Coursera Regression Model Course Project

Chuk Yong

31/12/2016

## Coursera Regression Model Course Project

Motor Trend is a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

“Is an automatic or manual transmission better for MPG” "Quantify the MPG difference between automatic and manual transmissions"

Manual Transmission is better for MPG than Automatic Transmission. The difference at first look was 7mpg or about 30% more fuel efficient in favour of Manual Transmission. However, on further anaylsis by including other relevent factors, the difference turn out to be less. The MPG difference between them, taking account for factors like cylinder, horse-power, displacement and weight, is only 1.8MPG or about 5%.

## A study of MPG between Automatic and Manual Transmission

Load dataset

data(mtcars)

Transform some variables into factor form

mtcars$cyl <- factor(mtcars$cyl)  
mtcars$vs <- factor(mtcars$vs)  
mtcars$am <- factor(mtcars$am,labels=c("Auto","Man"))  
mtcars$gear <- factor(mtcars$gear)  
mtcars$carb <- factor(mtcars$carb)

## Exploratory Data Analysis

mpgAuto <- subset(mtcars, mtcars$am == "Auto", select = c(mpg))  
mpgMan <- subset(mtcars, mtcars$am == "Man", select = c(mpg))  
t.test(mpgAuto$mpg, mpgMan$mpg)

##   
## Welch Two Sample t-test  
##   
## data: mpgAuto$mpg and mpgMan$mpg  
## t = -3.7671, df = 18.332, p-value = 0.001374  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -11.280194 -3.209684  
## sample estimates:  
## mean of x mean of y   
## 17.14737 24.39231

We can see that: 1. Average mpg for Automatic transmission is 17.15. 2. Average mpg for Manual transmission is 24,39. 3. p-value = 0.001374

Clearly, Manual transmission has a much better mpg and the test is significant (p-value < 0.01).

We can now proceed with regression modelling to see how well the is the fit.

fit <- lm(mpg ~ am, data = mtcars)  
summary(fit)

##   
## Call:  
## lm(formula = mpg ~ am, data = mtcars)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.3923 -3.0923 -0.2974 3.2439 9.5077   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 17.147 1.125 15.247 1.13e-15 \*\*\*  
## amMan 7.245 1.764 4.106 0.000285 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.902 on 30 degrees of freedom  
## Multiple R-squared: 0.3598, Adjusted R-squared: 0.3385   
## F-statistic: 16.86 on 1 and 30 DF, p-value: 0.000285

As confimed by the coefficients - Manual transmission is 7.245mpg better than Automatic transmission. p-value is much less than 0.01 and so they are significant. However, looking at R-squared and Adjusted R-squared, the value of 0.36 and 0.33 tells us that there is room for improvement.

## Quantifying the MPG difference between Automatic and Manual Transmission

So far, we know than there is a significant difference between Automatic and Manual Transmission. However, there might be other factors at play that may impact MPG. In order to make a better decision on what other variables could possible impact mpg, we created a pair-plot, Fig 1, in Appendix A.

From the pair-plot, we can see that there are possible correlations between cyl - cylinder, disp - displacement, hp - horse power and wt - weight.

## Multivariable Regression Study

We will fit in the variables and investigate if introduction them improve on our original fit.

A step by step approach is applied. We add the variables one at a time and evaluate the R.Squared value to see if we are making an improvement. While R.Squared may not tell the complete story, it is a start. Other diagnostic methods will be applied on our final model.

Fig 2 shows the result of R.Sqaured after adding the variables.

Final regression with 4 added variables: cyl, disp, hp and wt.

fit5 <- lm(mpg ~ am+cyl+disp+hp+wt, data = mtcars)  
summary(fit5)

##   
## Call:  
## lm(formula = mpg ~ am + cyl + disp + hp + wt, data = mtcars)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.9374 -1.3347 -0.3903 1.1910 5.0757   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 33.864276 2.695416 12.564 2.67e-12 \*\*\*  
## amMan 1.806099 1.421079 1.271 0.2155   
## cyl6 -3.136067 1.469090 -2.135 0.0428 \*   
## cyl8 -2.717781 2.898149 -0.938 0.3573   
## disp 0.004088 0.012767 0.320 0.7515   
## hp -0.032480 0.013983 -2.323 0.0286 \*   
## wt -2.738695 1.175978 -2.329 0.0282 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.453 on 25 degrees of freedom  
## Multiple R-squared: 0.8664, Adjusted R-squared: 0.8344   
## F-statistic: 27.03 on 6 and 25 DF, p-value: 8.861e-10

Now we can see that although Manual transmission still shows a better MPG vs Automatic, the difference is greatly reduced. At 1.8MPG over Automatic transmission, the difference is now only 5%.

Fig 2 in Appendix A shows the R.Sqaured value as we add the variables into the regression. We can see from the graph that all the variables improve the fit.

Finally, we use diagnostic plots as a verification of our findings. Fig 3 in Appendix A shows 4 diagnostic plots.

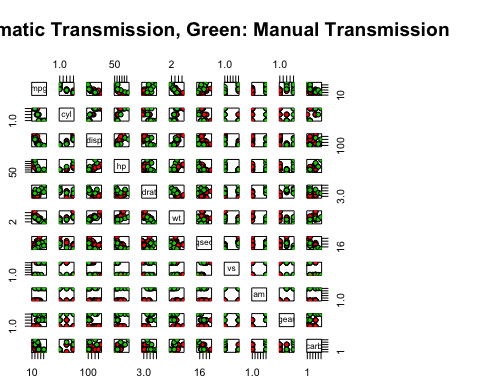
1. Residual vs fitted. The points are scattered and our line runs fairly straight. This indicates very little biasness in our fit.
2. Normal Q-Q. Shows an upward slope as expected.
3. Both Scale-Location and Residual-vs-Leverage show scattered points with no particular outstanding outliers.

## Conclusion

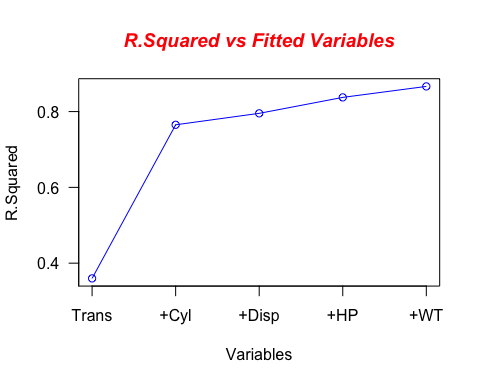
We can conclude from our study that there is a significant enough difference between Automatic and Manual Transmission in a car's MPG. Manual Transmission being more economical. However, taking into consideration of a car's cylinder, displacement, horse power and weight, the difference we not as big as we first expected.

# Appendix A

### Fig 1 - Pair plots betweem all variables in mtcars dataset



### Fig 2 - Plot of how R.Sqaured perform when we add in our variables



X-axis indicates the variables we added (cumulative) to the regression.

### Fig 3 - Diagnostic Plots

