Auto-mpg dataset Mileage per gallon performances of various cars

Members: Toby (Yan) Liu, Wei Jiang, Chao Zhang, Chen Liu

**Section One: Data Description**

**1.1 Introduction**:

The dataset was used in the 1983 American Statistical Association Exposition. (c) Date: July 7, 1993. It is selected with the aim of predicting fuel economy of cars, the attribute "mpg".

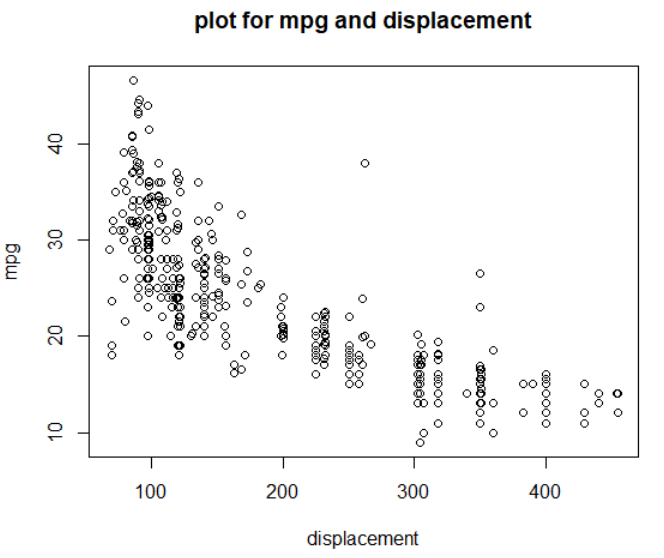
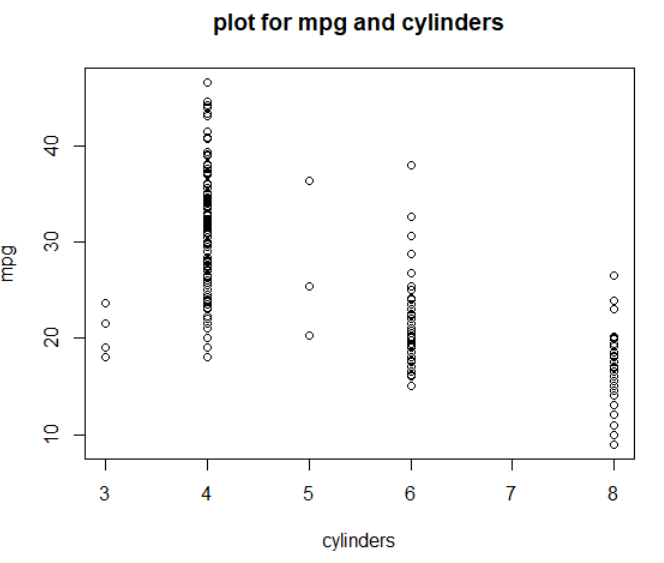
**Source**: Our dataset, which was originally taken from the StatLib library, can be found on Kaggle. [Auto-mpg dataset Mileage per gallon performances of various cars https://www.kaggle.com/uciml/autompg-dataset](https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset)

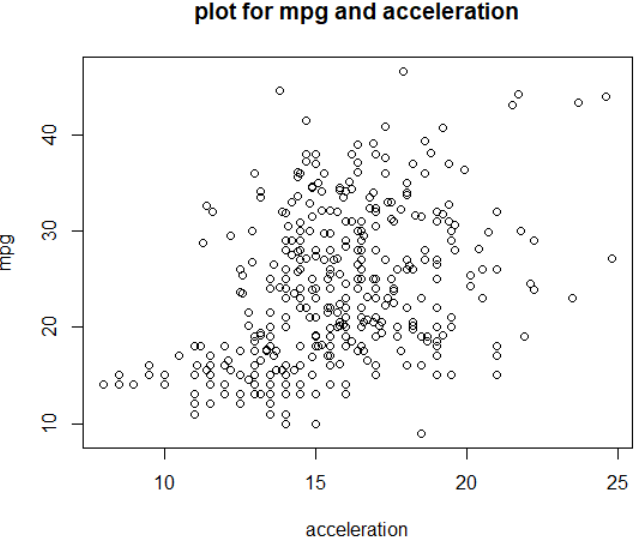
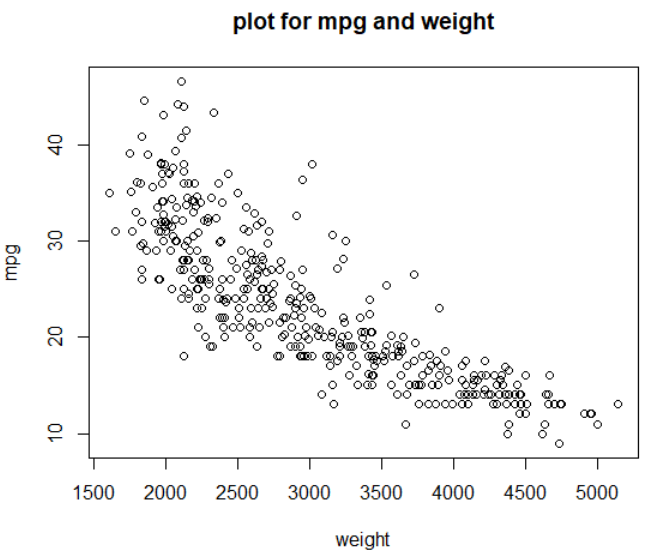
**Data size**: 398 observations in 1 response and 6 predictors.

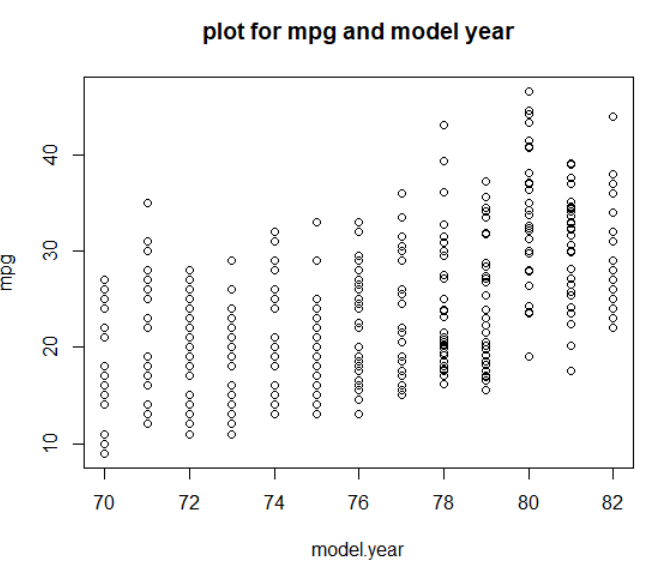
**1.2 Variable Description**:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Variable Name** | **Description** | **Summary** |
| **Response** | mpg | Fuel economy of cars: Miles per gallon | Min. : 9.00 Mean :23.51 Max. :46.60 |
| **Predictor X1** | cylinders | Discrete: number of cylinders, from 3 to 8 | Min. :3.000 Mean :5.455 Max. :8.000 |
| **Predictor X2** | displacement | Continuous in cc | Min. : 68.0 Mean :193.4 Max. :455.0 |
| **Predictor X3** | weight | Continuous in lb | Min. :1613 Mean :2970 Max. :5140 |
| **Predictor X4** | acceleration | Continuous: the acceleration time it takes from 0 to 60 mph, in seconds | Min. : 8.00 Mean :15.57 Max. :24.80 |
| **Predictor X5** | model.year | Discrete: from 70 to 82 | Min. :70.00 Mean :76.01 Max. :82.00 |
| **Predictor X6** | origin | Categorial: 1 = U.S., 2 = Europe, 3 = Japan | Min. :1.000 Mean :1.573 Max. :3.000 |

