Analysing Fuel consumption of mtcars data set in R database packages

Araks Stepanyan

8/10/2017

Executive Summary

This document is the final project of Coursera Regression Models course by Johns Hopkins University. Using R's *mtcars* data set, the main goal is to answer the following two questions. 1) Is an automatic or manual transmission better for MPG(miles per gallon)? 2) Can we quantify the MPG difference between automatic and manual transmissions?

We find that although 95% of the times manual transission is resulting on average 3.64 to 10.85 more miles per gallon, after we adjust for Number of cylinders, Gross horsepower and Weight, the difference between manual and automatic (although still positive) is not significant any longer.

Exploratory Data Analysis

Let's load the *mtcars* data set and look at it's first 3 rows.

```
##               mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4      21   6  160 110  3.9 2.620 16.46  0  1    4    4
## Mazda RX4 Wag  21   6  160 110  3.9 2.875 17.02  0  1    4    4
```

```
## [1] "Observations: 32 , Variables: 11"
```

Relationship between mpg and am (Appendix, Figure 1). Manual transmission (am = 1) seems to be more effective as it results in more miles per gallon then automatic transmission (am = 0).

Pairwise relationships (Appendix, Figure 2). 1) Variables **cyl**, **vs**, **am**, **gear**, and **carb** have only few levels, so we will treat them as factors. 2) **disp** and **cyl** are highly correlated (because dispacement divided by number of cyliders is the mass of one cylinder), so we will include only one of them in our regression model.

Initial Regression (*mpg* is the outcome, *am* is the only regressor)

```r
fit_initial <- lm(mpg ~ factor(am), mtcars)
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## factor(am)1  7.244939   1.764422  4.106127 2.850207e-04
```

```
## [1] "Adjusted R Squared: 0.34"
```

```
## [1] "Confidence Interval:( 3.64 , 10.85 )"
```

When we ignore other effects, 1) mean MPG in automatic setting is 17.15 and mean MPG in manual setting is 17.15 + 7.24 = 24.39. 2) 95% of the times the manual transmission results on average 3.64 to 10.85 more miles per gallon. 3) 34% of the variation in mpg is explained by our model. We will try to do better.

Nested ANOVA

From ANOVA results below we see that the additions of *cyl*, *disp*, *hp*, and *wt* variables are significant so we will include them in our final model (we will not include *disp* as it is highly correlated with *cyl*). As the additions of *drat*, *qsec*, *vs*, *gear* and *carb* are not significant, than we will omit these variable.

```
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(am)
## Model 2: mpg ~ factor(am) + factor(cyl) + disp
## Model 3: mpg ~ factor(am) + factor(cyl) + disp + hp
## Model 4: mpg ~ factor(am) + factor(cyl) + disp + hp + drat
## Model 5: mpg ~ factor(am) + factor(cyl) + disp + hp + drat + wt
## Model 6: mpg ~ factor(am) + factor(cyl) + disp + hp + drat + wt + qsec
## Model 7: mpg ~ factor(am) + factor(cyl) + disp + hp + drat + wt + qsec +
##     factor(vs) + factor(gear) + factor(carb)
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     27 230.46  3    490.44 20.3665 1.512e-05 ***
## 3     26 183.04  1     47.42  5.9078   0.02809 *
## 4     25 182.38  1      0.66  0.0820   0.77855
## 5     24 150.10  1     32.28  4.0216   0.06331 .
## 6     23 141.21  1      8.89  1.1081   0.30916
## 7     15 120.40  8     20.80  0.3240   0.94399
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Regression with Four Regressors

```
fit <- lm(mpg ~ factor(am) + factor(cyl) + hp + wt, mtcars)
```

```
##                  Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  33.70832390 2.60488618 12.940421 7.733392e-13
## factor(am)1   1.80921138 1.39630450  1.295714 2.064597e-01
## factor(cyl)6 -3.03134449 1.40728351 -2.154040 4.068272e-02
## factor(cyl)8 -2.16367532 2.28425172 -0.947214 3.522509e-01
## hp           -0.03210943 0.01369257 -2.345025 2.693461e-02
## wt           -2.49682942 0.88558779 -2.819404 9.081408e-03
```

