STOR 455 Homework #5

20 points - Due Tuesday 10/17 at 12:30pm

Theory Part

1. True or False: For a regression with 2 predictors, the VIF of the two predictors can be different.

True, because VIF can be different for each predictor in a multiple regression since the correlation between each predictor and the other predictor can be different, leading to different levels of multicollinearity.

2. True or False: Mallows' C_p depends only on the predictors in the model.

False, Cp depends on the larger pool of predictors as well as the set being considered.

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.
```

```r
StateHW5 = "NC"

# Creates a dataframe with the number of each model for sale in North Carolina
Vehicles = as.data.frame(table(UsedCars$Model[UsedCars$State==StateHW5]))

# 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==StateHW5)

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

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

Questions

Q1

Add a column of *logPrice* as the (natural) logarithm of the prices. Construct a model using two predictors (*Year* and *Mileage*) with *logPrice* as the response variable and provide the summary output. Comment on the diagnostic plots.

In the Residual versus Fitted Value plot, the relationship appears towards linearity, suggesting that the linearity assumption may 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. In terms of independence, logPrice and Year+Mileage appear to be largely independent. Also, the normal Q-Q plot reveals that they follow the path and don't have a weak or heavy tails,hence the normal distribution assumption for the errors. However, there are two extreme outliers that affects the weight of the tail in the QQ plot making them seem less normal with them making the curve look heavy.

