Transmission and fuel consumption analysis

Executive summary

The main goal of this analysis is using the `mtcars` dataset to answer two questions:

- is an automatic or manual transmission better for `mpg`?
- quantify the `mpg` difference between automatic and manual transmissions;

`mpg` being *miles per gallon*, an indicator of fuel consumption. The results are as follows: the main effect of transmission, `am`, on consumption, is relevant: the `mpg` mean for automatic transmission car is 17.1474 while the `mpg` mean for manual transmission car is 24.3923, *0* being automatic transmission, and *1* the manual one. It seems that cars with manual transmission has less fuel consumption on average, but analyzing data and hearing the domain experts opinion, I have detected a possible confounding variable, weight - `wt`. As you will see in the rest of the analysis, automatic tranmission cars tend to weigh more, and weight is able to explain all the relationship between `mpg` and `am`. At the end of the analysis I'll try to idenfity a parsimonious model to predict `mpg`.

Exploratory data analysis

The `mtcars` dataset is composed by 32 observations and 11 variables, for details please look at `?mtcars` on the R console. There are no missing values. At the appendix you find the plot *Scatterplot matrix*, extremely useful to orient the modeling. Infact you can see that the graph representing `wt` and `am` shows how the two groups are almost not overlapped. This, in addition to domain experts opinions, prompted me to adjust the relationship between `mpg` and `am` for `wt` at first.

Adjustement

Let's remind the main effect of `am` on `mpg`:the `mpg` mean for automatic transmission car is 17.1474 while the `mpg` mean for manual transmission car is 24.3923. This difference in means is also statistically significant. Infact if we perform a one-way ANOVA with `aov` function we see that *F- statistic p.value* is 0.0002, and *etaSquared* is 0.3598. But if we adjust the relationship for `wt`, things change a lot. You can see this looking at the Appendix at plot *mpg vs am adjusted for wt*

If you look at this plot, the horizontal lines represent the main effect and the blue line, manual transmission, has a highest `mpg` mean. But adjusting for weight reduce incredibly the gap in the means and reverse the sign of the difference. It seems that transmission is highly associated with weight: if a car weighs less than 3 lbs, then in this sample you probably are going to have a manual transmission; so the relationship between `mpg`and `am`is very well explained by `wt` and if we account for it then knowing the type of transmission doesn't affect in terms of knowing `mpg`. But there are some problems: the relationship between `wt` and `mpg` doesn't seem linear: so I made a second plot with a log-transformation of the response, *log(mpg) vs am adjusted for wt*. Thing are better and results don't change too much. Another issue is that no points overlap for a particular level of *X*: so we heavily rely on the model when we say that, having accounted for `wt` (for a certain level of *X*), the difference between manual or automatic transmission don't affect mpg, because we can't see in this sample different transmissions for the same weight.

Let's see statistically what we showed in plots. First I will show the coefficients for the model, and then I will perform an ANOVA to demonstrate that accounting for `wt` makes `am` not necessary. Here are the coefficients for `lm(log(mpg) ~ wt + factor(am), mtcars)`

```
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  3.89834    0.13567 28.7345 7.374e-23
## wt          -0.28699    0.03501 -8.1978 4.870e-09
## factor(am)1 -0.04307    0.06865 -0.6274 5.353e-01
```

According to this model (remember the log transformation of the response), for a unit increase of `wt`, the geometric mean of `mpg` is multiplied by $e^{\beta X_{wt}}$, 0.7505 (holding `am` constant). So it decreases, as expected. Furthermore, holding `wt` constant, moving from automatic to manual transmission change the intercept of a multiplying factor of 0.9578. We should be more accurate having some confidence intervals:

```
##                2.5 %   97.5 %
## (Intercept)  3.6209  4.17582
## wt          -0.3586 -0.21539
## factor(am)1 -0.1835  0.09733
```

