On Transmission Types and Mileage

Riccardo Finotello

24 June 2020

Summary

In this report we investigate the impact of gear transmission type on petrol mileage of several types of cars. The objective is to study what kind of transmission has the best impact on mileage (measured in miles per gallon) and quantify the difference between the two types.

Exploratory Data Analysis

We first load the *mtcars* dataset and look at its description and summary:

```
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 ...
```

As we can see, the dataset is mainly composed of numerical variables and factors which can be used for regression.

We will be interested in the `mpg` variable as dependent variable mainly as a function of the transmission type `am` which is a binary variable describing *automatic* ($am = 0$) or *manual* ($am = 1$) transmission type. In Figure 1 we show the distribution of the dependent variable (mpg) as a function of the two types of transmission we intend to probe.

In Figure 2 we show the correlation matrix of the variables together with their hierarchical structure. There are clear indications of interdependence between several variables.

Regression Analysis

We first try, as a base model, to predict the mileage using the most correlated variables (we do not fit the intercept as, in this context, it has no physical meaning):

```
fit.corr1 <- lm(mpg ~ cyl + disp + hp + wt + factor(cyl) -1, data = mtcars)
summary(fit.corr1)$coef
```

```
##                  Estimate Std. Error    t value     Pr(>|t|)
```

```
## cyl           4.031147191 0.51252890  7.8652095 2.429017e-08
## disp          0.004198868 0.01291717  0.3250609 7.477362e-01
## hp           -0.023516504 0.01221631 -1.9250086 6.523450e-02
## wt           -3.428625786 1.05545537 -3.2484801 3.193940e-03
## factor(cyl)4 19.877815737 1.42348475 13.9641930 1.358311e-13
## factor(cyl)6  8.349510384 1.58854211  5.2560837 1.711711e-05
```

As we can see most factors are relevant for the fit and we will try to keep them in further analysis. However the power output of the cars seems to irrelevant. We further investigate the impact of *hp* on the model by comparing the results when we remove it from the fit:

```
fit.corr2 <- lm(mpg ~ cyl + disp + wt + factor(cyl) -1, data = mtcars)
anova(fit.corr1, fit.corr2)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ cyl + disp + hp + wt + factor(cyl) - 1
## Model 2: mpg ~ cyl + disp + wt + factor(cyl) - 1
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     26 160.13
## 2     27 182.95 -1   -22.822 3.7057 0.06523 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This shows that the variable does indeed play a role in the model.

We finally investigate the impact of the transmission type on the model (we intentionally left it out from the previous fits in order to have a good background model to use for inference):

```
fit <- lm(mpg ~ cyl + disp + wt + hp + factor(cyl) + factor(am), data = mtcars)
summary(fit)$coef
```

```
##                  Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  36.582057350 3.76205614  9.7239531 5.624062e-10
## cyl          -0.679445322 0.72453735 -0.9377644 3.573375e-01
## disp          0.004087893 0.01276729  0.3201848 7.514890e-01
## wt           -2.738694608 1.17597755 -2.3288664 2.824553e-02
## hp           -0.032480178 0.01398322 -2.3227963 2.862128e-02
## factor(cyl)6 -1.777175912 1.22266200 -1.4535300 1.585156e-01
## factor(am)1   1.806099494 1.42107933  1.2709350 2.154510e-01
```

In this case we decide to fit the intercept which represents the base estimate for automatic ($am = 0$) transmission. With a p-value of 0.9 and 25 degrees of freedom we can also say that manual transmission ($am = 1$) has an average better impact on the petrol mileage of 1.8 miles per gallon with a confidence interval of

```
confint(fit)[8,]
```

```
##     2.5 %    97.5 %
## -1.120668  4.732867
```

In Figure 3 we finally show the diagnostic plots related to the fit we performed showing that the distribution of the residuals is indeed uncorrelated and no outlier issues are present.

Conclusion

From the analysis it seems that manual transmission cars have a better impact on petrol mileage which can be quantified in (1.8 ± 1.4) miles per gallon.

Appendix

In this appendix we show most plots and figures associated with the analysis.

The objective of the analysis is the prediction of petrol mileage as a function of the type of the transmission. We show the bounding boxes of such prediction as a reference:

```
boxplot(mpg ~ factor(am, labels = c("automatic", "manual")), data = mtcars,
        xlab = "Transmission type",
        ylab = "Miles per Gallon",
        main = "Mileage per Transmission Type"
       )
```

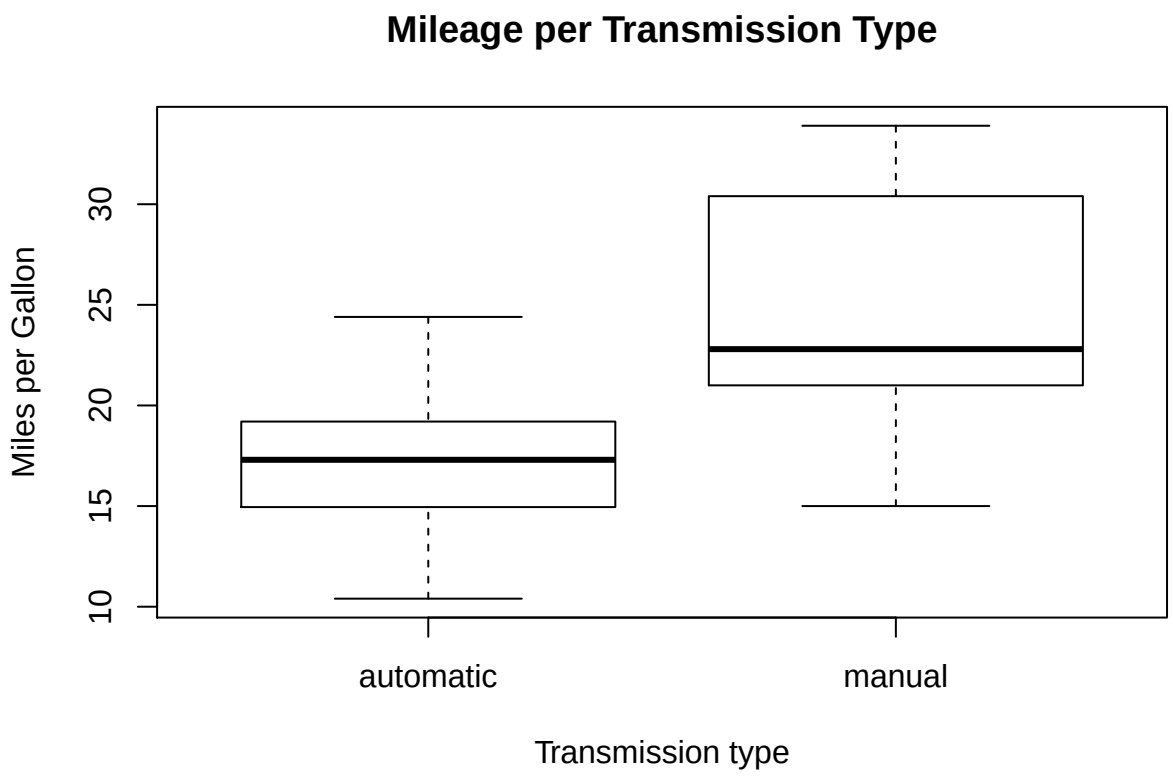

Figure 1: Boxplot of mileage vs type of transmission

We can then graphically show the correlation matrix of the dataset.

```
library(lattice)
```

```
## Warning: package 'lattice' was built under R version 3.6.3
```

```
levelplot(cor(mtcars))
```

We finally show the residual plots related to the analysis:

```
par(mfrow=c(2,2))
plot(fit)
```

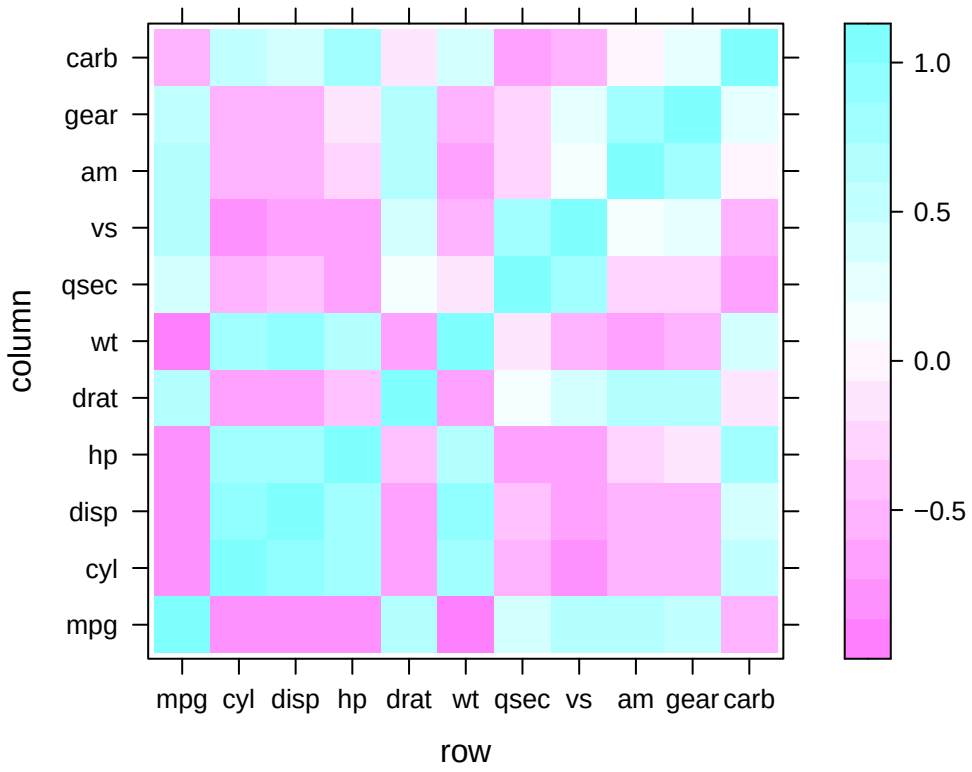

Figure 2: Correlation matrix.

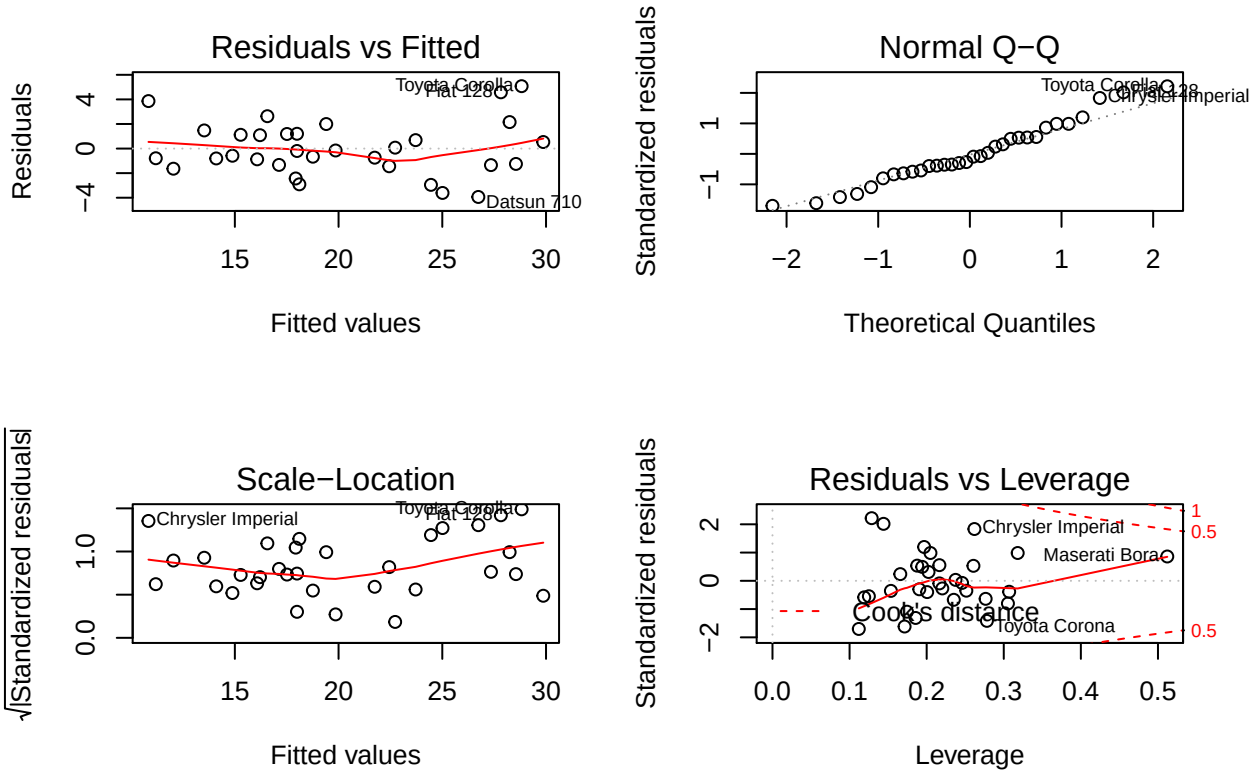

Figure 3: Diagnostic plots of the fit.