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

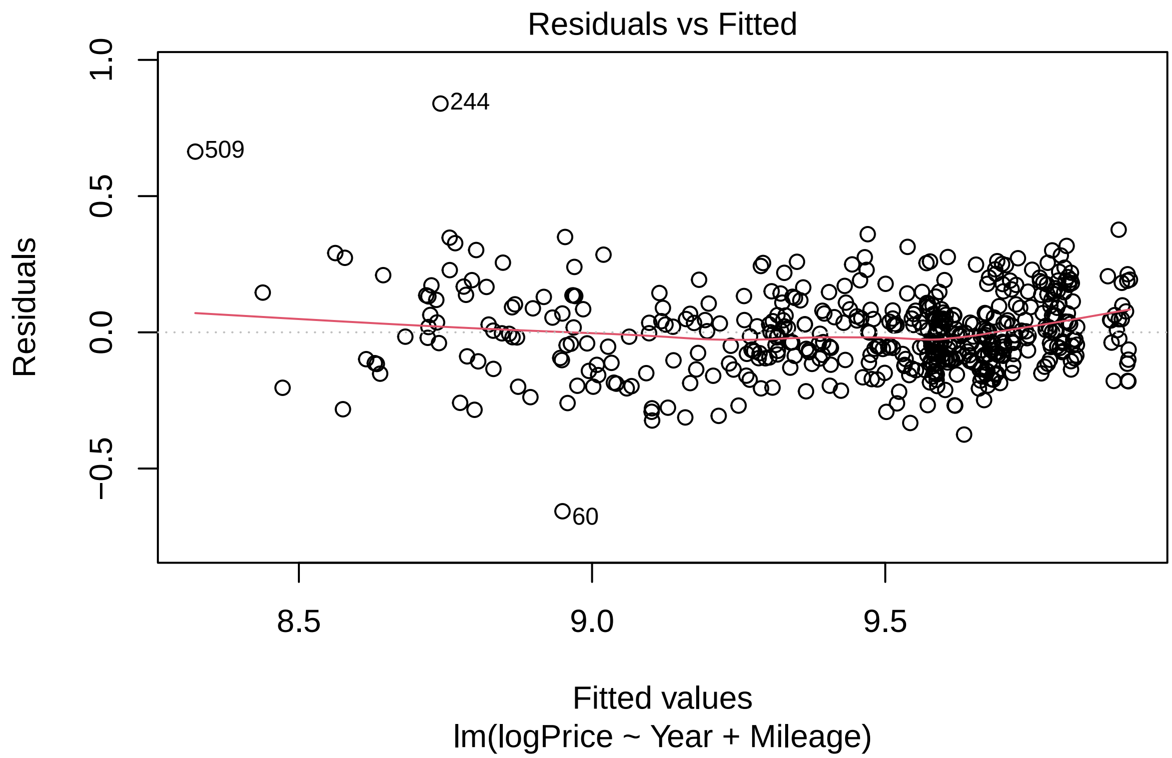
```
##
## Attaching package: 'olsrr'

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

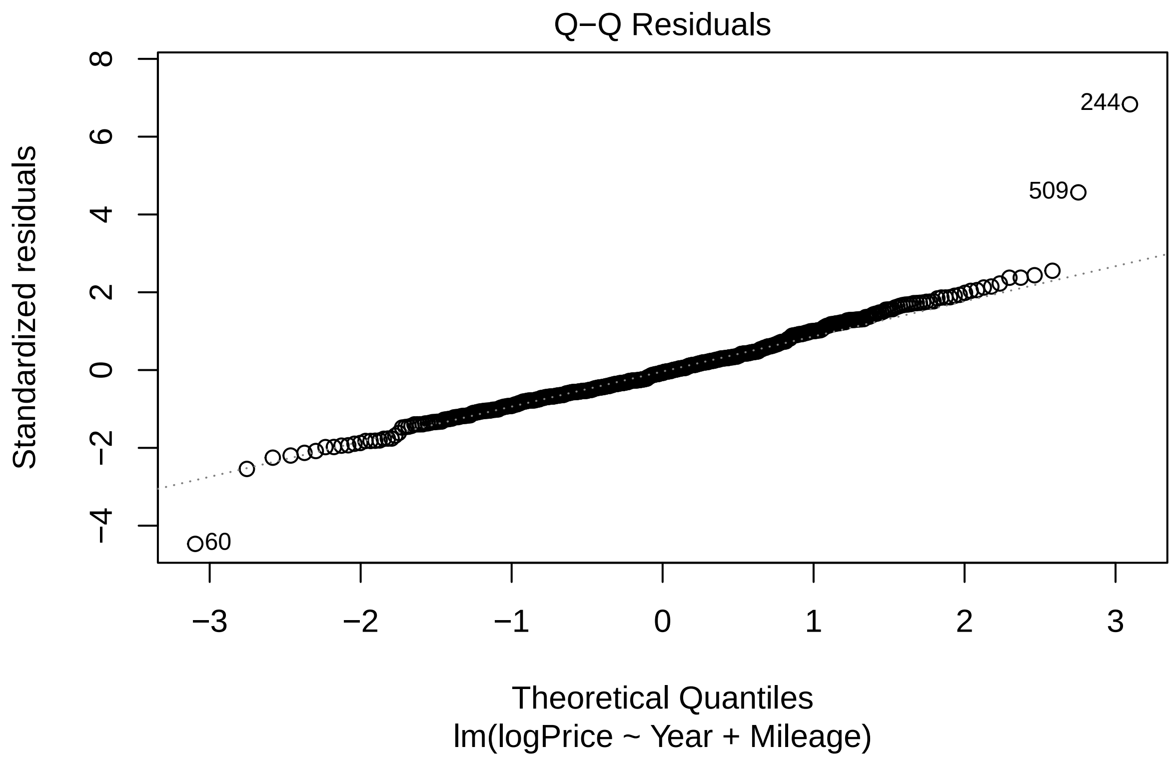
```r
MyVehicles$logPrice <- log(MyVehicles$Price)
mod1 = lm(logPrice~Year+Mileage, data = MyVehicles)
summary(mod1)
```

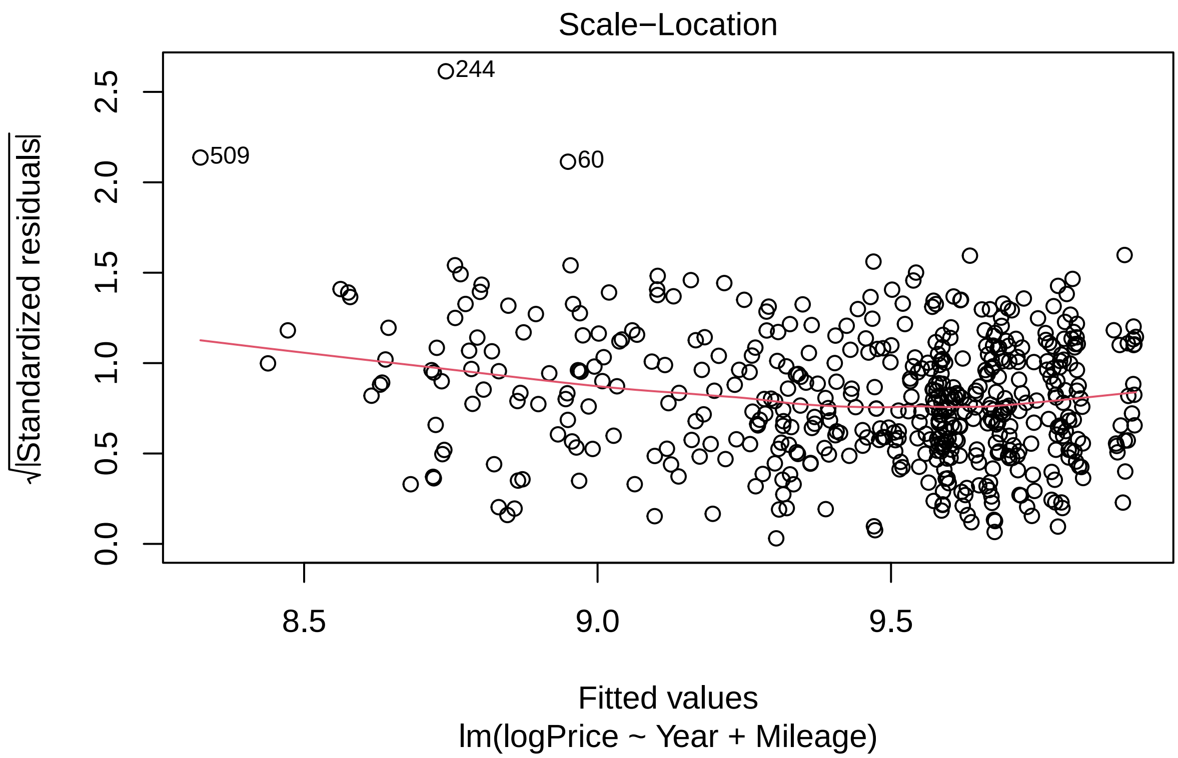
```
##
## Call:
## lm(formula = logPrice ~ Year + Mileage, data = MyVehicles)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.65679 -0.09531 -0.01057  0.08452  0.83959
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.648e+02  7.561e+00 -21.801   <2e-16 ***
## Year         8.664e-02  3.750e-03  23.101   <2e-16 ***
## Mileage     -2.359e-06  2.383e-07  -9.902   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1479 on 506 degrees of freedom
## Multiple R-squared:  0.8318, Adjusted R-squared:  0.8312
## F-statistic:  1251 on 2 and 506 DF,  p-value: < 2.2e-16
```

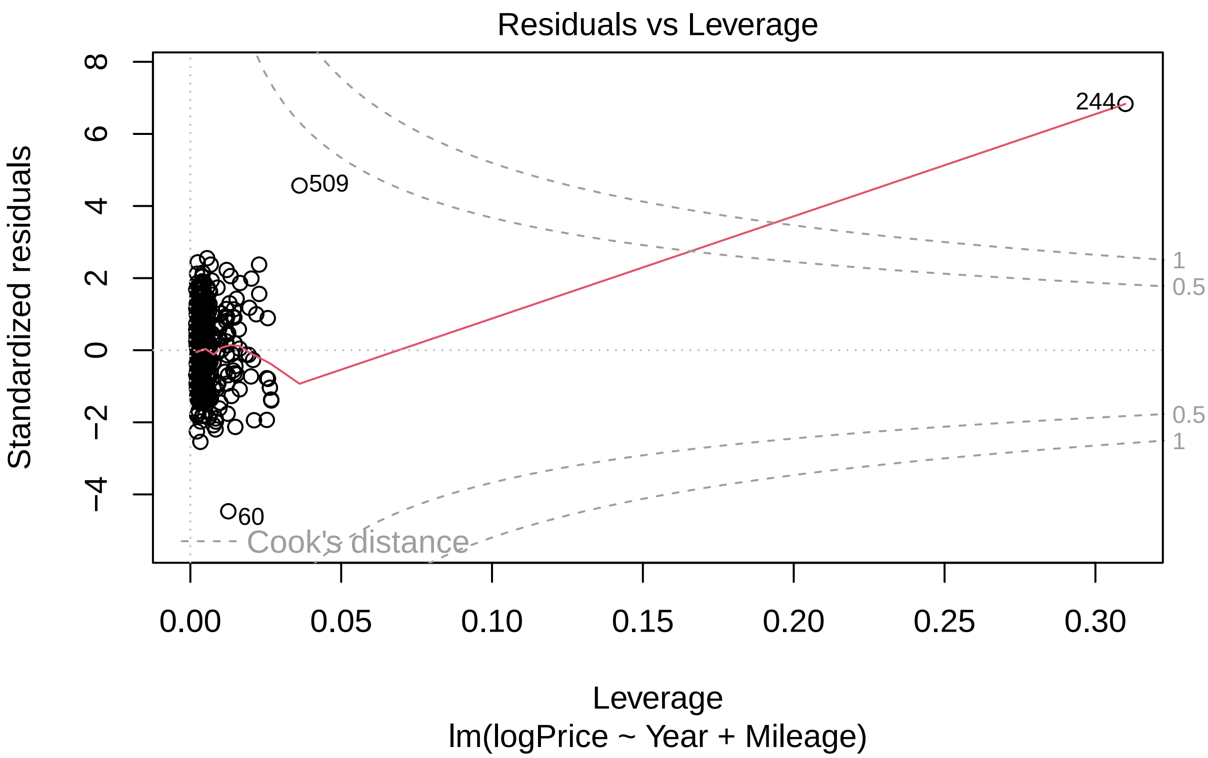
```r
plot(mod1)
```



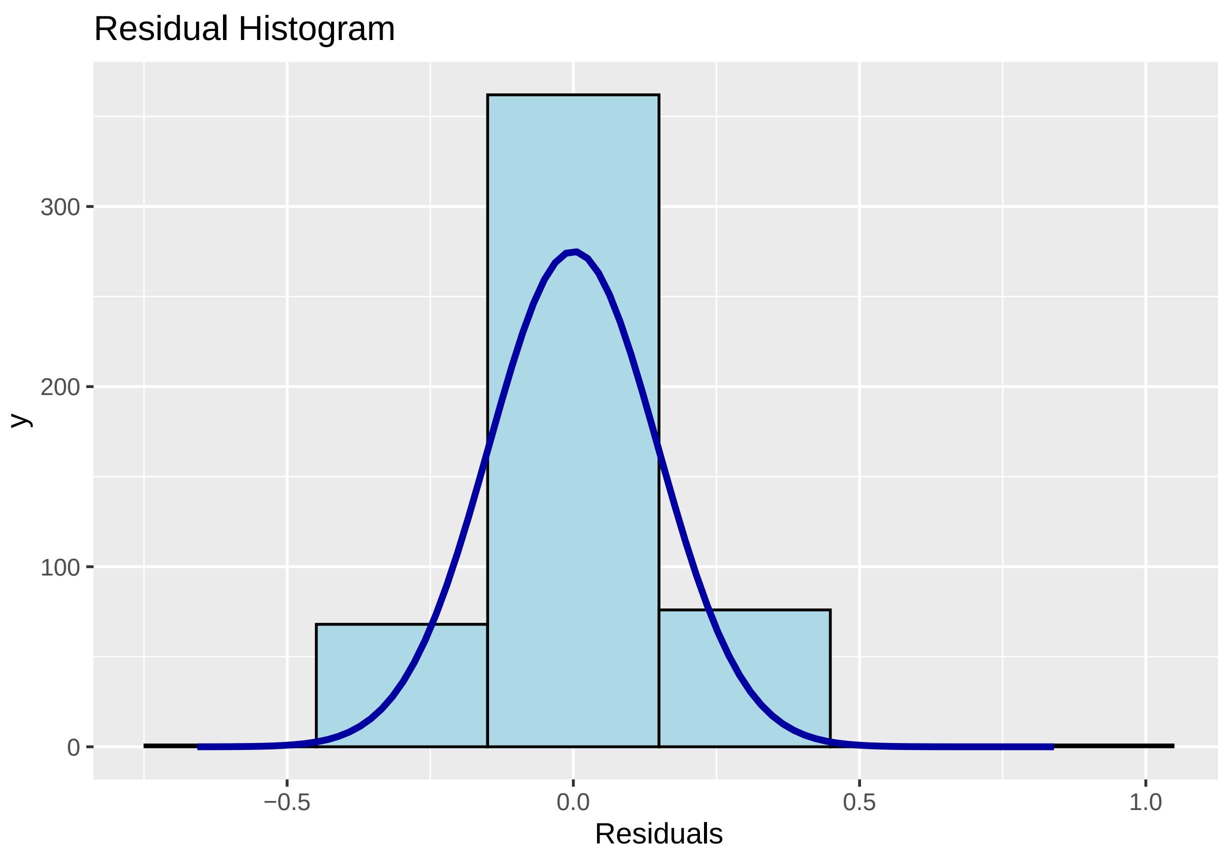
Residuals vs Fitted
Residuals
1.0
0.5
0.0
−0.5
8.5
9.0
9.5
244
509
60
Fitted values
lm(logPrice ~ Year + Mileage)


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(logPrice ~ Year + Mileage)
8
6
4
2
0
−2
−4
−3
−2
−1
0
1
2
3
244
509
60


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
244
509
60
8.5
9.0
9.5
Fitted values
lm(logPrice ~ Year + Mileage)