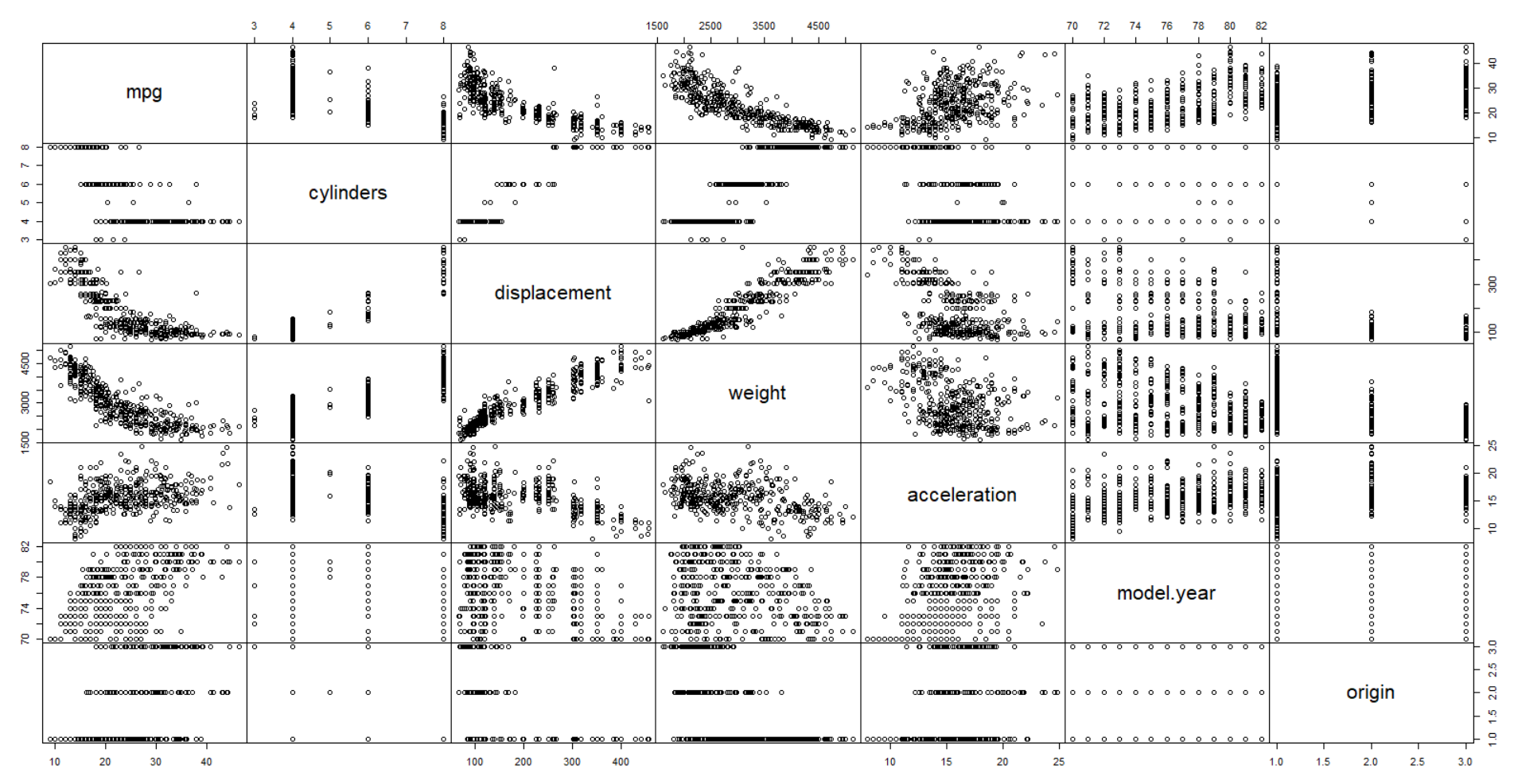
**1.3 Scatter Plot:**







**1.4 Pairs Plot:**



Note: Indicated by pairwise plot, there exists polynomial relationship between mpg and displacement, weight; check y transformation for acceleration.

**1.5 Correlation Matrix:**

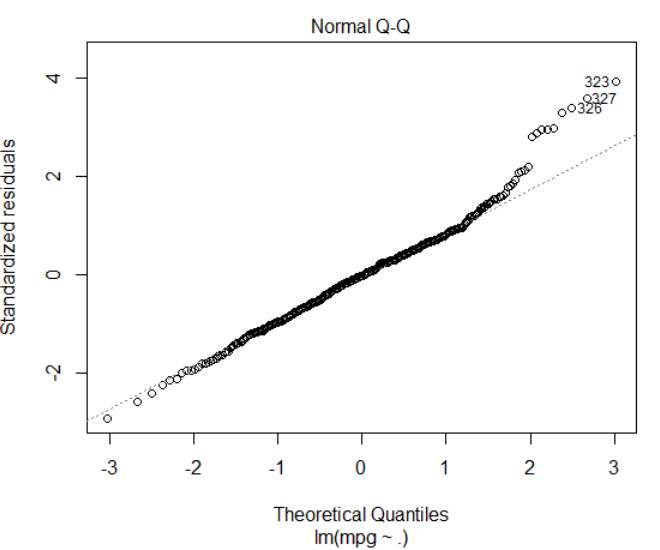
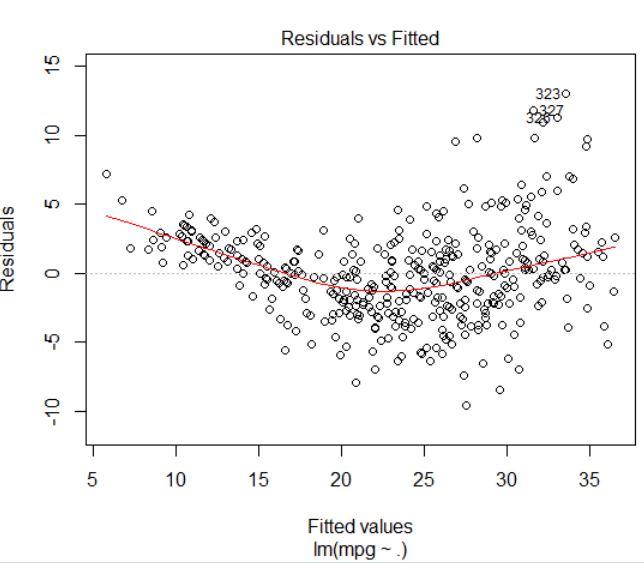
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | mpg | cylinders | displacement | weight | acceleration | model.year | origin |
| mpg | 1 | -0.775 | -0.804 | -0.832 | 0.42 | 0.579 | 0.563 |
| cylinders | -0.775 | 1 | 0.951 | 0.896 | -0.505 | -0.349 | -0.563 |
| displacement | -0.804 | 0.951 | 1 | 0.933 | -0.544 | -0.37 | -0.609 |
| weight | -0.832 | 0.896 | 0.933 | 1 | -0.417 | -0.307 | -0.581 |
| acceleration | 0.42 | -0.505 | -0.544 | -0.417 | 1 | 0.288 | 0.206 |
| model.year | 0.579 | -0.349 | -0.37 | -0.307 | 0.288 | 1 | 0.181 |
| origin | 0.563 | -0.563 | -0.609 | -0.581 | 0.206 | 0.181 | 1 |

**Section Two: Formulizing Model**

**2.1 First attempt: fit all and check plot**

**Related R code:**

Cars.noNA <-na.omit(Cars)  
head(Cars.noNA)  
#needs=c('mpg','cylinders','displacement','horsepower','weight','acceleration','model.year')  
Newdata <- Cars.noNA[c('mpg','cylinders','displacement','weight','acceleration','model.year','origin')]  
#as.numeric(levels(Newdata$horsepower))  
head(Newdata)  
summary(Newdata)  
pairs(Newdata,gap=0, cex.labels=2)  
cor(Newdata)MLRfull <- lm(mpg~.,data=Newdata)  
plot(MLRfull)

****

From residual plot, we can see the error is non-constant, as yhat increase, residuals variance increase, residuals show clear pattern. QQ-normality plot shows residuals are almost normal, with extreme outliers on the right tail. Both plots suggested there exist outliers and model need transformation.

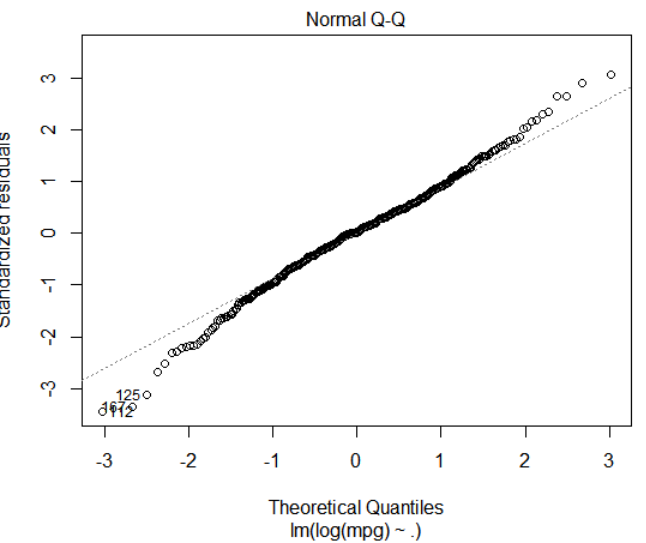
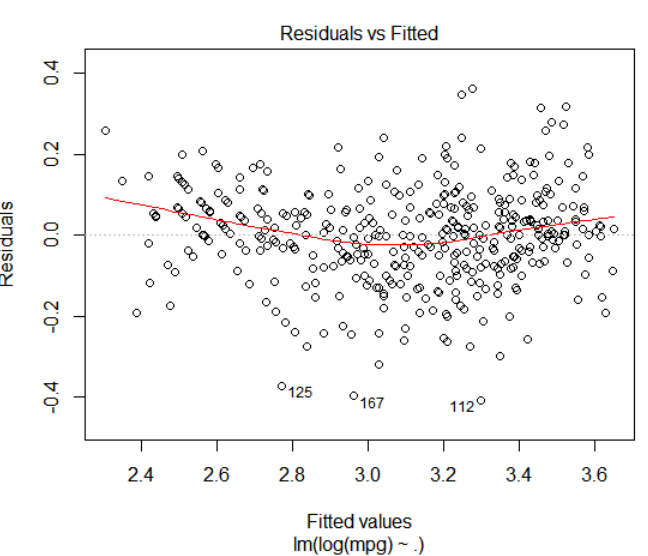
**2.2 Second attempt: transformation, higher terms and interaction**

