Regression Models - Course Project

Gildardo Rojas Nandayapa

Saturday, October 25, 2014

Data Science Specialization, Johns Hopkins University

Executive Summary

Motor Trend, a leading magazine about the automobile industry is interested in looking at a data set of a collection of cars, exploring the relationship between a set of variables and miles per gallon (MPG). There is particular interest in the following two questions:

1. Is an automatic or manual transmission better for MPG?
2. What is the difference in MPG between automatic and manual transmissions?

Using linear regression analysis, we can determine that there is a signficant difference between the mean MPG for automatic and manual transmission cars. Manual transmissions achieve a higher value of MPG compared to automatic transmission.

Transmission type is relevant, but not an unique factor predicting the mpg outcome as we may see in the following analysis.

The Data

The data was extracted from the 1974 Motor Trend US magazine, it comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models).

Basic summary of the data.

```
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 ...
```

```
summary(mtcars)
```

```
##       mpg             cyl             disp             hp       
##  Min.   :10.4   Min.   :4.00   Min.   : 71.1   Min.   : 52.0  
##  1st Qu.:15.4   1st Qu.:4.00   1st Qu.:120.8   1st Qu.: 96.5  
##  Median :19.2   Median :6.00   Median :196.3   Median :123.0  
##  Mean   :20.1   Mean   :6.19   Mean   :230.7   Mean   :146.7  
##  3rd Qu.:22.8   3rd Qu.:8.00   3rd Qu.:326.0   3rd Qu.:180.0  
##  Max.   :33.9   Max.   :8.00   Max.   :472.0   Max.   :335.0  
##       drat             wt             qsec             vs        
##  Min.   :2.76   Min.   :1.51   Min.   :14.5   Min.   :0.000  
##  1st Qu.:3.08   1st Qu.:2.58   1st Qu.:16.9   1st Qu.:0.000  
##  Median :3.69   Median :3.33   Median :17.7   Median :0.000  
##  Mean   :3.60   Mean   :3.22   Mean   :17.8   Mean   :0.438  
##  3rd Qu.:3.92   3rd Qu.:3.61   3rd Qu.:18.9   3rd Qu.:1.000  
##  Max.   :4.93   Max.   :5.42   Max.   :22.9   Max.   :1.000  
##        am              gear            carb      
##  Min.   :0.000   Min.   :3.00   Min.   :1.00  
##  1st Qu.:0.000   1st Qu.:3.00   1st Qu.:2.00  
##  Median :0.000   Median :4.00   Median :2.00  
##  Mean   :0.406   Mean   :3.69   Mean   :2.81  
##  3rd Qu.:1.000   3rd Qu.:4.00   3rd Qu.:4.00  
##  Max.   :1.000   Max.   :5.00   Max.   :8.00
```

The data set contains the following variables:**mpg** - Miles/(US) gallon, **cyl** - Number of cylinders, **disp** - Displacement (cu.in.), **hp** - Gross horsepower, **drat** - Rear axle ratio, **wt** - Weight (lb/1000), **qsec** 1/4 mile time, **vs** V/S, **am** - Transmission (0 = automatic, 1 = manual), **gear** - Number of forward gears, **carb** Number of carburetors.

Exploratory Analysis

To find relationships between all variables, data exploration can be extended visually using the pairs plot. By inspecting the plot we may notice variables related to mpg, though the impact varies in magnitude and slope. See **Appendix - Figure 1. Motor Trend Car Road Tests - Variable pairs plot.**

This study is focused on the effects of car transmission type over mpg efficiency, so a simple box plot may help to depict the difference between cars with automatic and manual transmission. The plot shows that manual transmissions have higher mpg. See **Appendix - Figure 2. MPG by Transmission Type - Box plot**

Statistical Inference

T-test Assuming a normal distribution, by performing a t-test we may observe that there is a significant difference between manual and automatic transmission types in the resulting mpg.

```
t.test(mpg ~ am, data = mtcars)
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## t = -3.767, df = 18.33, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.28  -3.21
```

```
## sample estimates:
## mean in group 0 mean in group 1
##           17.15           24.39
```

Here we may also observe that the difference in mpg for vehicles with manual over automatic transmission type is 7.24.

Regression Analysis

Linear Regression

First we attempt simple linear regression.

