RegressionModels_CourseProject

Joshua

23 September 2015

Regression Models - Course Project

Executive summary

This report explores the relationship between a set of variables and miles per gallon (MPG) from the mtcars data set and attempts to determine if an automatic or manual transmission better for MPG and quantifies the actual different in MPG.

Both the simple linear and multiple regression models indicate that manual transmission have on average a higher miles per gallon figure than automatic transmission. In the simple linear regression model between MPG and transmission type alone the increase in miles per gallon is approximately 7 miles per gallon. In the multiple regression model, considering the other predictors weight, horsepower and number of cylinders, the increase is approximately 1.8 MPG when switching from an automatic transmission to a manual one, with all other predictors held constant.

Exploratory data analysis and diagramatic plots are located in the Appendix of this document.

Data processing

We begin by loading the mtcars dataset and converting some variables to factors.

Analysis

Simple Linear Regression between mpg and transmission

Is there a different between mpg achieved with manual vs automatic transmissions? The boxplot in Figure 1 in the appendix clearly indicates a higher mean mpg figure when comparing automatic vs manual transmissions, on average about 7 more miles to the gallon. A t-test shows that the difference in means mpg is significant. (The p-value < 0.05).

```
## 
## Call:
## lm(formula = mpg ~ TransType, 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 ***
## TransTypeManual    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
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by TransType
## 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 in group Automatic    mean in group Manual
##                17.14737                24.39231
```

However we can also see that the adjust R-squared value of the simple linear regression model is 0.3385, meaning a different tranmission only explains 33.85% of the variance in mpg.

Multiple regression analysis

From the pairs plot in Figure 2, several other variables seem to have high correlation with mpg. To find the best model fit build an initial model including all the variables as predictors and perform a stepwise model selection to determine which are the predictors required for the best fit. The R function step can do this quite easily, it chooses a model based on AIC. (https://en.wikipedia.org/wiki/Akaike_information_criterion)

From the best model we see that in addition to transmission type, cyl, hp, and wt are all also important variables in predicting mpg. The adjusted R-squared value is .8401 which tells us that 84% of the variability is explained by this model. Comparing both the simple linear regression model and the best fit model, the p-value is very small, indicating that the two models are significantly different. The combination of weight, horsepower and transmission type explain 84% of the variability in mileage achieved per gallon.

```
## Analysis of Variance Table
##
## Model 1: mpg ~ TransType
## Model 2: mpg ~ cyl + hp + wt + TransType
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     26 151.03  4    569.87 24.527 1.688e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Diagnostics

Looking at a plot of the best model in Figure 3. we can see that: 1. The Residuals vs Fitted and Scale-Location plots do not have any systematic pattern. 2. The normal Q-Q plot shows the point generally clustered along the line, indicating the residuals are normally distributed. 3. There are a few

Influential points: Outlier points: Leverage:

Appendix

