Regression_Model_Coursera Project

Ch Sovan Krishna Patro

September 22, 2015

Executive summary

As per instruction the dataset to be used: mtcars (extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles. Particularly following two questions to be answered. (1)"Is an automatic or manual transmission better for MPG". (2)"Quantify the MPG difference between automatic and manual transmissions". For the 1st question I have used the t-test to find out the performance between cars with automatic and manual transmission. (Result: It's about 7 MPG more for cars with manual transmission than those automatic ones).For the 2nd question I have used backward regression model to find out the best suited model where I can get the highest adjusted R-squared value which will indicate the highest representation of Dependent variable by the Independent ones. First I will load the data set and change some of the variables from class of numeric to integer. I did some basic exploratory data analyses, which comes under the section Appendix

```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 3.2.2
```

```
require(lattice)
```

```
## Loading required package: lattice
```

```
data(mtcars)
str(mtcars) # getting the structure of the dataset
```

```
## '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 ...
```

```
# Converting non factor variables into Factor variables
mtcars$cyl <- as.factor(mtcars$cyl)
mtcars$vs <- as.factor(mtcars$vs)
mtcars$am <- as.factor(mtcars$am)
mtcars$gear <- as.factor(mtcars$gear)
mtcars$carb <- as.factor(mtcars$carb)
attach(mtcars) # attaching the dataset
```

```
## The following object is masked from package:ggplot2:
##
##     mpg
```

T-test

I will set the null hypothesis as the MPG of the automatic and manual transmissions are from the same population (assuming the MPG has a normal distribution). Two sample t-test will be used. Since the p-value is 0.00137, Null hypothesis will be rejected. So, the automatic and manual transmissions are from different populations. The mean for MPG of manual transmitted cars is about 7 more than that of automatic ones.

```
t_test <- t.test(mpg ~ am)
t_test$p.value
```

```
## [1] 0.001373638
```

Regression Analysis

To find out the best model I have used the backward regression model where the best model (Highest adjusted R-Squared value) will be considered. To save the space I have suppressed the out puts of unwanted ones and mentioned the best model explicitly. The reduced model had an adjusted R^2 of 0.8401 & 3 variables which have stars are have much siginificance than others. more stars more significance.

```
full.model <- lm(mpg ~ ., data = mtcars)
reduced.model <- step(full.model, direction = "backward")
```

```
summary(reduced.model)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -2.16368    2.28425  -0.947  0.35225
## hp          -0.03211    0.01369  -2.345  0.02693 *
## wt          -2.49683    0.88559  -2.819  0.00908 **
## am1          1.80921    1.39630   1.296  0.20646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

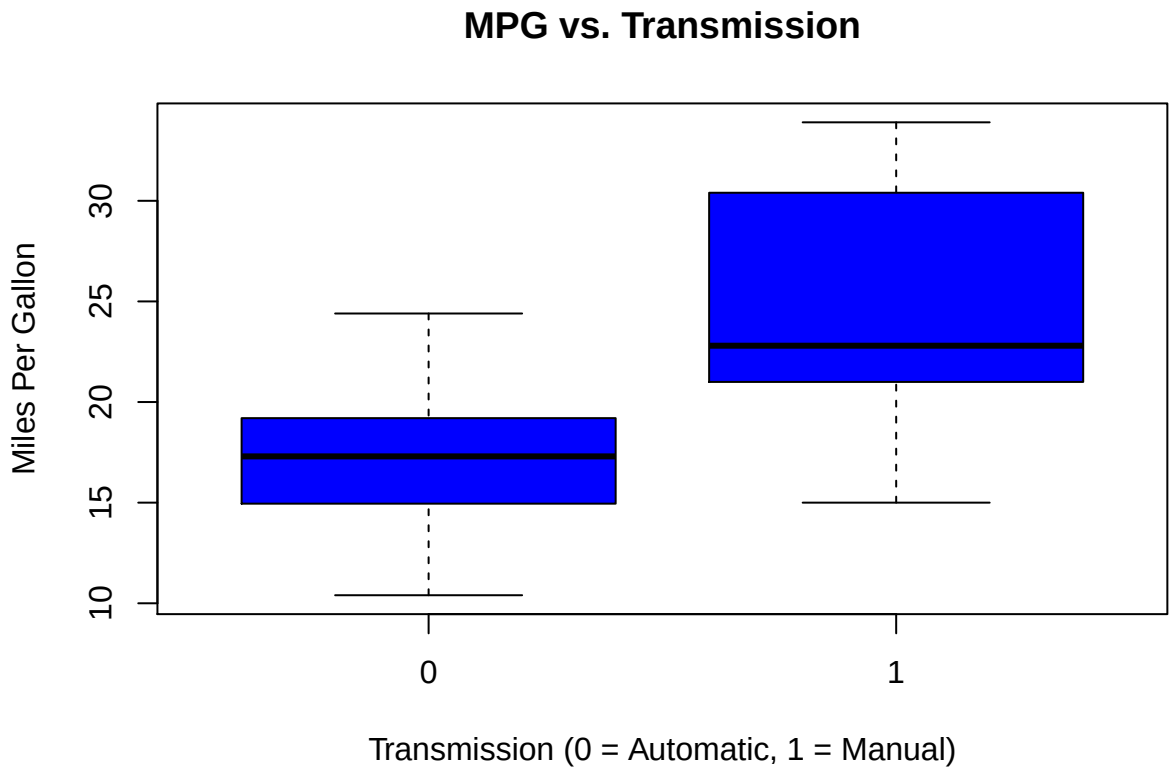
Residual Analysis & Diagnostics

Please refer to the Appendix: Figures section for the plots. According to the residual plots, we can verify the following underlying assumptions The Residuals vs. Fitted plot: Indicates no consistent pattern. The Normal Q-Q plot: Indicates that the residuals are normally. The Scale-Location plot: Confirms the constant variance assumption. The Residuals vs. Leverage: Tells that no outliers are present

Appendix: Figures

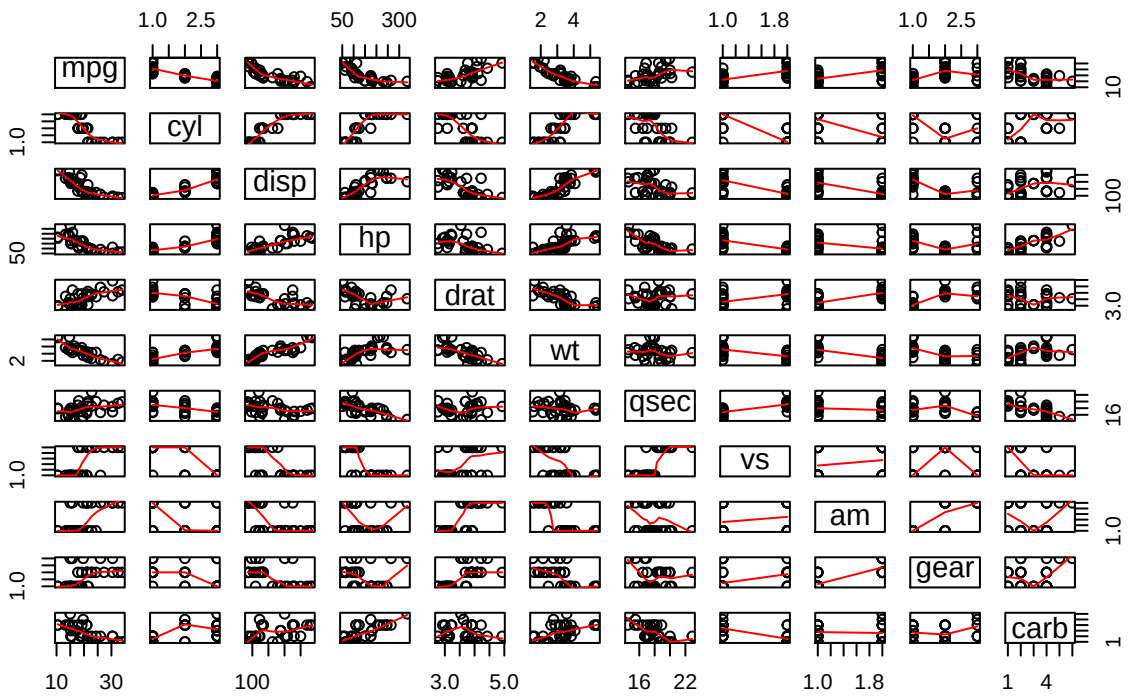
Boxplot of MPPG Vs Transmission