***2.2.1 Try to transform response (mpg)***. Consider possible transformation (noticeable non-constant variance, in order to assure the error assumption, try transfer Y)

#log y transformation for mpg~.  
AICp(lm(mpg~., data=Newdata)) #965.984

AICp(lm(log(mpg)~., data=Newdata)) #-1680.741

summary(lm(log(mpg)~., data=Newdata))  
Call:  
lm(formula = log(mpg) ~ ., data = Newdata)  
Residuals:  
 Min 1Q Median 3Q Max   
-0.40765 -0.07100 0.00232 0.06919 0.36239   
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 1.5363834 0.1489278 10.316 < 2e-16 \*\*\*  
cylinders -0.0217701 0.0114962 -1.894 0.059004 .   
displacement 0.0003928 0.0002587 1.518 0.129741   
weight -0.0002901 0.0000208 -13.947 < 2e-16 \*\*\*  
acceleration 0.0050651 0.0027812 1.821 0.069337 .   
model.year 0.0307413 0.0017728 17.341 < 2e-16 \*\*\*  
origin 0.0344066 0.0096830 3.553 0.000427 \*\*\*  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  
Residual standard error: 0.12 on 391 degrees of freedom  
Multiple R-squared: 0.8771, Adjusted R-squared: 0.8752   
F-statistic: 464.9 on 6 and 391 DF, p-value: < 2.2e-16



# try sqrt y

AICp(lm(sqrt(mpg)~., data=Newdata)) #-929.1965, we will follow log transformation

#log y transformation shows great improvement in variacne stability, smaller AIC also suggested a good fit.

**Conclusion:** Due to non-constant variance and pattern in residual, in order to meet the error assumption, log transfer y.

***2.2.2 Try to transform x***

AICp(lm(log(mpg)~displacement)) # -1375.69

AICp(lm(log(mpg)~(1/displacement))) # -858.5388

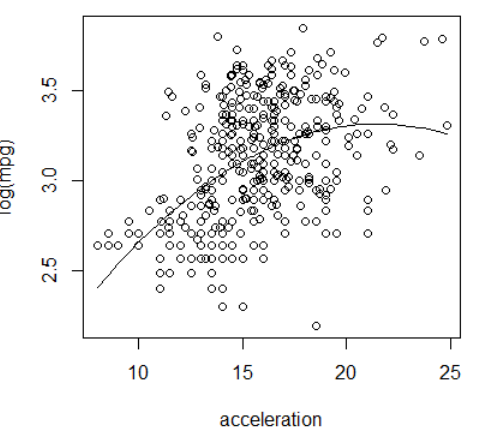
**Conclusion:** Nope, do not transfer x, at least not displacement

***2.2.3 Consider higher terms quadratic x^2***

#may exist polynomial transformation for log(mpg)~acceleration, but not strong

AICp(lm(log(mpg)~acceleration))# -944.7264

AICp(lm(log(mpg)~acceleration+I(acceleration^2)))#-957.1668

plot(y=log(mpg),x=acceleration)  
lines(sort(acceleration), fitted(lm(log(mpg)~acceleration+I(acceleration^2)))[order(acceleration)])   


summary(lm(log(mpg)~acceleration+I(acceleration^2)))  
Call:  
lm(formula = log(mpg) ~ acceleration + I(acceleration^2))  
Residuals:  
 Min 1Q Median 3Q Max   
-1.07126 -0.22527 -0.00066 0.21838 0.77803   
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 1.023320 0.331575 3.086 0.002170 \*\*   
acceleration 0.213095 0.041764 5.102 5.22e-07 \*\*\*  
I(acceleration^2) -0.004959 0.001298 -3.820 0.000155 \*\*\*  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  
Residual standard error: 0.2993 on 395 degrees of freedom  
Multiple R-squared: 0.2273, Adjusted R-squared: 0.2234   
F-statistic: 58.1 on 2 and 395 DF, p-value: < 2.2e-16

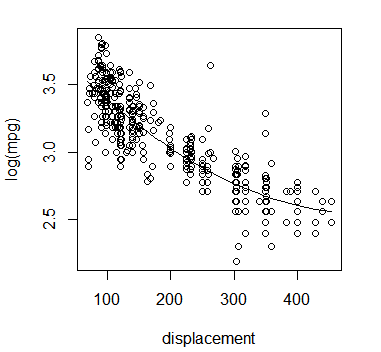
#may exist polynomial transformation for log(mpg)~displacement

AICp(lm(log(mpg)~displacement))#-1375.69

AICp(lm(log(mpg)~displacement+I(displacement^2)))# -1401.259

plot(y=log(mpg),x=displacement)

lines(sort(displacement),fitted(lm(log(mpg)~displacement+I(displacement^2)))[order(displacement)])



> summary(lm(log(mpg)~displacement+I(displacement^2)))

Call:

lm(formula = log(mpg) ~ displacement + I(displacement^2))

Residuals:

Min 1Q Median 3Q Max

-0.62116 -0.11040 0.00425 0.10885 0.78230

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.841e+00 4.202e-02 91.409 < 2e-16 \*\*\*

displacement -5.052e-03 4.346e-04 -11.625 < 2e-16 \*\*\*

I(displacement^2) 4.921e-06 9.245e-07 5.323 1.72e-07 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1713 on 395 degrees of freedom

Multiple R-squared: 0.7468, Adjusted R-squared: 0.7455

F-statistic: 582.6 on 2 and 395 DF, p-value: < 2.2e-16

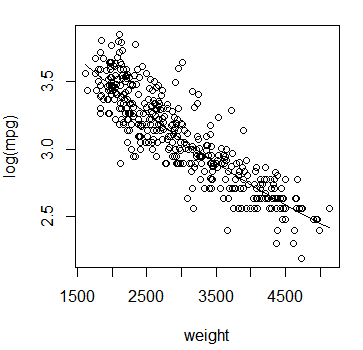
#may exist polynomial transformation for log(mpg)~weight

AICp(lm(log(mpg)~weight))# -1435.66

AICp(lm(log(mpg)~weight+I(weight^2)))#-1438.623

plot(y=log(mpg),x=weight)

lines(sort(weight), fitted(lm(log(mpg)~weight+I(weight^2)))[order(weight)])



> summary(lm(log(mpg)~displacement+I(displacement^2)))

Call:

lm(formula = log(mpg) ~ displacement + I(displacement^2))

Residuals:

Min 1Q Median 3Q Max

-0.62116 -0.11040 0.00425 0.10885 0.78230

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.841e+00 4.202e-02 91.409 < 2e-16 \*\*\*

displacement -5.052e-03 4.346e-04 -11.625 < 2e-16 \*\*\*

I(displacement^2) 4.921e-06 9.245e-07 5.323 1.72e-07 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1713 on 395 degrees of freedom

Multiple R-squared: 0.7468, Adjusted R-squared: 0.7455

F-statistic: 582.6 on 2 and 395 DF, p-value: < 2.2e-16

**Conclusion:** As indicated by pairwise plot, consider add acceleration^2, displacement^2, weight^2 to full model.

***2.2.4 Consider possible interactions***

From correlation matrix, there is strong correlation between cylinders, displacement and weight

vif(lm(mpg~., data=Newdata))  
# cylinders displacement weight acceleration model.year origin   
# 10.541162 20.057737 8.554710 1.621482 1.184443 1.662647   
max VIF\_displacement=20, excess of 10 is considered as indicator of multicolinearity

test interaction between cylinders, displacement and weight

summary(lm(mpg~displacement+cylinders+weight+displacement\*cylinders+displacement\*weight+cylinders\*weight))

Call:  
lm(formula = mpg ~ displacement + cylinders + weight + displacement \*   
 cylinders + displacement \* weight + cylinders \* weight)  
Residuals:  
 Min 1Q Median 3Q Max   
-13.2184 -2.5005 -0.3581 1.7894 17.9352   
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 4.901e+01 6.734e+00 7.278 1.87e-12 \*\*\*  
displacement -9.653e-02 3.923e-02 -2.461 0.01430 \*   
cylinders 1.916e+00 2.074e+00 0.924 0.35615   
weight -8.280e-03 3.017e-03 -2.745 0.00634 \*\*   
displacement:cylinders -1.856e-03 3.819e-03 -0.486 0.62730   
displacement:weight 2.541e-05 8.251e-06 3.079 0.00222 \*\*   
cylinders:weight -4.059e-04 6.721e-04 -0.604 0.54628   
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  
Residual standard error: 4.114 on 391 degrees of freedom  
Multiple R-squared: 0.7272, Adjusted R-squared: 0.723   
F-statistic: 173.7 on 6 and 391 DF, p-value: < 2.2e-16

# consider 3 way interaction and select interaction with most significant for further model

step(lm(mpg~displacement+cylinders+weight+displacement\*cylinders+displacement\*weight+cylinders\*weight+cylinders\*weight\*displacement))

Step: AIC=1128.09  
mpg ~ displacement + weight + displacement:weight  
 Df Sum of Sq RSS AIC  
<none> 6639.1 1128.1  
- displacement:weight 1 685.72 7324.8 1165.2  
Call:  
lm(formula = mpg ~ displacement + weight + displacement:weight)  
Coefficients:  
 (Intercept) displacement weight displacement:weight   
 5.395e+01 -7.936e-02 -8.998e-03 1.772e-05

# same as two way interaction significant results, suggest keep displacement\*weight

**Conclusion:** There is significant interaction between displacement, cylinders and weight, as collinearity test suggested, include displacement\*weight in final model.