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

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.392 -3.092 -0.297  3.244  9.508
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    17.15       1.12   15.25  1.1e-15 ***
## am              7.24       1.76    4.11  0.00029 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.9 on 30 degrees of freedom
## Multiple R-squared:  0.36,   Adjusted R-squared:  0.338
## F-statistic: 16.9 on 1 and 30 DF,  p-value: 0.000285
```

Observing the summary information, we can confirm that manual transmission vehicles have 7.24 MPG more than automatic ones. But this model doesn't provide complete information as the Multiple R-squared value of .3598 means the model can just explain 35.98% of the variance.

In order to fully understand the transmission type impact considering other variables, we need to create a multivariate model.

Building and selecting the Model

To start, we build the first model which considers all variables as predictors. Then we proceed to select the most significant predictors to build the best model.

The step function performs the best model selection, it calls lm repeatedly to build regression models selecting the most significant variables that can be considered as relevant mpg predictors, while discarding the less significant ones.

```r
firstmodel <- lm(mpg ~ ., data=mtcars)
bestmodel <- step(firstmodel, direction="both", trace=0)
summary(bestmodel)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.481 -1.556 -0.726  1.411  4.661
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.618      6.960    1.38  0.17792
## wt            -3.917      0.711   -5.51    7e-06 ***
## qsec           1.226      0.289    4.25  0.00022 ***
## am             2.936      1.411    2.08  0.04672 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.46 on 28 degrees of freedom
## Multiple R-squared:  0.85,   Adjusted R-squared:  0.834
## F-statistic: 52.7 on 3 and 28 DF,  p-value: 1.21e-11
```

```r
anova(firstmodel,bestmodel)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
## Model 2: mpg ~ wt + qsec + am
##   Res.Df RSS Df Sum of Sq    F Pr(>F)
## 1     21 148
## 2     28 169 -7     -21.8 0.44   0.86
```

By displaying the model summary, we can see that Weight and Accelaration are also relevant to predict mpg. Here the Multiple R-squared value of .8497 means the model can now explain 84.97% of the mpg variance, making it a better predictive model. We also computed the analysis of variance (or deviance) tables for one the previously fitted models (firstmodel and bestmodel).

Analysis of residuals

Now with the model selection done, we proceed with the analysis of residuals.

While reading the following observations, please see **Figure 3 - Residual plots.**

Observations

Residual vs Fitted plot The independence condition may be confirmed with the randomly scatterd points that this plot shows.

Normal Q-Q plot Normal distribution of residuals can be confirmed as most points fall in the line.

Scale-Location plot Constant variance can be observed as points are scattered in a constant band.

Residuals vs Leverage plot In this plot we can observe some outliers or leverage points, the points in the top right may indicate values with increased leverage of outliers.

Now we will also perform of regression diagnostics of the best model to find leverage points and any potential problems with the model.The function hatvalues() is used to find data points with most leverage and dfbetas() is used to find the data points that have bigger influence in the model coefficients.

```
leverage <- hatvalues(bestmodel)
tail(sort(leverage),3)
```

```
##   Chrysler Imperial Lincoln Continental           Merc 230
##              0.2296              0.2642             0.2970
```

```
influential <- dfbetas(bestmodel)
tail(sort(influential[,4]),3)
```

```
##     Toyota Corona          Fiat 128 Chrysler Imperial
##            0.4050            0.4766            0.5626
```

Conclusions

Based on the analysis and models built, we can depict the following conclusions:

- MT Vehicles are better to get more MPG compared with AT vehicles.
- MT vehicles have in average 7.24 MPG more than AT vehicles.
- Transmission Type has relevance as a predictor for MPG result, but there are other variables like Weight and Acceleration (wt and qseq) that also have strong influence predicting the mpg outcome.
- The model generated in this analysis has an 85% of accuracy adjusted to the above mentioned variables.

Apendix

Figure 1. Motor Trend Car Road Tests - Variable pairs plot

