STA 302 Assignment3

Multiple Linear Regression on Car Price

Weitong Luo
1002226432

1.Introduction

Dataset: hybrid_reg.csv(a3data.csv)

Source: D-J. Lim, S.R. Jahromi, T.R. Anderson, A-A. Tudorie (2014).

"Comparing Technological Advancement of Hybrid Electric Vehicles (HEV) in Different Market Segments," Technological Forecasting & Social Change,

http://dx.doi.org/10.1016/j.techfore.2014.05.008

Description:

Prices (MSRP, in 2013 $) for 154 hybrid models as dependent variable.

And indenpendent variable:

year: model year,

accelrate: acceleration rate,

mpg: fuel economy,

mpgmpge: max of MPG and MPGe for fully electric (plug-is). MPGe = 33.7*drive range/ battery capacity.

the range of each predictor are below:

```
      year        accelrate          mpg            mpgmpge            msrp       
 Min.   :1997   Min.   : 6.29   Min.   :17.00   Min.   : 17.00   Min.   : 11849  
 1st Qu.:2008   1st Qu.: 9.52   1st Qu.:26.00   1st Qu.: 26.00   1st Qu.: 24995  
 Median :2011   Median :11.63   Median :33.00   Median : 33.64   Median : 31950  
 Mean   :2010   Mean   :11.96   Mean   :34.80   Mean   : 38.45   Mean   : 39319  
 3rd Qu.:2013   3rd Qu.:13.47   3rd Qu.:41.26   3rd Qu.: 43.00   3rd Qu.: 49650  
 Max.   :2013   Max.   :20.41   Max.   :72.92   Max.   :100.00   Max.   :118544  
```

I did not use model id, car class and class id, since there are no relationship with the price.

People are wondering what really matters the price of a car model, So am I. Therefore, out of curiosity, I am doing this project to do the multiple linear regression test and residual test to predict the price of a car. I am going to use **correlation, residual plots, normal Q~Q plots, coxbox, log transformation, confidence interval and predicted interval** to test the price. Doing this project may help me and who is going to read it become more reasonable when purchasing a car.

2.Analysis

By paring all the factors, we can get the pairwise scatter and correlation plot in the **appendix page 1**. We can see that:

1. The following predictors have strong evidence of relationship and correlation with price:

    accelerate rate, r = 0.6956 and p-value is 1.916e-23 < 0.0001;

    fuel economy (mpg), r = -0.5318 and p-value is 1.507e-12 < 0.0001;

    max of MPG and MPGe for fully electric (plug-is) (mpgmpge), r = -0.3722 and p-value is 2.162e-06 < 0.0001;

2. The following predictor(s) have moderate evidence of relationship with price:

    model year, r= 0.2098 and p-value = 0.009251;

3. The following predictors have strong evidence of relationship with each other:

    model year and accelerate rate, r = 0.3594 and p-value is 5.046e-06 < 0.0001;

    accelerate rate and mpg, r = -0.5061 and p-value is 2.504e -11 < 0.0001;

    accelerate rate and mpgmpge, r= -0.3989 and p-value is 3.276e -07;

    mpg and mpgmpge, r = 0.6676 and p-value is 4.216e-21 < 0.0001;

4. The following predictors have moderate evidence of relationship with each other:

    year and mpg, r= -0.1699 and p-value is 0.03572;

    year and price, r= 0.2098 and p-value is 0.009251;

5. The following predictor has barely evidence of relationship with each other:

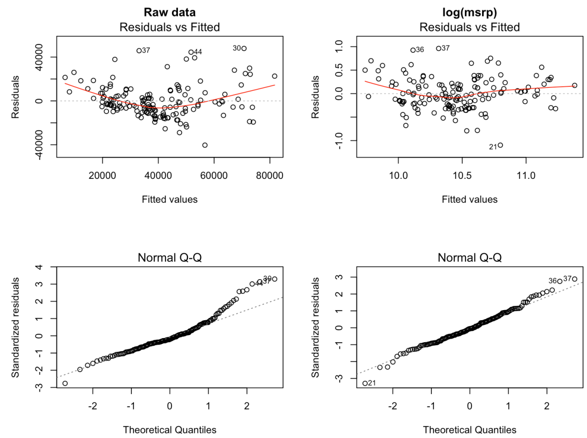
    year and mpgmpge, r= 0.005486 and p-value is 0.9463;

We can see that accelerate rate, mpg, mpgmpge and model year appear to have linear relationship with price. But due to the large different price and other factors, p-value and correlation value are relative low.

