STOR 455 Homework #4

20 points - Due Thursday 10/5 at 12:30pm

Theory Part

Below is an ANOVA table of a simple linear model. Complete this table by filling in missing values.

	Df	Sum of Squares	Mean of Squares	F value
Model	1	4.260	4.260	20.882
Residuals	212	42.974	0.204	20.882
Total	213	47.234	0.223	20.882

Computing Part

Instructions: 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}**.

```
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.
```

```
StateHW4 = "NC"

# Creates a dataframe with the number of each model for sale in North Carolina
Vehicles = as.data.frame(table(UsedCars$Model[UsedCars$State==StateHW4]))

# 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==StateHW4)

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

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

Questions

Q1

Construct a model using two predictors *Year* and *Mileage* with *Price* as the response variable and provide the summary output. Comment on the diagnostic plots.

In the Residual versus Fitted Value plot, the relationship appears to deviate from linearity, suggesting that the linearity assumption may not hold. However, the residuals exhibit a zero-mean, indicating an approximate balance in positive and negative errors. From the histogram, it's evident that the distribution is roughly normal with an Uniform spread. While the Scale-Location plot displays a bell-shaped pattern, implying reasonably constant variance, the linearity assumption still raises concerns. In terms of independence, Price and Year+Mileage appear to be largely independent. Yet, the normal Q-Q plot reveals deviations in the upper tail portion (more degree of freedoms may cause this), indicating a potential violation of the normal distribution assumption for the errors. Further analysis and potentially nonlinear modeling may be warranted to enhance the model's fit.

```r
#
library(olsrr)
```

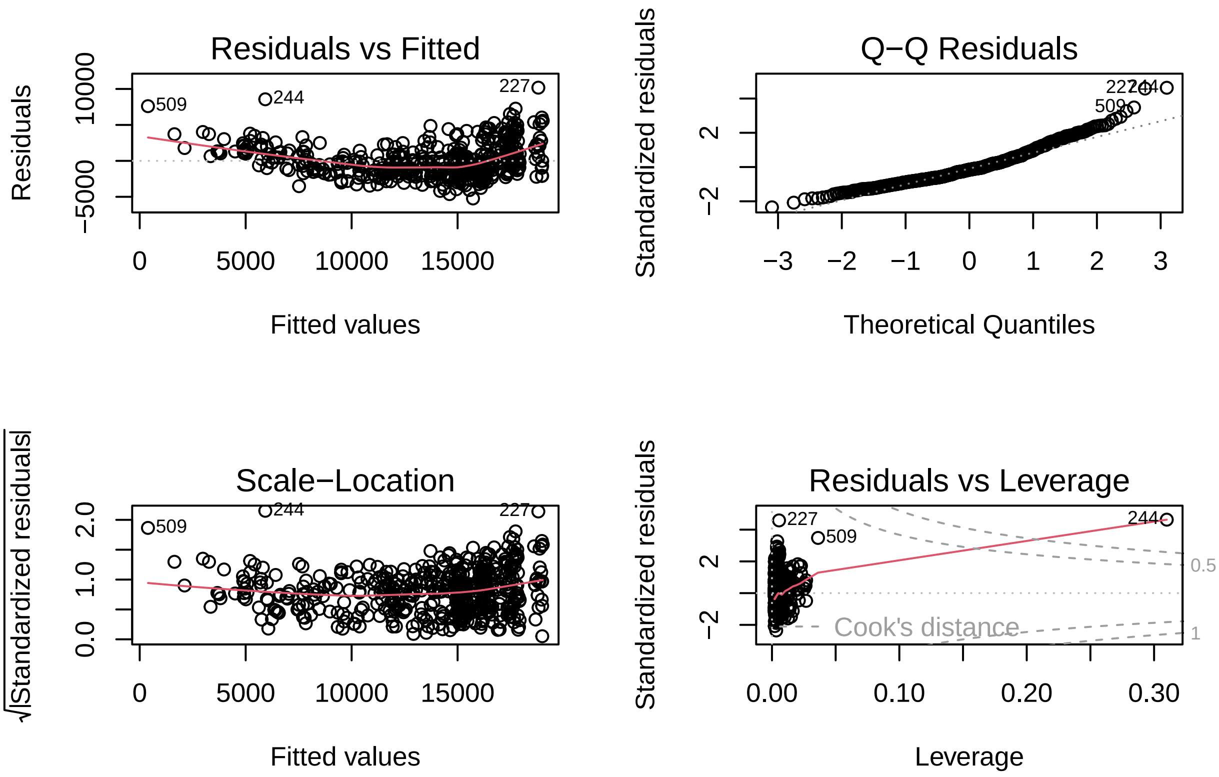
```
##
## Attaching package: 'olsrr'

