocGood ShelveLocMedium             Age        UrbanYes
##   -0.0954334977    4.8538485985    1.9681961372   -0.0461324381    0.1268438370
```

Answer: When a competitors price goes up one unit, Sales for the dealer we are predicting go up. Meanwhile, if the dealer we are predicting see a raise in prices, sales for that dealer go down.

Because we know that CompPrice's coefficent is 0.0926752906 and Price's coefficent is -0.0954334977, we know that for each unit that CompPrice goes up, Sales will go up by 0.0926752906 units.

Q6(c) [**2 points**] Which predictor(s) are not statistically significant in the model?

Answer: Population and UrbanYes

Q6(d) [**2 points**] What is the R^2 of `lm_full`? What are the (in-sample) MAE and RMSE of `lm_full`?

```r
#making a new model for the original scale
lm_full <- lm(Sales ~ CompPrice + Income + Advertising + Population + Price + ShelveLoc + Age + Urban,
              data = Carseats)

summary(lm_full)
```

```
##
## Call:
## lm(formula = Sales ~ CompPrice + Income + Advertising + Population +
##     Price + ShelveLoc + Age + Urban, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7868 -0.6988  0.0160  0.6786  3.2736
##
```

```
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      5.3037420  0.5263440  10.077  < 2e-16 ***
## CompPrice        0.0926753  0.0041490  22.337  < 2e-16 ***
## Income           0.0157694  0.0018400   8.570 2.44e-16 ***
## Advertising      0.1136057  0.0080241  14.158  < 2e-16 ***
## Population       0.0003255  0.0003627   0.897    0.370    
## Price           -0.0954335  0.0026720 -35.716  < 2e-16 ***
## ShelveLocGood    4.8538486  0.1530955  31.705  < 2e-16 ***
## ShelveLocMedium  1.9681961  0.1259228  15.630  < 2e-16 ***
## Age             -0.0461324  0.0031829 -14.494  < 2e-16 ***
## UrbanYes         0.1268438  0.1129471   1.123    0.262    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.02 on 390 degrees of freedom
## Multiple R-squared:  0.8726, Adjusted R-squared:  0.8697 
## F-statistic: 296.9 on 9 and 390 DF,  p-value: < 2.2e-16
```

```r
r_2_two <- summary(lm_full)$r.squared

accuracy(lm_full$fitted.values, Carseats$Sales)
```

```
##                     ME     RMSE      MAE MPE MAPE
## Test set -3.087808e-17 1.006678 0.805898 Inf  Inf
```

```r
MAE <- 0.805898
RMSE <- 1.006678
```

Answer: The R^2 of lm_full is 0.8726193 MAE: 0.805898 RMSE: 1.006678