Regression Model Course Project

Stanley Kan Chuen LAM

January 16, 2018

Executive Summary

In this project, we are interested in exploring the relationship between a set of variables and miles per gallon (MPG) outcome. In particular, we'd like to focus in the following two questions : 1. "Is an automatic or manual transmission better for MPG"? 2. "Quantify the MPG difference between automatic and manual transmissions".

Here is the URL for more information of the dataset, mtcars : https://www.rdocumentation.org/packages/datasets/versions/3.4.3/topics/mtcars

Exploratory Data Analysis

Fig-1 in the Appendix shows the boxplot of the MPG for "Manual" & "Automatic" transmission. We see that the values of MPG for "Manual" transmission is higher than that of "Automatic". Let's look at the mean MPG for "Automatic" and for "Manual" transmission :

```
# Note that mtcars is replaced by MT
round(c(mean(MT$mpg[MT$am == "Manu"]),mean(MT$mpg[MT$am == "Auto"])),2)
```

```
## [1] 24.39 17.15
```

The mean MPG for "Manual" transmission is higher than that of "Automatic" transmission by 7.22 miles per gallon. Let's do t-test to see if mean is higher.

```
t.test(MT$mpg[MT$am == "Manu"], MT$mpg[MT$am == "Auto"], alternative = "greater")$p.value
```

```
## [1] 0.0006868192
```

P-value is far less than 0.05 and so we should accept the alternative hypothesis that the mean MPG for "Manual" transmission is higher than that of "Automatic". Next we shall quantify the difference by taking into account other adjustments factors that can be explained by other variables.

Multivariate Linear Regression Models

Fig-2 & Fig-3 in the Appendix shows the pair plots with correlation among numerical variables and the categorical variables. It is found that weigth(wt) & number of cylinder(cyl) have relatively higher correlation with mpg. Furthermore, let's analysis of variance using aov functin :

```
summary(aov(mpg~.,MT))
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## cyl          2  824.8   412.4  51.377 1.94e-07 ***
## disp         1   57.6    57.6   7.181   0.0171 *
## hp           1   18.5    18.5   2.305   0.1497
## drat         1   11.9    11.9   1.484   0.2419
## wt           1   55.8    55.8   6.950   0.0187 *
## qsec         1    1.5     1.5   0.190   0.6692
## vs           1    0.3     0.3   0.038   0.8488
## am           1   16.6    16.6   2.064   0.1714
```

```
## gear         2    5.0     2.5   0.313   0.7361
## carb         5   13.6     2.7   0.339   0.8814
## Residuals   15  120.4     8.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now let's pick low P-value and high correlation variables, weight(wt) and cylinder(cyl) for adjustment the basic model, lm(mpg~am). Furthermore, we'd include an interaction term to see if it can raise the adjustment R-squared.

```
modelam  <- lm(mpg~am,MT)
model1   <- lm(mpg~am+cyl+wt,MT)
model2   <- lm(mpg~am+cyl+wt+am*wt,MT)
```

Now let's apply nested likelihood ratio tests (ANOVA) to help find the best model

ANOVA for modelam, model1 & model2 :

```
anova(modelam, model1, model2)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl + wt
## Model 3: mpg ~ am + cyl + wt + am * wt
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     27 182.97  3    537.93 33.7850 4.031e-09 ***
## 3     26 137.99  1     44.98  8.4744  0.007296 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion

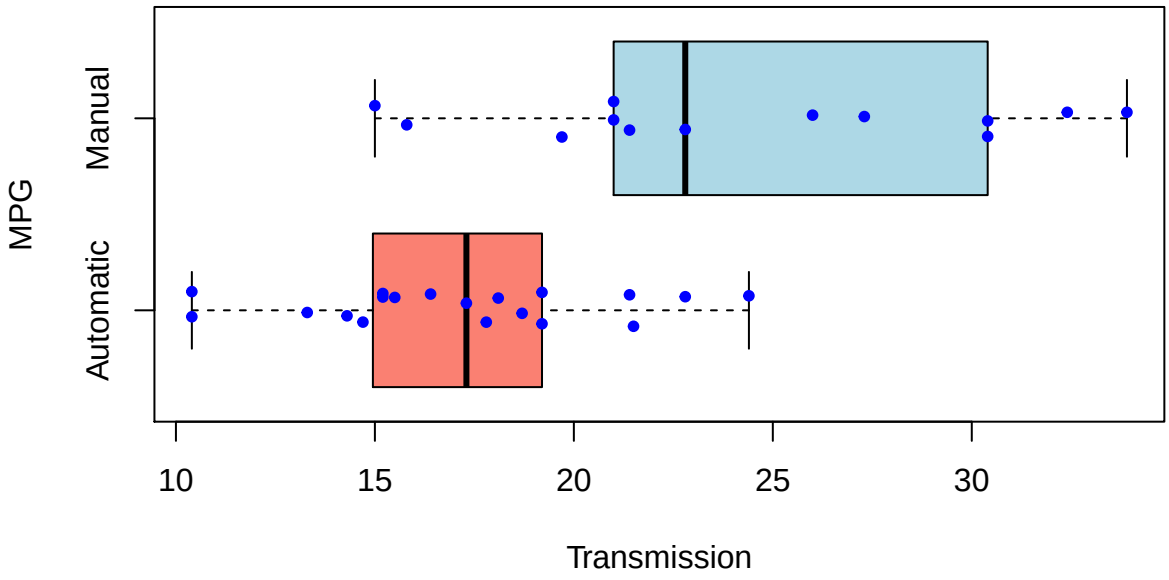
From the summaries of all models in the Appendix, we see that Model2 (Terms : am, cyl, wt, am:wt) has the highest adjusted R-squared value of 85.4%. That means model2 explains 85.4% of the variation in Miles per Gallon(mpg), leaving 14.6% uncertainty. Here it make good sense that increase in weight(wt) reduces the mpg. Here are the resulting coefficients of model2 :

```
## (Intercept)      amManu     cyl6cyl     cyl8cyl          wt   amManu:wt
##   29.774836   11.568790   -2.709777   -4.776110   -2.398713   -4.067981
```

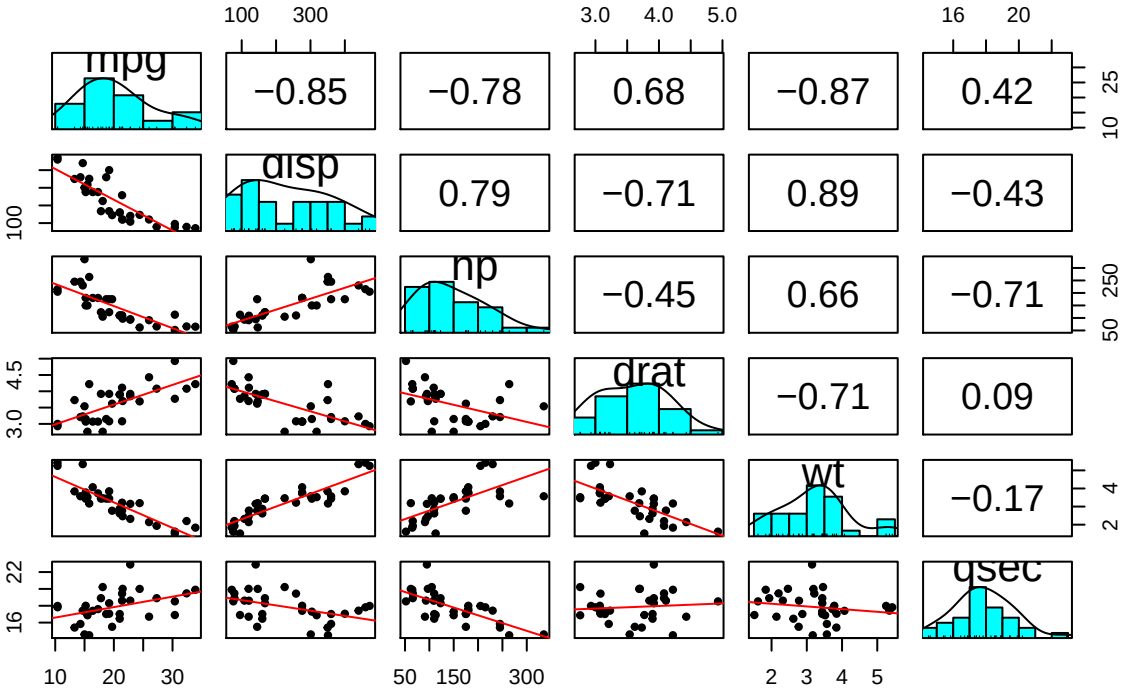
In the Appendix, the Residuals vs Fitted plot shows a mostly smooth residuals development, without any specific patterns that would indicate heteroskedasticity (non constant variance) or other model fitness issues. The Normal QQ plot shows a satisfactory residuals normality. The Scale-Location plot shows nothing realy spectacular about the standardized residuals. And the Residuals vs Leverage plot shows no systematic pattern either although that some residuals are shown to have relatively increased leverage.