```
library(ggplot2)
```

Figure 1: Box plot of mpg by transmission type

```
fig1<-ggplot(mtcars, aes(factor(TransType), mpg, fill=factor(TransType))) +
  geom_boxplot() +
  scale_colour_discrete(name = "Type") +
  scale_fill_discrete(name="Type", breaks=c("0", "1"),labels=c("Automatic", "Manual")) +
  scale_x_discrete(breaks=c("0", "1"), labels=c("Automatic", "Manual")) +
  xlab("Transmission Type: Automatic, Manual")
fig1
```

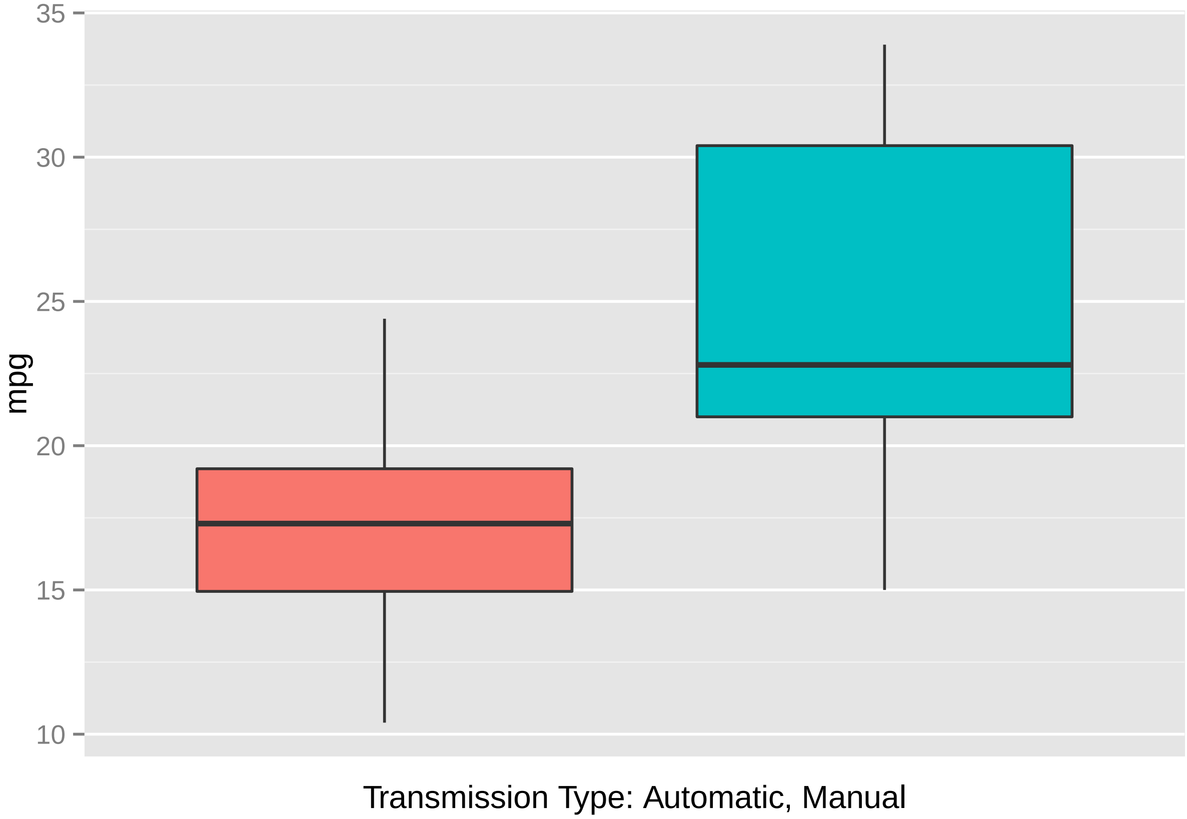
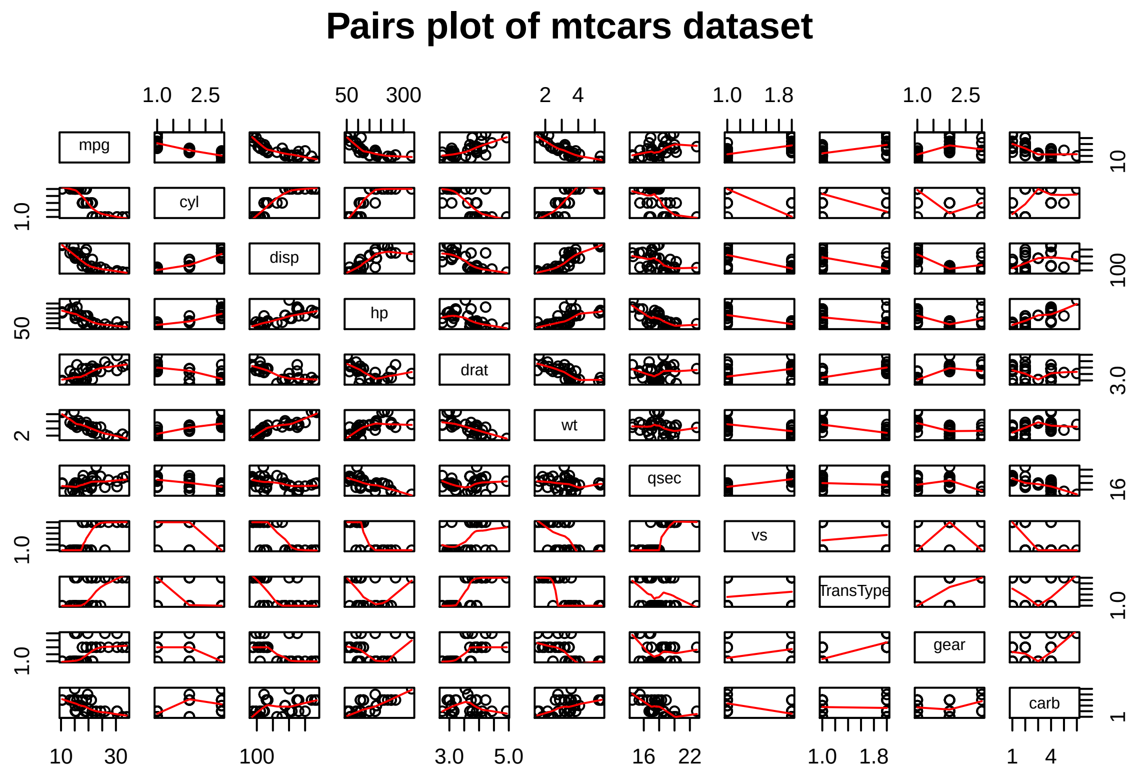


Figure 2: Pairs plot