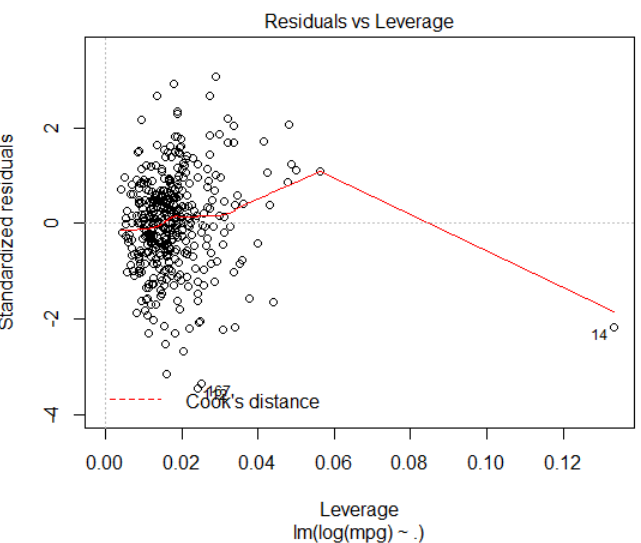
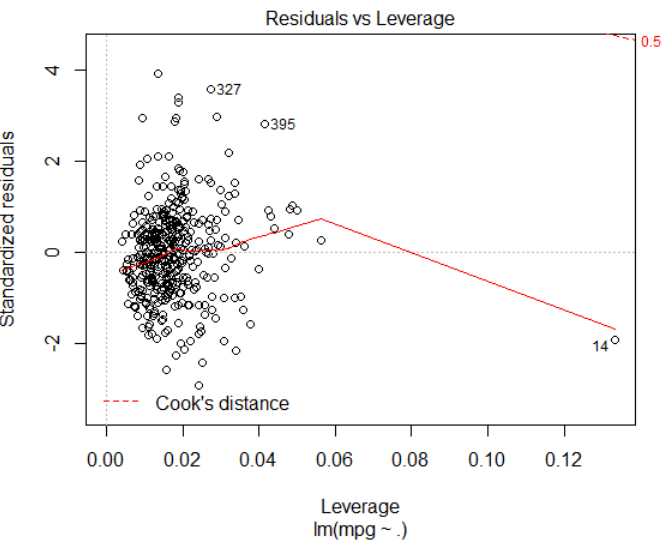
***2.2.5 Consider remove outliers***

Though Cook’s distance indicates no major outliers, there exist outliers with minor influential on regression model.

Test outliers influential in regression

summary(influence.measures(lm(mpg~., data=Newdata)))  
Potentially influential observations of lm(formula = mpg ~ ., data = Newdata) :

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | dfb.1\_ | dfb.cyln | dfb.dspl | dfb.wght | dfb.accl | dfb.mdl. | dfb.orgn | dffit | cov.r | cook.d | hat |
| 7 | 0.05 | -0.13 | 0.16 | -0.07 | -0.02 | -0.02 | 0.05 | 0.21 | 1.06\_\* | 0.01 | 0.05 |
| 14 | -0.1 | 0.26 | -0.65 | 0.62 | -0.14 | 0.01 | -0.12 | -0.76\_\* | 1.10\_\* | 0.08 | 0.13\_\* |
| 29 | 0.01 | 0.04 | -0.06 | 0.07 | 0.03 | -0.05 | 0 | 0.12 | 1.06\_\* | 0 | 0.04 |
| 112 | -0.24 | 0.16 | 0.04 | -0.15 | 0.22 | 0.18 | -0.16 | -0.46\_\* | 0.90\_\* | 0.03 | 0.02 |
| 167 | 0.01 | -0.23 | -0.01 | 0.25 | 0.05 | -0.05 | 0.05 | -0.39 | 0.94\_\* | 0.02 | 0.02 |
| 212 | 0.01 | -0.03 | 0.06 | -0.06 | 0.01 | 0.01 | -0.01 | -0.08 | 1.06\_\* | 0 | 0.04 |
| 223 | -0.01 | 0.02 | -0.01 | 0.01 | 0.01 | 0 | 0 | 0.02 | 1.06\_\* | 0 | 0.04 |
| 245 | -0.09 | 0.05 | 0.13 | -0.22 | 0.37 | 0.02 | 0.05 | 0.46\_\* | 0.85\_\* | 0.03 | 0.02 |
| 271 | 0.05 | 0.15 | -0.11 | -0.03 | 0.05 | -0.06 | -0.24 | -0.33 | 0.92\_\* | 0.02 | 0.02 |
| 300 | -0.03 | -0.03 | 0.02 | 0.02 | 0.07 | 0 | 0.02 | 0.09 | 1.07\_\* | 0 | 0.05 |
| 301 | -0.03 | 0.04 | 0 | -0.02 | 0.04 | 0.01 | 0 | 0.06 | 1.08\_\* | 0 | 0.06\_\* |
| 310 | -0.05 | 0.02 | -0.05 | -0.02 | -0.15 | 0.14 | -0.03 | 0.29 | 0.88\_\* | 0.01 | 0.01 |
| 323 | -0.22 | 0.06 | 0.05 | -0.08 | 0.12 | 0.16 | 0.31 | 0.47\_\* | 0.78\_\* | 0.03 | 0.01 |
| 326 | -0.19 | 0.04 | 0.12 | -0.18 | 0.38 | 0.11 | 0.05 | 0.48\_\* | 0.84\_\* | 0.03 | 0.02 |
| 327 | -0.25 | 0.03 | 0.09 | -0.08 | 0.53 | 0.09 | 0.09 | 0.61\_\* | 0.83\_\* | 0.05 | 0.03 |
| 328 | -0.19 | 0.16 | -0.2 | 0.18 | 0.14 | 0.09 | 0.05 | 0.39 | 0.89\_\* | 0.02 | 0.02 |
| 330 | -0.09 | 0.04 | 0.04 | -0.13 | -0.14 | 0.16 | 0.19 | 0.41\_\* | 0.89\_\* | 0.02 | 0.02 |
| 335 | 0.02 | 0.13 | 0.08 | -0.22 | 0.26 | -0.09 | -0.12 | -0.40\_\* | 0.97 | 0.02 | 0.03 |
| 388 | -0.28 | -0.14 | 0.37 | -0.32 | 0.21 | 0.32 | 0 | 0.52\_\* | 0.89\_\* | 0.04 | 0.03 |
| 395 | -0.29 | 0 | 0.22 | -0.22 | 0.51 | 0.17 | 0.1 | 0.59\_\* | 0.92\_\* | 0.05 | 0.04 |

****

Notice log y transformation (right as transformed, left as original) cannot reduced extreme outlying observation 14, influential test indicates observation 14 could be a outlier with considerable influential in regression, remove observation 14.

# Other potential outliers? check influence on log regression model

summary(influence.measures(lm(log(mpg)~., data=Newdata)))  
Potentially influential observations of lm(formula = log(mpg) ~ ., data = Newdata) :