Appendix

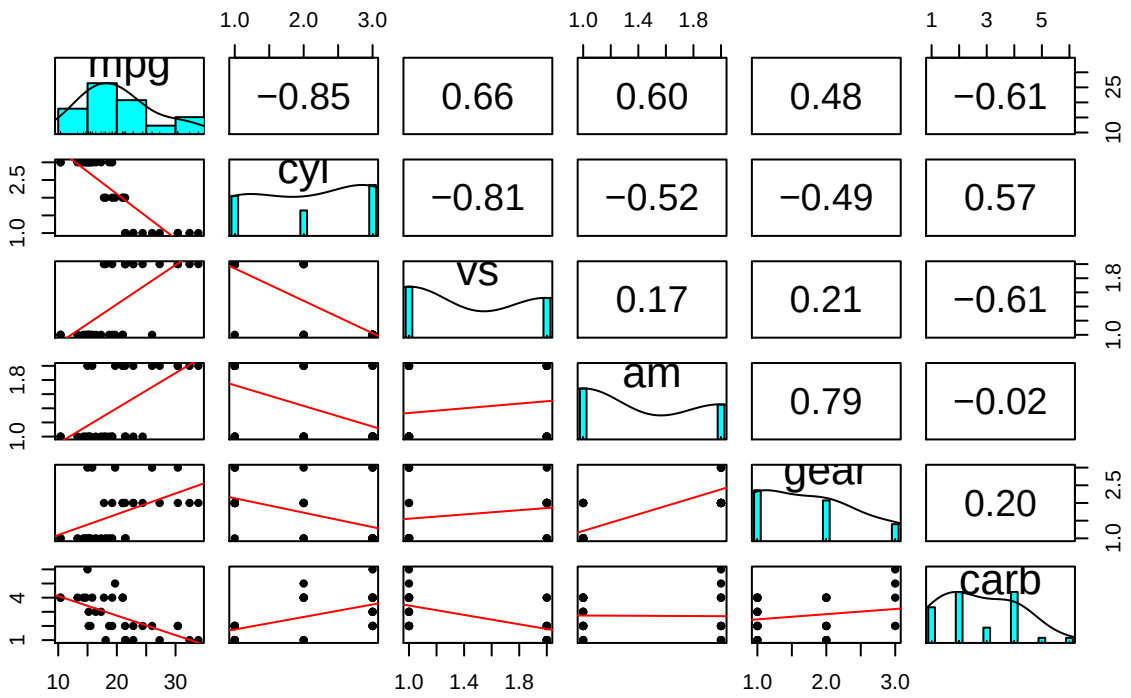
Fig−1 : Miles per Gallon(MPG) by Transmission Type



