SUMMER 2016 REGRESSION ANALYSIS PROJECT

SAMI R. AL-MUALEM / SAMIALMUALEM@GMAIL.COM / +639176224065

1. Introduction

Fuel expense is always a major item in my budget and since I would like to buy a car soon but I do not have much knowledge on what factors to consider (since I'm not car savvy), I got interested in this subject.

This study is simply to construct a model for the car's mileage, measured in miles per gallon (mpg), based on the following preliminary variables: number of cylinders, engine displacement, power, weight, acceleration, and age of vehicle.

2. Executive Summary

In this project, we determine a model for the milage (in miles per gallon):

$$\text{mpg} = e^{\alpha+\beta_{cyl}\text{cyl}+\beta_{\text{disp}}\text{disp}+\beta_{\text{hp}}\text{hp}+\beta_{\text{weight}}\text{weight}+\beta_{\text{yrs}}\text{yrs}+\beta_{\text{cyl:disp}}\text{cyl}*\text{disp}}$$

and find the estimates

$$\begin{aligned}
\alpha &= +4.685 \\
\beta_{cyl} &= -8.909 * 10^{-2} \\
\beta_{disp} &= -2.340 * 10^{-3} \\
\beta_{hp} &= -2.125 * 10^{-3} \\
\beta_{weight} &= -2.239 * 10^{-4} \\
\beta_{yrs} &= -2.986 * 10^{-2} \\
\beta_{cyl:disp} &= +3.919 * 10^{-4}
\end{aligned}$$

where

(1) **mpg** - mileage in km/L, continuous (the explained variable)
(2) **cyl** - # of cylinders, integral
(3) **disp** - engine displacement in cubic inches/CID, continuous
(4) **hp** - horsepower, continuous
(5) **weight** - in lbs, continuous
(6) **acc** - acceleration, continuous; and
(7) **yrs** - age of vehicle in years, continuous

Multiple linear regression was done on log-MPG vs. all the listed variables and acceleration.

- We have found that the engine acceleration is not significant in this model.
- Also, we have checked the model's goodness of fit, homoskedasticity and its residues' normality, and have found no reasonable doubts to use this said model.
- Lastly, we have shown that using a same model but non-logarithmic on mpg will result to a heteroskedastic model which may contribute to errors in estimation of coefficients.

Date: 12 September 2016.

3. Data

We first load the data and summarize the fields

```
dfmpg = read.csv("auto-mpg.csv", header=TRUE, sep=",")
summary(dfmpg)
##       mpg              kpl               cyl             disp
##  Min.   : 9.00   Min.   : 3.826   Min.   :3.000   Min.   : 68.0
##  1st Qu.:17.00   1st Qu.: 7.227   1st Qu.:4.000   1st Qu.:105.0
##  Median :22.75   Median : 9.671   Median :4.000   Median :151.0
##  Mean   :23.45   Mean   : 9.967   Mean   :5.472   Mean   :194.4
##  3rd Qu.:29.00   3rd Qu.:12.328   3rd Qu.:8.000   3rd Qu.:275.8
##  Max.   :46.60   Max.   :19.810   Max.   :8.000   Max.   :455.0
##
##        hp            weight          acc            modelyr
##  Min.   : 46.0   Min.   :1613   Min.   : 8.00   Min.   :70.00
##  1st Qu.: 75.0   1st Qu.:2225   1st Qu.:13.78   1st Qu.:73.00
##  Median : 93.5   Median :2804   Median :15.50   Median :76.00
##  Mean   :104.5   Mean   :2978   Mean   :15.54   Mean   :75.98
##  3rd Qu.:126.0   3rd Qu.:3615   3rd Qu.:17.02   3rd Qu.:79.00
##  Max.   :230.0   Max.   :5140   Max.   :24.80   Max.   :82.00
##
##      origin           yrs              maker                     name
##  Min.   :1.000   Min.   : 1.00   ford      : 48   amc matador       :  5
##  1st Qu.:1.000   1st Qu.: 4.00   chevrolet: 44   ford pinto        :  5
##  Median :1.000   Median : 7.00   plymouth : 31   toyota corolla    :  5
##  Mean   :1.577   Mean   : 7.02   dodge     : 28   amc gremlin       :  4
##  3rd Qu.:2.000   3rd Qu.:10.00   amc       : 27   amc hornet        :  4
##  Max.   :3.000   Max.   :13.00   toyota    : 25   chevrolet chevette:  4
##                                  (Other)   :189   (Other)           :365
```

The following fields are to be used in this project.