Powering e to lower and upper bound of `am` coefficient we see that intercept could be multiplied by 0.8324 or 1.1022. So we don't even know if changing transmission, holding `wt` constant, decrease o increase `mpg`. Is `am` necessary for this model?

```
anova(lm(log(mpg) ~ wt, mtcars), lm(log(mpg) ~ wt + factor(am), mtcars))
```

Adding `am` in this case isn't necessary, because p.value of the F statistics is 0.5353

A linear model

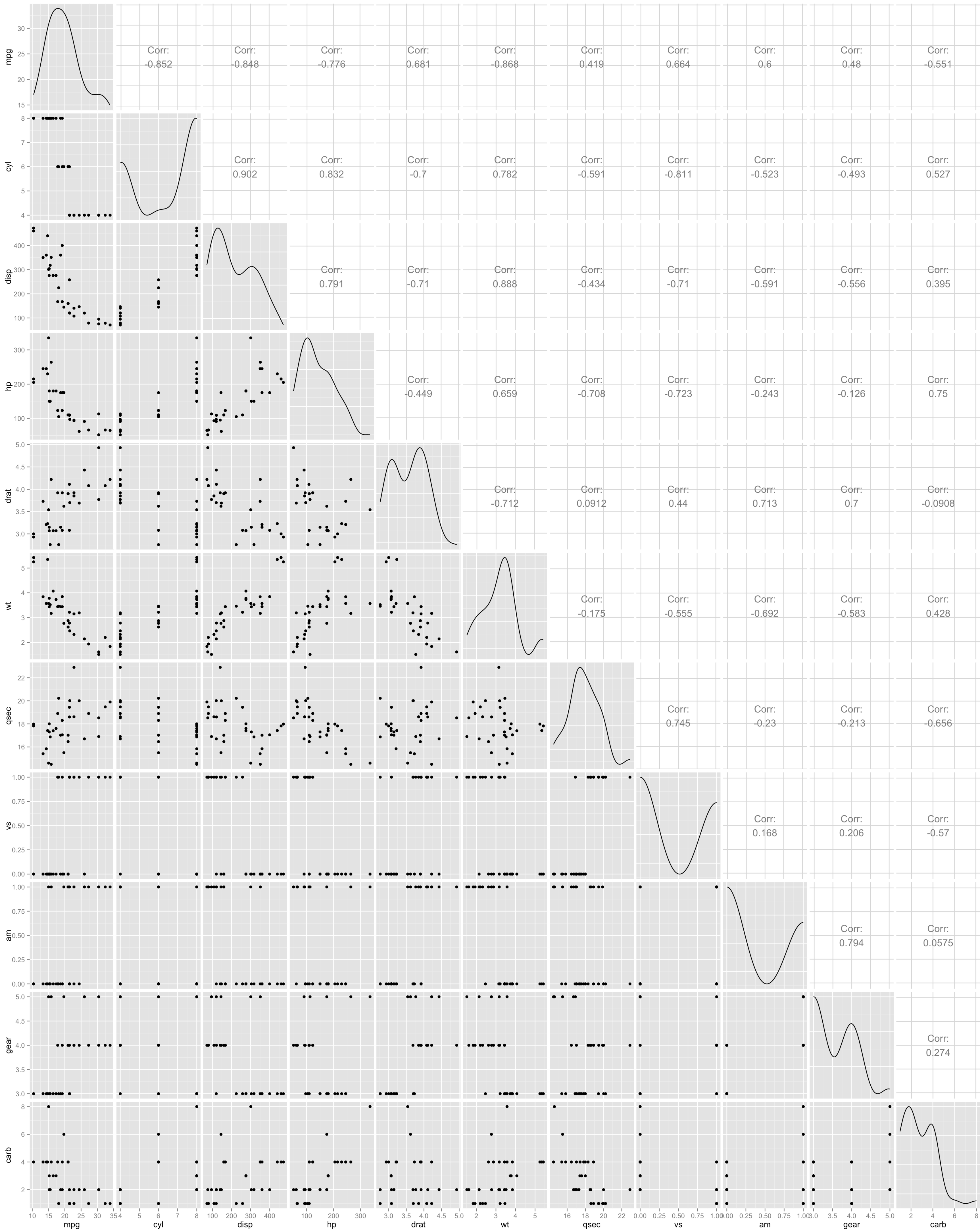
I would like to conclude this analysis fitting a linear model that, in according to parsimony principles, explain as much as possible `mpg` variance. I started with `wt` predictor, because R squared is high, 0.7976; after that I tried to include a third variable, looking for significant predictors. I have started with `hp`, which is significant (in the model I divided it by 100 for interpretability of coefficients), and after that no fourth variable seemed statistically necessary to the model. Here are the coefficients of `lm(log(mpg) ~ wt + I(hp/100), mtcars)`:

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)   3.8291    0.06868  55.752 4.716e-31
## wt           -0.2005    0.02718  -7.378 3.962e-08
## I(hp/100)    -0.1543    0.03879  -3.979 4.234e-04
```

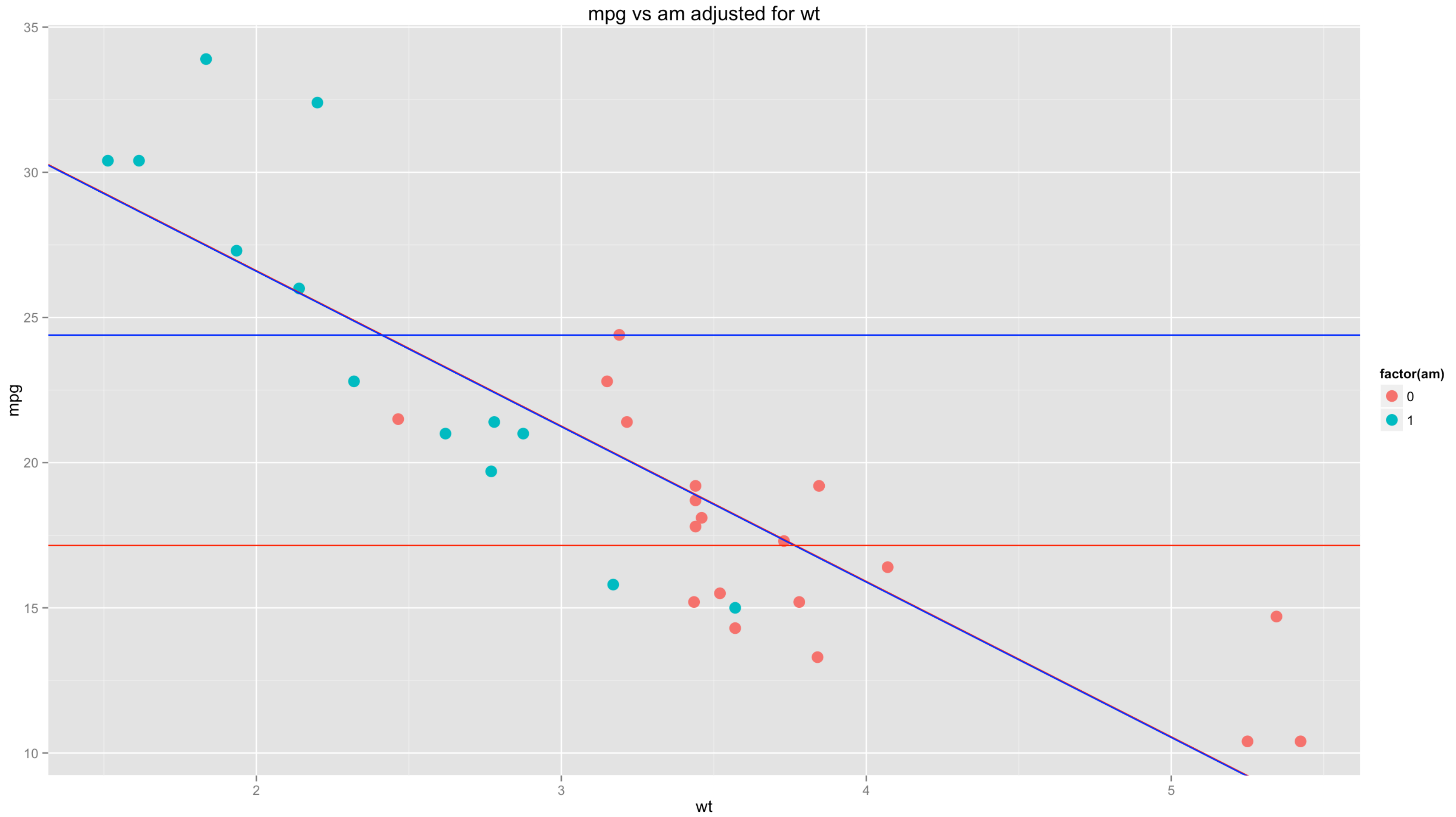
