Regression Model Course Project

Executive Summary

Based on our regression models and exploratory data analyses, we observe that, by switching from automatic to manual tansmission on average improve mpg. In particular, using our regression model with predictor variables am and hp, we observe that if hp is kept constant, by switching from automatic to manual transmission will improve the mpg by 5.277085.

Detail Analysis

The questions we are exploring in this course project are as follows:

- Is an automatic or manual transmission better for MPG?
- Quantifying how different is the MPG between automatic and manual transmissions?

First we convert the 'am' variables in 'mtcars' to factor variable for ease of analysis. We also convert other categorical variables, e.g. cyl, gear, carb and vs using factor command.

```
mtcars$am = as.factor(mtcars$am);levels(mtcars$am) = c("automatic", "manual")
mtcars$cyl = as.factor(mtcars$cyl);mtcars$gear = as.factor(mtcars$gear);mtcars$carb =
as.factor(mtcars$carb);mtcars$vs = as.factor(mtcars$vs)
```

Next we try to determine the difference in mean mpg for the cars with manual and automatic transmission. We also plot a box plot of am vs. mpg(see appendix), as well as observe the correlation between mpg and am variables.

```
mean(mtcars$mpg[mtcars$am=="manual"]) - mean(mtcars$mpg[mtcars$am=="automatic"])
```

```
## [1] 7.245
```

```
cor(as.numeric(mtcars$am), mtcars$mpg)
```

```
## [1] 0.5998
```

We observe, that the am and mpg variables are highly correlated (60%), as well as cars with manual transmission on average improves mpg by 7.24. Next, we will try to fit a linear regression model using mpg as the dependent variable and am variables as predictor.

```
modelFit = lm(mpg~am, data=mtcars);summary(modelFit)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.392 -3.092 -0.297  3.244  9.508
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    17.15       1.12   15.25  1.1e-15 ***
## ammanual        7.24       1.76    4.11  0.00029 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.9 on 30 degrees of freedom
## Multiple R-squared:  0.36,   Adjusted R-squared:  0.338
## F-statistic: 16.9 on 1 and 30 DF,  p-value: 0.000285
```

We observe from the summary of the regression model, both intercept and ammanual coefficient have very low p-values indicating both of them rejecting the Null hypothesis. The adjusted R-squared value indicating the model accouting for only 33.85% variation of the data. The standard error of 4.9 of the fitted line, 95% of the data values will fall within +/- 9.8 mpg of fitted line. From the above model, our conclusion is by changing the transmission of a car from 'automatic' to 'manual', on an average an mpg improvement of 7.245 can be expected, which is also the conclusion we observe from our earlier exploratory analysis. However, our analysis of the model is not complete, unless we looked into the residual errors. The residul plot indicates heteroscedascity of the fit. Therefore, probably some variable may be missing and we can probably perform a better fit by incorporating additionl variables. For further exploratory analysis, we look into the pair plots to see at a glance relationship among different variables (please see Appendix for this figure). By glancing into this figures, we observe multiple trends. We also observe, regardless of whether a car is manual or automatic, mpg drops as the number of cylinder, displacement, horse power or the weight of the car increases.

After a bit of trial and error, and playing with different combination, we found that a model with mpg as dependent variable and am and hp as predictor have a much better fit.

```
modelNew = lm(mpg ~am + hp, data=mtcars)
summary(modelNew)
```

```
##
## Call:
## lm(formula = mpg ~ am + hp, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.384 -2.264  0.137  1.697  5.866
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 26.58491    1.42509   18.65  < 2e-16 ***
## ammanual     5.27709    1.07954    4.89  3.5e-05 ***
## hp          -0.05889    0.00786   -7.50  2.9e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.91 on 29 degrees of freedom
## Multiple R-squared:  0.782,  Adjusted R-squared:  0.767
## F-statistic:   52 on 2 and 29 DF,  p-value: 2.55e-10
```

The above model has an adjusted R-square value of 76.7% and a standard error of 2.9 and the p-values for co-efficients are much less than 0.05. Compare to our earlier model, these values indicates a much better fit. We also tried other combinations, some of the combination can provide better R values, however the p-values were not good. The residual plots are shown in the appendix, which indicates that the residual is homoscedastic and randomly distributed. The equation of the above model is as follows:

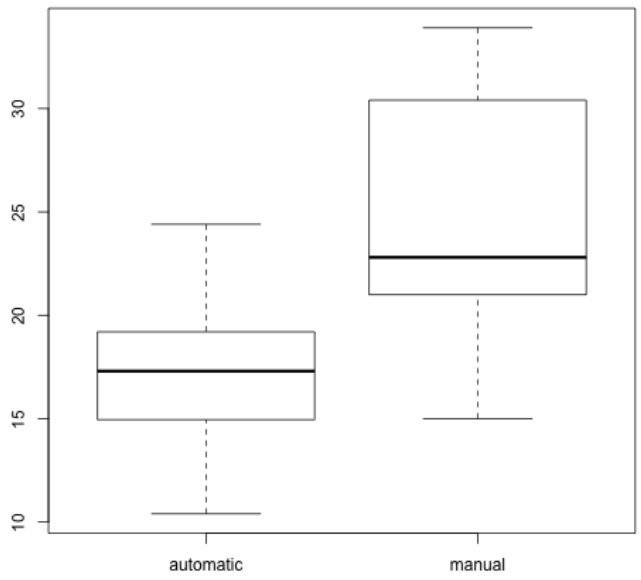
$$\text{mpg} = \beta_0 + \beta_1 \mathbb{I}(\text{am} = \text{"manual"}) + \beta_2 \text{hp}$$

Here, $\beta_0 = 26.584914, \beta_1 = 5.277085$ and $\beta_2 = -0.058888$. By using the two lines, same slope, the interpretation here is, by keeping hp, by switching from automatic to manual, a mpg increase of 5.277085 can be achieved, i.e. the line will shift up by 5.277085 (see appendix for this figure). We also tried to use the interaction term, however it has insignificant effect.

Appendix

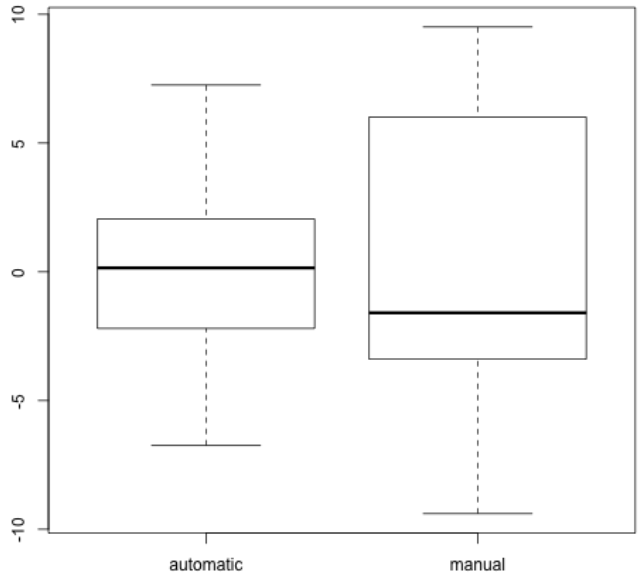
- Boxplot of am vs. mpg

```
plot(mtcars$am, mtcars$mpg)
```



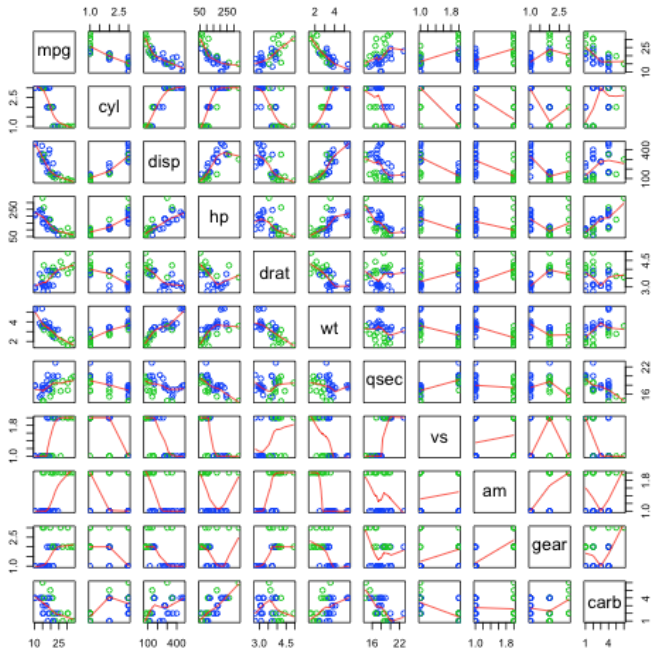
- Residual plot of first regression model (modelFit)

```
plot(factor(modelFit$fitted.values, labels = c('automatic', 'manual')), modelFit$residuals)
```



