Coursera - Regression Models course - Course Project 1

Martin Cote

May 24, 2015

Executive Summary

Using the "mtcars" dataset (a standard dataset provided with R), we try to answer, using statistical regresssion models, the following questions:

- Is an automatic or manual transmission better for MPG
- Quantify the MPG difference between automatic and manual transmissions

Pre-Processing

- Loading libraries (GGPLOT2 & CAR required);
- Loading the required data ("mtcars", provided with R); and
- Cleaning the data (i.e. switching 'am' to a factor variable - for specific processing only).

Note: All pre-processing is hidden.

Exploratory Data Analysis

Analysis: Is transmission (i.e. 'am', 0 = automatic and 1 = manual) a good predictor for the 'mpg'?

Using first a plot to demonstrate the relation and validate if there are, if any, possible relation between the two random variables (refer to figure 1).

Inference: Validating the hypothesis using a Student t-test

```
test <- t.test(mpg ~ am, data = mtcars_f, paired=FALSE, var.equal=FALSE)
test$p.value
```

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

```
test$conf.int
```

```
## [1] -11.280194  -3.209684
## attr(,"conf.level")
## [1] 0.95
```

We observe the p-value of the **Student's Test Distribution** is far below 0.05 which indicates there is a solid relation between Transmission (i.e. 'am') and MPG (i.e. 'mpg'), giving a confidence interval of -11 to -3 MPG at 95% (i.e. a decrease).

Regression Models

Linear Model between Transmission and MPG

First, we will validate the linear model between the two desired variables to study.

```
fit <- lm(mpg ~ am, data = mtcars_f)
summary(fit)$coef # printing out only the coefficients
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## amManual     7.244939   1.764422  4.106127 2.850207e-04
```

As we can see, there is indeed a linear relationship (which we can be observed in figure 2 as well). However, to investigate the validity of this linear model, we investigate the residuals (via plot) and confirm that the variability is much greater for the manual transmission (see firgure 3), which could potentially further investigation required (see *selecting the linear model* section).

Quantifying the differences

Using the linear model (without intercept for quick validation), we can obtain the differences of the impacts on both:

```
fit.nointer <- lm(mpg ~ am - 1, data = mtcars_f)
summary(fit.nointer)$coef # printing out only the coefficients
```

```
##             Estimate Std. Error  t value     Pr(>|t|)
## amAutomatic 17.14737   1.124603 15.24749 1.133983e-15
## amManual    24.39231   1.359578 17.94109 1.376283e-17
```

On average, a manual car will provide a ~7 MPG difference avantage over an automatoic (~24 MPG versus ~17 MPG).

Diagnostics

Since data entry was performed by hand, we use the handy "leverage measures" to validate if any data is an outlier. The test confirms that none of the *hatvalues* are too different (i.e. 2, 3 times greater than the mean - '0' means 'FALSE'):

```
## [1] 0
```

Further investigations: selecting models

Since the residual investigation confirmed that other variables might be at play, it is suggested to look at other models. Looking at all variables within a linear model, we can see that 'qsec' and 'wt' seems to have an impact (**note**: hidden for quicker read and brief).

Since these variables have an impact, we then study a second regression model:

```
summary(lm(mpg ~ am + wt + qsec, data = mtcars))$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.617781  6.9595930  1.381946 1.779152e-01
## am           2.935837  1.4109045  2.080819 4.671551e-02
## wt          -3.916504  0.7112016 -5.506882 6.952711e-06
## qsec         1.225886  0.2886696  4.246676 2.161737e-04
```

As we can observe from the coefficients, 'wt' (the weight) seems to have a greater impact overall although all of them have a solid relation using the p-values calculated.

Appendix

Figure 1 - Exploratory Analysis (MPG by Transmission)