From the box-cox graph in the **appendix page 2**, we can find out that lambda is more close to zero than one, so we can do a log transformation on the linear regression model.

model name	R^2	SSE	MSE	AIC
raw data	0.53	3.2783e+10	2.2150e+08	3381.152
log(price)	0.5236	17.0384	0.1151	110.3653

Further, we use R^2 (in log scale), or SSE (in original scale), MSE or AIC as criteria, we observed that the log model has similar R^2 with raw data and smaller SSE and AIC. Therefore the log model is chosen as a better model because it satisfies the model assumptions and appears to be linear with constant variance over most of the interval.



From the residual plot, we can see that the residual plots of raw data values change with the fitted values change, hence the residual has relationship with fitted value, which is not appreciate. Regarding the log(msrp) model, we can see that residual points spread out evenly, it has better constant-variance and linearity.

From the Normal Q-Q plot, we can see that the log(msrp) has a more fitted line to the fitted line than the raw data. So it has a better normally of error comparing to the raw date's Normal Q-Q plots. Therefore, after taking log transformation, we can get a better model rather than the raw data.

Then I checked 95% level confidence interval for each beta in model after log transformation.

Assume H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$,
H_1: not all β_k in H_0 $=0$;

year: (-0.016637612 , 0.017787807), since 0 is in the interval, we don't have evidence to support the alternative hypothesis that the slope is different from 0. Fail to reject null hypothesis, hence the data **do not** give evidence of a linear relationship between **year** and **price**.

accelerate rate: (0.070792137, 0.116663572), since 0 is not in the interval, we have evidence to support the alternative hypothesis that the slope is different from 0. Reject null hypothesis, hence the data give evidence of a linear relationship between **accelerate rate** and **price**.

fuel economy: (-0.020506301, -0.006196473), since 0 is not in the the interval, we have evidence to support the alternative hypothesis that the slope is different from 0. Reject null hypothesis, hence the data give evidence of a linear relationship between **fuel economy** and **price**.

max of MPG and MPGe for fully electric (plug-is): (-0.001816077, 0.006349735), since 0 is in the interval, we don't have evidence to support the alternative hypothesis that the slope is different from 0. Fail to reject null hypothesis, hence the data **do not** give evidence of a linear relationship between **max of MPG and MPGe for fully electric (plug-is)** and **price**.

Therefore log(price) has strong linear relationship with accelerate rate and fuel economy, which is $\widehat{price} = 8.5565533 + 0.0937279 accelrate - 0.0133514 mpg$

Finally I want to introduce a new car model Camry 2017 which has 8.0 accelerate rate and 29 mpg, to predict its price. The confidence interval of log(price) for the new model is (9.94459, 10.3452); the predict interval for the new model is (9.445117, 10.84467); We can find that predict interval is wider than the confidence interval, and the price of this new Camry 2017 is 26780 USD and log(26780) = 10.195 which is in both predict interval and confidence interval. So the model is fitted.

3.Conclusion

From this project we can conclude:

By paring each factors price has strong correlation with accelerate rate, fuel economy and max of MPG and MPGe for fully electric (plug-is); and moderate correlation with year of the model. Also the latest model always has a better accelerate rate and fuel economy.

And by boxcox, I found that we can do a log transformation on price. By comparing R-square, SSE, MSE, AIC we can see that after doing log transformation, model may have less residuals.

And after comparing residual plots and QQ plots, using a log transformation on the raw data may get a better constant-variance, linearity and normality of the errors about model on price and accelerate rate, fuel economy and max of MPG and MPGe for fully electric(plug-is).

