Coursera - Regression Models course - Course Project 1

Martin Cote

May 24, 2015

Executive Summary

The following report is part of the Regression Models course on Coursera.org. Using the "mtcars" dataset (a standard dataset provided with R from the **Motor Trend** US magazine in 1974), 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

Validating the inference

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).

Validating the hypothesis using a Student t-test

```
t.test(mpg ~ am, data = mtcars_f, paired=FALSE, var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## 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
```

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').

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)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars_f)
##
## 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 ***
## amManual       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
```

As we can see, there is indeed a linear relationship (which we can be observed in figure 2 as well). However, since the variable is not continuous, we investigate the residuals (via plot) and confirm that the variability is much greater for the manual transmission (see firgure 3).

Selecting the models

Although we know that the transmission does have an impact on the MPG, we want to know if there are any other random variables part of the dataset that could influence (as much as, lower or higher) the MPG value observed.

The first validation we perform is to confirm if any variables are dropped as par of the Variance Inflation Factors:

```
fit <- lm(mpg ~ ., data = mtcars) # Use ?influence.measures for more information
vif(fit)
```

```
##       cyl      disp        hp      drat        wt      qsec        vs
## 15.373833 21.620241  9.832037  3.374620 15.164887  7.527958  4.965873
##        am      gear      carb
##  4.648487  5.357452  7.908747
```

No random variables from the dataset are dropped, hence all of them could potentially add information to the linear model and the model selected.

Second, we validate the correlation for each variable and MPG as well as completing a first linear model with all variables included and no *intercept*.

```
cor(mtcars$mpg, mtcars$am)
```

```
## [1] 0.5998324
```

```
cor(mtcars$mpg, mtcars$cyl)
```

```
## [1] -0.852162
```

```
cor(mtcars$mpg, mtcars$disp)
```

```
## [1] -0.8475514
```

```
cor(mtcars$mpg, mtcars$hp)
```

```
## [1] -0.7761684
```

```
cor(mtcars$mpg, mtcars$drat)
```

```
## [1] 0.6811719
```

```
cor(mtcars$mpg, mtcars$wt)
```

```
## [1] -0.8676594
```

```
cor(mtcars$mpg, mtcars$qsec)
```

```
## [1] 0.418684
```

```
cor(mtcars$mpg, mtcars$vs)
```

```
## [1] 0.6640389
```

```
cor(mtcars$mpg, mtcars$gear)
```

```
## [1] 0.4802848
```

```
cor(mtcars$mpg, mtcars$carb)
```

```
## [1] -0.5509251
```

```r
summary(lm(mpg ~ . -1, data = mtcars))
```

```
##
## Call:
## lm(formula = mpg ~ . - 1, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7721 -1.6249  0.1699  1.1068  4.4666
##
## Coefficients:
##      Estimate Std. Error t value Pr(>|t|)
## cyl   0.35083    0.76292   0.460   0.6501
## disp  0.01354    0.01762   0.768   0.4504
## hp   -0.02055    0.02144  -0.958   0.3483
## drat  1.24158    1.46277   0.849   0.4051
## wt   -3.82613    1.86238  -2.054   0.0520 .
## qsec  1.19140    0.45942   2.593   0.0166 *
## vs    0.18972    2.06825   0.092   0.9277
## am    2.83222    1.97513   1.434   0.1656
## gear  1.05426    1.34669   0.783   0.4421
## carb -0.26321    0.81236  -0.324   0.7490
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.616 on 22 degrees of freedom
## Multiple R-squared:  0.9893, Adjusted R-squared:  0.9844
## F-statistic:   203 on 10 and 22 DF,  p-value: < 2.2e-16
```

Since the correlation values vary from -1 to 1, 0 being the weakest correlation, along with the interpretation of the first linear model, we include the following random variables:

- am (transmission)
- cyl (number of cylinders)
- disp (displacement)
- hp (horsepower)
- drat (rear axle ratio)
- wt (weight)
- vs (V/S)
- carb (number of carburetors)

The regression model then becomes:

```r
fit.all <- lm(mpg ~ am + cyl + disp + hp + drat + wt + vs + carb, data = mtcars)
summary(fit.all)
```

```
##
## Call:
## lm(formula = mpg ~ am + cyl + disp + hp + drat + wt + vs + carb,
##     data = mtcars)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9540 -1.5150 -0.2235  1.6171  5.1623
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.769832   9.884770   3.113   0.0049 **
## am           2.129301   1.827570   1.165   0.2559
## cyl         -0.554718   0.930564  -0.596   0.5569
## disp         0.008244   0.016788   0.491   0.6280
## hp          -0.023384   0.021243  -1.101   0.2824
## drat         0.764939   1.607323   0.476   0.6386
## wt          -2.783772   1.555905  -1.789   0.0868 .
## vs           1.191481   1.932518   0.617   0.5436
## carb        -0.320044   0.697347  -0.459   0.6506
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.614 on 23 degrees of freedom
## Multiple R-squared:  0.8604, Adjusted R-squared:  0.8118
## F-statistic: 17.72 on 8 and 23 DF,  p-value: 3.965e-08
```

Conclusion

Appendix

Figure 1 - Exploratory Analysis (MPG by Transmission)

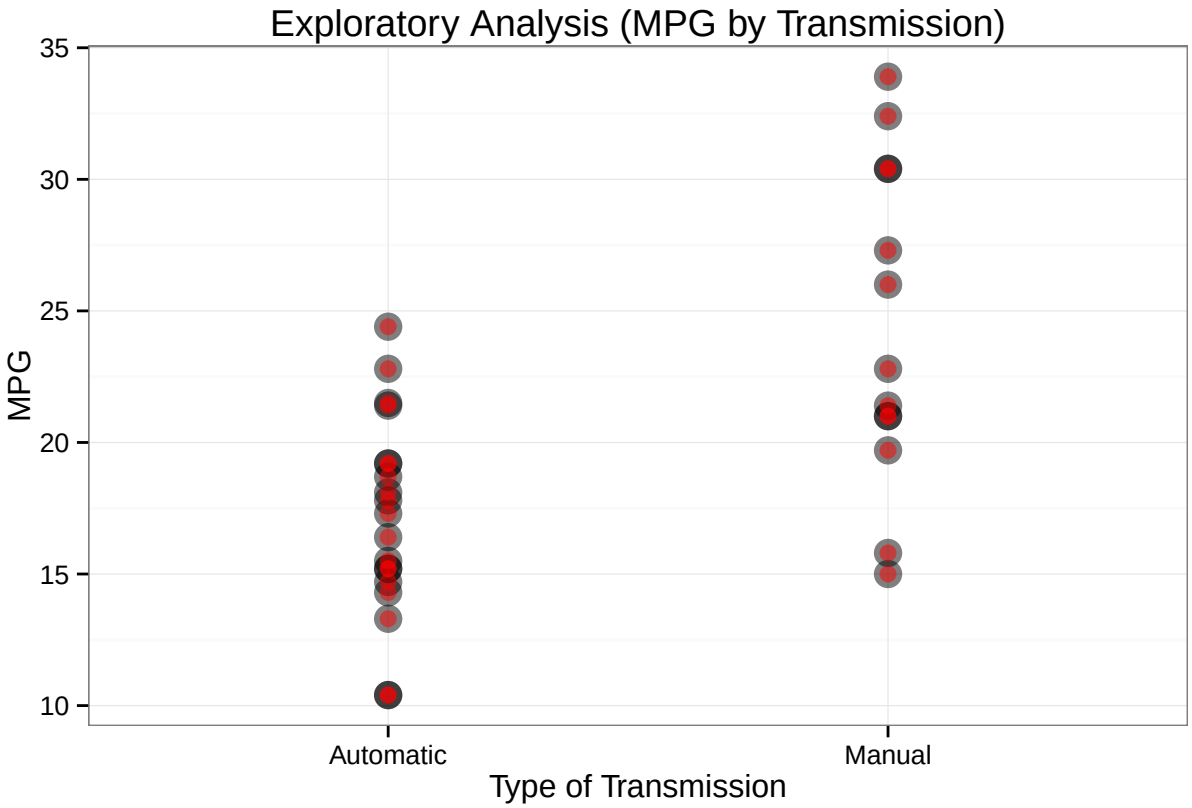

Figure 2 - Linear Regression Model for MPG by Transmission

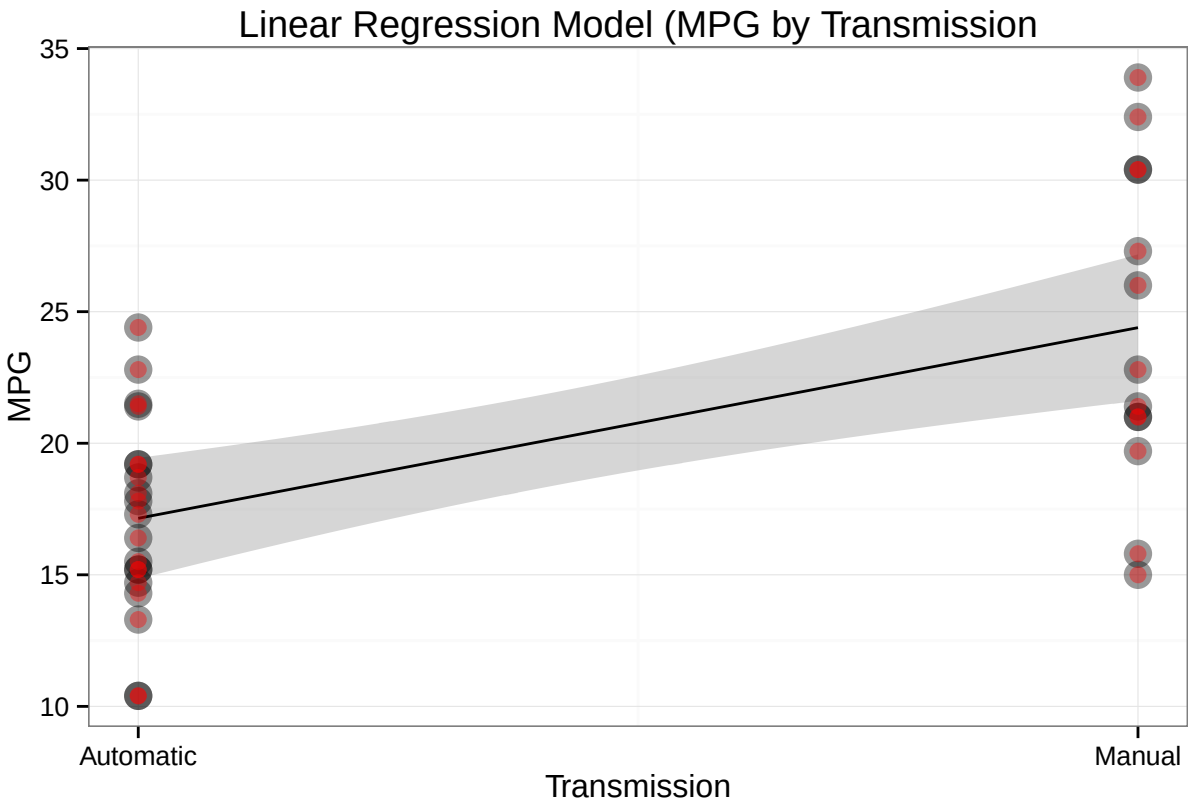

Figure 3 - Verifying variability within the residual