```
ols_plot_resid_hist(mod1)
```



Q2

Add two columns of *Mileage2* and *Mileage3* as Mileageˆ2 and Mileageˆ3 respectively. Construct a model using four predictors (*Year*,*Mileage*, *Mileage2* and *Mileage3*) with *logPrice* as the response variable and provide the summary output. Call this model *Full*. Comment on the diagnostic plots.

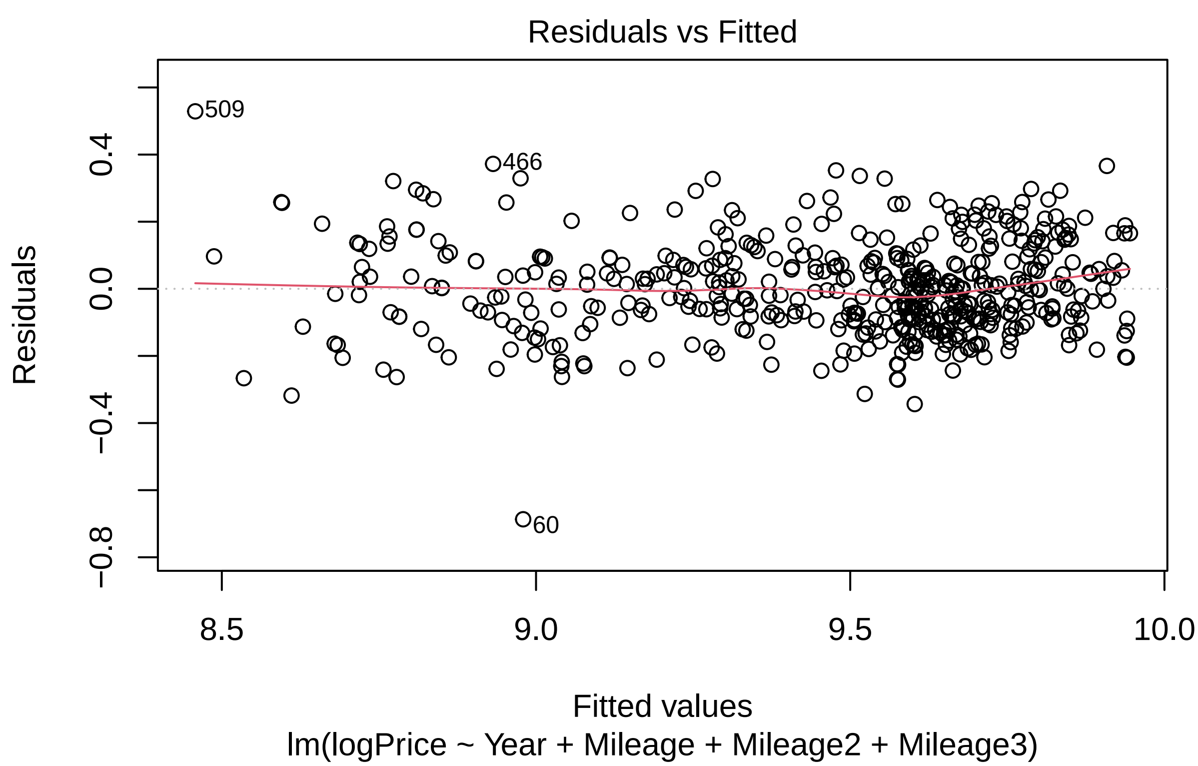
Similar to the above, the Residual versus Fitted Value plot, the relationship appears towards linearity, suggesting that the linearity assumption holds. However, the residuals exhibit a zero-mean, indicating an approximate balance in positive and negative errors. From the histogram, it's evident that the distribution is normal with an Uniform spread. While the Scale-Location plot displays a bell-shaped pattern