Coursera Regression Models Course Project

E121977

Sunday, February 28, 2016

Excecutive Summary

In this report, we analyze the `mtcars` data set exploring the relationship between a set of variables and miles per gallon (MPG). The data provided includes fuel consumption, 10 aspects of automobile design and performance for 32 automobiles manufactured between 1973 and 1974. Regression models and exploratory data analyses are used to investigate how **automatic (AT)** and **manual (MT)** transmissions affect **MPG**. T-tests show the performance difference between autos with AT and MT revealing a 7+ MPG advantage for autos with a MT. Linear regression models are applied with the highest Adjusted R-squared value selected. When weight and 1/4 mile time are held constant, MT cars are more efficient on average better than AT cars. Autos that are lighter in weight with a manual transmission and cars that are heavier in weight with an automatic transmission will have higher MPG values.

Data Preprocessing and Exploratory Analysis

The data set `mtcars` is loaded and variables are transformed from `numeric` to `factor` class.

```
library(ggplot2)
data(mtcars)
mtcars$cyl <- as.factor(mtcars$cyl)
mtcars$vs <- as.factor(mtcars$vs)
mtcars$am <- factor(mtcars$am)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
attach(mtcars)
```

Basic exploratory data analyses is conducted. Refer to the Appendix section for all plots. The box plot, shows that autos with a MT yields higher values of MPG in general. The pair graph shows higher correlations between the "wt", "disp", "cyl" and "hp" variables.

Inference

Inferences are made based on a two sample T-test. The null hypothesis will be "the MPG of the AT and MT are from the same population (assuming the MPG has a normal distribution) utilizing the two sample T-test.

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

Since the p-value is 0.001373, we reject the null hypothesis. This tells us the AT autos and MT autos are from different populations. The mean/average for AT (group 0) is 17.15 miles per gallon and the mean/average for MT(group 1) is 24.40 miles per gallon. This initially reveals a ~7.25 MPG advantage for MT autos.

Regression Analysis

Now we will build linear regression models based on the different variables in order to determine the best model fit and show comparisons with the base model which we will obtain using "anova".

```
initialModel <- lm(mpg ~ ., data=mtcars)
bestModel <- step(initialModel)
```

The best model based on the above computations consists of variables `cyl`, `wt`, and `hp` as confounders and `am` as the independent variable. This is the model with the highest Adjusted R-squared value.

Looking further into the selected model details:

```
summary(bestModel)
```

This model configuration is "mpg ~ cyl + hp + wt + am". The Residual standard error is 2.41 on 26 degrees of freedom. The Adjusted R-squared value is 0.8401, which means that the model can explain about 84% of the variance of the MPG variable. All of the coefficients are significant yielding at 0.05 significance level.

Fitting the base model with only `am` as the predictor variable coupled with the best model chosen above.

```
amBaseModel <- lm(mpg ~ am, data=mtcars)
anova(amBaseModel, bestModel)
```

Reviewing the results, the p-value is highly significant. This indicates the null hypothesis should be rejected and that the confounder variables `cyl`, `hp`, and `wt` do not contribute to the accuracy of the model.

Residual Analysis and Diagnostics

"Figure 3* in the appendix displays the residual plot of our regression analysis showing the following observations:

- Randomness of the scatter plot points on the "Residuals vs. Fitted" indicate the suspected independence.
- The "Normal Q-Q"" plot shows that the points fall mostly on the line. The residuals are normally distributed.
- The "Scale-Location" scatter plot shows points scattered in a band patter, indicating constant variance.
- There are distinct outlier points in the top right-hand corner of each of these plots.

Investigating the top three outliers in each case of influence measurements:

```
outliers <- hatvalues(bestModel)
tail(sort(outliers),3)
```

```
##       Toyota Corona Lincoln Continental       Maserati Bora
##           0.2777872           0.2936819           0.4713671
```

```
influence <- dfbetas(bestModel)
tail(sort(influence[,6]),3)
```

```
## Chrysler Imperial          Fiat 128     Toyota Corona
##         0.3507458         0.4292043         0.7305402
```

The influence analysis performed was accurate, since the same autos are shown in the residual plots.

Conclusion (completed and proofed)

Given the analysis above, our best fit model shows that autos with MT perform better in MPG compared with vehicles with AT. Also, the MPG will decrease by a factor of 2.5 for every 1,000 pounds increase in weight (wt). There is also a slight decrease in MPG as horsepower (HP) increases. As the number of cylinders increase from 4 to 6 and 6 to 8 we see a decrease in miles per gallon by a factor of 3 and 2.2, respectively.

Appendix: Figures

Figure 1 - Pair Graph: Motor Trend Car Road Tests

```
pairs(mtcars, panel=panel.smooth, main="Pairs Plot for mtcars Data Set")
```