(1) **mpg** - mileage in km/L, continuous (the explained variable)
(2) **cyl** - # of cylinders, integral
(3) **disp** - engine displacement in cubic inches/CID, continuous
(4) **hp** - horsepower, continuous
(5) **weight** - in lbs, continuous
(6) **acc** - acceleration, continuous; and
(7) **yrs** - age of vehicle in years, continuous

The original data is from UCI Machine Learning Repository, Auto-MPG Data [Lichman, 2013] which contains items (1) - (6) above. Furthermore, **age**, the age of the vehiclein years, was derived as

$$age = 1983 - \text{model year}$$

the latter term being a part of the original data set. (The data was collected in 1983.) Lastly, six rows were deleted since they have null values of horsepower, leaving us with a sample size of n=392. All the other data columns are not used for simplicity.

4. Linear Model

The study is conducted at the 99% confidence. Using the sample data, we get the estimates for the coefficients of each variables in the linear equation for log(mpg) in terms of cyl, $disp$, hp, $weight$, acc, and yrs. Since we suspect that the relationship of mileage with engine displacement might change as the number of cylinders of the engine changes, that is, that cyl interacts with disp, we include interaction of these two. We therefore have the following model:

$$\log(\text{mpg}) = \alpha + \beta_{cyl}\text{cyl} + \beta_{\text{disp}}\text{disp} + \beta_{\text{hp}}\text{hp} + \beta_{\text{weight}}\text{weight} + \beta_{\text{yrs}}\text{yrs} + \beta_{\text{cyl:disp}}\text{cyl} * \text{disp}$$

Note that we have transformed mpg and instead considered its logarithm. This is because of the heteroskedasticity that we experience when we use mpg instead. We will elaborate on this observation later in Section 4. Residual Analysis and Goodness of Fit.

Using R [2016], we arrive at the following output.

```
lm1 = lm(log(mpg) ~ (cyl + disp)^2 + hp + weight + acc + yrs, data=dfmpg)
summary(lm1)
##
## Call:
## lm(formula = log(mpg) ~ (cyl + disp)^2 + hp + weight + acc +
##     yrs, data = dfmpg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45506 -0.06823  0.00411  0.06201  0.40163
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.685e+00  9.524e-02  49.194  < 2e-16 ***
## cyl         -8.909e-02  1.513e-02  -5.890 8.46e-09 ***
## disp        -2.340e-03  4.765e-04  -4.911 1.34e-06 ***
## hp          -2.125e-03  4.969e-04  -4.276 2.41e-05 ***
## weight      -2.239e-04  2.339e-05  -9.572  < 2e-16 ***
## acc         -1.880e-03  3.434e-03  -0.548    0.584
## yrs         -2.986e-02  1.770e-03 -16.867  < 2e-16 ***
## cyl:disp     3.919e-04  6.073e-05   6.453 3.30e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1156 on 384 degrees of freedom
## Multiple R-squared:  0.8866,Adjusted R-squared:  0.8845
## F-statistic: 428.7 on 7 and 384 DF,  p-value: < 2.2e-16
```

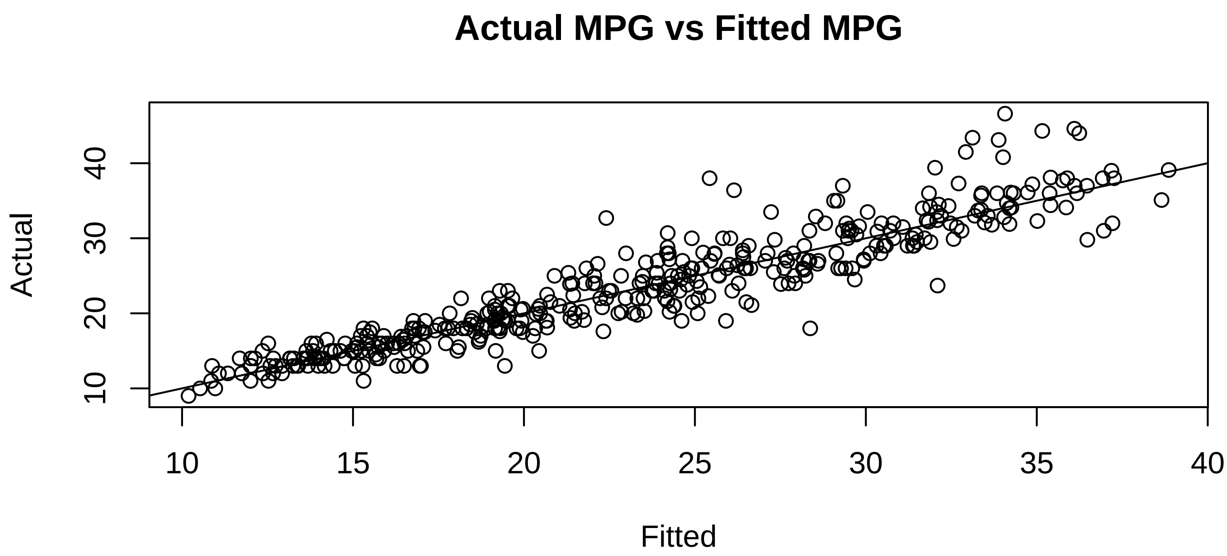
We make the following observations:

- From the value of the R^2 above, we see that this model explains about 89% of the variation.
- Note that all coefficients, including the intercept and the interaction term for cyl and $disp$, are statistically-significant at the 95% level except for acc, the acceleration, since they all have p-values very less than 0.01.
- We also note that, as expected, the mileage decreases as each of cyl, $disp$, hp, $weight$, and yrs increases. Further, each unit increase in cyl (or in $disp$) increases the slope of $disp$ (cyl, respectively) by about 0.039%.

5. Residual Analysis and Goodness of Fit

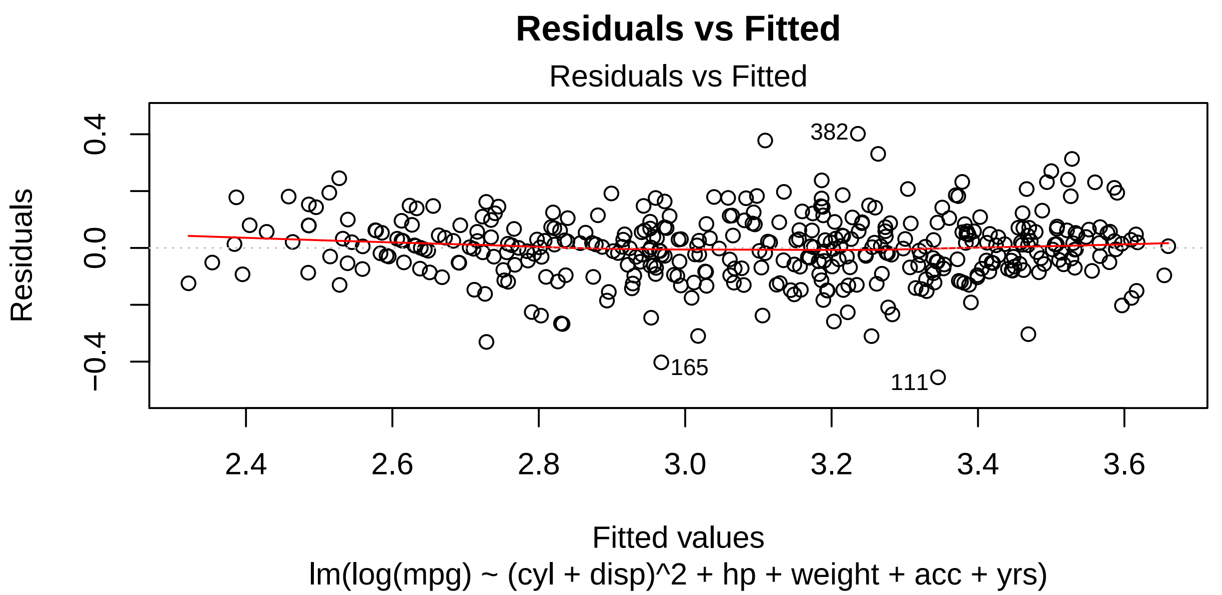
Let us plot the actual *mpg* values vs. fitted *mpg* values per sample point over the diagonal $y = x$. From this, we see that the fit is reasonable.

```
plot(x=exp(lm1$fitted.values), y=dfmpg$mpg, xlab='Fitted', ylab='Actual',
  main='Actual MPG vs Fitted MPG')
abline(a=0,b=1)
```