```
require(stats)
require(graphics)
pairs(mtcars, pane=panel.smooth, main="Motor Trend Car Road Tests", col=3 + (swiss$Catholic>50))
```

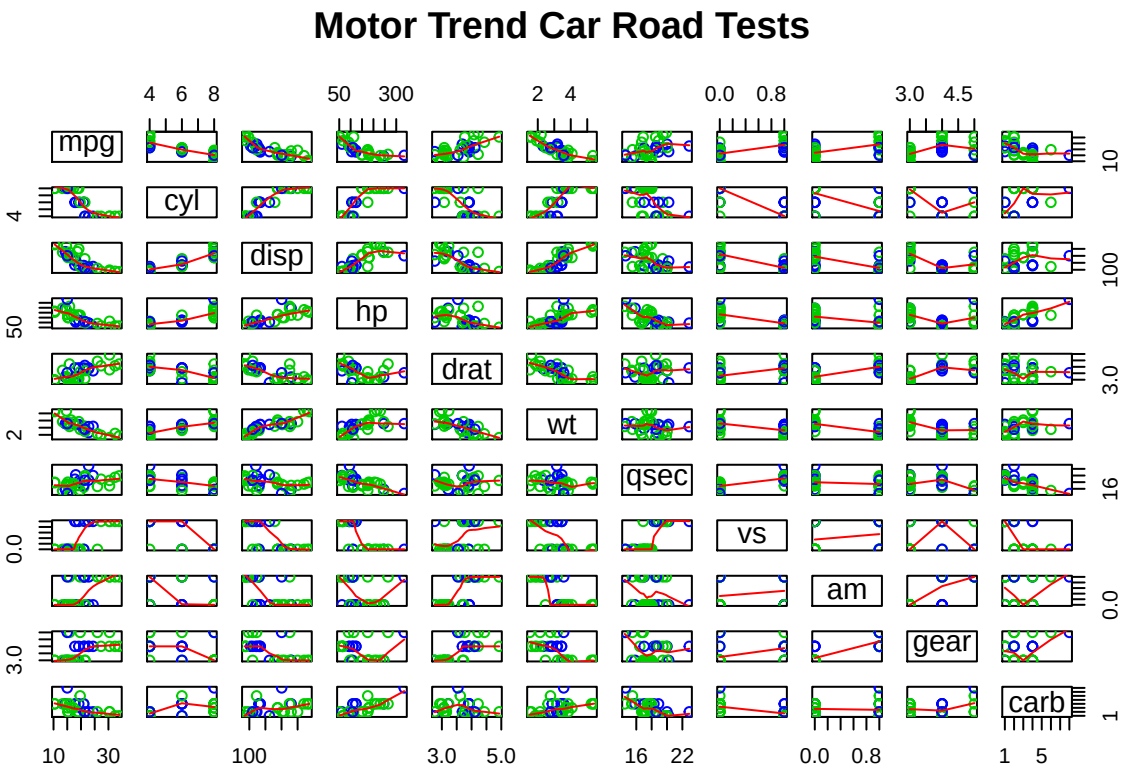


Figure 2. MPG by Transmission Type - Box plot

