STOR 455 Homework #3

20 points - Due Thursday 09/21 at 12:30pm

Directions: You may (and should) collaborate with other students. However, you must complete the assignment by yourself. You should complete this assignment in an R Notebook, including all calculations, plots, and explanations. Make use of the white space outside of the R chunks for your explanations rather than using comments inside of the chunks. For your submission, you should knit the notebook to PDF (it is usually smoother first knit to Word then save the file as pdf) and submit the file to Gradescope. The submitted PDF should not be longer than 20 pages.

Situation: Suppose that you are interested in purchasing a used vehicle. How much should you expect to pay? Obviously the price will depend on the type of vehicle that you get (the model) and how much it's been used. For this assignment you will investigate how the price might depend on the vehicle's year and mileage.

Data Source: To get a sample of vehicles, begin with the UsedCars CSV file (posted on Sakai). The data was acquired by scraping TrueCar.com for used vehicle listings on 9/24/2017 and contains more than 1.2 million used vehicles. For this assignment you will choose a vehicle *Model* from a US company for which there are at least 100 of that model listed for sale in North Carolina. Note that whether the companies are US companies or not is not contained within the data. It is up to you to determine which *Make* of vehicles are from US companies. After constructing a subset of the UsedCars data under these conditions, check to make sure that there is a reasonable amount of variability in the years for your vehicle, with a range of at least six years.

Directions: The code below should walk you through the process of selecting data from a particular model vehicle of your choice. Each of the following two R chunks begin with {r, eval=FALSE}. eval=FALSE makes these chunks not run when I knit the file. **Before you knit these chunks, you should revert them to {r}**.

```r
library(readr)

# This line will only run if the UsedCars.csv is stored in the same directory as this notebook!
UsedCars <- read_csv("UsedCars.csv")
```

```
## Rows: 1048575 Columns: 9
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): City, State, Vin, Make, Model
## dbl (4): Id, Price, Year, Mileage
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
StateHW3 = "NC"

# Creates a dataframe with the number of each model for sale in North Carolina
Vehicles = as.data.frame(table(UsedCars$Model[UsedCars$State==StateHW3]))
```

```r
# Renames the variables
names(Vehicles)[1] = "Model"
names(Vehicles)[2] = "Count"

# Restricts the data to only models with at least 100 for sale
# Vehicles from non US companies are contained in this data
# Before submitting, comment this out so that it doesn't print while knitting
Enough_Vehicles = subset(Vehicles, Count>=100)
Enough_Vehicles
```

```
##                Model Count
## 21         200Limited   191
## 34                  3   477
## 74                  5   174
## 130         AcadiaAWD   103
## 131         AcadiaFWD   259
## 139            Accord   776
## 141        AccordEX-L   132
## 149         Altima2.5   779
## 153         Altima4dr   131
## 245       CamaroCoupe   322
## 247          Camry4dr   106
## 251           CamrySE   133
## 284     ChallengerR/T   123
## 309  CherokeeLatitude   108
## 315             Civic   509
## 324           CivicLX   135
## 355      ColoradoCrew   112
## 384            Cooper   237
## 394       Corvette2dr   101
## 405            CR-VEX   127
## 406          CR-VEX-L   231
## 407            CR-VLX   115
## 423          Cruze1LT   120
## 434        CruzeSedan   185
## 438               CTS   132
## 464           DartSXT   124
## 500           EdgeSEL   205
## 504        Elantra4dr   178
## 508         ElantraSE   164
## 521    EnclaveLeather   144
## 545        EquinoxAWD   129
## 546        EquinoxFWD   454
## 550                ES   220
## 563         EscapeFWD   219
## 568          EscapeSE   230
## 570    EscapeTitanium   133
## 573              ESES   109
## 598   ExplorerLimited   138
## 603       ExplorerXLT   258
## 606          F-1502WD   225
## 607          F-1504WD   623
## 613       F-150Lariat   142
```

```
## 623          F-150XLT   332
## 685    FocusHatchback   161
## 689          FocusSE   181
## 690        FocusSedan   195
## 707          ForteLX   115
## 734          FusionSE   414
## 737    FusionTitanium   115
## 754              G37   124
## 801            Grand  1066
## 874               IS   158
## 876            Jetta   115
## 902       LaCrosseFWD   109
## 962         Malibu1LT   121
## 973          MalibuLS   121
## 974          MalibuLT   243
## 997           Mazda3i   128
## 1062      Mustang2dr   138
## 1070  MustangFastback   152
## 1071       MustangGT   151
## 1102     OdysseyEX-L   176
## 1109        OptimaEX   142
## 1111        OptimaLX   317
## 1161    PatriotSport   132
## 1166       PilotEX-L   122
## 1244             Ram   289
## 1305          RogueS   149
## 1307         RogueSV   148
## 1311           Rover   190
## 1316              RX   237
## 1318            RXRX   119
## 1352           Santa   386
## 1367        SedonaLX   111
## 1372         SentraS   149
## 1375        SentraSV   159
## 1389          Sierra   770
## 1390        Silverado  1807
## 1410       Sonata2.4L   224
## 1411        Sonata4dr   208
## 1428        SorentoLX   263
## 1431            Soul+   114
## 1433    SoulAutomatic   155
## 1463        SRXLuxury   109
## 1476     Suburban4WD   166
## 1479            Super   428
## 1483        Tacoma4WD   127
## 1488        Tahoe2WD   103
## 1490        Tahoe4WD   217
## 1506       TerrainFWD   212
## 1540            Town   250
## 1544         Transit   159
## 1548      TraverseFWD   162
## 1577           Tundra   109
## 1607            Versa   114
## 1625         Wrangler   604
```

```
## 1731            Yukon   176
## 1734         Yukon4WD   135
```

```r
# Delete the ** below and enter the model that you chose from the Enough_Vehicles data.
ModelOfMyChoice = "Civic"

# Takes a subset of your model vehicle from North Carolina
MyVehicles = subset(UsedCars, Model==ModelOfMyChoice & State==StateHW3)

# Check to make sure that the vehicles span at least 6 years.
range(MyVehicles$Year)
```

```
## [1] 2005 2017
```