to the right, implying somewhat constant variance. In terms of independence, logPrice and Year+Mileage+Mileage2+Mileage3 appear to be largely independent. Also, the normal Q-Q plot reveals that they follow the path and don't have a weak or heavy tails,hence the normal distribution assumption for the errors.Also, there does seem to be some outliers but are not affecting the models as much because the assumption still holds, but may have an affect on the qq plot with further investigation.

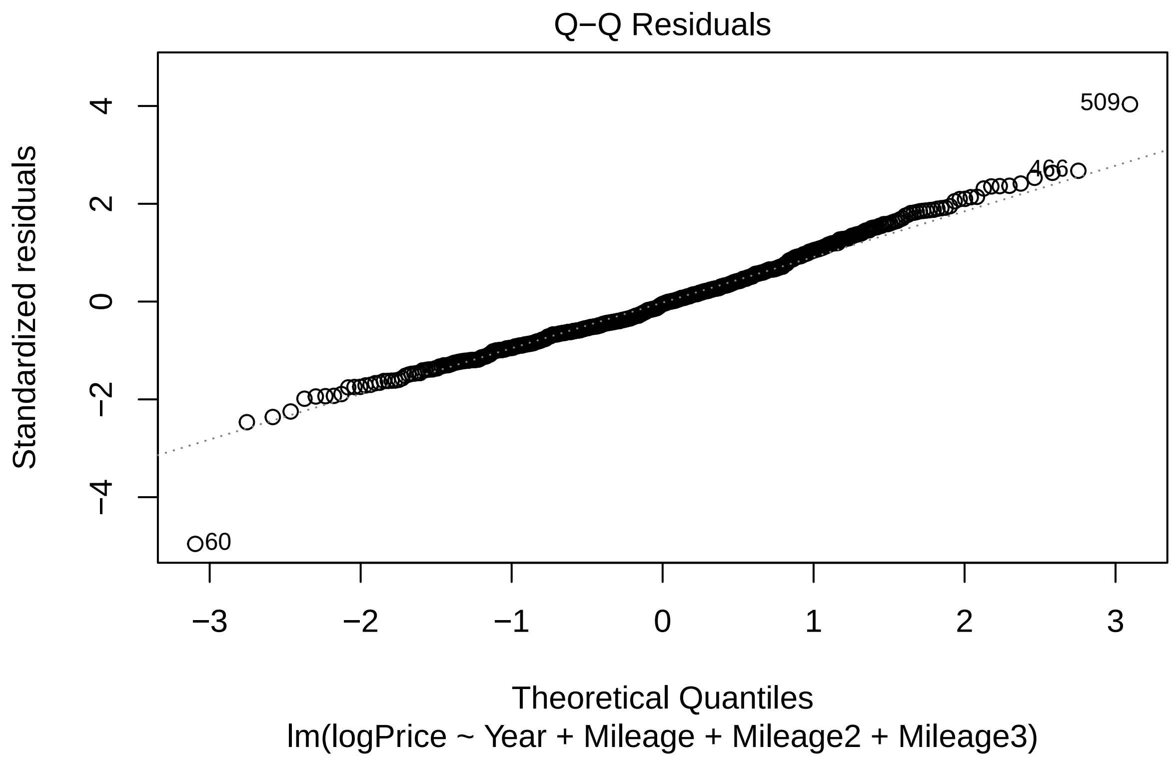
```
#
library(olsrr)
MyVehicles$Mileage2 <- MyVehicles$Mileage^2
MyVehicles$Mileage3 <- MyVehicles$Mileage^3
Full = lm(logPrice~Year+Mileage+Mileage2+Mileage3, data=MyVehicles)
summary(Full)
```

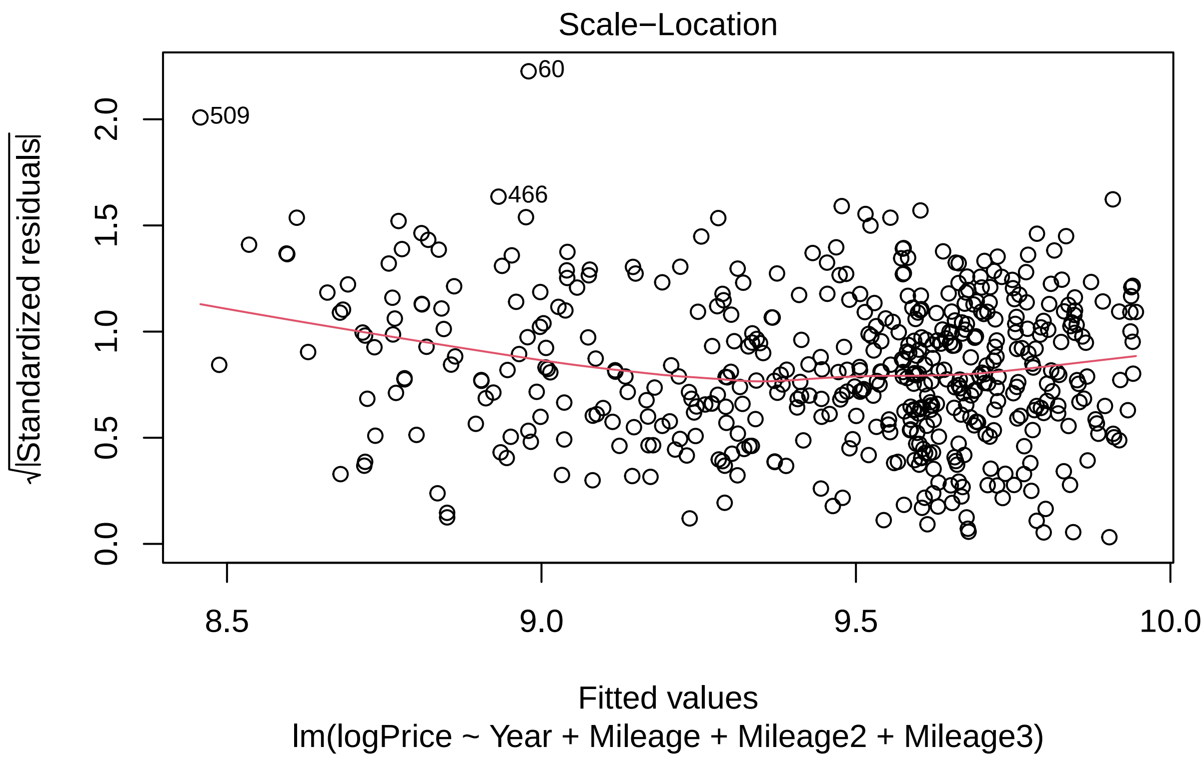
```
##
## Call:
```