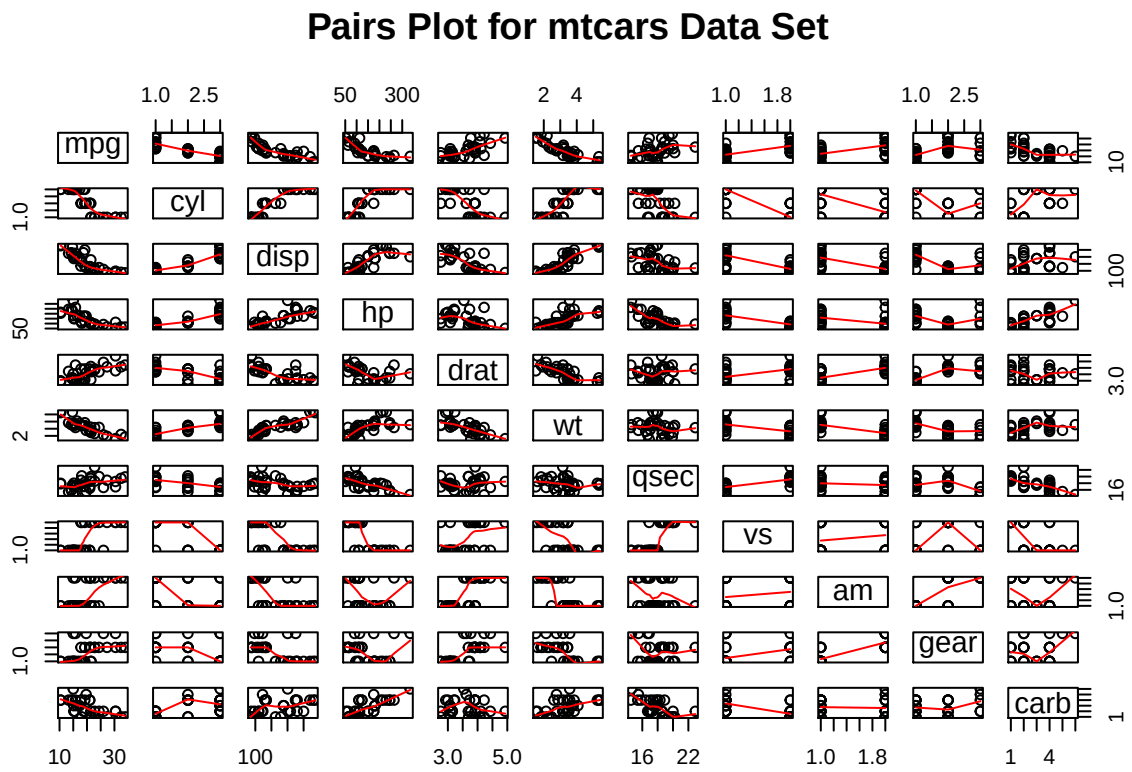
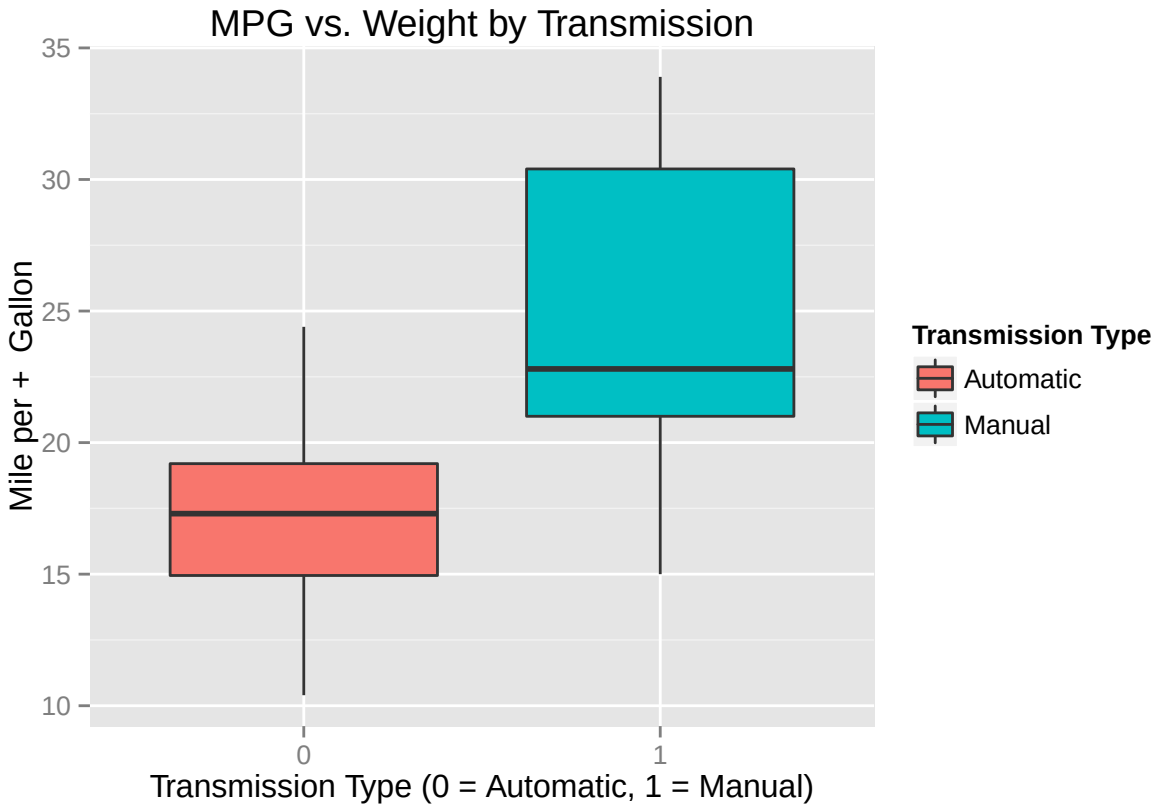


Figure 2 - Boxplot: MPG vs. Transmission

```
p <- ggplot(mtcars, aes(factor(am), mpg))
p + geom_boxplot(aes(fill = factor(am))) + xlab("Transmission Type (0 = Automatic, 1 = Manual)") +
        ylab("Mile per +  Gallon") + ggtitle("MPG vs. Weight by Transmission") +
        scale_fill_discrete(name="Transmission Type", labels=c("Automatic", "Manual"))
```



3. Residual Plots

```
residualData <- lm(mpg ~ cyl + hp + wt + am, data=mtcars)
par(mfrow = c(2, 2))
plot(residualData)
```