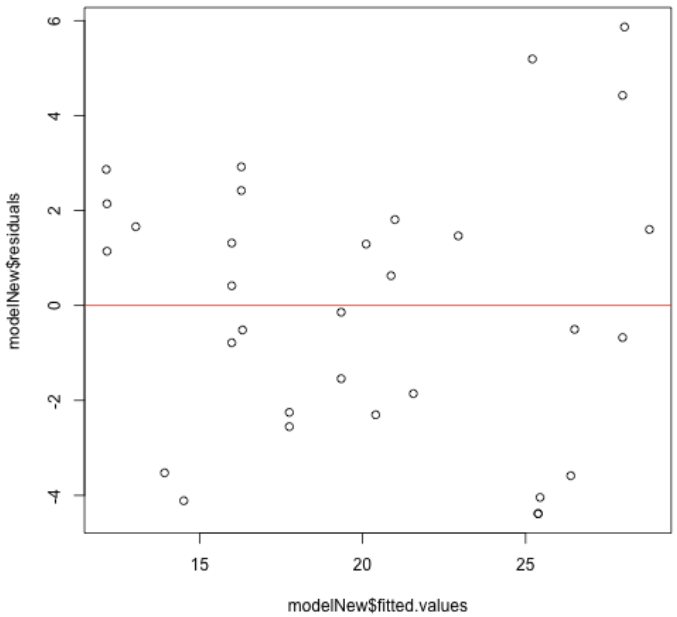
- Pair plot

```
pairs(mtcars, panel=panel.smooth,col = 3+(mtcars$am=="automatic"))
```



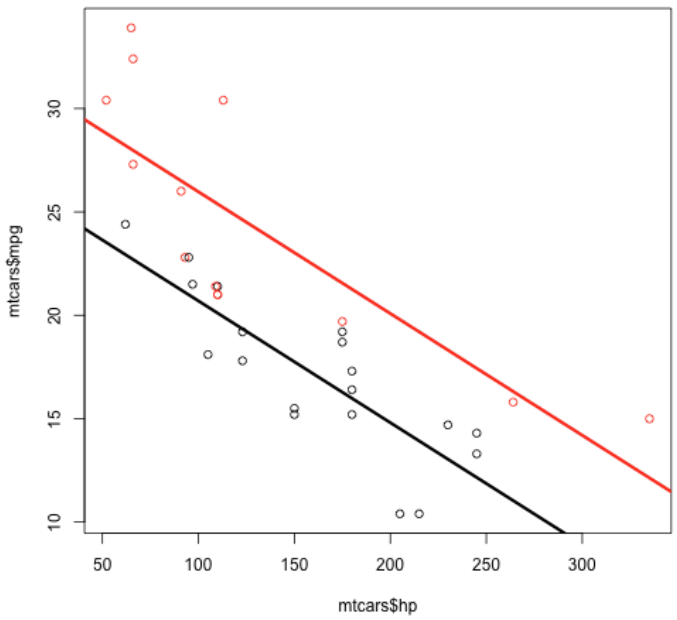
- Residual plot of second regression model (modelNew)

```
plot(modelNew$fitted.values, modelNew$residuals)
abline(h=0,col='red')
```



- Two line, same slope interpretation of second regressiong model (modelNew)

```
plot(mtcars$hp, mtcars$mpg, col=((mtcars$am=="manual")*1+1))
abline(modelNew$coefficients[1],modelNew$coefficients[3],col="black",lwd=3)
abline(modelNew$coefficients[1]+modelNew$coefficients[2],modelNew$coefficients[3],col="red",lwd=3)
```


mtcars$mpg
mtcars$hp
10
15
20
25
30
50
100
150
200
250
300