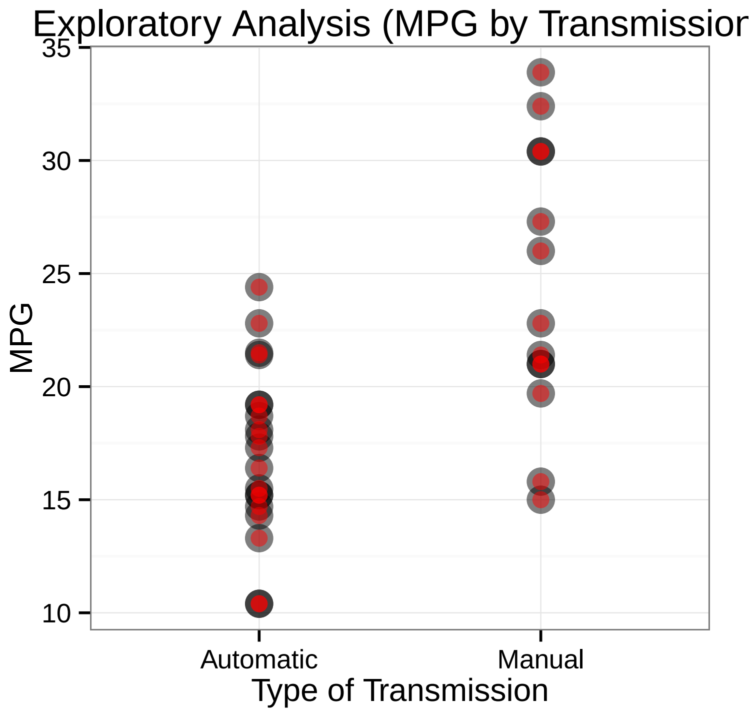


Figure 2 - Linear Regression Model for MPG by Transmission

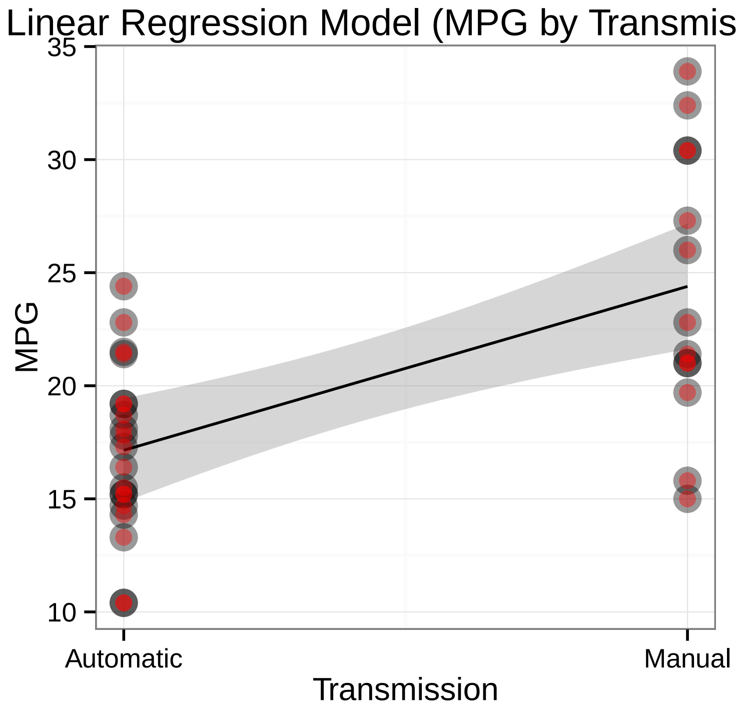


Figure 3 - Verifying variability within the residual

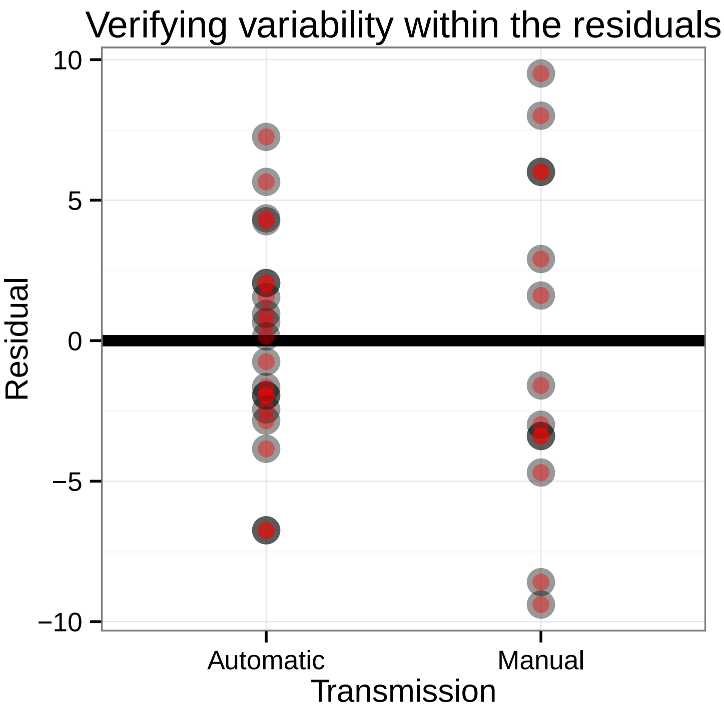
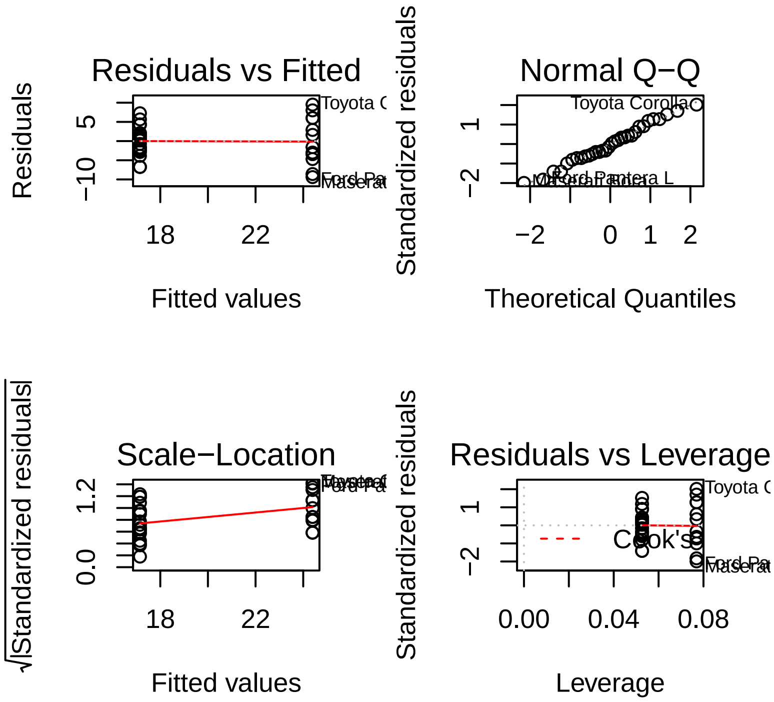


Figure 4 - Diagnostics


Residuals vs Fitted
Residuals
5
−10
18
22
Fitted values
Toyota C
Ford Pa
Maserat
Normal Q−Q
Standardized residuals
1
−2
−2
0
1
2
Theoretical Quantiles
Toyota Corolla
Ford Pantera L
Maserati Bora
Scale−Location
√|Standardized residuals|
1.2
0.0
18
22
Fitted values
Residuals vs Leverage
Standardized residuals
1
−2
0.00
0.04
0.08
Leverage
Toyota C
Ford Pa
Maserat
Cook's