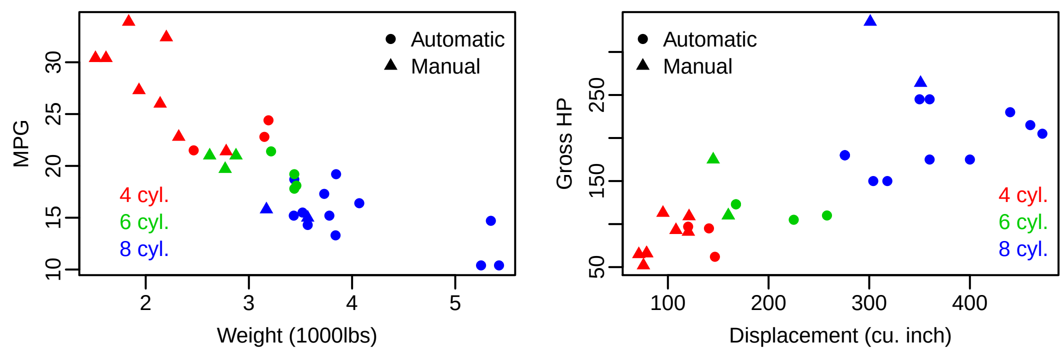
The effect of the transmission on fuel consumption

Executive summary

This report uses data extracted from the 1974 Motor Trend US magazine that comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models). This data is used to estimate the effect of the type of transmission (automatic vs. manual) on the milage (MPG). The results suggest that manual transmission allows for a higher MPG. However this result only becomes statistically significant after a model adjustment that attributes more explanatory power to the transmission.

Introduction and exploratory data analysis

The starting point for this analysis is a simple theory based approach. Based on intuition it seems reasonalbe to assume that the mpg of a car is mainly determined by its weight and its power. Furthermore it is likely that cars with more cylinders have a lager displacement and therefore also more hp. The following short exploratory data analysis verifies these assumptions.



The left part of above figure plots the MPG against the weight using colors to show the number of cylinders an symbols to represent the transmission. The plot provides a couple of interesting insights: (1.1) There is a clear negative relationship between MPG and weight, (1.2) there is a clear positive relationship between the number of cylinders and the weight of a car and (1.3) automatic transmission seems to be used almost exclusively in heavy cars with 8 cylinder engines whereas manual transmissions are mainly used in lighter cars with 4 cylinder engines. The right part shows that (2.1) there is a strong linear positive relationship between HP and displacement (2.2) engines with more cylinders clearly also provide more HP and (2.3) apart from two excpetions all high-powered cars use an automatic transmission.

The initial model

The exploratory data analysis prooves that a model using the weight and some variable measuring the power of a car is a reasonable starting point. However (2.1) and (2.2) suggest that a model containing all of the three power-relatded variables (`cyl, disp, hp`) will suffer from multicollinearity. Therefore the initial model will only use cyl as a factor-varible. In addition (1.3) suggests that including weight in the model will suppress any potential impact of the transmission on `mpg`. The initial model for this analysis is therefore `mpg ~ am + cyl + wt` (whereas `am` and `cyl` a treated as factor variables). The estimation results for this regression can be found in the appendix and show that the coefficients on both `wt` and `cyl` are statistically significant on a 1% significance level both coefficients feature expected signs, i.e. higher `wt` implies lower `mpg` and more cylinders imply lower `mpg` as well. Furthermore we cannot reject the hypothesis that the coefficient on `am` is 0.

Strategy for model enhancement and selection

From the exploratory data analysis we have seen that automatic transmissions are mainly used in heavy, high-powered cars. We cannot say anything about the causal direction of the correlation but theory suggests that an automatic transmission is heavier than a manual transmission which will result in a lower `mpg`. However, in the initial model this lower `mpg` is attributed to the variable `wt` and not to `am`. Therefore my approach is to adjust `wt` to attribute any weight-related effects of the transmission to `am`. This is done by an auxialliary regression where I regress `wt` on `am` and some other explanatory variables from the `mtcars` dataset. I used likelihood ratio tests to determine the correct model for this auxilliary regression. The estimation of the auxilliary regression shows that, when controlling for other factors, cars with automatic transmissions tend to be 700lbs heavier than cars with manual transmissions. Details on the results including residual plots and diagnostics can be found in the appendix. The residual plot and the diagnostics suggest that the model is valid.

To retrieve the adjusted model I subtract these 700lbs from all cars with a automatic transmission and store the new weight in a variable `mwt`: `mtcars$mwt <- mtcars$wt + (1-mtcars$am)*-0.70885`.

The adjusted base model is `mpg~am+cyl+mwt`. Starting from this base model I gradually add the variables `hp`, `drat` and `vs` select the final model based on a likelihood ratio test. The likelihood ratio test suggests that `hp` should be added to the model. The detailed results can be found in the appendix.