```
fig2<-pairs(mtcars,panel=panel.smooth,col=9,main="Pairs plot of mtcars dataset")
```

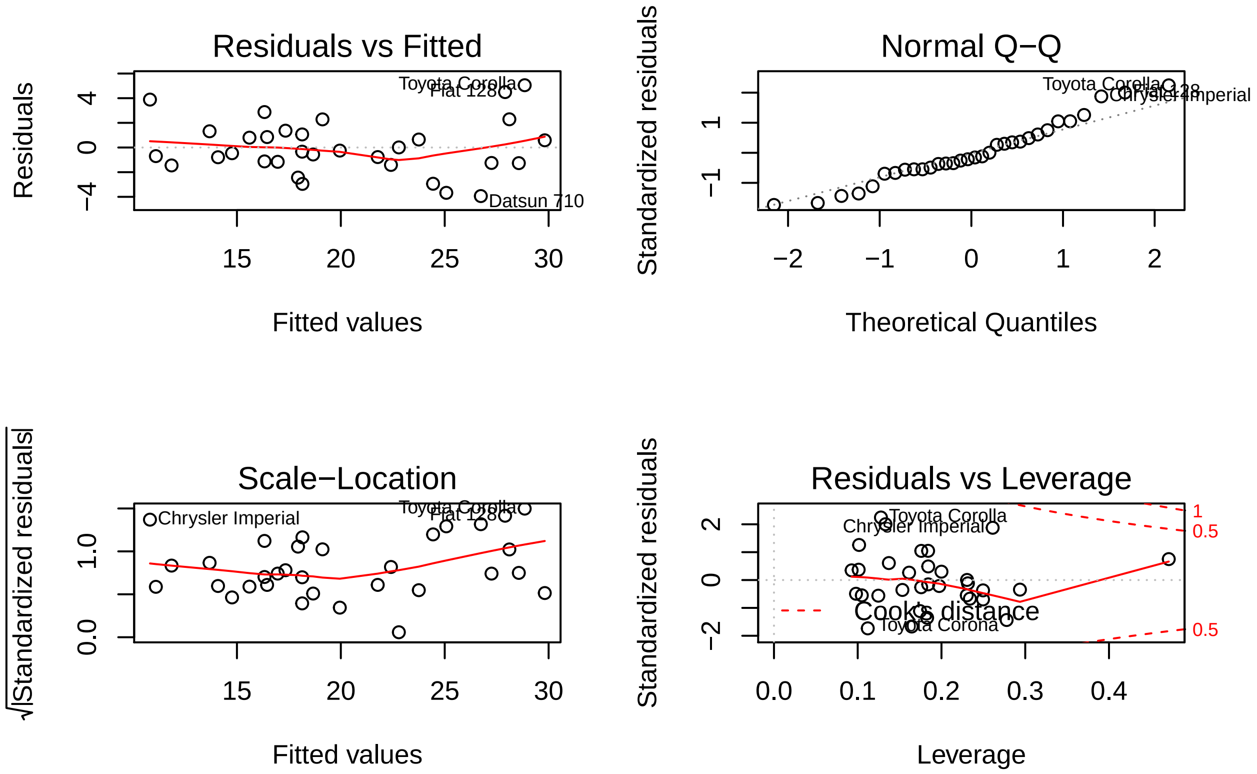


```
fig2
```

```
## NULL
```

Figure 3: Plot of best fit model

```
par(mfrow =c(2,2))
plot(bestmodel)
```


Residuals vs Fitted
Residuals
Fitted values
−4
0
4
15
20
25
30
Toyota Corolla
Fiat 128
Datsun 710
Normal Q−Q
Standardized residuals
Theoretical Quantiles
−1
1
−2
−1
0
1
2
Toyota Corolla
Fiat 128
Chrysler Imperial
Scale−Location
√|Standardized residuals|
Fitted values
0.0
1.0
15
20
25
30
Chrysler Imperial
Toyota Corolla
Fiat 128
Residuals vs Leverage
Standardized residuals
Leverage
−2
0
2
0.0
0.1
0.2
0.3
0.4
Toyota Corolla
Chrysler Imperial
Toyota Corona
Cook's distance
1
0.5
0.5