Regression Models Course Project - MTCARS Evaluation

Nikolaos Perdikis

7 8 2019

Executive Summary

MTCARS data set analysis: MPG and type of transmission Evaluate the relationship between consumption of the engine, and type of transmission Attempt to answer two (2) questions:
1. "Is an automatic or manual transmission better for MPG"
2. "Quantify the MPG difference between automatic and manual transmissions"

Prepare the R enviornment, by loading libraries and data:

```
## Loading required package: magrittr
```

Format of the data:

Format: A data frame with 32 observations on 11 (numeric) variables

Data Evaluation

As a first step, we perform what is called a "normality test". This is because, all our subsequent tests expect a normally distributed sample population. We expect to see a p-value larger than 0.05 to accept the NULL hypothesis (stated as **"The samples come from a Normal Distribution"**)

```
shapiro.test(mtcars$mpg)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mtcars$mpg
## W = 0.94756, p-value = 0.1229
```

Returing a p-value of 0.1228814 we can continue our investigation, taking for granted that **our sample is normal**. See Figure #1 in the Appendix for a display of the data

Analysis

The simplest test to perform, is to aggregate the consumption of all automatic and all manual cars, and compare their respective MPG means

```
aggregate(mpg~am, data = mtcars, mean)
```

```
##       am      mpg
## 1   Auto 17.14737
## 2 Manual 24.39231
```

The result of the function tells us that vehicles with manual transmission have a better MPG compared to vehicles equipped with automatic transmission, more than 7 MPG. Next, we will quantify the significance of this information using a T-test:

```
trans_auto <- mtcars[mtcars$am=="Auto",]
trans_manual <- mtcars[mtcars$am=="Manual",]
t.test(trans_auto$mpg,trans_manual$mpg)
```

```
##
##  Welch Two Sample t-test
##
## data:  trans_auto$mpg and trans_manual$mpg
## 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 of x mean of y
##  17.14737  24.39231
```

The absence of 0 from the 95% confidence interval tells us that there is a significant difference (not 0) in the mean MPG between automatic and manual transmission. What we have not researched so far, is how significant is the contribution of the transmission to our model, and if we need to evaluate other regressors as well.

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

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

From the output of our model, we focus on the value of **R-squared**, which is the "coverage" of our regressors towards the predicted value. In this case, we infer that the transmission type, accounts for ~36% of the consumption of the vehicle. This means that, by only accounting for the transmission, we leave outside 2/3 of the regressors that would be necessary for accurate prediction of the vehicle's mileage.

Multi-variate model

And this is exactly what this next model does:

```
fit <- betterFit <- lm(mpg~am + cyl + disp + wt, data = mtcars)

anova(init,fit)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl + disp + wt
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     26 182.87  4    538.03 19.124 1.927e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using anova to compare the two models, we have a p-value of 1.927e-07 when we add further regressors to the model, such as number of cylinders, displacement and weight of the vehicle. Meaning, a much better model fit would include all above parameters.