```
## [1] "Adjusted R Squared:  0.84"
```

This model is fitting better, it explains 84% of variation in mpg. But the transmission coeffitient (manual - automatic) is not significant anymore.

We could do farther regressions with interaction terms but this is already enough to see that the type of transmission is not so important when we account for number of cylinders, gross horsepower and weight.

Residuals (Appendix, Figure 3)

Residuals vs Fitted. There isn't any distinctive pattern, meaning that there isn't much relationship left out in the residuals. **Normal Q-Q**. The points are mostly close to the line (a good sign). **Scale-Location**. Similar to the first plot with standardized residuals (they appear randomly spread). **Residuals vs Leverage**. There are no influential cases as, there is no point outside of the Cook's distance.

Appendix

Figure 1. Manual Transmission Yields Higher Miles Per Gallon
(Ignoring Other Effects)

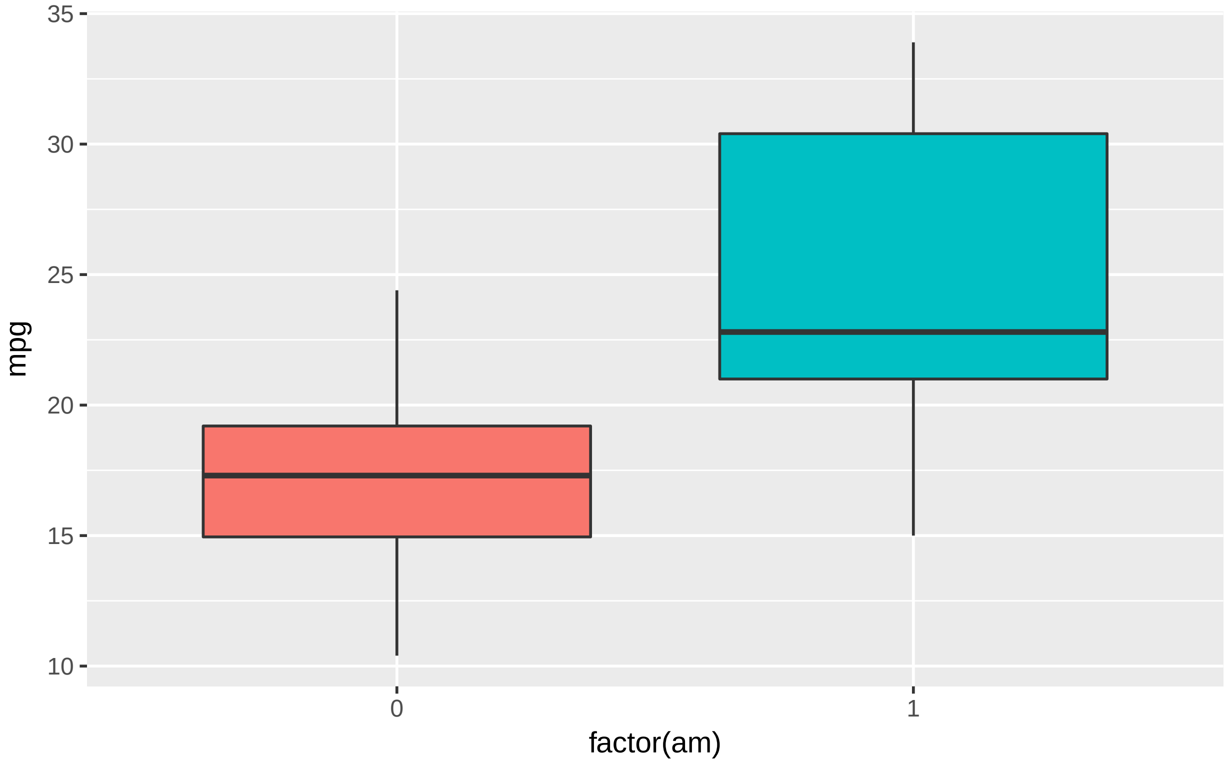


Figure 2. Pairwise Relaitionships

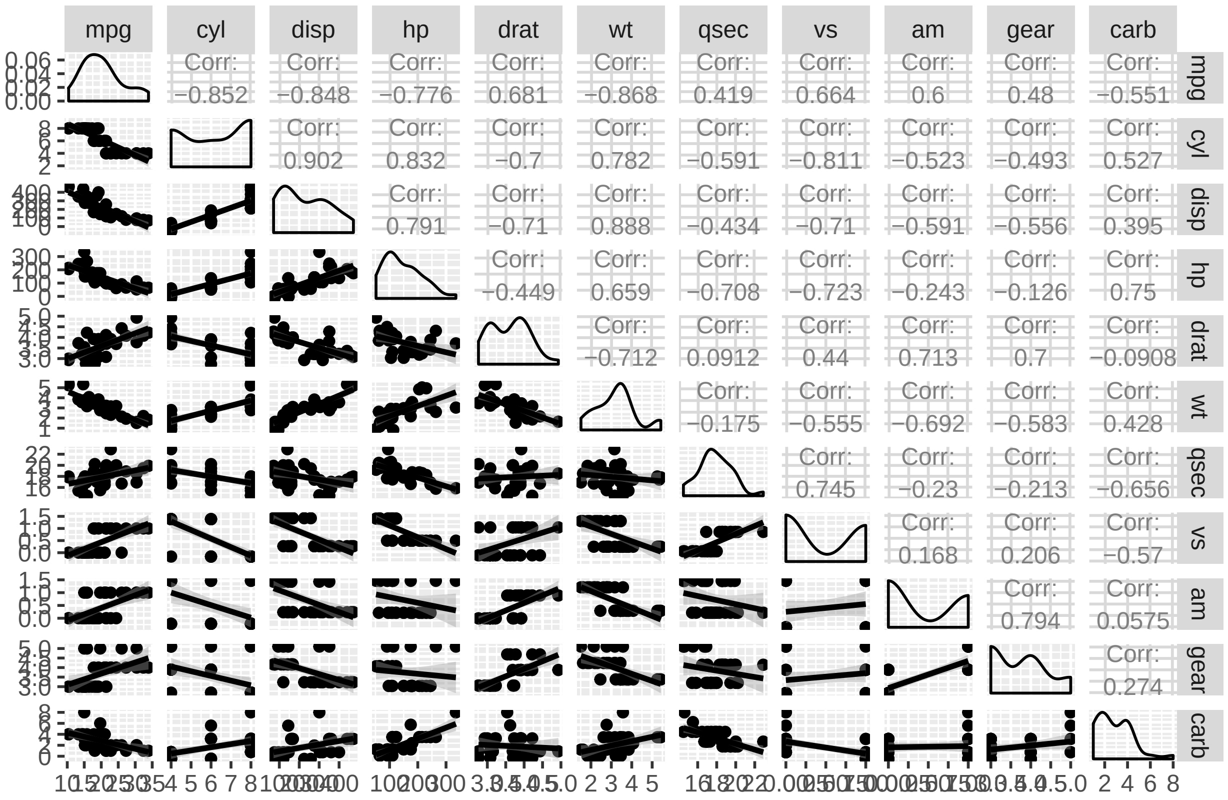
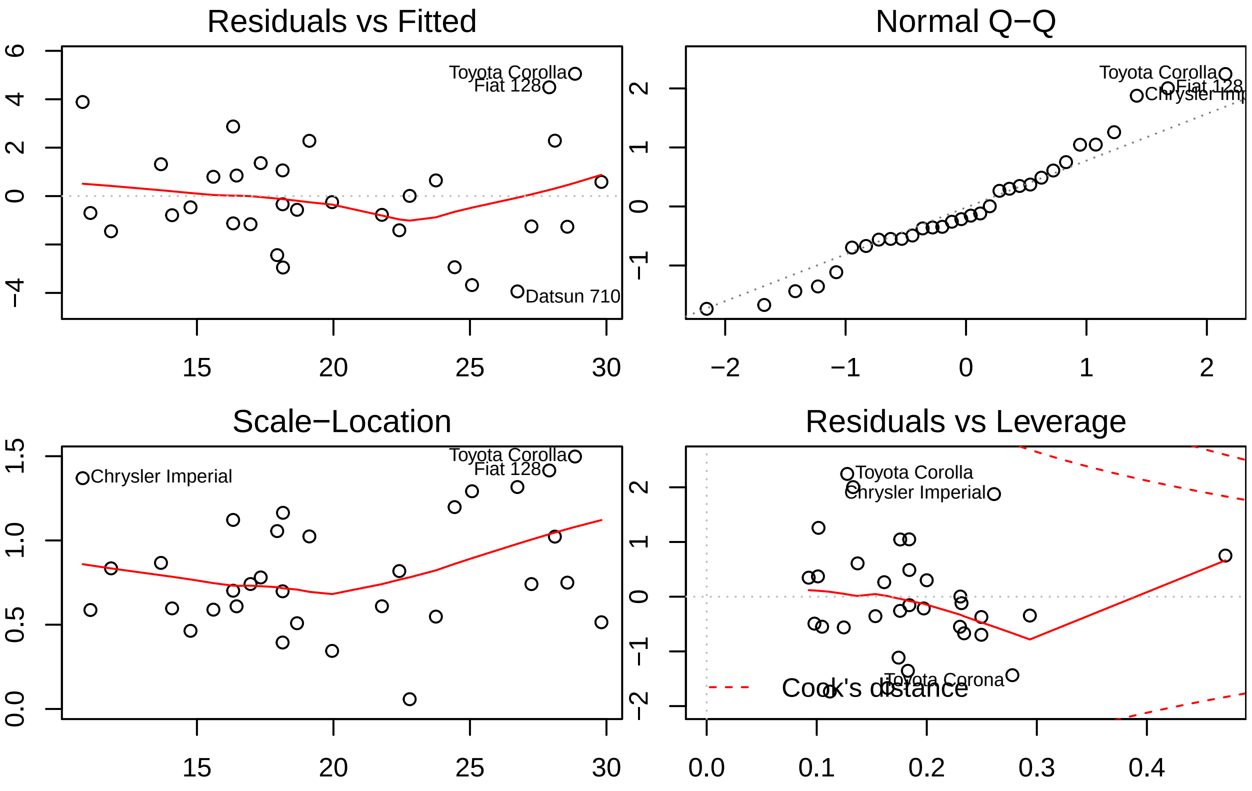


Figure 3. Residuals and Leverage Points


Residuals vs Fitted
Toyota Corolla
Fiat 128
Datsun 710
Normal Q−Q
Toyota Corolla
Fiat 128
Chrysler Imp
Scale−Location
Chrysler Imperial
Toyota Corolla
Fiat 128
Residuals vs Leverage
Toyota Corolla
Chrysler Imperial
Toyota Corona
Cook's distance