Car Modelling

Martin J Page

28/05/2020

Note: this report is not divided into a main body and an appendix. Rather the code and figures are intercalated with the write-up for easier reading.

Executive Summary

In this analysis we try to answer two key questions:
1. Is an automatic or manual transmission better for MPG, and
2. Quantify the MPG difference between automatic and manual transmissions.

We find that while looking at the relationship between mile per gallon (`mpg`) and transmission (`am`) in isolation suggests that there might be difference in mileage efficiency between automatic and manual cars, when we include other variables it actually seems that the relationship between mileage efficiency and the specifications of the car is not driven primarily by transmission type.

Exploring the Data

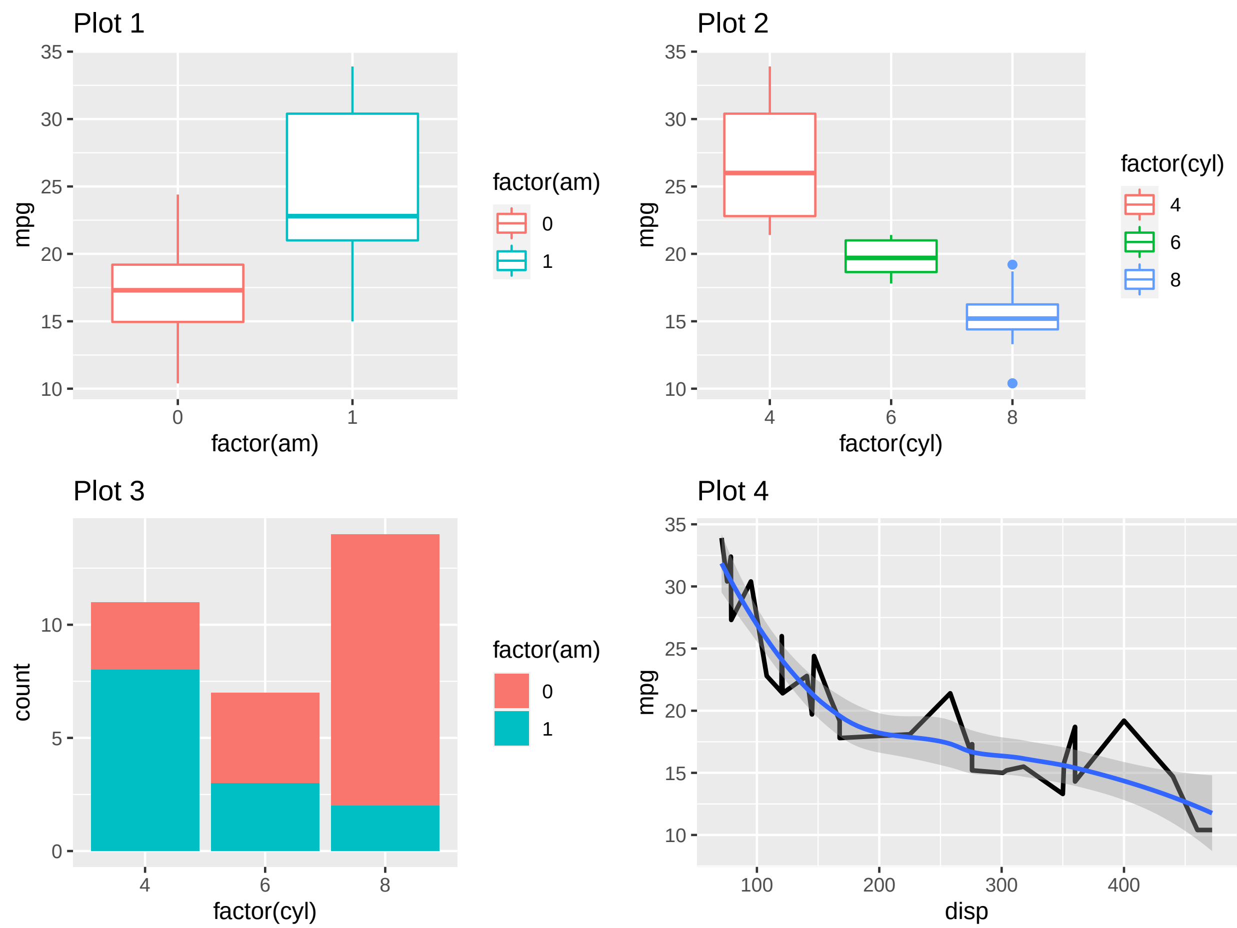
Load the data `mtcars` and have a look at the structure. `help(mtcars)` describes the variables.

```
data(mtcars)
str(mtcars)
```