Fig−2 : Pair plot with correlation among numerical variables

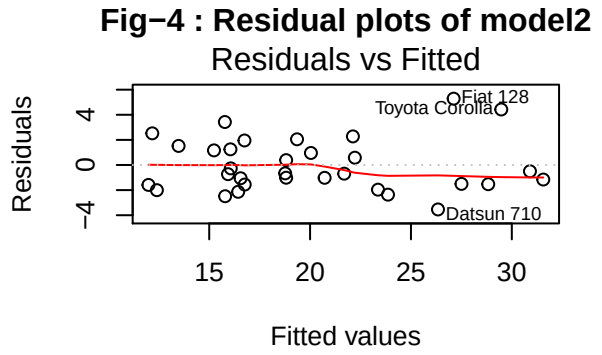


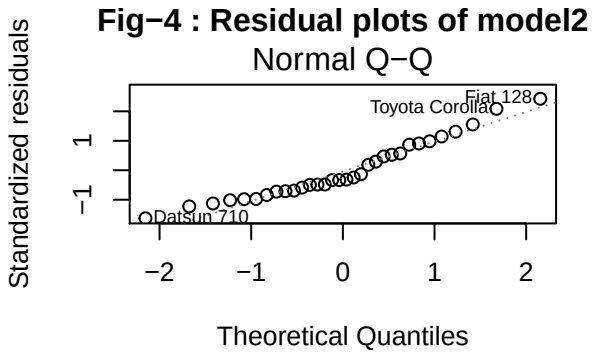
Fig−3 : Pair plot with correlation among categorical variables

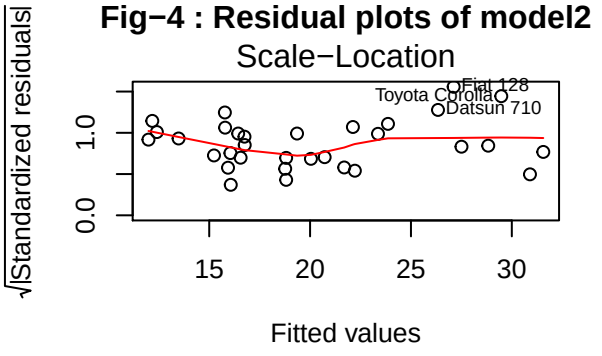


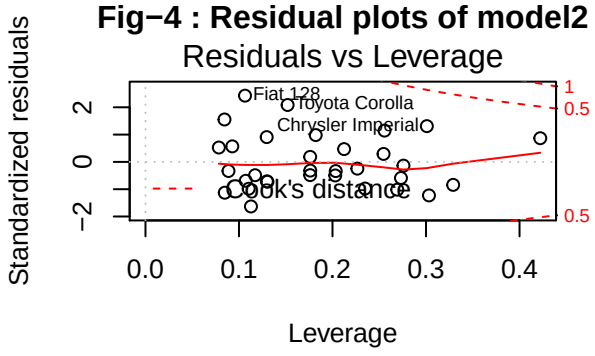

```
summary(model2)
```

```
##
## Call:
## lm(formula = mpg ~ am + cyl + wt + am * wt, data = MT)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5409 -1.5377 -0.6783  1.3160  5.2831
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   29.775      2.840  10.483 7.87e-11 ***
## amManu        11.569      4.088   2.830  0.00885 **
## cyl6cyl       -2.710      1.357  -1.996  0.05647 .
## cyl8cyl       -4.776      1.556  -3.070  0.00496 **
## wt            -2.399      0.844  -2.842  0.00860 **
## amManu:wt     -4.068      1.397  -2.911  0.00730 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.304 on 26 degrees of freedom
## Multiple R-squared:  0.8775, Adjusted R-squared:  0.8539
## F-statistic: 37.23 on 5 and 26 DF,  p-value: 4.743e-11
```


Fig−4 : Residual plots of model2
Residuals vs Fitted
Residuals
Fitted values
Fiat 128
Toyota Corolla
Datsun 710


Fig−4 : Residual plots of model2
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Fiat 128
Toyota Corolla
Datsun 710


Fig−4 : Residual plots of model2
Scale−Location
√|Standardized residuals|
Fitted values
Fiat 128
Toyota Corolla
Datsun 710


Fig−4 : Residual plots of model2
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
Fiat 128
Toyota Corolla
Chrysler Imperial