Regression Models Course Project

joannanw

Sunday, April 26, 2015

## Regression Models Course Project

## Executive Summary

Motor Trend aims to analyze the relationship between a set of variables and miles per gallon (MPG) from the data set "mtcars". "mtcars" comprises of 32 vehicle make and 11 variables such as transmission type, weight, and cylinder. The analysis aims to answer whether an automatic or manual transmission is better for fuel consumption, and quantify the miles per gallon (MPG) difference in a sensible manner. This paper first uses exploratory data analysis to identify how transmission type affects the MPG. This is followed by a hypothesis testing to show that there is a difference between automatic and manual transmission performance on MPG.A regression analysis is then performed to identify the variables that are strongly correlated. These results are finally used to create a regression model that significantly explains the variance of MPG.

## Analysis

First, the data set is loaded and observed.

data(mtcars)  
str(mtcars)

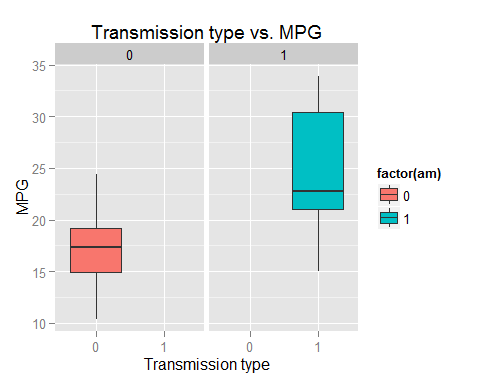
## 'data.frame': 32 obs. of 11 variables:  
## $ mpg : num 21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...  
## $ cyl : num 6 6 4 6 8 6 8 4 4 6 ...  
## $ disp: num 160 160 108 258 360 ...  
## $ hp : num 110 110 93 110 175 105 245 62 95 123 ...  
## $ drat: num 3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...  
## $ wt : num 2.62 2.88 2.32 3.21 3.44 ...  
## $ qsec: num 16.5 17 18.6 19.4 17 ...  
## $ vs : num 0 0 1 1 0 1 0 1 1 1 ...  
## $ am : num 1 1 1 0 0 0 0 0 0 0 ...  
## $ gear: num 4 4 4 3 3 3 3 4 4 4 ...  
## $ carb: num 4 4 1 1 2 1 4 2 2 4 ...

head(mtcars,3)

## mpg cyl disp hp drat wt qsec vs am gear carb  
## Mazda RX4 21.0 6 160 110 3.90 2.620 16.46 0 1 4 4  
## Mazda RX4 Wag 21.0 6 160 110 3.90 2.875 17.02 0 1 4 4  
## Datsun 710 22.8 4 108 93 3.85 2.320 18.61 1 1 4 1

Next, we will check if the transmission type (automatic or manual) affects the miles per gallon (MPG) variable. In here, am = 0 is automatic and am = 1 is manual.

library(ggplot2)  
ggplot(data=mtcars, aes(x=as.factor(am), y=mpg, fill = factor(am))) +  
geom\_boxplot(notch = F, show\_guide = TRUE) +  
facet\_grid(. ~ am) +  
xlab("Transmission type") +  
ylab("MPG") +  
ggtitle("Transmission type vs. MPG")



The boxplot above shows that there is a difference in MPG between automatic and manual transmissions. To explore this further, a hypothesis testing is performed. The null hypothesis indicates no performance difference in MPG between automatic and manual cars. The alternative hypothesis indicates that there is a difference in MPG between the 2 transmission types.

testTransmission <- t.test(mtcars[mtcars$am == 0,]$mpg, mtcars[mtcars$am == 1,]$mpg)

This hypothesis test yields p value of 0.0013736, which is smaller than 0.05. This allows us to reject the null hypothesis and conclusively determine that there is a different in MPG between different transmission types.

## Regression Modeling

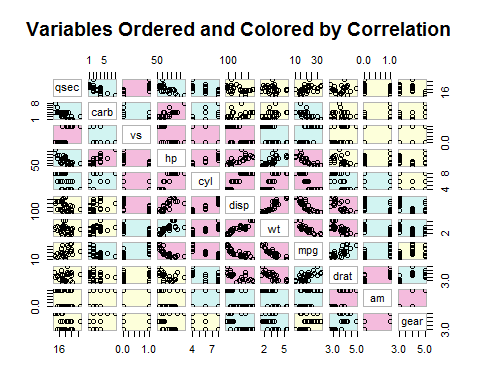
The first step to modeling the regression equation is to identify variables that are significantly correlated to one another. The scatterplot below is ordered and colored by correlation. The ones closer to the diagonal have a stronger correlation.

library(gclus)

## Warning: package 'gclus' was built under R version 3.1.3

## Loading required package: cluster

dta <- mtcars # get data   
dta.r <- abs(cor(dta)) # get correlations  
dta.col <- dmat.color(dta.r) # get colors  
# reorder variables so those with highest correlation  
# are closest to the diagonal  
dta.o <- order.single(dta.r)   
cpairs(dta, dta.o, panel.colors=dta.col, gap=.5,  
main="Variables Ordered and Colored by Correlation" )



Through the plot, we are able to surmise 3 variables with strong correlation: cyl, disp and wt. Using this, we are able to create a combination of regression models:

* Model 1: cyl
* Model 2: disp
* Model 3: cyl + disp
* Model 4: wt
* Model 5: wt + cyl
* Model 6: wt + disp
* MOdel 7: cyl + disp + wt

fitCyl <- lm(mpg ~ cyl, mtcars)  
fitDisp <- lm(mpg ~ disp, mtcars)  
fitCylDisp <- lm(mpg ~ cyl + disp, mtcars)  
fitWt <- lm(mpg ~ wt, mtcars)  
fitWtCyl <- lm(mpg ~ wt + cyl, mtcars)  
fitWtDisp <- lm(mpg ~ wt + disp, mtcars)  
fitAll <- lm(mpg ~ cyl + disp + wt, mtcars)

To ascertain how much the model impacts the MPG performance, the Adj. R Squared value is studied. This value is the percentage of which a variance in the MPG could be explained by the model.

The Adj. R Squared values are: \* Model 1: 0.7170527 \* Model 2: 0.7089548 \* Model 3: 0.7429841 \* Model 4: 0.7445939 \* Model 5: 0.8185189 \* Model 6: 0.7658223 \* Model 7: 0.8146721

The highest Adj. R Squared values belonging to Model 5 is picked.

summary(fitWtCyl)$coefficients

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 39.686261 1.7149840 23.140893 3.043182e-20  
## wt -3.190972 0.7569065 -4.215808 2.220200e-04  
## cyl -1.507795 0.4146883 -3.635972 1.064282e-03

The coefficients of Model 5 indicate that assuming weight stays the same, every increase in cylinder results in decrease of MPG by 1.501 units. Assuming cylinder stays the same, every increase in weight results in decrease of MPG by 3.19 units.

As the transmission type also affects the MPG performance, let us explore how its addition to the model affect the Adj. R Squared value.

fitNew <- lm(mpg ~ wt + cyl + am, mtcars)

## Conclusion

This new model results in Adj. R Squared value of 0.8121603. This new value is very close to Model 5 which is 0.8185189. This indicates that although the transmission type is a predictor of MPG performance, there are other variables that predict MPG at a higher percentage. Therefore, we should select Cylinder and Weight as the variables in the regression model.

To further study the results, the residual plots are explored that show accuracy of independent assumptions, normality of the residual distribution, randomness, and insignificant outliers.

par(mfrow = c(2,2))  
plot(fitWtCyl)

