Effect of autumatic vs. manual transmission on MPG

Rok Bohinc

June 28, 2019

Summary

In this work I investigate the relationship between the mpg and automatic vs. manual transmission of the "mtcars" data set in R. The instructions for the task reads: You work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

1. Is an automatic or manual transmission better for MPG?
2. Quantify the MPG difference between automatic and manual transmissions!

I first do an exploratory data analysis, then I perform a t-test to answer question 1, last I perform a multivariate fit to answer question 2.

Exploratory analysis

In the mtcars data we are interested in two variables "mpg" - Miles/(US) gallon and "am" - Transmission (0 = automatic, 1 = manual). First of all lets check if there seems to be any difference between mpg the two transmission modes. In Figure 1 I compare histograms for the manual and automatic transmission mode. So we can see that the mpg is higher for manual transmission 24.4 $\pm$ 6.2 than for the automatic 17.1$\pm$ 3.8. This would indicate that manual transmission is better for the mpg than automatic.

Question 1

We can verify if manual transmission is significantly better for the mpg than the automatic transmission by performing a t-test for unpaired data and unequal variances (we saw in the previous section that the variances are clearly different for the two modes) and an error 1 type rate of 5%:

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

The associated p-value for the test above is 0.1% which is less than set type 1 error rate and we can therefore reject the null hypothesis that the means are equal. From this we can already answer the first question: **Manual transmission is better for the mpg**.

Qestion 2

In order to quantify the transmission mode effect on mpg we look at which variables can additionally affect mpg. To identify which variables to additionally include into the model I look at the absolute correlation matrix (see Figure 2). I am especially interested in correlation coefficients of "am" and "mpg"" variables with the other variables.

```
correlat <- as.data.frame(round(abs(cor(mtcars)),2))[c(1,9),]
correlat[order(correlat$mpg, decreasing = TRUE),]
```

```
##     mpg  cyl disp   hp drat   wt qsec   vs  am gear carb
## mpg 1.0 0.85 0.85 0.78 0.68 0.87 0.42 0.66 0.6 0.48 0.55
## am  0.6 0.52 0.59 0.24 0.71 0.69 0.23 0.17 1.0 0.79 0.06
```

In the fit I want to adjust for variables which are **not** especially correlated to "am" and are correlated to "mpg". We see that that the "hp" variable meets this criteria the best as it has a big influence on "mpg" but is not correlated to "am". Because of the strong correlation with "mpg", I probably do want to include variables "wt", "cyl" and "disp", although they all have a relative big correlation with "am", so the adjustment for any of these variables might result in an insignificant coefficient of "am". Variables "wt", "cyl" and "disp" are all correlated (correlation between 0,78 and 0.89), so it is probably fine if I include only one of them. Variable "wt" has the biggest correlation with "mpg", and therefore I choose this variable.

Fitting

Below I consider several models to fit "mpg" with "am" with adjustments.

```
fit0 <- lm(data=mtcars, mpg ~ .)
fit1 <- lm(data=mtcars, mpg ~ am)
fit2 <- lm(data=mtcars, mpg ~ am + hp)
fit3 <- lm(data=mtcars, mpg ~ am + hp + wt)
anova(fit1, fit2, fit3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + hp
## Model 3: mpg ~ am + hp + wt
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     29 245.44  1    475.46 73.841 2.445e-09 ***
## 3     28 180.29  1     65.15 10.118  0.003574 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the table above I see that in addition to "am", "hp" and "wt" have a significant effect on the fit. The inclusion of the additional fitting parameter does not result in a significant improvement of the model. I however cannot show the results because of the restrictions on the length of the report. In Figure 3 I show residual and Q-Q plots for model 3 and model 0 where we adjust for all parameters. Both residual plots exhibit a mild non-ideal V-type trend, which means that there is likely another unknown variable responsible for the observed trend. The Q-Q plot for model 3 perhaps just passes the thick pencil approach.

Interpretation