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | dfb.1\_ | dfb.cyln | dfb.dspl | dfb.wght | dfb.accl | dfb.mdl. | dfb.orgn | dffit | cov.r | cook.d | hat |
| 14 | -0.12 | 0.29 | -0.74 | 0.7 | -0.16 | 0.01 | -0.14 | -0.86\_\* | 1.08\_\* | 0.1 | 0.13\_\* |
| 96 | 0 | -0.06 | 0.05 | 0.01 | 0 | 0 | 0.02 | 0.08 | 1.06\_\* | 0 | 0.04 |
| 109 | -0.07 | 0.01 | -0.03 | 0.01 | -0.15 | 0.15 | -0.2 | -0.32 | 0.94\_\* | 0.01 | 0.02 |
| 112 | -0.28 | 0.19 | 0.05 | -0.17 | 0.26 | 0.21 | -0.19 | -0.55\_\* | 0.84\_\* | 0.04 | 0.02 |
| 125 | -0.06 | -0.06 | -0.14 | 0.22 | 0.08 | 0.01 | -0.01 | -0.40\_\* | 0.87\_\* | 0.02 | 0.02 |
| 156 | 0.03 | 0.08 | -0.24 | 0.19 | -0.32 | 0.03 | 0.01 | -0.39 | 0.91\_\* | 0.02 | 0.02 |
| 167 | 0.02 | -0.32 | -0.02 | 0.35 | 0.07 | -0.06 | 0.06 | -0.54\_\* | 0.85\_\* | 0.04 | 0.02 |
| 212 | 0.01 | -0.03 | 0.07 | -0.07 | 0.01 | 0.01 | -0.01 | -0.09 | 1.06\_\* | 0 | 0.04 |
| 216 | 0.08 | -0.14 | 0.01 | 0.08 | -0.02 | -0.06 | 0 | -0.25 | 0.94\_\* | 0.01 | 0.01 |
| 245 | -0.07 | 0.03 | 0.09 | -0.16 | 0.26 | 0.01 | 0.03 | 0.33 | 0.94\_\* | 0.02 | 0.02 |
| 271 | 0.05 | 0.15 | -0.11 | -0.03 | 0.05 | -0.06 | -0.23 | -0.32 | 0.92\_\* | 0.01 | 0.02 |
| 300 | -0.06 | -0.07 | 0.05 | 0.05 | 0.15 | 0.01 | 0.05 | 0.19 | 1.06\_\* | 0.01 | 0.05 |
| 301 | -0.12 | 0.16 | -0.02 | -0.08 | 0.18 | 0.05 | 0 | 0.26 | 1.06\_\* | 0.01 | 0.06\_\* |
| 310 | -0.04 | 0.01 | -0.03 | -0.02 | -0.11 | 0.1 | -0.02 | 0.21 | 0.95\_\* | 0.01 | 0.01 |
| 323 | -0.15 | 0.04 | 0.03 | -0.05 | 0.08 | 0.11 | 0.21 | 0.31 | 0.91\_\* | 0.01 | 0.01 |
| 326 | -0.13 | 0.03 | 0.08 | -0.12 | 0.25 | 0.08 | 0.04 | 0.32 | 0.94\_\* | 0.01 | 0.02 |
| 327 | -0.18 | 0.02 | 0.07 | -0.06 | 0.39 | 0.07 | 0.06 | 0.45\_\* | 0.92\_\* | 0.03 | 0.03 |
| 328 | -0.2 | 0.16 | -0.21 | 0.18 | 0.14 | 0.09 | 0.05 | 0.4 | 0.89\_\* | 0.02 | 0.02 |
| 335 | 0.02 | 0.13 | 0.09 | -0.22 | 0.27 | -0.09 | -0.12 | -0.41\_\* | 0.97 | 0.02 | 0.03 |
| 365 | -0.28 | 0.02 | 0.26 | -0.26 | 0.31 | 0.22 | 0.09 | 0.46\_\* | 0.99 | 0.03 | 0.05 |
| 388 | -0.29 | -0.14 | 0.38 | -0.32 | 0.22 | 0.33 | 0 | 0.53\_\* | 0.88\_\* | 0.04 | 0.03 |

summary(lm(log(mpg)~., data=Newdata))$fstatistic  
# Fvalue numdf dendf   
# 464.863 6.000 391.000   
summary(lm(log(mpg)~., data=Newdata))$adj.r.squared  
#adjusted R square 0.8751642  
AICp(lm(log(mpg)~., data=Newdata)) #-1680.741

summary(lm(log(mpg)~., data=Newdata, subset=c(-14,-167,-327, -388)))$fstatistic  
# Fvalue numdf dendf   
# 492.1659 6.0000 387.0000  
summary(lm(log(mpg)~., data=Newdata, subset=c(-14,-167,-327, -388)))$adj.r.squared  
#adjusted R square 0.8823352  
AICp(lm(log(mpg)~., data=Newdata, subset=c(-14,-167,-327, -388))) #-1693.884

**Conclusion:** Compared potential outliers’ influence on regression model, suggest remove outlying observation 14, 167, 327, 388 due to largest DIFFT, cov.r and cook’s value.

**2.3 Third attempt: consider reduction of variables**

#Original dataset

summary(lm(mpg~., data=Newdata)

Call:  
lm(formula = mpg ~ ., data = Newdata)  
Residuals:  
 Min 1Q Median 3Q Max   
-9.5773 -2.1658 -0.0487 1.8191 13.0069   
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) -1.976e+01 4.140e+00 -4.772 2.58e-06 \*\*\*  
cylinders -3.848e-01 3.196e-01 -1.204 0.2294   
displacement 1.705e-02 7.192e-03 2.371 0.0182 \*   
weight -6.958e-03 5.783e-04 -12.031 < 2e-16 \*\*\*  
acceleration 1.520e-01 7.732e-02 1.966 0.0500 \*   
model.year 7.659e-01 4.928e-02 15.541 < 2e-16 \*\*\*  
origin 1.369e+00 2.692e-01 5.086 5.69e-07 \*\*\*  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  
Residual standard error: 3.336 on 391 degrees of freedom  
Multiple R-squared: 0.8206, Adjusted R-squared: 0.8178   
F-statistic: 298 on 6 and 391 DF, p-value: < 2.2e-16

# AICp=965.984

**2.4 Model selection**

***2.4.1 Model selection with log y transformation, I(acceleration^2) , I(displacement^2), I(weight^2) and interaction between displacement\*weight***

MLRnull <- lm(log(mpg)~1,data=Newdata)  
MLRfull<-lm(log(mpg)~.+I(acceleration^2)+I(displacement^2)+I(weight^2)+displacement\*weight,data=Newdata)  
step(MLRfull,data=Newdata, direction="backward")

Start: AIC=-1719.1

log(mpg) ~ cylinders + displacement + weight + acceleration +

model.year + origin + I(acceleration^2) + I(displacement^2) +

I(weight^2) + displacement \* weight

Df Sum of Sq RSS AIC

- cylinders 1 0.0038 5.0157 -1720.8

- I(weight^2) 1 0.0051 5.0171 -1720.7

- I(displacement^2) 1 0.0173 5.0292 -1719.7

- origin 1 0.0205 5.0325 -1719.5

<none> 5.0120 -1719.1

- displacement:weight 1 0.0552 5.0671 -1716.7

- acceleration 1 0.0880 5.0999 -1714.2

- I(acceleration^2) 1 0.1122 5.1242 -1712.3

- model.year 1 4.5579 9.5698 -1463.7

Step: AIC=-1720.8

log(mpg) ~ displacement + weight + acceleration + model.year +

origin + I(acceleration^2) + I(displacement^2) + I(weight^2) +

displacement:weight

Df Sum of Sq RSS AIC

- I(weight^2) 1 0.0069 5.0226 -1722.3

- I(displacement^2) 1 0.0140 5.0297 -1721.7

- origin 1 0.0174 5.0331 -1721.4

<none> 5.0157 -1720.8

- displacement:weight 1 0.0562 5.0720 -1718.4

- acceleration 1 0.0849 5.1006 -1716.1

- I(acceleration^2) 1 0.1090 5.1247 -1714.2

- model.year 1 4.5580 9.5737 -1465.5

Step: AIC=-1722.26

log(mpg) ~ displacement + weight + acceleration + model.year +

origin + I(acceleration^2) + I(displacement^2) + displacement:weight

Df Sum of Sq RSS AIC

- I(displacement^2) 1 0.0075 5.0301 -1723.7

- origin 1 0.0166 5.0392 -1722.9

<none> 5.0226 -1722.3

- acceleration 1 0.0897 5.1123 -1717.2

- I(acceleration^2) 1 0.1133 5.1359 -1715.4

- displacement:weight 1 0.1762 5.1988 -1710.5

- model.year 1 4.6784 9.7010 -1462.3

Step: AIC=-1723.66

log(mpg) ~ displacement + weight + acceleration + model.year +

origin + I(acceleration^2) + displacement:weight

Df Sum of Sq RSS AIC

- origin 1 0.0130 5.0431 -1724.6

<none> 5.0301 -1723.7

- acceleration 1 0.0857 5.1158 -1718.9

- I(acceleration^2) 1 0.1117 5.1418 -1716.9

- displacement:weight 1 0.4601 5.4902 -1690.8

- model.year 1 4.6732 9.7033 -1464.2

Step: AIC=-1724.63

log(mpg) ~ displacement + weight + acceleration + model.year +

I(acceleration^2) + displacement:weight

Df Sum of Sq RSS AIC

<none> 5.0431 -1724.6

- acceleration 1 0.0924 5.1356 -1719.4

- I(acceleration^2) 1 0.1176 5.1607 -1717.5

- displacement:weight 1 0.5710 5.6141 -1683.9

- model.year 1 4.6733 9.7165 -1465.6

Call:

lm(formula = log(mpg) ~ displacement + weight + acceleration +

model.year + I(acceleration^2) + displacement:weight, data = Newdata)

Coefficients:

(Intercept) displacement weight acceleration

2.204e+00 -2.093e-03 -3.800e-04 -5.011e-02

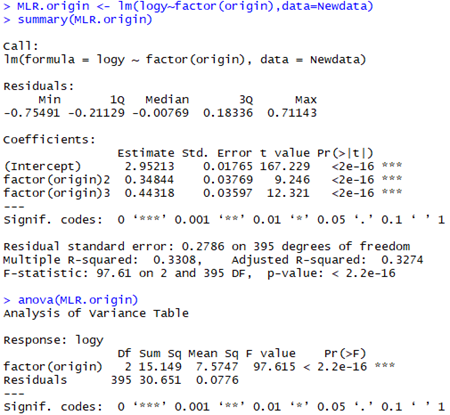
model.year I(acceleration^2) displacement:weight

3.216e-02 1.676e-03 5.293e-07

Though smallest AIC suggested final model as log(mpg) ~ displacement + weight + acceleration +   
 model.year + I(acceleration^2) + displacement:weight, we’d like to keep origin (AICp without origin -1724.63 is almost same as AICp with origin -1723.66). The reason we decided to keep origin in the model is based on common knowledge, we all believe Japanese cars have better mpg performance European and American cars due to its fuel efficiency and lightweight. We’d like to perform one-way anova test to prove that.