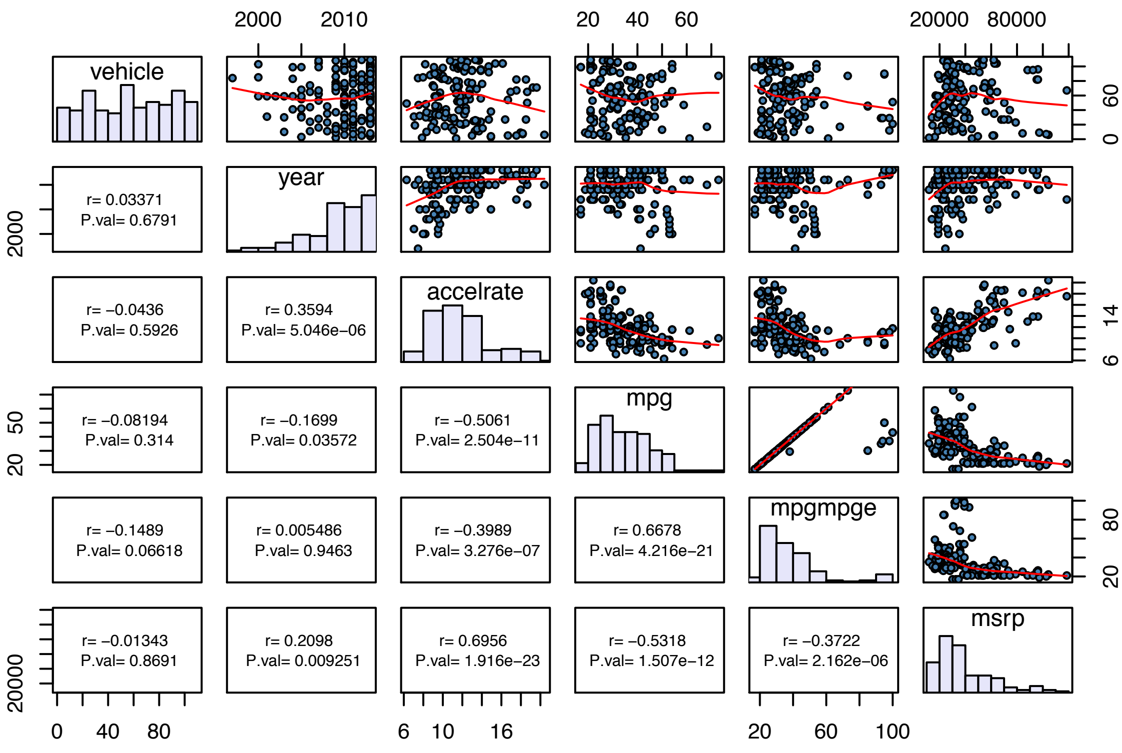
Additionally I checked the confidence interval for each factor, and find that accelerate rate and full economy have strong linear relationship with price. Because 0 is not in there's interval shows there could not be zero linear relationship with price, which means there is linear relationship with price.

Finally, I tried to predict a new car model's price, and I found out that predict interval is wider than confidence interval and the price of the new model is in the both intervals which supports my final model is a better model than the raw data again.

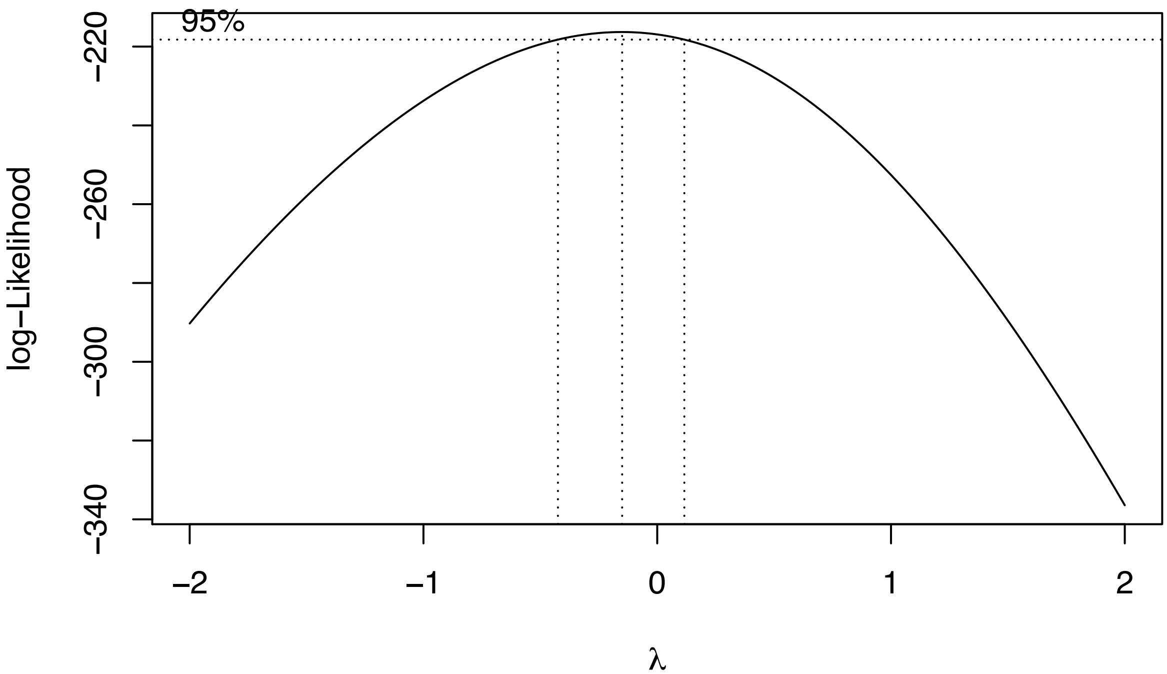
After doing this project, we can use this multiple linear regression to check if a model's price is reasonable regarding its model year accelerate rate, fuel economy or may predict a new model's retail price with given stats. It helps people making better decision when purchase a car, and also set a possible range for the company set their price for the coming models.