```
round(summary(fit3)$coef,3)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   34.003      2.643  12.867    0.000
## am             2.084      1.376   1.514    0.141
## hp            -0.037      0.010  -3.902    0.001
## wt            -2.879      0.905  -3.181    0.004
```

From the table above we see that mean mpg for automatic transmission is estimated to be 34 holding other variables constant while the increase in the mean of the mpg going from automatic to manual transmission is estimated to be 2.1 holding other variables constant. So even with the adjustments it seems that the manual transmission is better for mpg. The inclusion of the am parameter in the fitting model is however not significant as the p-value is 0.14. The model however strongly supports the inclusion of "hp" and "wt".

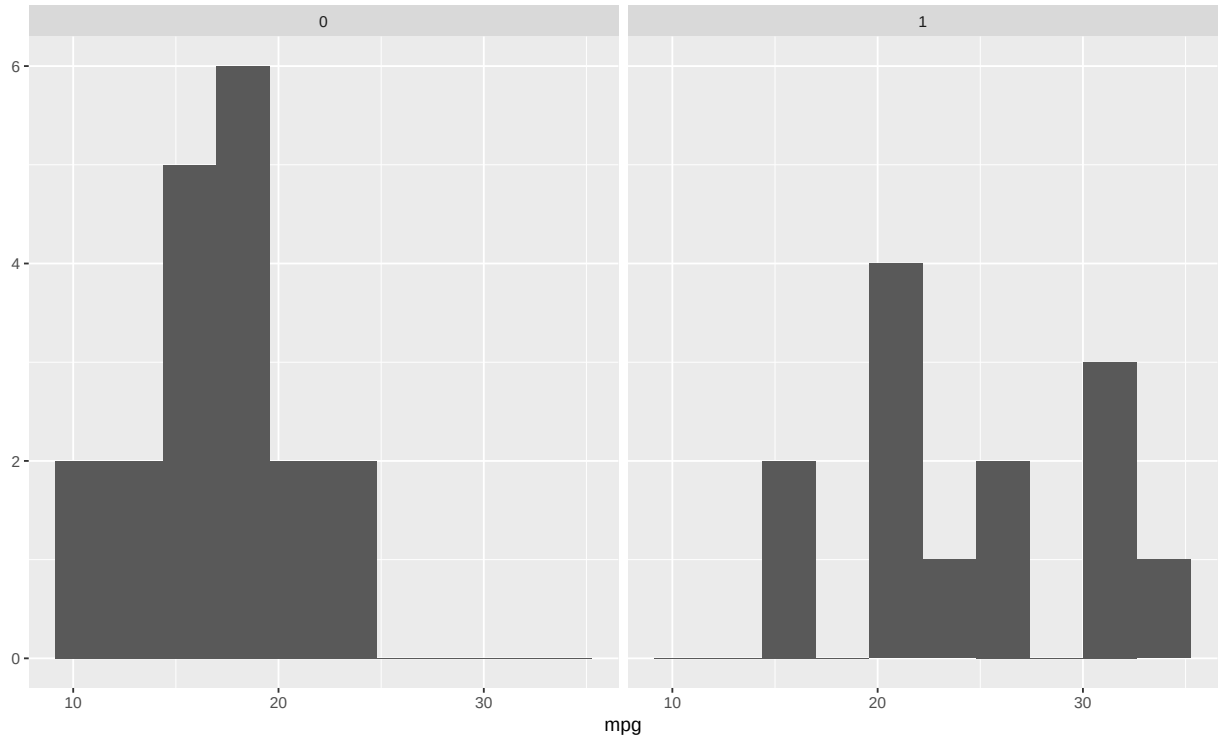

Figure 1: Histograms of mpg for automatic transmission (left) and manual transmission (right)

Appendix

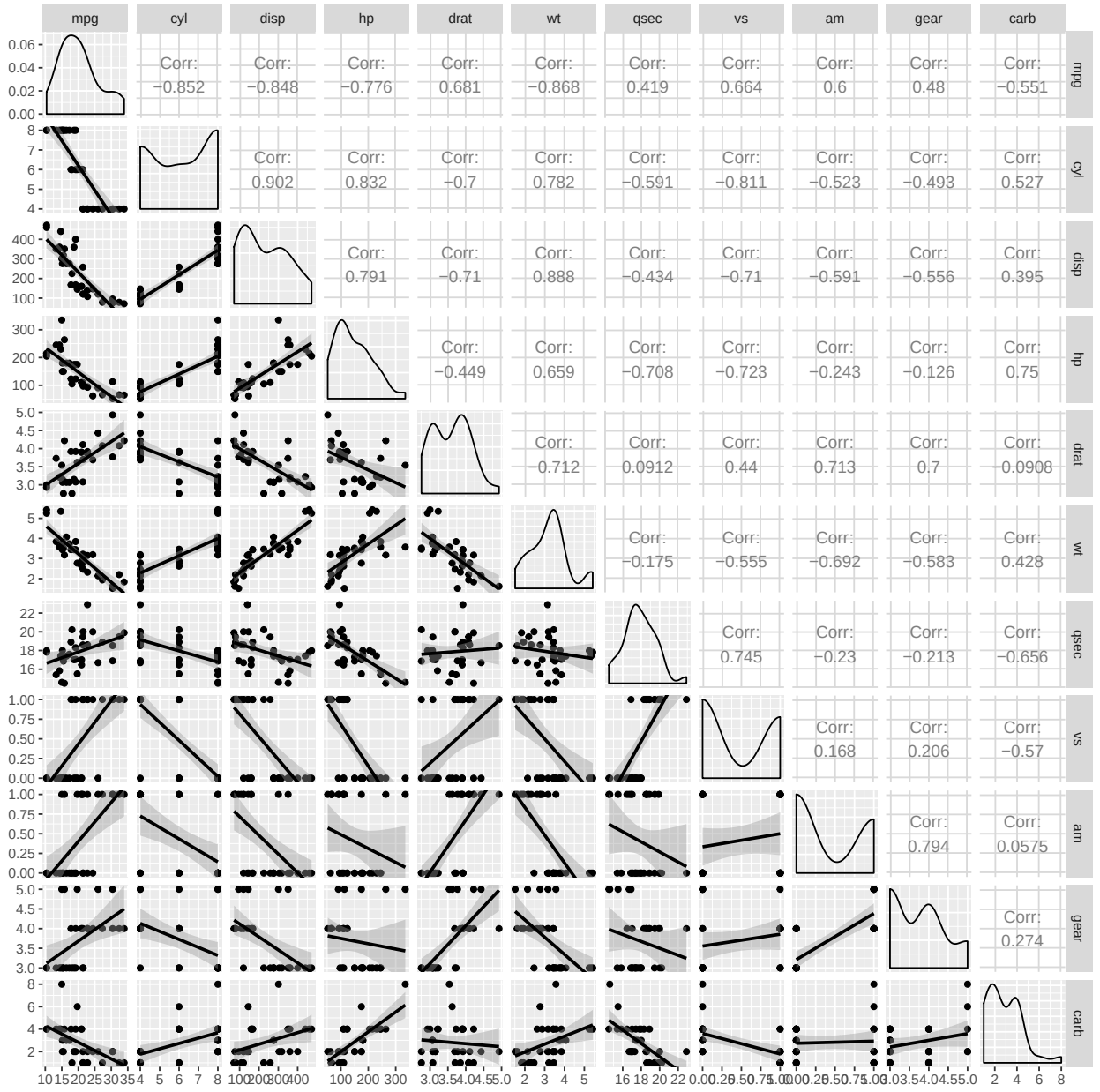

Figure 2: Pairs plot of the selected variables and the corresponsing correlation coefficients

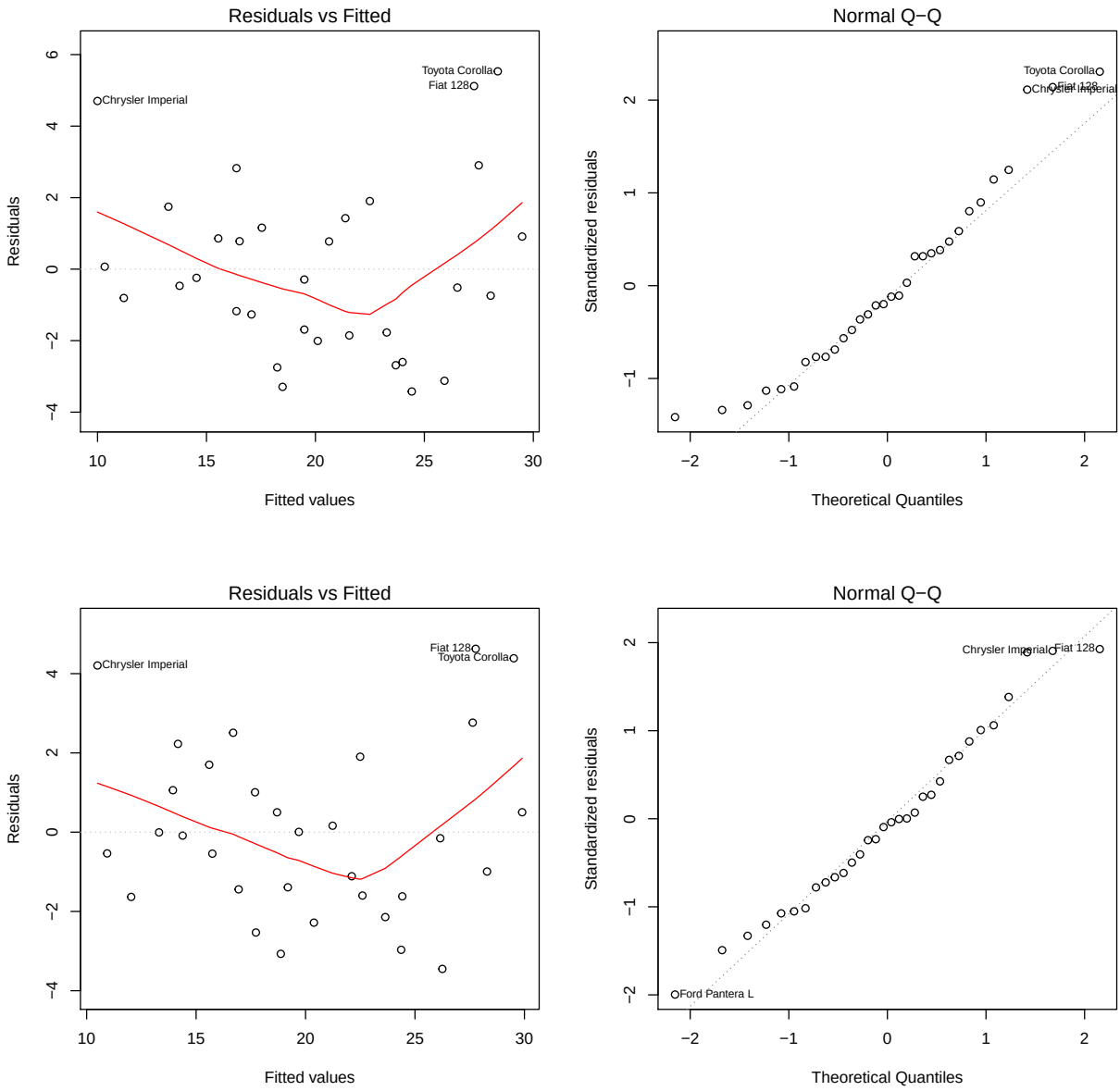

Figure 3: Residual plot (left) and Q-Q plot (rigt) for model 3 (top) and model 0 (bottom).