```
boxplot(mpg ~ am, data=mtcars, ylab="Miles per Gallon", xlab="Transmission Type (0=Automatic, 1=Manual)"
```

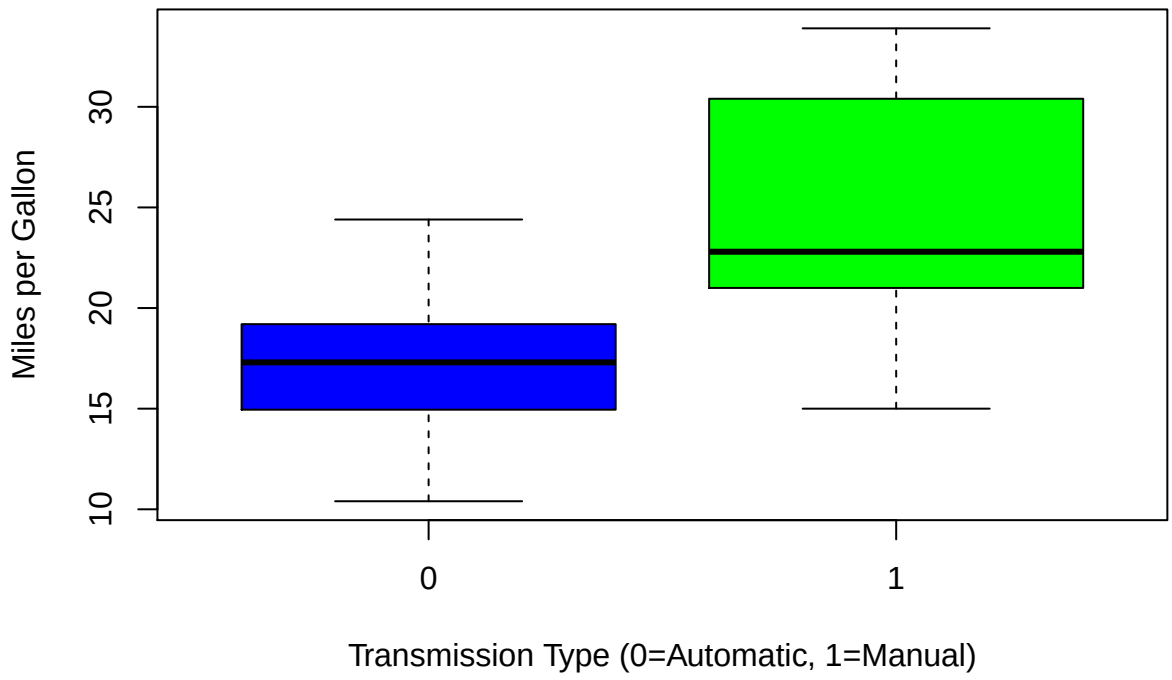
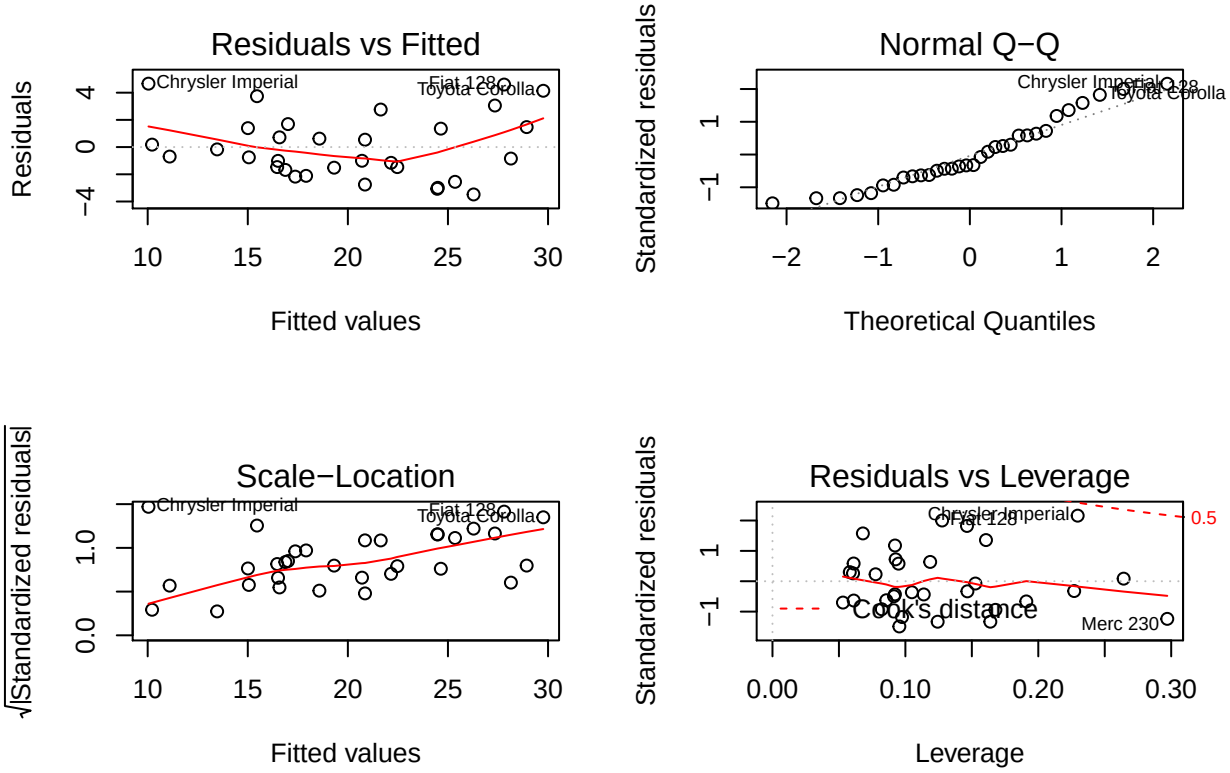


Figure 3. Residual Plots

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


Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Residuals vs Leverage
Leverage
Cook's distance
0.5
Merc 230