As many may have heard, it is likely that the difference of cars’ origin plays a role in predicting cars’ fuel economy. For instance, nowadays, U.S. cars tend to have a reputation with lower mpg, while Japanese cars tend to have higher. Thus, it is natural that we want to find out if this reputation can be justified in our dataset.

Let us take a one-way ANOVA test to check the effect of car origins across regions, considering “origin” as the only factor here:



From the output, we see very low P-value (2.2 x 10-16). It means there is significant difference in cars’ mpg attribute across those origins. Also, the coefficients of origin = 2 and origin = 3 are both > 0 (0.348 and 0.443, respectively). Since U.S. cars are considered as reference, this result is consistent with the “common knowledge” that U.S. cars tend to have lower mpg performance, compared with European and Japanese cars.

All in all, we decide to keep the categorical predictor, “origin”, in the model, regardless of P-value.

**2.4.2 Conclusion:** **Suggested model with reduced variables**

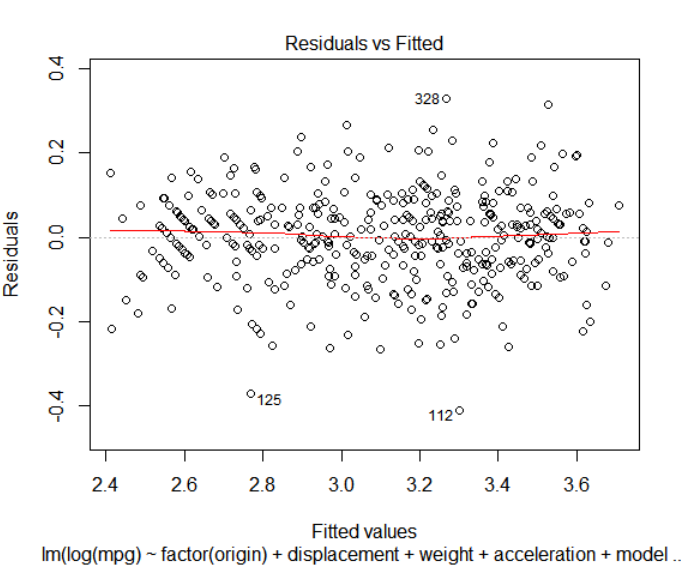
***summary(lm(log(mpg) ~ factor(origin)+displacement + weight + acceleration + model.year + I(acceleration^2) + displacement:weight, data = Newdata, subset=c(-14,-167,-327, -388)))***Call:  
lm(formula = log(mpg) ~ factor(origin) + displacement + weight +   
 acceleration + model.year + I(acceleration^2) + displacement:weight,   
 data = Newdata, subset = c(-14, -167, -327, -388))  
Residuals:  
 Min 1Q Median 3Q Max   
-0.41045 -0.06355 0.00599 0.06321 0.32882   
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 2.113e+00 2.046e-01 10.327 < 2e-16 \*\*\*  
factor(origin)2 3.601e-02 1.965e-02 1.833 0.0676 .   
factor(origin)3 2.270e-02 1.889e-02 1.202 0.2302   
displacement -1.681e-03 3.739e-04 -4.494 9.25e-06 \*\*\*   
weight -3.818e-04 2.613e-05 -14.612 < 2e-16 \*\*\*   
acceleration -4.133e-02 1.920e-02 -2.153 0.0319 \*   
model.year 3.191e-02 1.671e-03 19.099 < 2e-16 \*\*\*   
I(acceleration^2) 1.385e-03 5.707e-04 2.427 0.0157 \*   
displacement:weight 4.691e-07 8.298e-08 5.654 3.06e-08 \*\*\*   
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  
Residual standard error: 0.1095 on 385 degrees of freedom  
Multiple R-squared: 0.8964, Adjusted R-squared: 0.8943   
F-statistic: 416.5 on 8 and 385 DF, p-value: < 2.2e-16

AICp(lm(log(mpg) ~ factor(origin)+displacement + weight + acceleration + model.year + I(acceleration^2) + displacement:weight, data = Newdata, subset=c(-14,-167,-327, -388))) #-1734.029

**Section Three : Model Assumption**

**3.1 constant variance, error mean is 0:**

plot(log(mpg) ~ factor(origin)+displacement + weight + acceleration + model.year + I(acceleration^2) + displacement:weight, data = Newdata, subset=c(-14,-167,-327, -388))

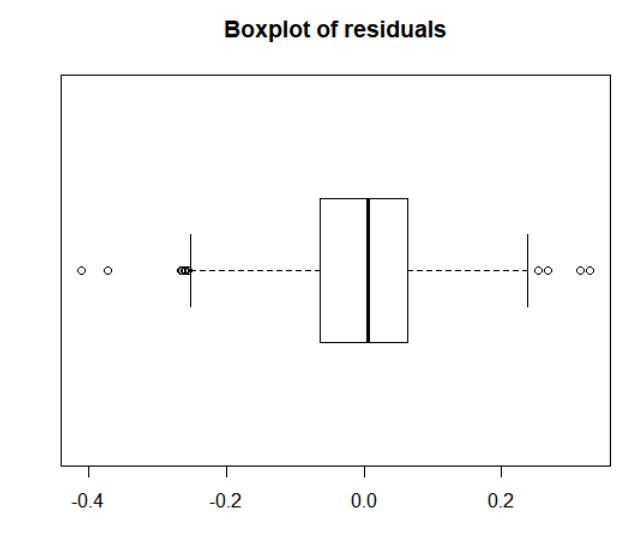
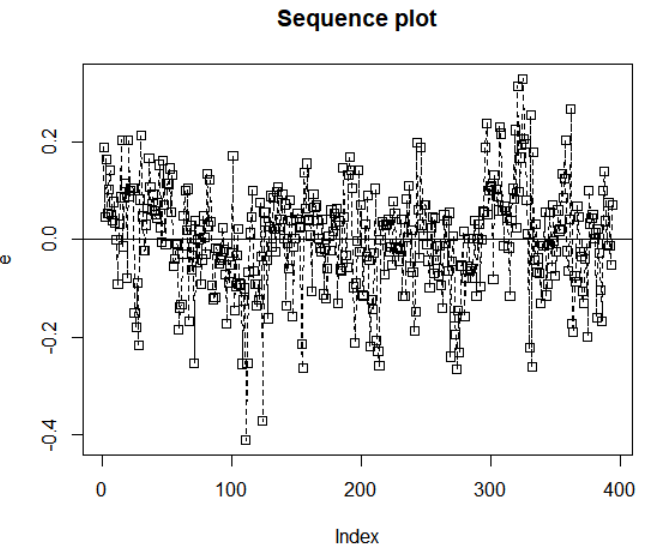


From residual plot against fitted value, we can see all errors are centered around 0, with mean of the distribution of errors as 0. And the residuals seems to be constant as fitted value increase. Also, there is no pattern for residuals, errors all seems to be randomly distributed, no independence on fitted values.

Further check on residual

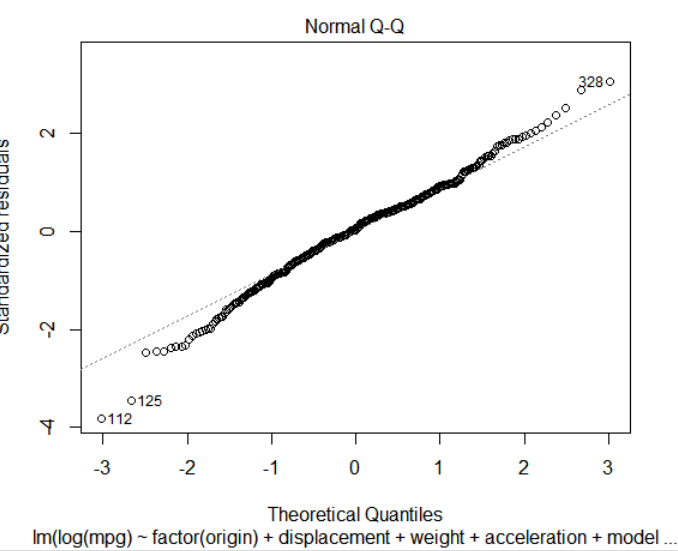
# residual plot and boxplot

plot( e, main="Sequence plot", pch=0)  
lines( e, main="Sequence plot", lty=2)  
abline(h=0)



boxplot(e, horizontal = TRUE, main="Boxplot of residuals")