Questions

Q1

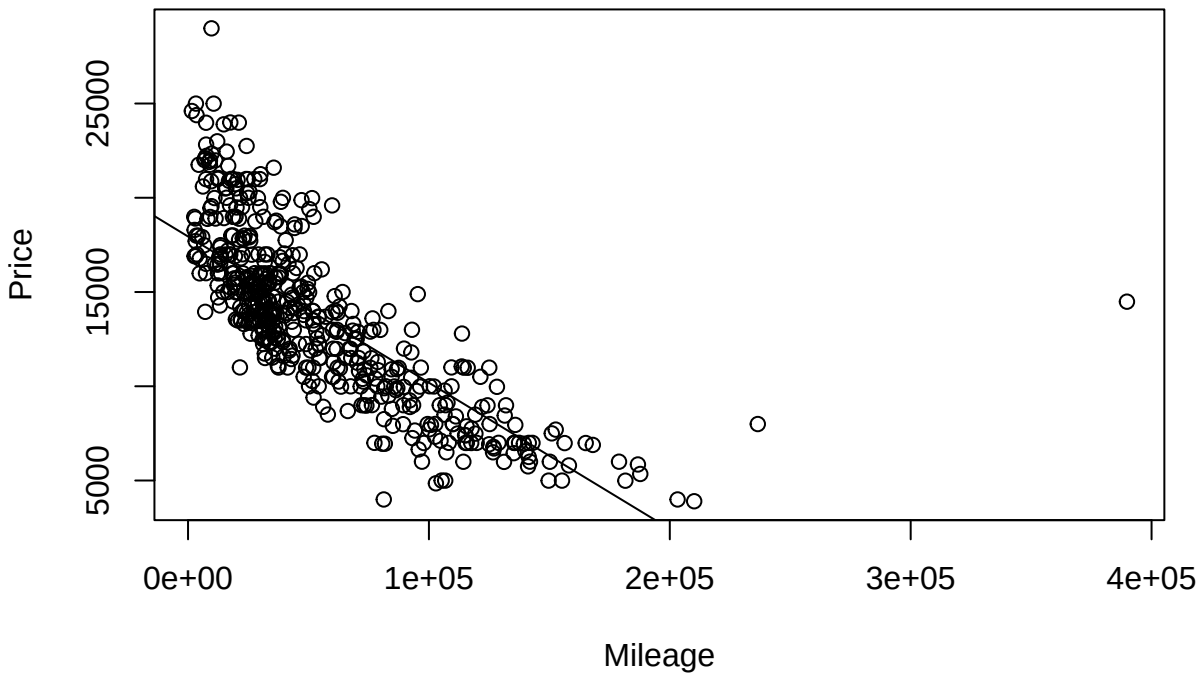
Calculate the least squares regression line that best fits your data using **Mileage** as the predictor and **Price** as the response. Produce a scatterplot of the relationship with the regression line on it. Interpret (in context) what the slope estimate tells you about prices and mileages of your used vehicle model. Explain why the sign (positive/negative) makes sense.