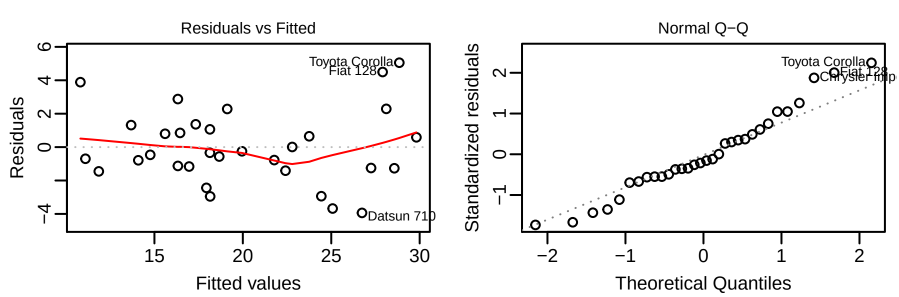
Estimating the final model and interpreting the results

The final model used to estimate the impact of the transmission type on the mileage of the car is therefore `mpg~am+cyl+mwt+hp`. The estimastion results are given below.

```
Call:
lm(formula = mpg ~ factor(am) + factor(cyl) + mwt + hp, data = mtcars)

Residuals:
   Min     1Q Median     3Q    Max 
-3.939 -1.256 -0.401  1.125  5.051 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   31.9384     2.0794   15.36   1.5e-14 ***
factor(am)1    3.5791     1.1392    3.14    0.0042 ** 
factor(cyl)6  -3.0313     1.4073   -2.15    0.0407 *  
factor(cyl)8  -2.1637     2.2843   -0.95    0.3523    
mwt           -2.4968     0.8856   -2.82    0.0091 ** 
hp            -0.0321     0.0137   -2.35    0.0269 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.41 on 26 degrees of freedom
Multiple R-squared:  0.866, Adjusted R-squared:  0.84 
F-statistic: 33.6 on 5 and 26 DF,  p-value: 1.51e-10
```

The estimation results from my final model report a positive coefficient on `am`, i.e. a manual transmission is better for MPG. Recall that `am` is 0 for cars with an automatic transmission and 1 for cars with a manual transmission. The p-value for `am` is `0.004`, therefore the coefficient is considered statistically significant on a 1% significance level. Additionally the estimation results support the hypothesis that a higher weight, more hp and also more cylinders yield a lower MPG.

The value of the coefficient on `am` is `3.58`. Therefore, holding the other factors constant, the MPG of a car with a manual transmission is 3.58 higher than the MPG of a car with an automatic transmission. To quantify the uncertainty in this conclusion I compute the 95% confidence interval. The 95% confidence interval for `am` is [1.2373, 5.9208]. Therefore we can say that we are 95% confident that the true effect of `am` is in this range. Furthermore the R^2 shows that the model explains 86,6% of the variation in MPG.



The residual plot shows no systematic trend in the residuals. Fuetheremore the Q-Q plot does not indicate a non-normal distribution of the residuals. The appendix shows additional diagnostics for the estimation. Looking at the Cook's distance and the residual vs leverage plot in the appendix one can tell none of the single observations has a critical influence on the model. All in all the diagnostics give indication that the chosen model is valid.

Appendix

Correlation matrix

```
       mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
mpg   1.00 -0.85 -0.85 -0.78  0.68 -0.87  0.42  0.66  0.60  0.48 -0.55
cyl  -0.85  1.00  0.90  0.83 -0.70  0.78 -0.59 -0.81 -0.52 -0.49  0.53
disp -0.85  0.90  1.00  0.79 -0.71  0.89 -0.43 -0.71 -0.59 -0.56  0.39
hp   -0.78  0.83  0.79  1.00 -0.45  0.66 -0.71 -0.72 -0.24 -0.13  0.75
drat  0.68 -0.70 -0.71 -0.45  1.00 -0.71  0.09  0.44  0.71  0.70 -0.09
wt   -0.87  0.78  0.89  0.66 -0.71  1.00 -0.17 -0.55 -0.69 -0.58  0.43
qsec  0.42 -0.59 -0.43 -0.71  0.09 -0.17  1.00  0.74 -0.23 -0.21 -0.66
vs    0.66 -0.81 -0.71 -0.72  0.44 -0.55  0.74  1.00  0.17  0.21 -0.57
am    0.60 -0.52 -0.59 -0.24  0.71 -0.69 -0.23  0.17  1.00  0.79  0.06
gear  0.48 -0.49 -0.56 -0.13  0.70 -0.58 -0.21  0.21  0.79  1.00  0.27
carb -0.55  0.53  0.39  0.75 -0.09  0.43 -0.66 -0.57  0.06  0.27  1.00
```