```
with(mtcars, plot(mpg ~ am, col = "Blue",
                  xlab= "Transmission (0 = Automatic, 1 = Manual)",
                  ylab= "Miles Per Gallon", main= "MPG vs. Transmission"))
```



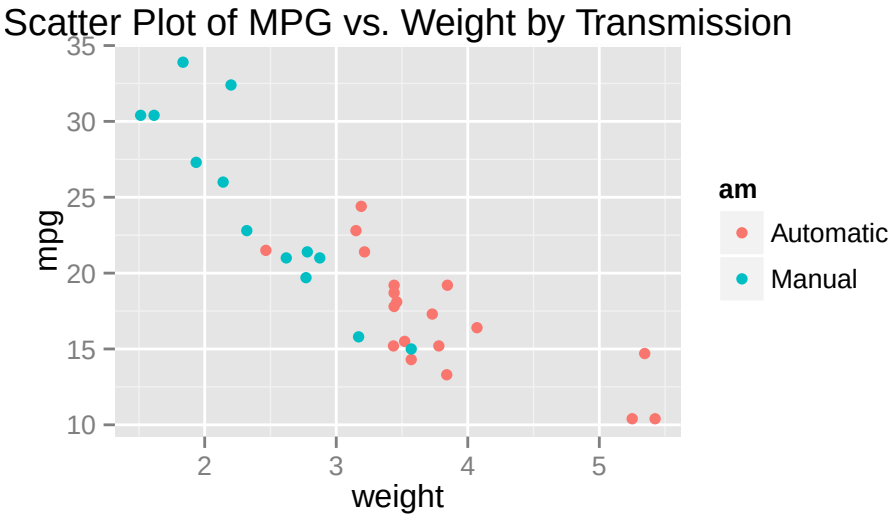
Pair Graph for motor trend car road tests

```
pairs(mtcars, panel=panel.smooth, main="Pair Graph of Motor Trend Car Road Tests")
```



Scatter plot of MPG Vs Weight by transmission

```
g <- ggplot(mtcars, aes(x=wt, y=mpg, group=am, color=am, height=2, width=2))
g <- g + geom_point() + scale_colour_discrete(labels=c("Automatic", "Manual")) + xlab("weight")
g <- g+ ggtitle("Scatter Plot of MPG vs. Weight by Transmission")
plot(g)
```



Residual plots for the regression model

```
par(mfrow=c(2, 2))
    plot(reduced.model)
```