The slope -7.757e-02 means that as the mileage increases, the value of the vehicle decreases. It makes sense because as you are using the car more the orginial value decreases making the sign negative instead of positive. Positive would imply that the care price increases with mileage, which is false.

```
#
md1 <-lm(Price~Mileage, data = MyVehicles)
summary(md1)
```

```
##
## Call:
## lm(formula = Price ~ Mileage, data = MyVehicles)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7647.6 -1807.3  -519.1  1283.3 26776.7
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.795e+04  2.048e+02   87.63   <2e-16 ***
## Mileage     -7.757e-02  2.922e-03  -26.54   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2888 on 507 degrees of freedom
## Multiple R-squared:  0.5815, Adjusted R-squared:  0.5807
## F-statistic: 704.6 on 1 and 507 DF,  p-value: < 2.2e-16
```

```
plot(Price~Mileage, data = MyVehicles)
abline(md1)
```

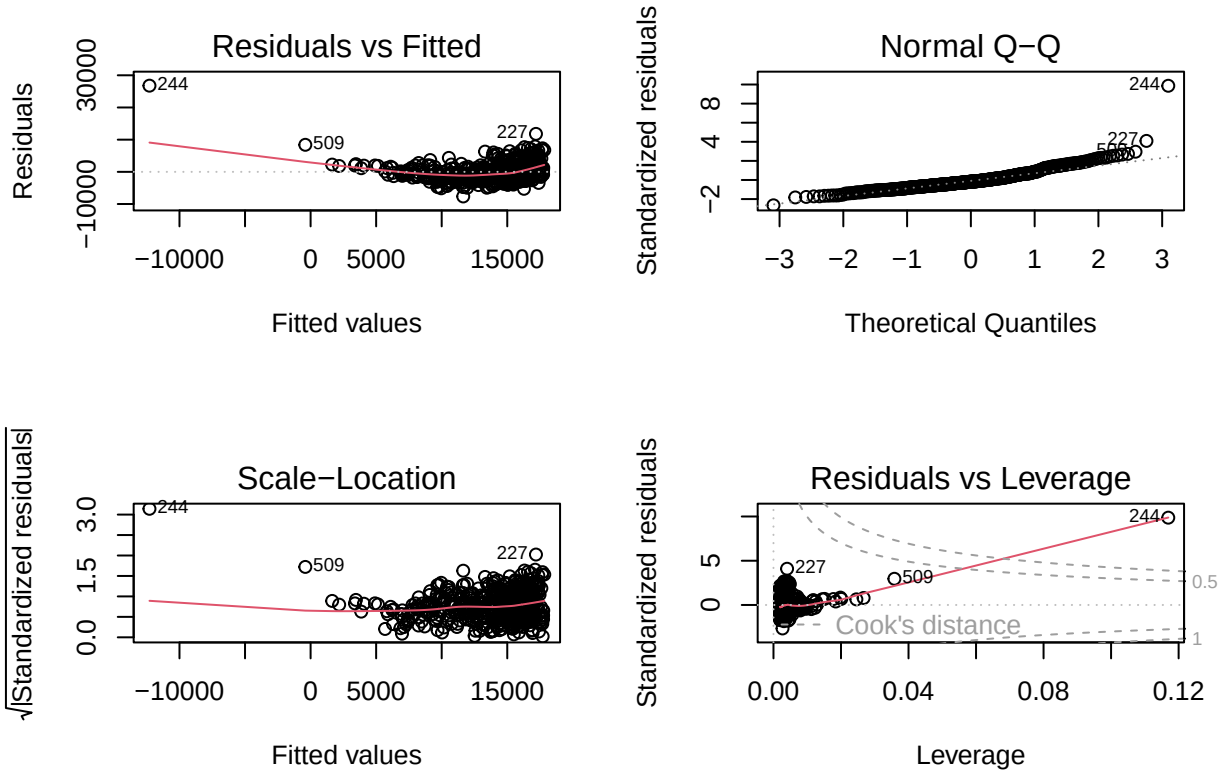


Q2

Produce appropriate residual plots and comment on how well your data appear to fit the conditions for a linear model. Don't worry about doing transformations at this point if there are problems with the conditions.

The data appears to fit the conditons okay for a linear model. We can see the assumption from the residuals vs fitted graph which shows the residuals are close to the red line. Also, from the QQ plot we see that the data is normal and the scale-location indicating the standarized residuals. The standarized residuals are pretty close to each other with couple of outliers, hence the linear model is good/okay fit.

```
#
par(mfrow = c(2,2))
plot(md1)
```



Q3

Find the five vehicles in your sample with the largest residuals (in magnitude - positive or negative). For these vehicles, find their standardized and studentized residuals. Based on these specific residuals, would any of these vehicles be considered outliers? Based on these specific residuals, would any of these vehicles possibly be considered influential on your linear model?