Appendix

```
a3 <- read.csv("/Users/tonyluo/UOFT/STA302/A3/a3data.csv",header=T)


mycor <- function(a3){
panel.hist <- function(x, ...){
   usr <- par("usr"); on.exit(par(usr))
   par(usr = c(usr[1:2], 0, 1.5) )
   h <- hist(x, plot = FALSE)
   breaks <- h$breaks; nB <- length(breaks)
   y <- h$counts; y <- y/max(y)
   rect(breaks[-nB], 0, breaks[-1], y, col="lavender", ...)
   }
panel.cor <- function(x, y, digits=4, prefix="", cex.cor, ...){
   usr <- par("usr");
   on.exit(par(usr))
   par(usr = c(0, 1, 0, 1))

   txt1 <- format( cor(x,y), digits=digits )
   txt2 <- format(cor.test(x,y)$p.value , digits=digits)
 text(0.5,0.5, paste("r=",txt1, "\n P.val=",txt2), cex=0.8)
   }
 pairs(a3, lower.panel=panel.cor, cex =0.7, pch = 21, bg="steelblue",
          diag.panel=panel.hist, cex.labels = 1.1,
          font.labels=0.9, upper.panel=panel.smooth)
}
mycor(a3)
```



```r
library(MASS)

model1 <- lm(msrp~year+accelrate+mpg+mpgmpge, data = a3) #setup a linear regression model using price a
bc=boxcox(model1,lambda=seq(-2,2,by=0.01)) #box-cox
```




```r
model2 <- lm(log(msrp)~year+accelrate+mpg+mpgmpge, data = a3)#setup a log transformation on model1.
summary(model1)
```

```
##
## Call:
## lm(formula = msrp ~ year + accelrate + mpg + mpgmpge, data = a3)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -40356  -9225  -2894   6527  47834
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 629176.14  765711.36   0.822  0.41258
## year          -311.22     382.07  -0.815  0.41662
## accelrate     4338.14     509.10   8.521 1.67e-14 ***
## mpg           -525.87     158.82  -3.311  0.00117 **
## mpgmpge         53.00      90.63   0.585  0.55959
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14880 on 148 degrees of freedom
## Multiple R-squared:   0.53,  Adjusted R-squared:  0.5173
## F-statistic: 41.72 on 4 and 148 DF,  p-value: < 2.2e-16
```

```r
anova(model1)
```

```
## Analysis of Variance Table
```

```
##
## Response: msrp
##            Df     Sum Sq    Mean Sq F value    Pr(>F)
## year        1 3.0696e+09 3.0696e+09  13.858 0.0002796 ***
## accelrate   1 3.0806e+10 3.0806e+10 139.076 < 2.2e-16 ***
## mpg         1 3.0137e+09 3.0137e+09  13.605 0.0003161 ***
## mpgmpge     1 7.5747e+07 7.5747e+07   0.342 0.5595881
## Residuals 148 3.2783e+10 2.2150e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(model2)
```

```
##
## Call:
## lm(formula = log(msrp) ~ year + accelrate + mpg + mpgmpge, data = a3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.09702 -0.21818 -0.01726  0.20079  0.96076
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.5565533 17.4565299   0.490 0.624744
## year         0.0005751  0.0087103   0.066 0.947447
## accelrate    0.0937279  0.0116064   8.076 2.16e-13 ***
## mpg         -0.0133514  0.0036207  -3.688 0.000317 ***
## mpgmpge      0.0022668  0.0020661   1.097 0.274361
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3393 on 148 degrees of freedom
## Multiple R-squared:  0.5236, Adjusted R-squared:  0.5107
## F-statistic: 40.67 on 4 and 148 DF,  p-value: < 2.2e-16
```

```r
anova(model2)
```

```
## Analysis of Variance Table
##
## Response: log(msrp)
##            Df  Sum Sq Mean Sq  F value    Pr(>F)
## year        1  2.4195  2.4195  21.0162 9.616e-06 ***
## accelrate   1 14.5266 14.5266 126.1813 < 2.2e-16 ***
## mpg         1  1.6425  1.6425  14.2669 0.0002292 ***
## mpgmpge     1  0.1386  0.1386   1.2037 0.2743607
## Residuals 148 17.0384  0.1151
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

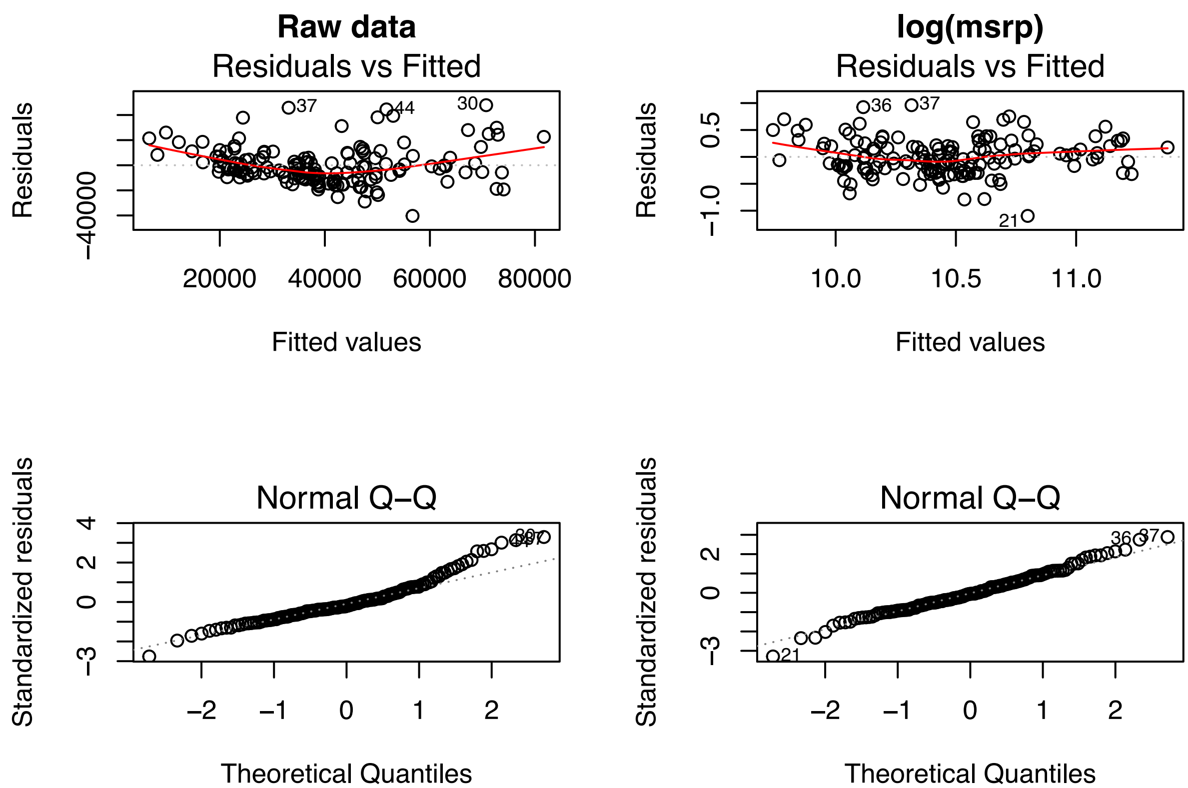
```r
AIC(model1)
```