## The following object is masked from 'package:datasets':
##
##     rivers
```

```r
modq1 = lm(Price~Mileage+Year, data=MyVehicles)
summary(modq1)
```

```
##
## Call:
## lm(formula = Price ~ Mileage + Year, data = MyVehicles)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5227.9 -1589.0  -352.8  1182.2 10177.8
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.101e+06  1.138e+05 -18.460  < 2e-16 ***
## Mileage     -2.561e-02  3.587e-03  -7.139 3.29e-12 ***
## Year         1.051e+03  5.646e+01  18.617  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2227 on 506 degrees of freedom
## Multiple R-squared:  0.7517, Adjusted R-squared:  0.7507
## F-statistic: 765.7 on 2 and 506 DF,  p-value: < 2.2e-16
```

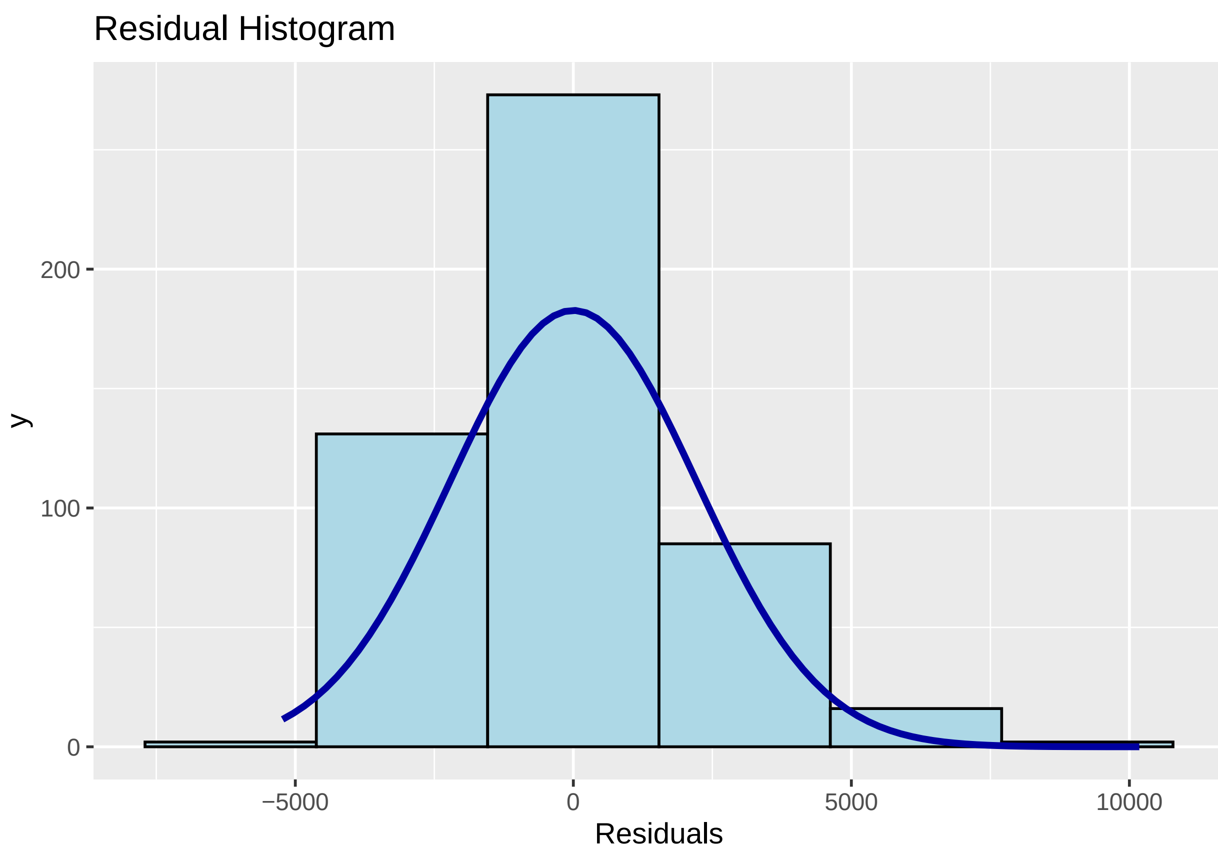
```r
par(mfrow=c(2,2))
plot(modq1)
```