The vehiciles with the highest values are 244,227,509,230, and 210. All of these would be considered outliers because they have values higher than abs(3). Therefore, these vehicles would affect my linear model, making it less accurate.

```
#
MyVehicles$resid <- resid(md1)
head(sort(abs(MyVehicles$resid), dec = TRUE), 5)
```

```
##       244       227       509       230       210
## 26776.726 11798.521  8396.789  7873.546  7678.204
```

```
rstandard(md1)[244]
```

```
##      244
## 9.867492
```

```r
rstandard(md1)[227]
```

```
##      227
## 4.094023
```

```r
rstandard(md1)[509]
```

```
##      509
## 2.961278
```

```r
rstandard(md1)[230]
```

```
##      230
## 2.731969
```

```r
rstandard(md1)[210]
```

```
##      210
## 2.663076
```

```r
rstudent(md1)[244]
```

```
##      244
## 10.96692
```

```r
rstudent(md1)[227]
```

```
##      227
## 4.159314
```

```r
rstudent(md1)[509]
```

```
##      509
## 2.984277
```

```r
rstudent(md1)[230]
```

```
##      230
## 2.749587
```

```r
rstudent(md1)[210]
```

```
##      210
## 2.679253
```

Q4

Determine the leverages for the vehicles with the five largest absolute residuals. What do these leverage values say about the potential for each of these five vehicles to be influential on your model?

These vehicles have the potential to have an influence on my model the most: 244 and 509.The high leverage model from the actual value from 2(p/n), indicates that removing that particular vehicle could significantly change the regression coefficients and the overall model.

```
#
2*(2/nrow(MyVehicles))
```

```
## [1] 0.007858546
```

```
hatvalues(md1)[244]
```

```
##       244
## 0.1169381
```

```
hatvalues(md1)[227]
```

```
##         227
## 0.004034541
```

```
hatvalues(md1)[509]
```

```
##        509
## 0.03582064
```

```
hatvalues(md1)[230]
```

```
##         230
## 0.003954575
```

```
hatvalues(md1)[210]
```

```
##         210
## 0.003121985
```

Q5

Determine the Cook's distances for the vehicles with the five largest absolute residuals. What do these Cook's distances values say about the influence of each of these five vehicles on your model?