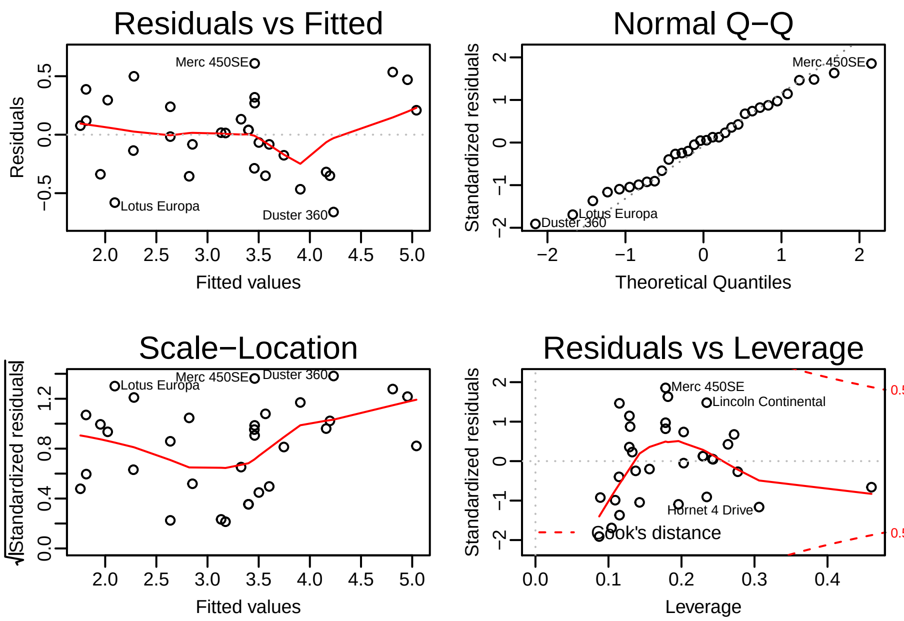
Estimation of the initial model

```
Call:
lm(formula = mpg ~ factor(am) + factor(cyl) + wt, data = mtcars)

Residuals:
   Min     1Q Median     3Q    Max
-4.490 -1.312 -0.504  1.416  5.776

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    33.754      2.813   12.00  2.5e-12 ***
factor(am)1     0.150      1.300    0.12   0.9089
factor(cyl)6   -4.257      1.411   -3.02   0.0055 **
factor(cyl)8   -6.079      1.684   -3.61   0.0012 **
wt             -3.150      0.908   -3.47   0.0018 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.6 on 27 degrees of freedom
Multiple R-squared:  0.838, Adjusted R-squared:  0.813
F-statistic: 34.8 on 4 and 27 DF,  p-value: 2.73e-10
```

Auxilliary regression to estimate impact of transmission on weight

```
Analysis of Variance Table

Model 1: wt ~ factor(cyl) + disp + factor(am)
Model 2: wt ~ factor(cyl) + disp + factor(am) + carb
Model 3: wt ~ factor(cyl) + disp + factor(am) + carb + hp
  Res.Df  RSS Df Sum of Sq    F Pr(>F)
1     27 4.51
2     26 3.41  1     1.093 8.96 0.0061 **
3     25 3.05  1     0.363 2.98 0.0968 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Call:
lm(formula = wt ~ factor(cyl) + disp + factor(am) + carb, data = mtcars)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-0.6616 -0.2956  0.0158  0.2463  0.6095

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.79024    0.24502    7.31  9.3e-08 ***
factor(cyl)6 -0.25131    0.23949   -1.05  0.30367
factor(cyl)8 -0.81067    0.38429   -2.11  0.04468 *
disp          0.00723    0.00139    5.19  2.1e-05 ***
factor(am)1  -0.70885    0.18478   -3.84  0.00072 ***
carb          0.16243    0.05630    2.89  0.00776 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.362 on 26 degrees of freedom
Multiple R-squared:  0.885, Adjusted R-squared:  0.863
F-statistic:   40 on 5 and 26 DF,  p-value: 2.11e-11
```



Steps to the final model

```
Analysis of Variance Table

Model 1: mpg ~ factor(am) + factor(cyl) + mwt
Model 2: mpg ~ factor(am) + factor(cyl) + mwt + hp
Model 3: mpg ~ factor(am) + factor(cyl) + mwt + hp + drat
Model 4: mpg ~ factor(am) + factor(cyl) + mwt + hp + drat + factor(vs)
  Res.Df RSS Df Sum of Sq    F Pr(>F)
1     27 183
2     26 151  1      31.9 5.35   0.03 *
3     25 151  1       0.2 0.04   0.85
4     24 143  1       7.5 1.25   0.27
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Diagnostics for the final model