```
## lm(formula = logPrice ~ Year + Mileage + Mileage2 + Mileage3,
##     data = MyVehicles)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68672 -0.09020 -0.00697  0.08498  0.52883
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.321e+02  8.230e+00 -16.047  < 2e-16 ***
## Year         7.041e-02  4.081e-03  17.254  < 2e-16 ***
## Mileage     -4.527e-06  7.291e-07  -6.208 1.12e-09 ***
## Mileage2     2.687e-12  5.943e-12   0.452   0.6513
## Mileage3     2.045e-17  1.220e-17   1.676   0.0943 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1396 on 504 degrees of freedom
## Multiple R-squared:  0.8507, Adjusted R-squared:  0.8496
## F-statistic: 718.1 on 4 and 504 DF,  p-value: < 2.2e-16
```

```r
plot(Full)
```

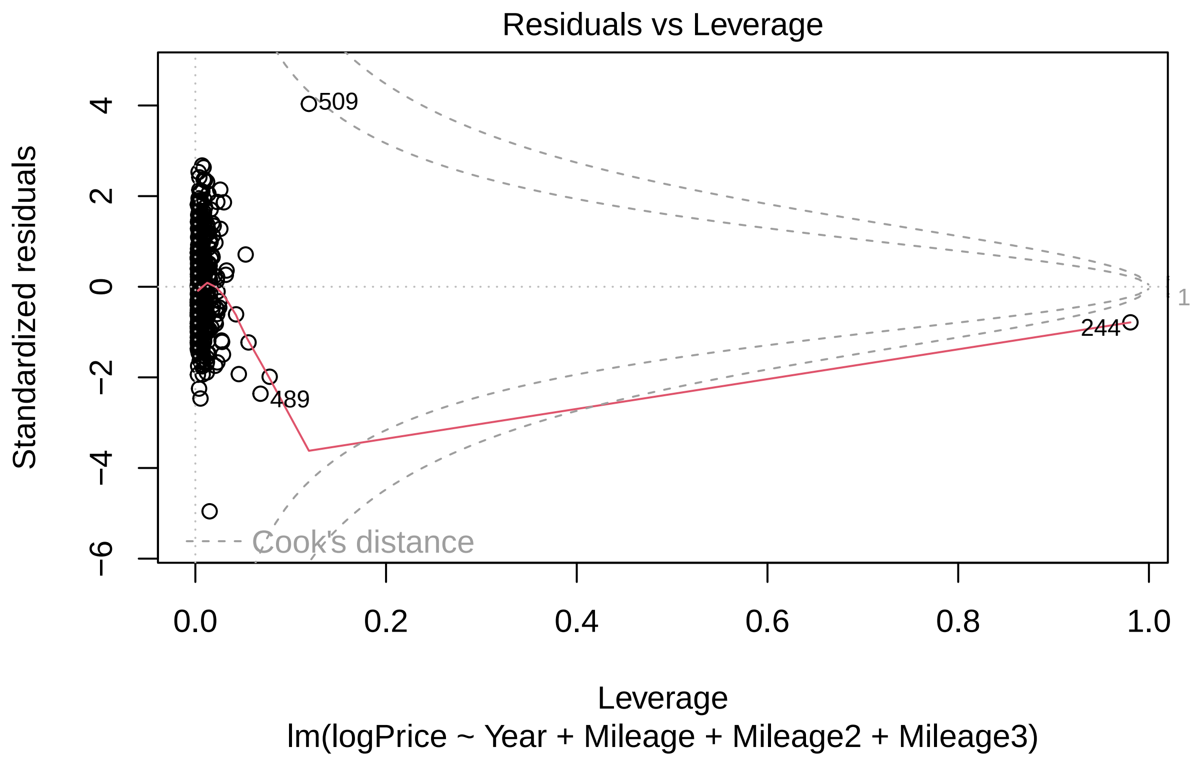



Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(logPrice ~ Year + Mileage + Mileage2 + Mileage3)
509
466
60
4
2
0
−2
−4
−3
−2
−1
0
1
2
3

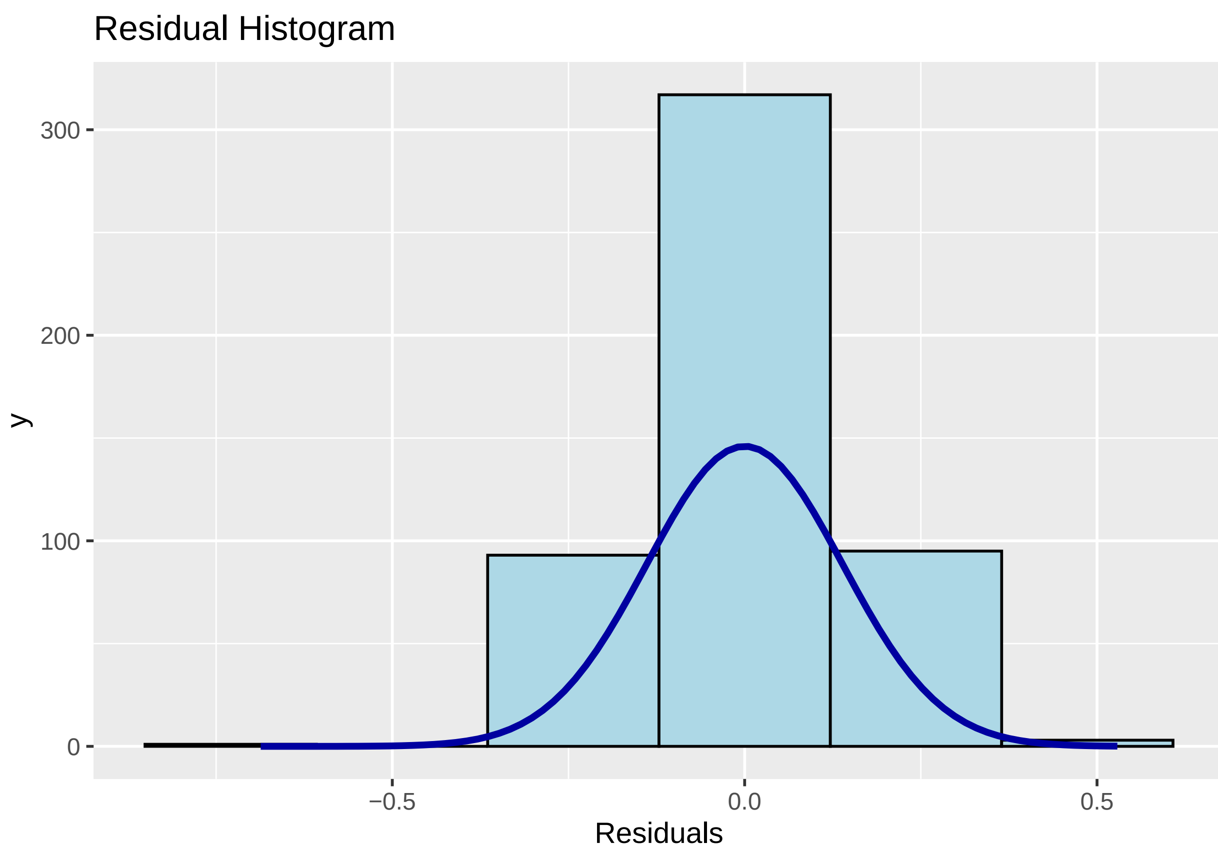