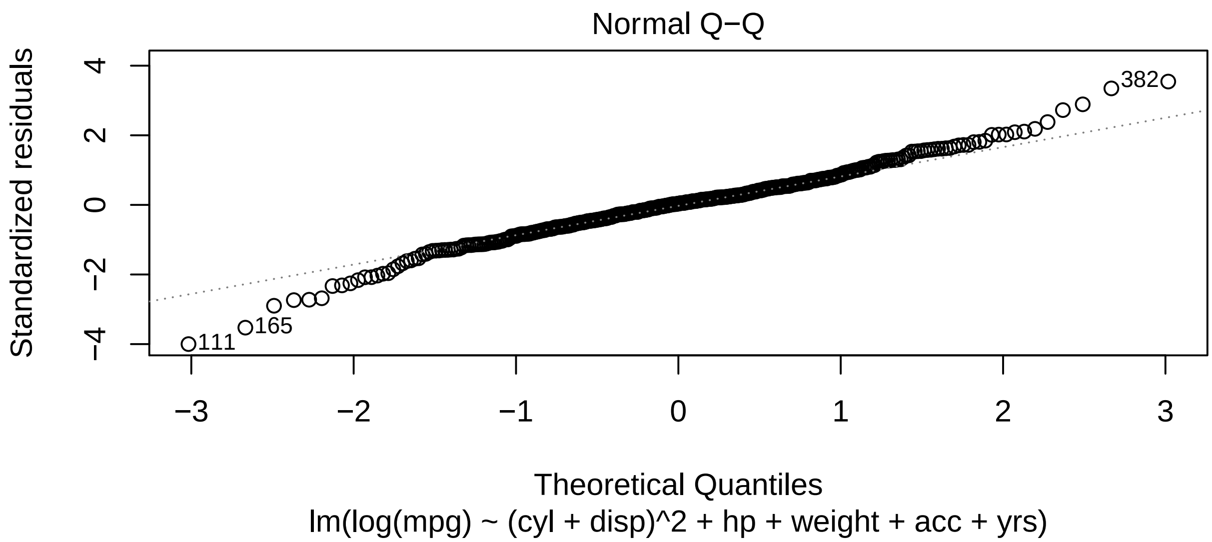
Equivalently, we plot the residuals vs. fitted log-MPG values and see that the fit is also reasonable esp. for lower fitted values.

```
plot(lm1, which=1, main='Residuals vs Fitted')
```



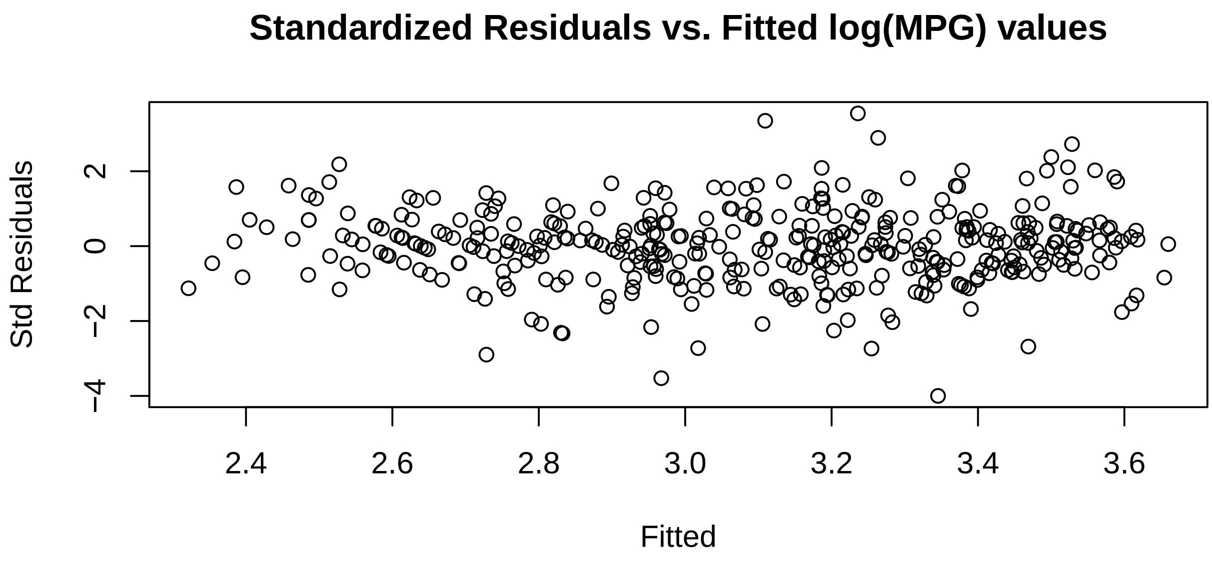
To check for normality of the log-MPG residuals, we look at the q-q plot below. We see that the fit is a bit heavy-tailed.

```
plot(lm1, which=2)
```