# residuals are randan, constant, normal with mean as 0. There are handful of outliers.

**3.2 normality check:**  


QQ Normality plots indicates that errors are all lined up in diagonal line, though there are couple of outliers but no significant skewness, which means errors have a normal distribution ∼ N(0, σ2). Proved by following shapiro test against residuals: small p-value as residuals are normal.

shapiro.test(residuals(lm(log(mpg) ~ factor(origin)+displacement + weight + acceleration + model.year + I(acceleration^2) + displacement:weight, data = Newdata, subset=c(-14,-167,-327, -388))))  
# Shapiro-Wilk normality test  
# data: residuals(lm(log(mpg) ~ factor(origin) + displacement + weight + acceleration + model.year + I(acceleration^2) + displacement:weight, data = Newdata, subset = c(-14, -167, -327, -388)))  
# W = 0.98969, p-value = 0.007198

# further normaility check

e.rank <- rank(e)

z <- (e.rank - 0.375) / (394 + 0.25)

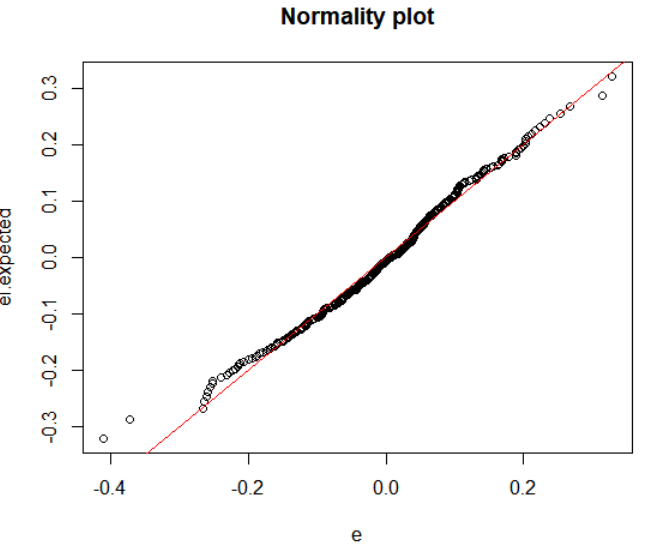
ei.expected <- sqrt(sum(e^2)/ (394 - 2 )) \* qnorm(z)

ei.expected

plot(e, ei.expected, main="Normality plot")

abline(a=0, b=1, col="red")

cbind(e, e.rank, ei.expected)



**3.3 Independence check:**

The errors associated to different observations are independent. check residuals against variables

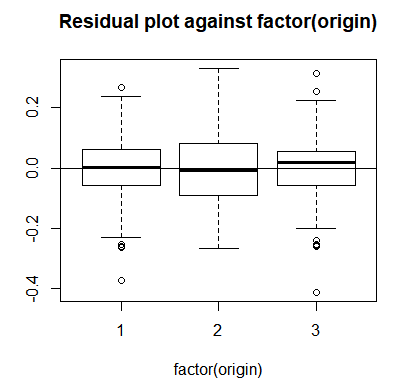
# residual plot against factor(origin)

x <- I(factor(origin)[c(-14, -167, -327, -388)])

e <- residuals(lm(log(mpg) ~ factor(origin) + displacement + weight + acceleration + model.year + I(acceleration^2) + displacement:weight, data = Newdata, subset = c(-14, -167, -327, -388)))

plot(x, e, main="Residual plot against factor(origin)" , xlab='factor(origin)')

abline(h=0)



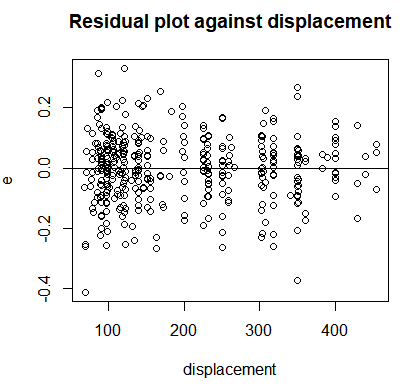
# residual plot against displacement

x <- I(displacement[c(-14, -167, -327, -388)])

e <- residuals(lm(log(mpg) ~ factor(origin) + displacement + weight + acceleration + model.year + I(acceleration^2) + displacement:weight, data = Newdata, subset = c(-14, -167, -327, -388)))

plot(x, e, main="Residual plot against displacement" , xlab='displacement')

abline(h=0)



From the residual plot against displacement, all errors are centered around zero within a band of (-0.4, 0.4). Error seems to be constant and randomly distributed as weight increased. Errors are independent against displacement.

# residual plot against weight

x <- I(weight[c(-14, -167, -327, -388)])

e <- residuals(lm(log(mpg) ~ factor(origin) + displacement + weight + acceleration + model.year + I(acceleration^2) + displacement:weight, data = Newdata, subset = c(-14, -167, -327, -388)))

plot(x, e, main="Residual plot against weight" , xlab='weight')

abline(h=0)



From the residual plot against weight, all errors are centered around zero within a band of (-0.4, 0.4). Error seems to be constant and randomly distributed as weight increased.Errors are independent against weight.

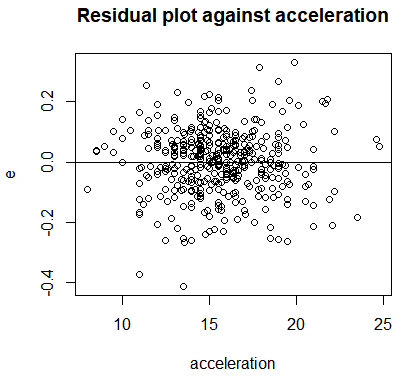
# residual plot against acceleration

x <- I(acceleration[c(-14, -167, -327, -388)])

e <- residuals(lm(log(mpg) ~ factor(origin) + displacement + weight + acceleration + model.year + I(acceleration^2) + displacement:weight, data = Newdata, subset = c(-14, -167, -327, -388)))

plot(x, e, main="Residual plot against acceleration" , xlab='acceleration')

abline(h=0)



From the residual plot against acceleration, all errors are centered around zero within a band of (-0.4, 0.4). Error seems to be constant and randomly distributed as acceleration increased.Errors are independent against acceleration.

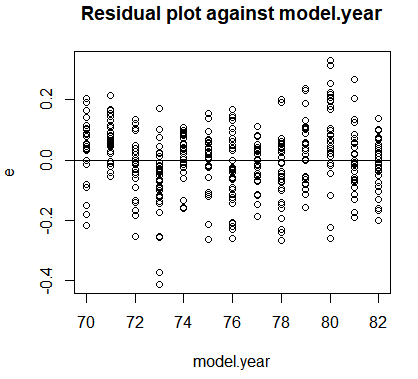
# residual plot against model.year

x <- I(model.year[c(-14, -167, -327, -388)])

e <- residuals(lm(log(mpg) ~ factor(origin) + displacement + weight + acceleration + model.year + I(acceleration^2) + displacement:weight, data = Newdata, subset = c(-14, -167, -327, -388)))

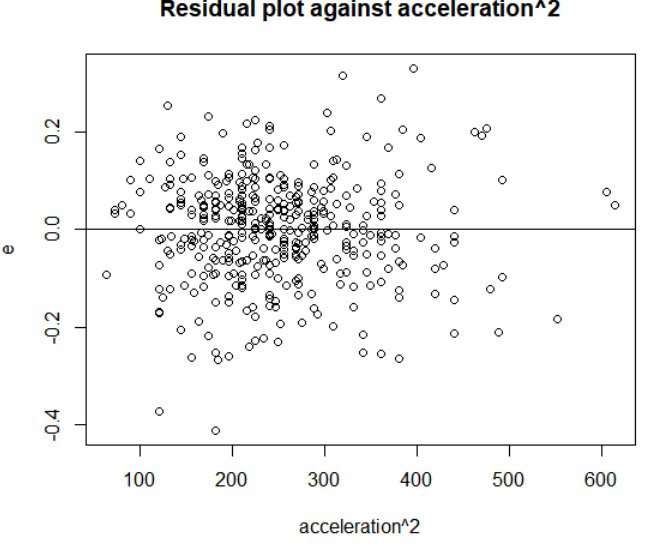
plot(x, e, main="Residual plot against model.year" , xlab='model.year')

abline(h=0)



# There is no pattern in residual plots against model.year, all errors are centered around 0 within a band of +/-0.4, errors seem to be constant and randomly distributed as model.year increased, errors are independent against model.year.

