Regression Models: Week 4 Project

Author: Alexander Johannes

Date: 6 September 2021

Executive Summary

This report explores the following two points using the `mtcars` dataset from base R:

a) "Is an automatic or manual transmission better for MPG?"; and
b) "Quantify the MPG difference between automatic and manual transmissions"

The analysis was conducted using several regression models and suggests that we can be confident that manual transmission is better for MPG based on this dataset.

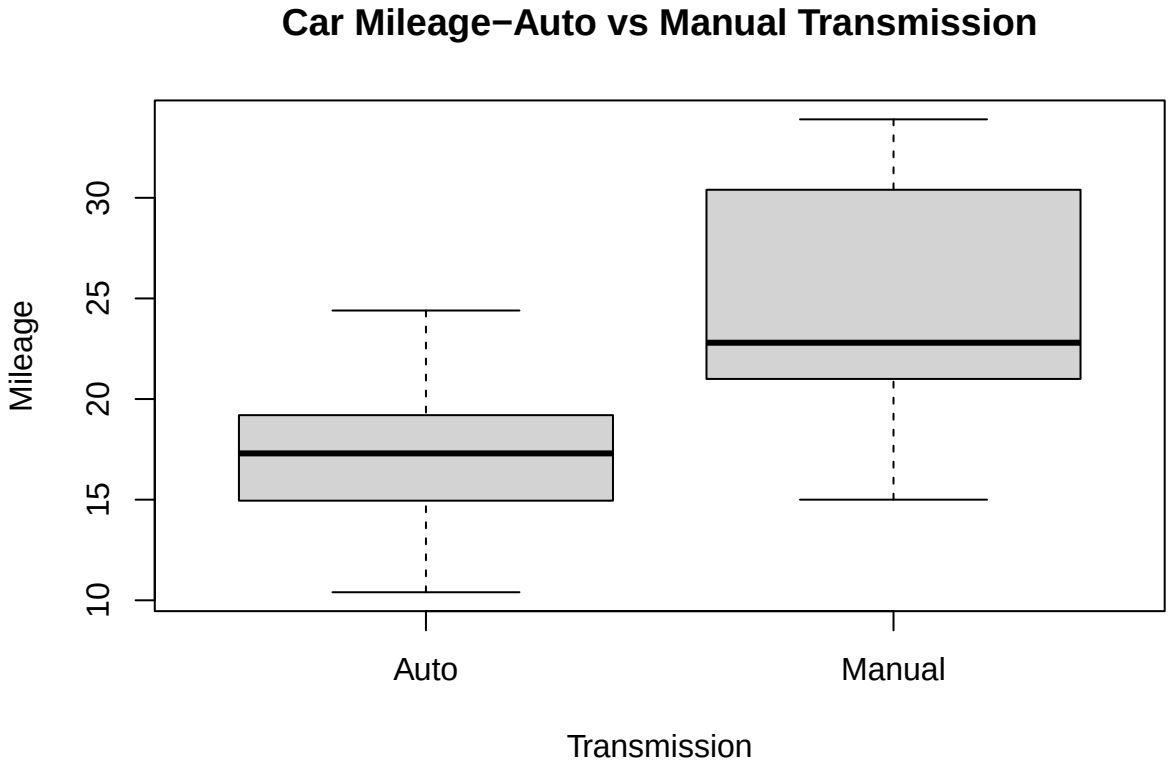
Part I: Automatic vs Manual Transmission

First, we load the dataset and conduct some basic exploratory analysis.

```
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 ...
```

Next, we explore the transmission factor variable `am`:



Consequently, our initial conclusion to part a) is that a car with manual transmission does have better fuel consumption. Now, we need to quantify the extent to which we believe this to be the case.

To do this, we fit a series of regression models and conduct a rudimentary process of model selection.

```
fit1 <- lm(mpg ~ factor(am)-1, data = mtcars)
summary(fit1)$coef
```

```
##             Estimate Std. Error  t value     Pr(>|t|)
## factor(am)0 17.14737   1.124603 15.24749 1.133983e-15
## factor(am)1 24.39231   1.359578 17.94109 1.376283e-17
```

This model indicates that mileage (mpg) increases with both transmission types, but that manual has the more pronounced benefit. The p values associated with both factors are below the .05 significance level, so we have a strong degree of belief that this is the case.

Next, we generate the correlation matrix of non-factor variables to get of a sense of how the variables interact, and which variables might be worth including in another model.

```
my_data <- mtcars[, c(1,3,4,5,6,7)]
corrMat <- cor(my_data)
round(corrMat, 2)
```

```
##        mpg  disp    hp  drat    wt  qsec
## mpg   1.00 -0.85 -0.78  0.68 -0.87  0.42
## disp -0.85  1.00  0.79 -0.71  0.89 -0.43
## hp   -0.78  0.79  1.00 -0.45  0.66 -0.71
## drat  0.68 -0.71 -0.45  1.00 -0.71  0.09
## wt   -0.87  0.89  0.66 -0.71  1.00 -0.17
## qsec  0.42 -0.43 -0.71  0.09 -0.17  1.00
```