We are interested in the relationship between transmission (`am`) and miles per gallon (`mpg`). We also look at the relationships between some other variables and `mpg`, which might in fact influence the relationship between `am` and `mpg`.

```
p1 <- ggplot(mtcars, aes(x = factor(am), y = mpg, col = factor(am))) + geom_boxplot() +
  ggtitle("Plot 1")
p2 <- ggplot(mtcars, aes(x = factor(cyl), y = mpg, col = factor(cyl))) + geom_boxplot() +
  ggtitle("Plot 2")
p3 <- ggplot(mtcars, aes(x = factor(cyl), fill = factor(am))) + geom_bar() +
  ggtitle("Plot 3")
p4 <- ggplot(mtcars, aes(x = disp, y = mpg)) + geom_line(lwd=1) + geom_smooth() +
  ggtitle("Plot 4")

grid.arrange(p1, p2, p3, p4, ncol = 2)
```



From Plot 1, it seems that there might be some relationship between `am` (0 = automatic and 1 = manual) and `mpg`. However, other variables such as the number of cylinders (`cyl`) and the displacement (`disp`) also seem to be related to miles per gallon (`mpg`). Moreover, these variables might also be assicated with our `am` variable, as seen in Plot 3 where cars with automatic transmission might tend to have more cylinders.

We now look at which variables are correlated to `mpg` and which variables are correlated to `am` before we start building our model.

```
correlations_mpg <- sapply(mtcars[,-1], function(col) cor(mtcars$mpg, col))
mpg_vars <- which(abs(correlations_mpg) >= 0.8)
mpg_vars
```

```
 cyl disp   wt
   1    2    5
```

These variables have a strong correlation with `mpg` and are of interest to include in our model.

```
correlations_am <- sapply(mtcars[,-9], function(col) cor(mtcars$am, col))
correlations_am
```

```
        mpg         cyl        disp          hp        drat          wt
 0.59983243 -0.52260705 -0.59122704 -0.24320426  0.71271113 -0.69249526
       qsec          vs        gear        carb
-0.22986086  0.16834512  0.79405876  0.05753435
```

None of these variables have correlations that meet the $\pm$ 0.8 cutoff.

Fitting Models

We consider the relationship between `am` and `mpg`, while also taking into account the potential influence of cyl, disp, wt, which are highly correlated with the two main variables in our model. Omitting these variables might result in bias in the coefficients of the regressors which are correlated with these variables. We will construct our models in a nested manner and check using ANOVA that adding each regressors results in a significant addition to the model. Adding irrelevant regressors can cause the model to tend towards a perfect fit by increasing the standard errors of the other regressors. We check that the residuals are normally distributed in line with the assumptions of the ANOVA test.

```
mtcars$am <- factor(mtcars$am)
mtcars$cyl <- factor(mtcars$cyl)
lm1 <- lm(mpg ~ am - 1, data = mtcars)
lm2 <- lm(mpg ~ am + cyl - 1, data = mtcars)
resd1 <- shapiro.test(lm1$residuals)
resd2 <- shapiro.test(lm2$residuals)
test1 <- anova(lm1, lm2)
round(test1$`Pr(>F)`,4)
```

```
[1] NA  0
```

Adding `cyl` is a significant addition to the model.

```
lm3 <- update(lm2, mpg ~ am + cyl + disp - 1)
resd3 <- shapiro.test(lm3$residuals)
test2 <- anova(lm1, lm2, lm3)
round(test2$`Pr(>F)`,4)
```

```
[1]    NA 0.000 0.056
```

Adding `disp` is an irrelevant addition to the model. We do not include it.

```
lm4 <- update(lm2, mpg ~ am + cyl + wt - 1)
resd4 <- shapiro.test(lm4$residuals)
test3 <- anova(lm1, lm2, lm4)
round(test3$`Pr(>F)`,4)
```

```
[1]     NA 0.0000 0.0018
```

Adding `wt` is a signifiant addition to the model.

```
lm5 <- update(lm4, mpg ~ am + cyl + wt + am*cyl - 1)
resd5 <- shapiro.test(lm5$residuals)
test4 <- anova(lm1, lm2, lm4, lm5)
round(test4$`Pr(>F)`, 4)
```