```
## [1] 3381.152
```

```r
AIC(model2)
```

```
## [1] 110.3653
```

```r
par(mfrow=c(2,2))
plot(model1,which=1,main="Raw data") #plot the residual plot for raw data.
plot(model2,which=1,main="log(msrp)") #plot the residual plot for model after log transformation.

plot(model1,which=2) #plot the Normal Q-Q plot for raw data.
plot(model2,which=2) #plot the Normal Q-Q plot for model after log transformation.
```




```r
confint(model2, level = 0.95) #check confidence interval for each beta in model after log transformatio
```

```
##                    2.5 %       97.5 %
## (Intercept) -25.939688101 43.052794694
## year         -0.016637612  0.017787807
## accelrate     0.070792137  0.116663572
## mpg          -0.020506301 -0.006196473
## mpgmpge      -0.001816077  0.006349735
```

```r
newX=list(year = 2017, accelrate=5, mpg=50, mpgmpge = 50)#setup a new car model.
predict(model2, newdata=newX, interval = "confidence") #check the confidence interval for the new model
```

```
##        fit      lwr      upr
## 1 9.630936 9.398697 9.863175
```

```r
predict(model2, newdata=newX, interval = "predict")#check the predict interval for the new model.
```

```
##        fit      lwr      upr
## 1 9.630936 8.921357 10.34052
```