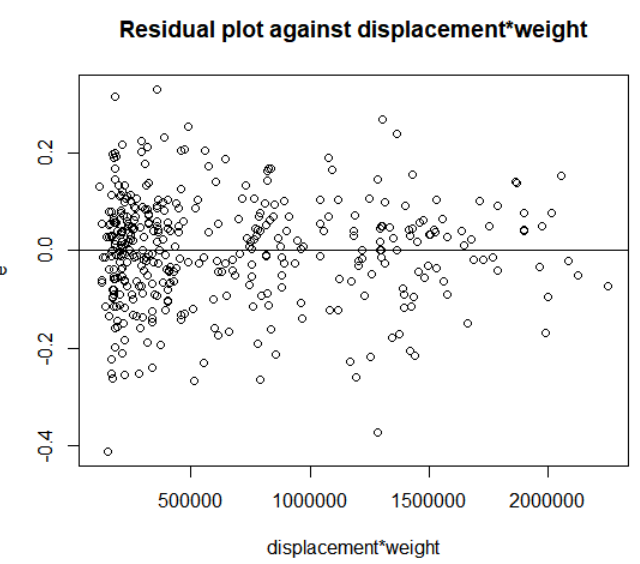
# residual plot against acceleration^2  
x <- I(acceleration[c(-14, -167, -327, -388)]^2)  
e <- residuals(lm(log(mpg) ~ factor(origin) + displacement + weight + acceleration + model.year + I(acceleration^2) + displacement:weight, data = Newdata, subset = c(-14, -167, -327, -388)))  
plot(x, e, main="Residual plot against acceleration^2" , xlab='acceleration^2')  
abline(h=0)



# There is no pattern in residual plots against acceleration^2, all errors are centered around 0 within a band of +/-0.4, errors seem to be constant and randomly distributed as acceleration^2 increased, errors are independent against acceleration^2.

# residual plot against displacement\*weight

x <- displacement[c(-14, -167, -327, -388)]\*weight[c(-14, -167, -327, -388)]  
e <- residuals(lm(log(mpg) ~ factor(origin) + displacement + weight + acceleration + model.year + I(acceleration^2) + displacement:weight, data = Newdata, subset = c(-14, -167, -327, -388)))  
plot(x, e, main="Residual plot against displacement\*weight", xlab='displacement\*weight')  
abline(h=0)



# There is no pattern in residual plots against interaction, all errors are centered around 0 within a band of +/-0.4, errors seem to be constant and randomly distributed as interaction increased, errors are independent against interaction.

**Section Four: Model Test**

**4.1 Final model Hypothesis testing steps and results**

#anova regression test for fitting

anova(lm(log(mpg) ~ factor(origin)+displacement + weight + acceleration + model.year + I(acceleration^2) + displacement:weight, data = Newdata, subset=c(-14,-167,-327, -388)))  
Analysis of Variance Table  
Response: log(mpg)  
 Df Sum Sq Mean Sq F value Pr(>F)   
factor(origin) 2 14.8543 7.4272 619.4372 < 2.2e-16 \*\*\*  
displacement 1 18.3490 18.3490 1530.3339 < 2.2e-16 \*\*\*  
weight 1 1.9957 1.9957 166.4445 < 2.2e-16 \*\*\*  
acceleration 1 0.0663 0.0663 5.5319 0.019176 \*   
model.year 1 4.1800 4.1800 348.6186 < 2.2e-16 \*\*\*  
I(acceleration^2) 1 0.1188 0.1188 9.9108 0.001771 \*\*   
displacement:weight 1 0.3833 0.3833 31.9661 3.061e-08 \*\*\*  
Residuals 385 4.6162 0.0120   
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

F-statistic: 416.5 on 8 and 385 DF, p-value: < 2.2e-16

Hypothesis: H0 : Beta1=Beta2=....=Betap=0 vs Hα : at least one Beta ≠ 0

Test statistic: 

FS=MSR/MSE=Fp-1,N-p

Fs ≈ 416.5

Decision rule: Fs > F(.95,8, 385) =2.224; pvalue =2.2\*10^-16 < 0.05

Conclusion: Reject H0, conclude alternative hypothesis, there is at least one Beta not 0. Large F value indicates regression relationship is significant.

# R^2 and adjusted R^2

Multiple R-squared: 0.8964, Adjusted R-squared: 0.8943

Both value are relatively high, close to 1, higher R^2 suggest model is considerably good.

**4.2 Statistic test**  
Test hypothesis about each coefficient, list confidence interval

factor(origin)2 3.601e-02 1.965e-02 1.833 0.0676 .   
factor(origin)3 2.270e-02 1.889e-02 1.202 0.2302   
displacement -1.681e-03 3.739e-04 -4.494 9.25e-06 \*\*\*   
weight -3.818e-04 2.613e-05 -14.612 < 2e-16 \*\*\*   
acceleration -4.133e-02 1.920e-02 -2.153 0.0319 \*   
model.year 3.191e-02 1.671e-03 19.099 < 2e-16 \*\*\*   
I(acceleration^2) 1.385e-03 5.707e-04 2.427 0.0157 \*   
displacement:weight 4.691e-07 8.298e-08 5.654 3.06e-08 \*\*\*

**T-test for factor of origin**

Hypothesis: H0: 𝛾origin2 = 𝛾origin3 = 0. VS Ha: at least one of the coefficients ≠ 0.

Test statistic:

TSorigin2 = 1.833, P-value2 = 0.0676

TSorigin3 = 1.202, P-value3 = 0.2302

Decision rule: if P-value < α, then reject the null hypothesis(H0); if P-value > α, the do not reject the null hypothesis. Here, α = 0.05

P-value3 > P-value2 > 0.05

Conclusion: statistically, the null hypothesis is not rejected. It means that, at the level of α = 0.05, the difference among the origins does not appear to play an important role in predicting cars’ mpg.

However, P-value2 = 0.0676, which is not a lot bigger than 0.05. And, more importantly, from general knowledge we know origin may be a factor in explaining cars’ mpg. Thus, to avoid loss of any potential important information, we would like to keep origin in the model.

**Interpretation:**

In this data, cars whose origin were Europe, on average, have higher mpg – exp(0.036) = 1.037 times of those whose origin were U.S., holding other variables constant.

In this data, cars whose origin were Japan, on average, have higher mpg – exp(0.023) = 1.023 times of those whose origin were U.S., holding other variables constant.

#displacement\_Beta =-1.681\*10^-03

T-test for predictor of displacement:

Hypothesis: H0 : Beta for displacement =0 vs Hα : Beta for displacement ≠ 0

Test statistic: 

Ts ≈ -4.494

Decision rule: |Ts| > t (.975, 385) = 1.966

Conclusion: Reject H0, there is a significant relationship between displacement and log(mpg).

Interpretation:

In this data, per each 1-unit increase in displacement, on average, we expect 1.681\*10^-03 decrease in response log(mpg).

#weight\_Beta =-3.818\*10^-4

T-test for predictor of weight:

Hypothesis: H0 : Beta for weight =0 vs Hα : Beta for weight ≠ 0

Test statistic: 

Ts ≈ -14.612

Decision rule: |Ts| > t (.975, 385) = 1.966

Conclusion: Reject H0, there is a significant relationship between weight and log(mpg).

Interpretation:

In this data, per each 1-unit increase in weight, on average, we expect 3.818×10^4 decrease in response log(mpg).

#acceleration\_Beta = -4.133\*10^-2

T-test for predictor of acceleration:

Hypothesis: H0 : Beta for acceleration =0 vs Hα : Beta for acceleration ≠ 0

Test statistic: 

Ts ≈ -2.153

Decision rule: |Ts| > t(.975, 385) =1.966

Conclusion: Reject H0, there is a significant relationship between acceleration and log(mpg).

Interpretation:

In this data, per each 1-unit increase in acceleration, on average, we expect 4.133×10^2 decrease in response log(mpg).

#model.year\_Beta = 3.191\*10^-02

T-test for predictor of model.year:

Hypothesis: H0 : Beta for model.year = 0 vs Hα : Beta for model.year ≠ 0

Test statistic: 

Ts ≈ 19.099

Decision rule: |Ts| > t(.975, 385) =1.966

Conclusion: Reject H0, there is a significant relationship between model.year and log(mpg).

Interpretation:

In this data, per each 1-unit increase in model.year, on average, we expect 3.191\*10^-02 increase in response log(mpg).

# acceleration^2\_ Beta=1.38\*10^-3

T-test for predictor of acceleration^2:

Hypothesis: H0 : Beta for acceleration^2 =0 vs Hα : Beta for acceleration^2 ≠ 0

Test statistic: 

Ts ≈ 2.427

Decision rule: |Ts| > t(.975, 385) =1.966

Conclusion: Reject H0, there is a significant quadratic relationship between acceleration^2 and log(mpg).

Interpretaion:

Per 1 unit increase in predictor acceleration^2, on average, we expect 1.38\*10^-3 increase in response log(mpg).

# displacement\*weight\_ Beta=4.69\*10^-7

T-test for predictor of displacement\*weight:

Hypothesis: H0 : Beta for displacement\*weight =0 vs Hα : Beta for displacement\*weight ≠ 0

Test statistic: 

Ts ≈ 5.654

Decision rule: |Ts| > t(.975, 385) =1.966

Conclusion: Reject H0, there is a significant relationship between interaction of displacement and weight and log(mpg).

Interpretaion:

Interaction term displacement\*weight points to the different effects of displacement on response log(mpg) depending on weight. The slope of displacement on log(mpg) for certain weight car gets 4.69\*10^-7 larger than weight increased by 1.