```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```




```r
ols_plot_resid_hist(Full)
```



Q3

Select a subset of predictors in *Full* using each of the four methods: all subsets, backward elimination, forward selection, and stepwise regression. Use Mallows' Cp (AIC) as the criterion.

```
#all sets
library(leaps)
all_sbset <- regsubsets(logPrice~Year+Mileage+Mileage2+Mileage3, data=MyVehicles)
sum.sets <- summary(all_sbset)
coef(all_sbset, id = which.min(sum.sets$cp))
```

```
##   (Intercept)          Year       Mileage      Mileage3
## -1.321841e+02  7.046620e-02 -4.231421e-06  2.577564e-17
```

```
#backward elimination
model.b.elim<- step(Full, direction = 'backward')
```

```
## Start:  AIC=-1999.24
## logPrice ~ Year + Mileage + Mileage2 + Mileage3
##
##            Df Sum of Sq    RSS     AIC
## - Mileage2  1    0.0040 9.8298 -2001.0
## <none>                  9.8259 -1999.2
## - Mileage3  1    0.0548 9.8806 -1998.4
```

```
## - Mileage   1    0.7514 10.5772 -1963.7
## - Year      1    5.8036 15.6295 -1765.0
##
## Step:  AIC=-2001.04
## logPrice ~ Year + Mileage + Mileage3
##
##            Df Sum of Sq     RSS     AIC
## <none>                   9.8298 -2001.0
## - Mileage3  1    1.2416 11.0714 -1942.5
## - Mileage   1    3.3050 13.1349 -1855.5
## - Year      1    5.8178 15.6476 -1766.4
```

```
#forward selection
null <-lm(logPrice~1, data =MyVehicles)
model.f.selct <- step(null, scope = list(lower=null, upper=Full), direction = "forward")
```

```
## Start:  AIC=-1039.1
## logPrice ~ 1
##
##            Df Sum of Sq    RSS     AIC
## + Year      1    52.613 13.217 -1854.3
## + Mileage   1    43.082 22.748 -1578.0
## + Mileage2  1    20.563 45.266 -1227.7
## + Mileage3  1     4.512 61.318 -1073.2
## <none>                  65.829 -1039.1
##
## Step:  AIC=-1854.34
## logPrice ~ Year
##
##            Df Sum of Sq    RSS     AIC
## + Mileage   1   2.14527 11.071 -1942.5
## + Mileage2  1   0.54767 12.669 -1873.9
## + Mileage3  1   0.08184 13.135 -1855.5
## <none>                  13.217 -1854.3
##
## Step:  AIC=-1942.49
## logPrice ~ Year + Mileage
##
##            Df Sum of Sq     RSS     AIC
## + Mileage3  1    1.2416  9.8298 -2001.0
## + Mileage2  1    1.1908  9.8806 -1998.4
## <none>                  11.0714 -1942.5
##
## Step:  AIC=-2001.04
## logPrice ~ Year + Mileage + Mileage3
##
##            Df Sum of Sq    RSS     AIC
## <none>                  9.8298 -2001.0
## + Mileage2  1 0.0039864 9.8259 -1999.2
```

```
#stepwise regression
model.stw.reg <- step(null, scope = list(lower=null, upper=Full), direction = "both")
```

```
## Start:  AIC=-1039.1
```

```
## logPrice ~ 1
##
##            Df Sum of Sq    RSS     AIC
## + Year      1    52.613 13.217 -1854.3
## + Mileage   1    43.082 22.748 -1578.0
## + Mileage2  1    20.563 45.266 -1227.7
## + Mileage3  1     4.512 61.318 -1073.2
## <none>                  65.829 -1039.1
##
## Step:  AIC=-1854.34
## logPrice ~ Year
##
##            Df Sum of Sq    RSS     AIC
## + Mileage   1     2.145 11.071 -1942.5
## + Mileage2  1     0.548 12.669 -1873.9
## + Mileage3  1     0.082 13.135 -1855.5
## <none>                  13.217 -1854.3
## - Year      1    52.613 65.829 -1039.1
##
## Step:  AIC=-1942.49
## logPrice ~ Year + Mileage
##
##            Df Sum of Sq     RSS     AIC
## + Mileage3  1    1.2416  9.8298 -2001.0
## + Mileage2  1    1.1908  9.8806 -1998.4
## <none>                  11.0714 -1942.5
## - Mileage   1    2.1453 13.2167 -1854.3
## - Year      1   11.6762 22.7477 -1578.0
##
## Step:  AIC=-2001.04
## logPrice ~ Year + Mileage + Mileage3
##
##            Df Sum of Sq     RSS     AIC
## <none>                   9.8298 -2001.0
## + Mileage2  1    0.0040  9.8259 -1999.2
## - Mileage3  1    1.2416 11.0714 -1942.5
## - Mileage   1    3.3050 13.1349 -1855.5
## - Year      1    5.8178 15.6476 -1766.4
```

Q4

Assess and compare the overall effectiveness of the four models (some or all of them may be identical).

-All Subsets: consideration of all possible predictor subsets, optimal but computationally intensive and prone to over fitting. -Backward Elimination:Starts with all predictors and iteratively removes the least significant ones, less computationally intensive than all subsets. -Forward Selection: Begins with an empty model and adds the most significant predictors, less computationally intensive and useful for a large number of predictors. -Step wise Regression:Combines forward and backward selection, adding/removing predictors based on a criterion, balancing efficiency and model performance. –In this situation,all the models strike a balance between effectiveness and computational easiness, making them commonly used for variable selection. The choice of method should align with the data set's specifics, hence all of these models may be identical with the three predictors (Year, Mileage, Mileage^3)

Q5

Suppose that you are interested in purchasing a car of this model that is from the year 2018 with 60K 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).

Using the confidence interval with 95% confidence I predict that the mean price of all cars with this model that is from the year 2018 with 60K miles is between $16717.29 and $18260.22. Also using the prediction interval with 95% confidence I predict that the price of a car with this model that is from the year 2018 with 60K miles is between $13236.09 and $23062.8

```r
#
Car = data.frame(Year = 2018, Mileage=60000, Mileage3=60000^3)
best_mod=lm(logPrice~Year+Mileage+Mileage3, data=MyVehicles)

#Confidence
exp(predict.lm(best_mod, Car, interval= "confidence", level = .95))
```

```
##        fit      lwr      upr
## 1 17471.73 16717.29 18260.22
```

```r
#Prediction
exp(predict.lm(best_mod, Car, interval= "prediction", level = .95))
```

```
##        fit      lwr     upr
## 1 17471.73 13236.09 23062.8
```