The cook distnace indciate how well the model fits, from the obeservation the vehicels aboce 1 are 244, which greatly affect my model, leading to a poor fit.

```
#
head(sort(cooks.distance(md1), decreasing =TRUE), n=5)
```

```
##        244        509        227        220        230
## 6.44686281 0.16289356 0.03394849 0.01509180 0.01481638
```

Q6

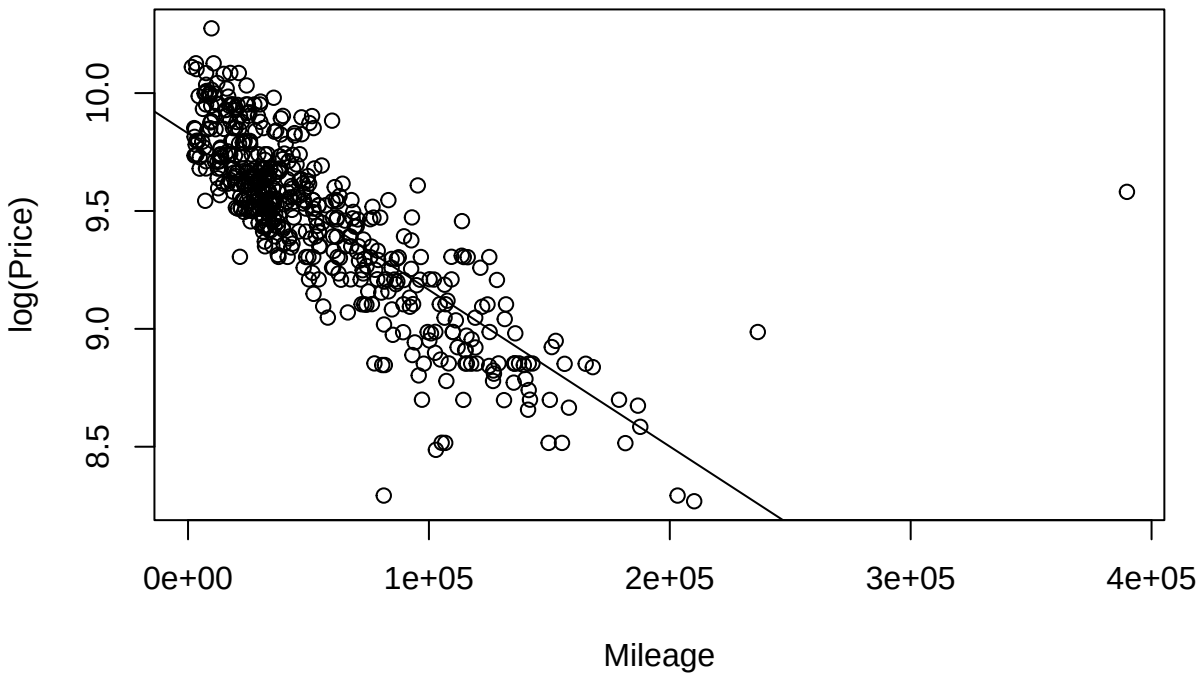
Experiment with some transformations to attempt to find one that seems to do a better job of satisfying the linear model conditions. Include the summary output for fitting that model and a scatterplot of the original data with this new model (which is likely a curve on the original data). Explain why you think that this transformation does or does not improve satisfying the linear model conditions.

Utilizing the different transformation model without putting one on the predictor value, I found that sqrt was the second best with the r-squred from .5807 to 0.7641. The best fit model was with logrithims because it riased the value of the r-square from .5807 to 0.789. Moreover, the values is model 2 with logs are closer to the line of best fit compared to model 1, which was the standard linear model. Therefore, the transformation does make the data fit the model more suitably.

```
#
md2<-lm(log(Price)~(Mileage), data=MyVehicles)
summary(md2)
```

```
##
## Call:
## lm(formula = log(Price) ~ (Mileage), data = MyVehicles)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.99669 -0.11723 -0.01559  0.11173  2.34082
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.829e+00  1.502e-02  654.36   <2e-16 ***
## Mileage     -6.642e-06  2.143e-07  -30.99   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2118 on 507 degrees of freedom
## Multiple R-squared:  0.6544, Adjusted R-squared:  0.6538
## F-statistic: 960.2 on 1 and 507 DF,  p-value: < 2.2e-16
```

```
plot(log(Price)~(Mileage), data=MyVehicles)
abline(md2)
```


log(Price)
8.5
9.0
9.5
10.0
0e+00
1e+05
2e+05
3e+05
4e+05
Mileage