```
mean(modq1$residuals)
```

```
## [1] 1.587423e-13
```

```
ols_plot_resid_hist(modq1)
```



Q2

Assess the importance of each of the predictors in the regression model - be sure to indicate the specific value(s) from the summary output you are using to make the assessments. Include hypotheses and conclusions in context.

Year: Null Hypothesis is the coefficient for the "Year" predictor is zero (Year has no effect on Price).Alternative Hypothesis is the coefficient for the "Year" predictor is not zero (Year has an effect on Price).P-value is below the .05 threshold (2.2e-16 is the p-value), we reject the null hypothesis, indicating that "Year" is a significant predictor of "Price.

Mileage: Null Hypothesis is the coefficient for the "Mileage" predictor is zero (Mileage has no effect on Price).Alternative Hypothesis is the coefficient for the "Mileage" predictor is not zero (Mileage has an effect on Price).p-value is below the threshold of .05 (2.2e-16 is the p-value), we reject the null hypothesis, indicating that "Mileage" is a significant predictor of "Price."

Therefore, For my model both p-values for the Year and Mileage predictors are well below 0.05, hence they are significant and useful in the model.

```
#
md1 = lm(Price~Year, data = MyVehicles)
md2 = lm(Price~Mileage, data = MyVehicles)
summary(md1)
```

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

```
## lm(formula = Price ~ Year, data = MyVehicles)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5609.8 -1636.2  -386.2  1280.6 10148.7
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.734e+06  7.484e+04  -36.53   <2e-16 ***
## Year         1.365e+03  3.718e+01   36.71   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2334 on 507 degrees of freedom
## Multiple R-squared:  0.7266, Adjusted R-squared:  0.7261
## F-statistic:  1348 on 1 and 507 DF,  p-value: < 2.2e-16
```

```r
summary(md2)
```

```
##
## 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
```

Q3

Assess the overall effectiveness of this model (with a formal test). Again, be sure to include hypotheses and the specific value(s) you are using from the summary output to reach a conclusion.

Null Hypothesis is all the coefficients for the predictors are zero, implying that none of the predictors have an effect on Price (the model has no explanatory power).Alternative Hypothesis is at least one of the coefficients for the predictors is not zero, implying that at least one predictor has an effect on Price (the model has explanatory power). The p-value is less than 0.05, we can conclude that the overall model is statistically significant and at least one predictor is important in predicting the response variable of Price. Therefore, for my model the p-value is small (2.2e-16), so I have evidence to support the alternative, that at least one of the coefficients is nonzero.

```r
#
anova(modq1)
```

```
## Analysis of Variance Table
##
## Response: Price
##            Df     Sum Sq    Mean Sq F value    Pr(>F)
## Mileage     1 5875478801 5875478801  1184.9 < 2.2e-16 ***
## Year        1 1718705142 1718705142   346.6 < 2.2e-16 ***
## Residuals 506 2509131066    4958757
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary (modq1)
```

```
##
## Call:
## lm(formula = Price ~ Mileage + Year, data = MyVehicles)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5227.9 -1589.0  -352.8  1182.2 10177.8
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.101e+06  1.138e+05 -18.460  < 2e-16 ***
## Mileage     -2.561e-02  3.587e-03  -7.139 3.29e-12 ***
## Year         1.051e+03  5.646e+01  18.617  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2227 on 506 degrees of freedom
## Multiple R-squared:  0.7517, Adjusted R-squared:  0.7507
## F-statistic: 765.7 on 2 and 506 DF,  p-value: < 2.2e-16
```

Q4

Compute and interpret the variance inflation factor (VIF) for your predictors.

The presence of multicollinearity in the model can be assessed using Variance Inflation Factor (VIF) values. Generally, a VIF exceeding 5 suggests substantial multicollinearity, while values below 5 indicate minimal multicollinearity. In this case, the VIF for the predictors is relatively small (2.53 for both), indicating a low concern for multicollinearity.

```r
#
library(car)
```

```
## Loading required package: carData
```

```r
vif(modq1)
```

```
##  Mileage     Year
## 2.533963 2.533963
```

Q5

Suppose that you are interested in purchasing a car of this model that is from the year 2017 with 50K miles. Determine each of the following: a 95% confidence interval for the mean price at this year and odometer reading, and a 95% prediction interval for the price of an individual car at this year and odometer reading. Write sentences that carefully interpret each of the intervals (in terms of car prices).

The confidence interval predicts the average price of cars from the year 2017 with 50k miles in the model from the year and odometer readings. On the other hand, the prediction interval forecasts the price of a specific car from the year 2017 with 50k miles in the model from the year and odometer readings.

```
#
oneCar = data.frame(Year = 2017, Mileage=50000)
predict.lm(modq1, oneCar, interval = "confidence", level=.95)
```

```
##        fit      lwr      upr
## 1 17779.29 17341.67 18216.92
```

```
predict.lm(modq1, oneCar, interval = "prediction", level=.95)
```

```
##        fit     lwr      upr
## 1 17779.29 13382.5 22176.09
```