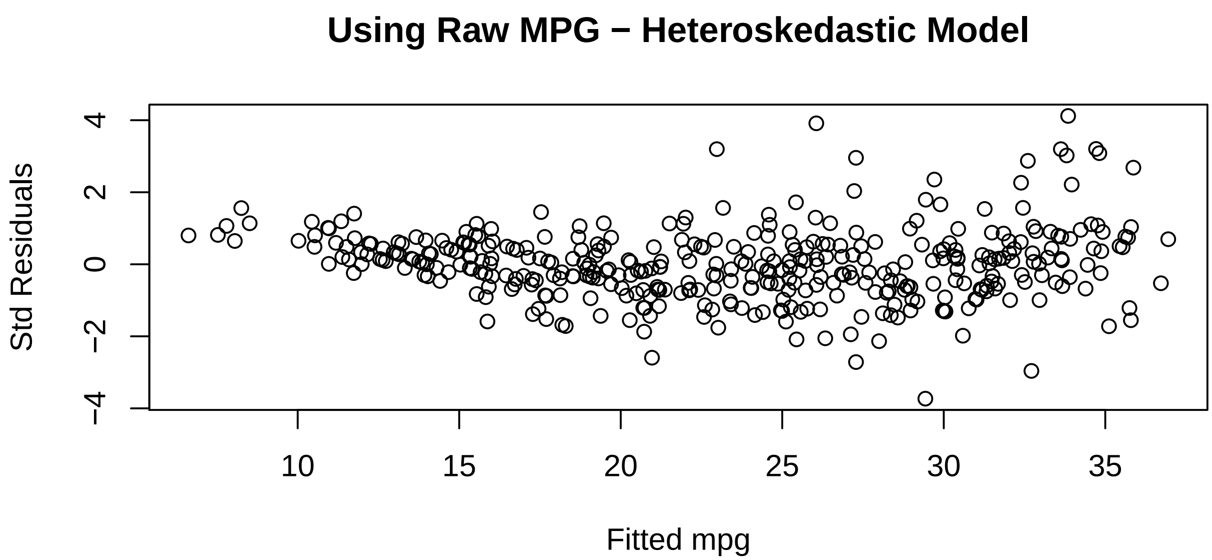
To check for heteroskedasticity, we plot the standardized residuals vs. fitted log-MPG values. We can see in the following plot that these is reasonable homoskedasticity since the spread is almost uniform for any level of fitted log(mpg)

```
lm1.stdres = rstandard(lm1)
plot(y=lm1.stdres, x=lm1$fitted.values, xlab='Fitted', ylab='Std Residuals',
  main='Standardized Residuals vs. Fitted log(MPG) values')
```



Earlier, we mentioned that we have used the log-transformed mpg instead of the raw mpg values due to homoskedasticity. Indeed, if we consider modelling mpg directly and then plot the standardized residuals vs. fitted mpg values, we get the following fan-shaped plot opening to the right. This indicates that higher fitted values are associated with higher variances.

```
lm2 = lm(mpg ~ (cyl + disp)^2 + hp + weight + acc + yrs, data=dfmpg)
lm2.stdres = rstandard(lm2)
plot(y=lm2.stdres, x=lm2$fitted.values, xlab='Fitted mpg', ylab='Std Residuals',
  main = 'Using Raw MPG - Heteroskedastic Model')
```



6. Conclusion

We have determined a model for the milage (in miles per gallon):

$$\text{mpg} = e^{\alpha+\beta_{cyl}\text{cyl}+\beta_{\text{disp}}\text{disp}+\beta_{\text{hp}}\text{hp}+\beta_{\text{weight}}\text{weight}+\beta_{\text{yrs}}\text{yrs}+\beta_{\text{cyl:disp}}\text{cyl}*\text{disp}}$$

where

$$\begin{aligned}
\alpha &= +4.685 \\
\beta_{cyl} &= -8.909 * 10^{-2} \\
\beta_{disp} &= -2.340 * 10^{-3} \\
\beta_{hp} &= -2.125 * 10^{-3} \\
\beta_{weight} &= -2.239 * 10^{-4} \\
\beta_{yrs} &= -2.986 * 10^{-2} \\
\beta_{cyl:disp} &= +3.919 * 10^{-4}
\end{aligned}$$

We have found that the engine acceleration is not significant in this model. Also, we have checked the model's goodness of fit, homoskedasticity and its residues' normality, and have found no reasonable doubts to use this said model. Lastly, we have shown that using a same model but non-logarithmic on mpg will result to a heteroskedastic model which may contribute to errors in estimation of coefficients.

References

M. Lichman. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`.

R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2016. URL `https://www.R-project.org/`.