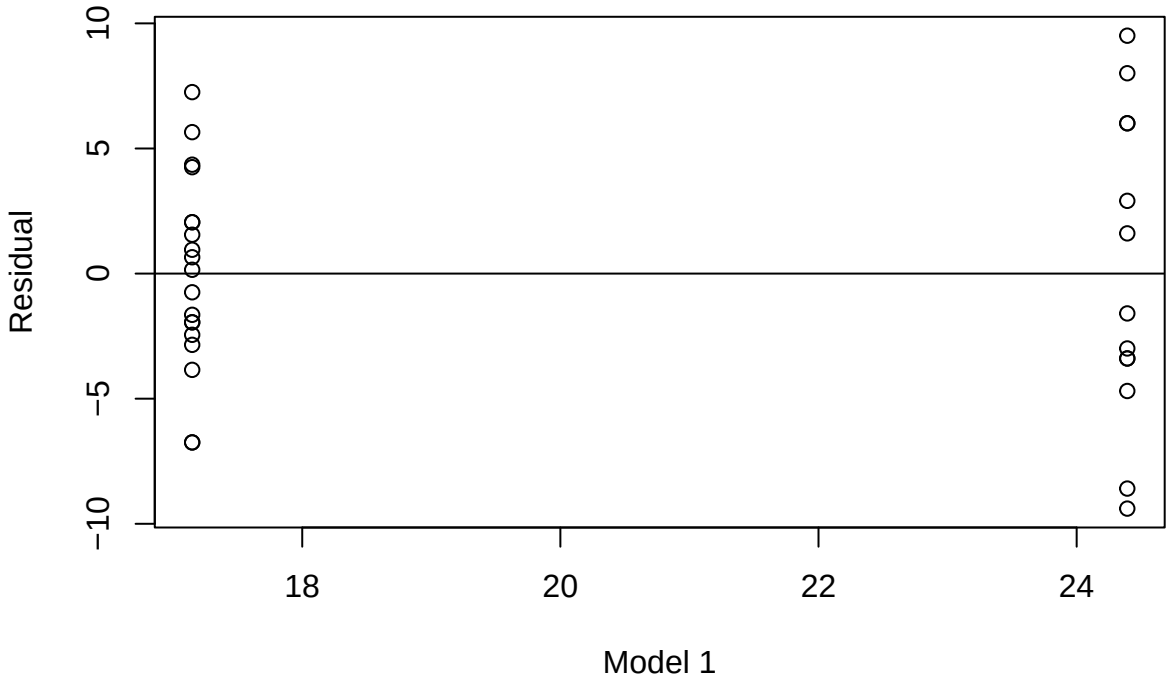
Analyzing this matrix, we conclude that a model using horsepower and weight could be more significant.

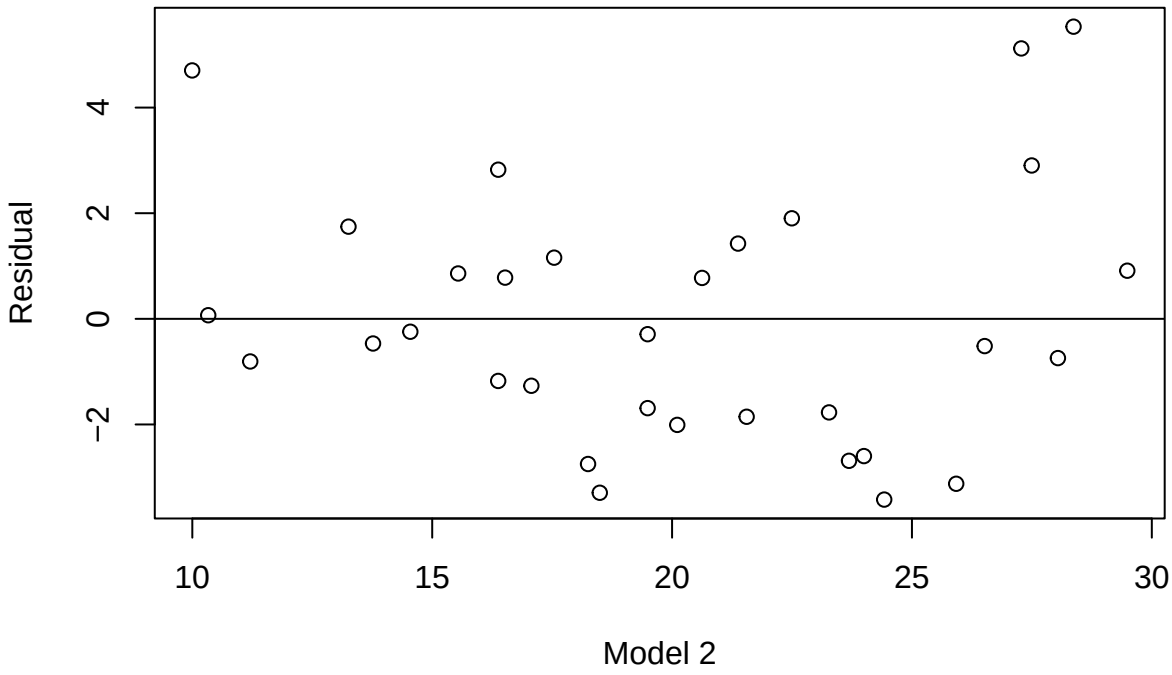
We generate this model and conduct a significance test.

```
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(am) - 1
## Model 2: mpg ~ hp + wt + factor(am) - 1
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     28 180.29  2    540.61 41.979 3.745e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see that this model is indeed more significant.

Finally, let us explore the residual plots of both the fitted models.





These plots both indicate that the residuals are approximately normal with mean 0 and constant variance, hence we are not concerned.

We therefore stick by our conclusions with confidence, given the test statistics we generated earlier in the analysis.