```
summary(fit)
```

```
##
## Call:
## lm(formula = mpg ~ am + cyl + disp + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5029 -1.2829 -0.4825  1.4954  5.7889
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.816067   2.914272  11.604 8.79e-12 ***
## amManual     0.141212   1.326751   0.106  0.91605
## cyl6        -4.304782   1.492355  -2.885  0.00777 **
## cyl8        -6.318406   2.647658  -2.386  0.02458 *
## disp         0.001632   0.013757   0.119  0.90647
## wt          -3.249176   1.249098  -2.601  0.01513 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.652 on 26 degrees of freedom
## Multiple R-squared:  0.8376, Adjusted R-squared:  0.8064
## F-statistic: 26.82 on 5 and 26 DF,  p-value: 1.73e-09
```

From the output above, we see we have achieved a coverage of more than 83%, using addditional regressors.

Conclusions

1. There is a significant difference in consumption, so Manual transmission is better for MPG than Automatic, keeping all other possible variables (weight, displacement, number of cylinders) constant
2. The difference in MPG is around 7.245 between automatic and manual cars, again, with the assumption that all other factors remain constant.
3. While there is a documented difference between automatic and manual cars, the fact that automatic cars are also cars of higher displacement/weight and number of cylinders, makes the isolated comparison of mileage and transmission type by far **not** the ideal model for any conclusions.

Appendix

The data:

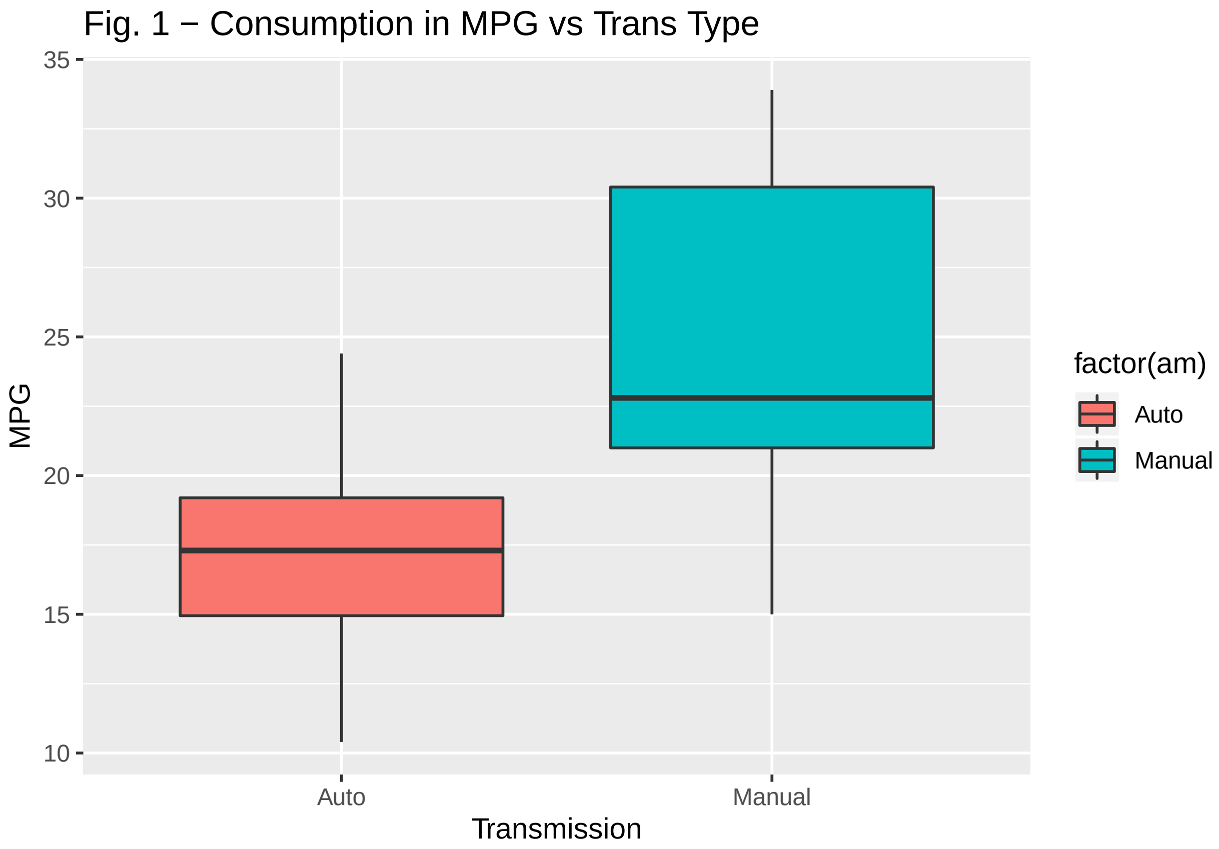
```
head(mtcars)
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs     am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0 Manual    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0 Manual    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1 Manual    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1   Auto    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0   Auto    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1   Auto    3    1
```

The graphic below is a visual representation of the MPG:

```
mtcars$am <- as.factor(mtcars$am)
levels(mtcars$am) <- c("Auto", "Manual")
plot_mpgam <- ggplot(data = mtcars, aes(x = factor(mtcars$am), y = mpg, fill = factor(am))) +
  geom_boxplot() +
  xlab("Transmission") +
  ylab("MPG") +
  ggtitle("Fig. 1 - Consumption in MPG vs Trans Type")

plot_mpgam
```



The following plot is to check the residuals, and therefore the fitment of the model:

```r
par(mfrow = c(2,2))
plot(fit)
title("Figure 2 - Model Fitment with additional regressors",outer=TRUE)
```

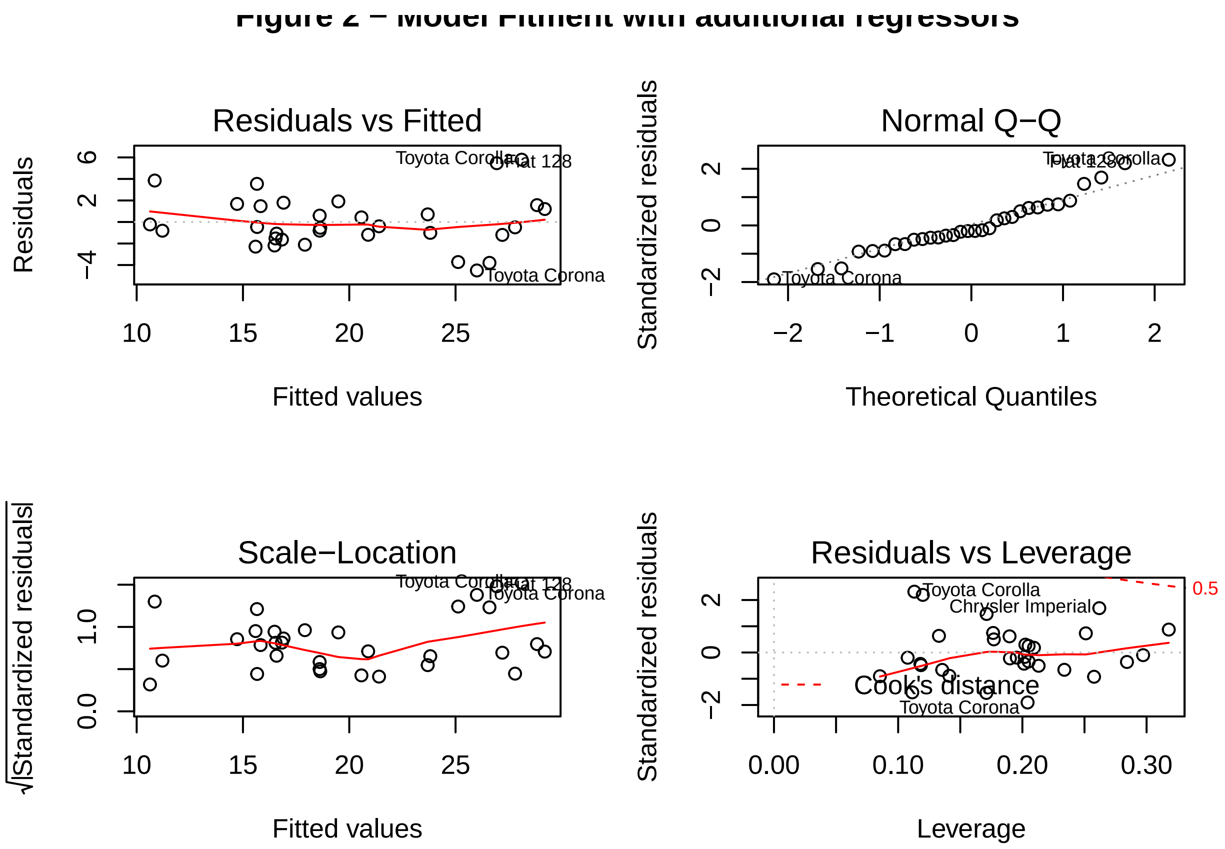


Figure 2 − Model Fitment with additional regressors
Residuals vs Fitted
Residuals
−4
2
6
10
15
20
25
Fitted values
Toyota Corolla
Fiat 128
Toyota Corona
Normal Q−Q
Standardized residuals
−2
0
2
−2
−1
0
1
2
Theoretical Quantiles
Toyota Corolla
Fiat 128
Toyota Corona
Scale−Location
√|Standardized residuals|
0.0
1.0
10
15
20
25
Fitted values
Toyota Corolla
Fiat 128
Toyota Corona
Residuals vs Leverage
Standardized residuals
−2
0
2
0.00
0.10
0.20
0.30
Leverage
Toyota Corolla
Chrysler Imperial
Toyota Corona
Cook's distance
0.5