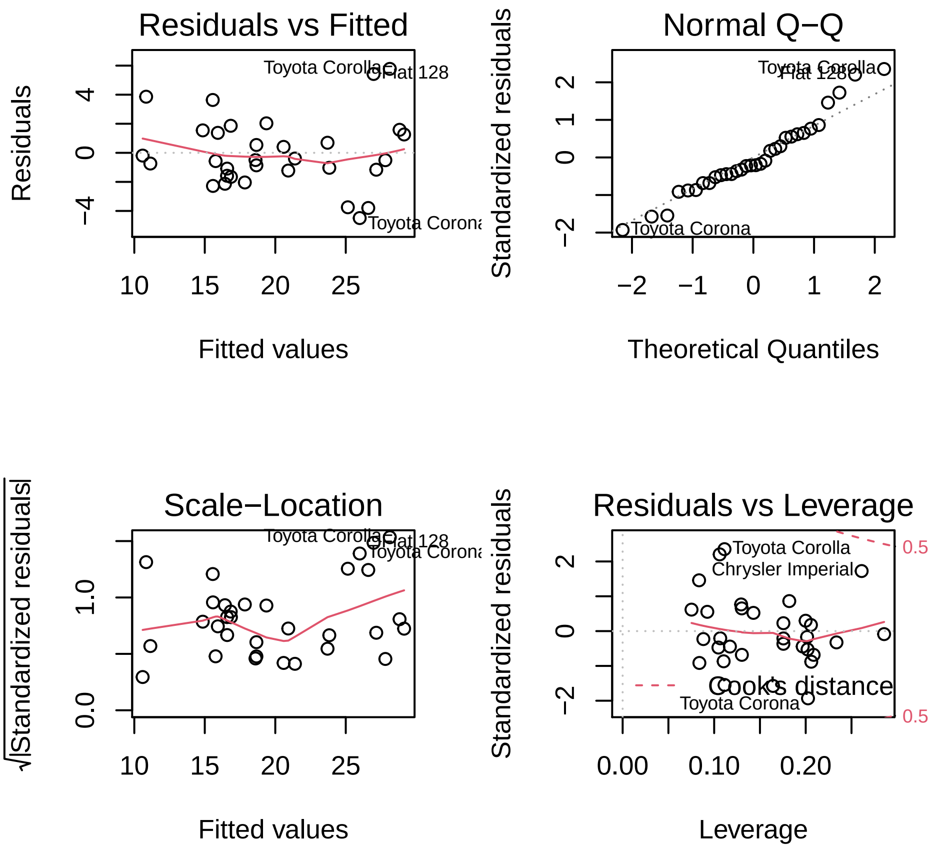
```
[1]     NA 0.0000 0.0016 0.2486
```

Adding an interaction term for `am` and `cyl` is an irrelevant addition to the model. We do not include it.

Therefore the final model `lm4` includes `am`, `cyl` and `wt` to predict `mpg`, so that we can hone in on the relatioship between `am` and `mpg` while keeping the other regressors fixed.

Diagnostics

```r
par(mfrow = c(2,2))
plot(lm4)
```




```r
lever <- hatvalues(lm4)
betas <- dfbetas(lm4)
```

Diagnostics show that our model is not being unduly influenced by outliers. The line of fitted values vs residuals runs close to the y = 0, as does the line of leverage vs standardised residuals. The QQ plot also runs close to the theortical identity line.

Conclusion: Coefficients and Uncertainty

Final model

```r
coef(lm4)
```

```
      am0       am1      cyl6      cyl8        wt 
33.753592 33.903695 -4.257319 -6.079119 -3.149598 
```

We can interpret the coefficients to say that a 4-cylinder car with automatic transmission has an average `mpg` of 33.75 versus a 4-cylinder car with automatic transmission that has an average `mpg` of 33.9. With a 6-cylinder car, these respective group averages decrease by 4.26 miles per gallon and with an 8-cylinder car they decrease by 6.08. For each increase in unit `wt` (i.e. 1000 lbs), the `mpg` decreases by 3.15.

```
cbind(round(summary(lm4)$coef,2), round(confint(lm4),2))
```

```
     Estimate Std. Error t value Pr(>|t|) 2.5 % 97.5 %
am0     33.75       2.81   12.00     0.00 27.98  39.53
am1     33.90       2.06   16.42     0.00 29.67  38.14
cyl6    -4.26       1.41   -3.02     0.01 -7.15  -1.36
cyl8    -6.08       1.68   -3.61     0.00 -9.53  -2.62
wt      -3.15       0.91   -3.47     0.00 -5.01  -1.29
```

The confidence intervals for `am0`, automatic tranmission of a 4-cylinder car, and for `am1`, manual transmission of a 4-cylinder car, clearly overlap. (Notes: the p-value represets the likelihood of observing a relationship between the predictor and response (`mpg`) due to chance. In this model, `cyl = 4` is the reference level).

```
lm6 <- lm(mpg ~ cyl + am + wt -1 , data = mtcars)
round(summary(lm6)$coef, 2)
```

```
     Estimate Std. Error t value Pr(>|t|)
cyl4    33.75       2.81   12.00     0.00
cyl6    29.50       3.31    8.90     0.00
cyl8    27.67       3.80    7.29     0.00
am1      0.15       1.30    0.12     0.91
wt      -3.15       0.91   -3.47     0.00
```

If we fit the model a bit differently where `am = 0` is set as the reference level instead of `cyl = 4`, we can see that the p-value for `am1` tells us that the mean is not different from the reference `am0`.

Comapre to the initial model

```
summary(lm1)$coef
```

```
    Estimate Std. Error  t value     Pr(>|t|)
am0 17.14737   1.124603 15.24749 1.133983e-15
am1 24.39231   1.359578 17.94109 1.376283e-17
```

Adding the additional variables in the model is very important. The initial model `lm1` that used only transmission (`am`) as the regressor suggested that transmission values might be different from each other. However, after including other regressors in the final model `lm4`, it is clear that other variables such as `cyl` might be stronger drivers of the relationship between different cars specifications and miles per gallon performance than is `am`.