R squared adjusted is 0.8691, and the variance inflation factors are low, 1.7666for `wt` and 1.7666for `hp`. In the appendix you can find the *diagnostic plot for mpg ~ wt + hp*, which shows that homoschedasticity and normality of residuals are not so well respected, but probably acceptable for our purposes (p.value of a *shapiro test for normality of residuals* is 0.2701). I decided to not remove outlier, because are due to specific properties of some units of the sample.

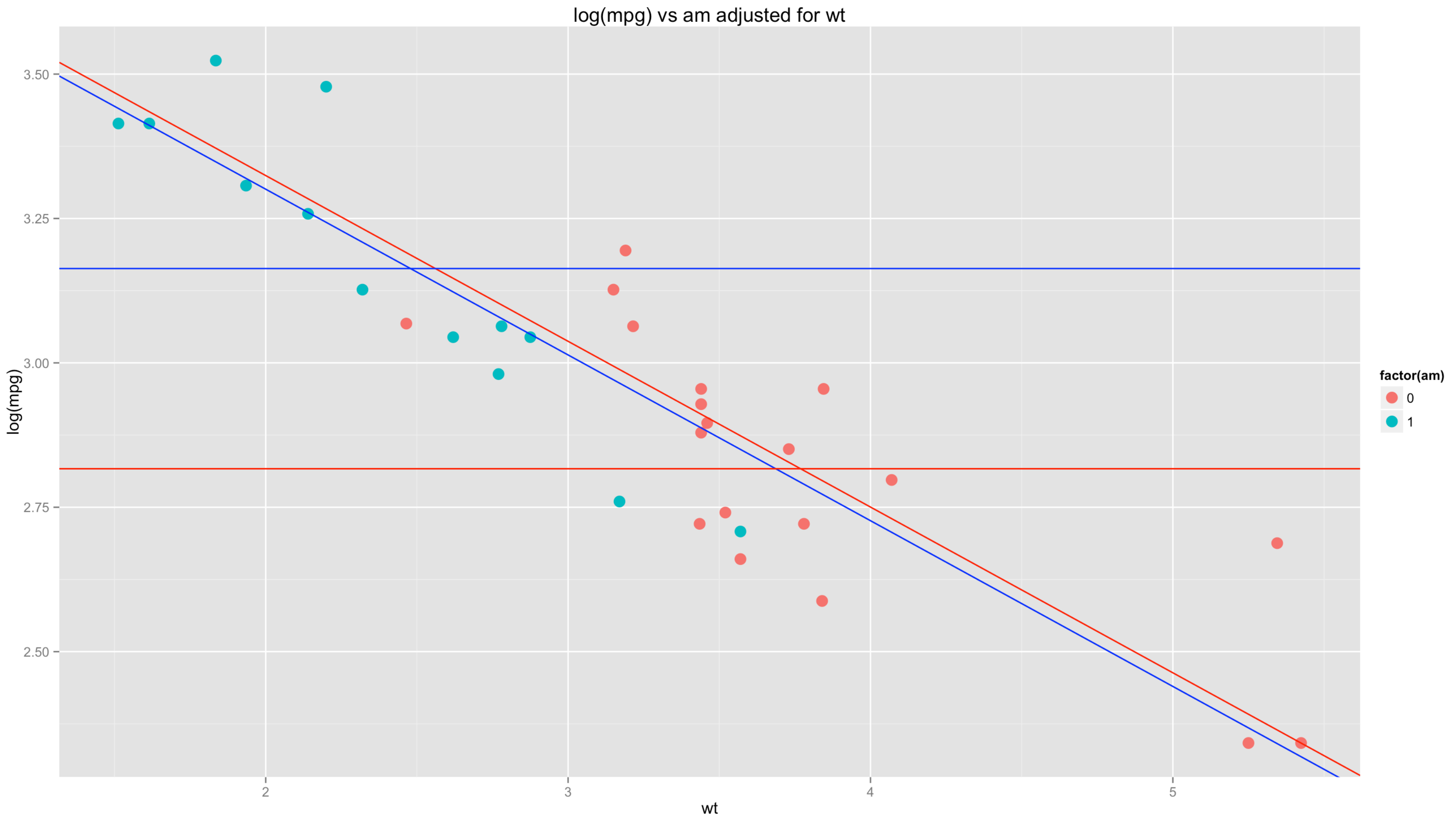
Appendix

Scatterplot matrix

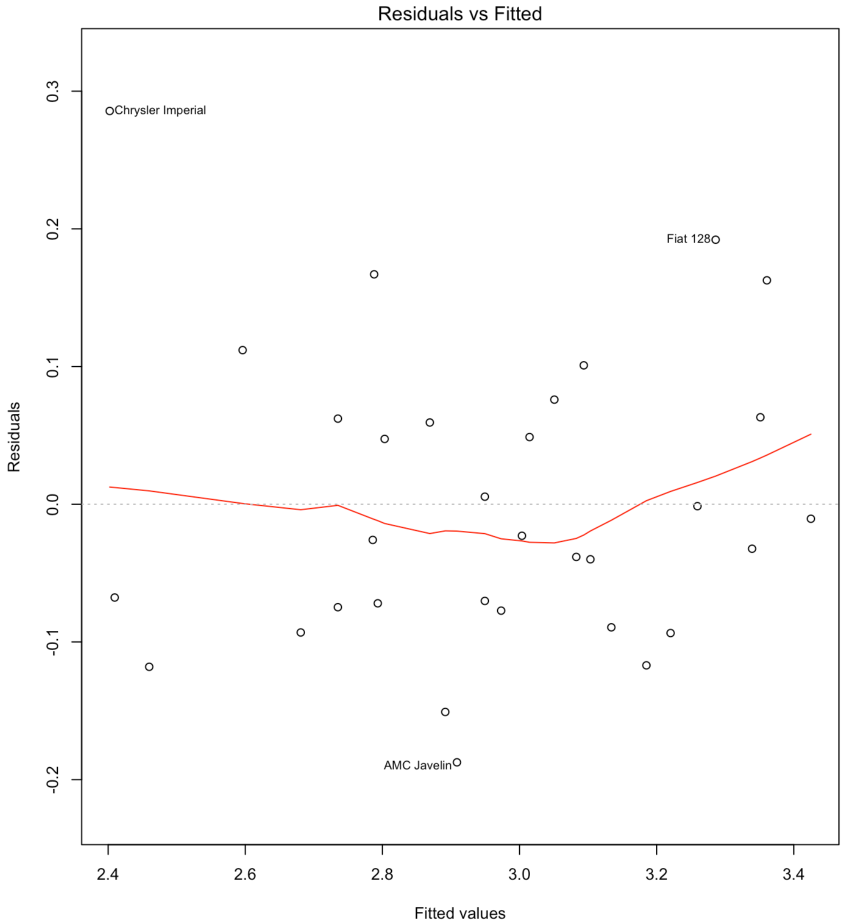


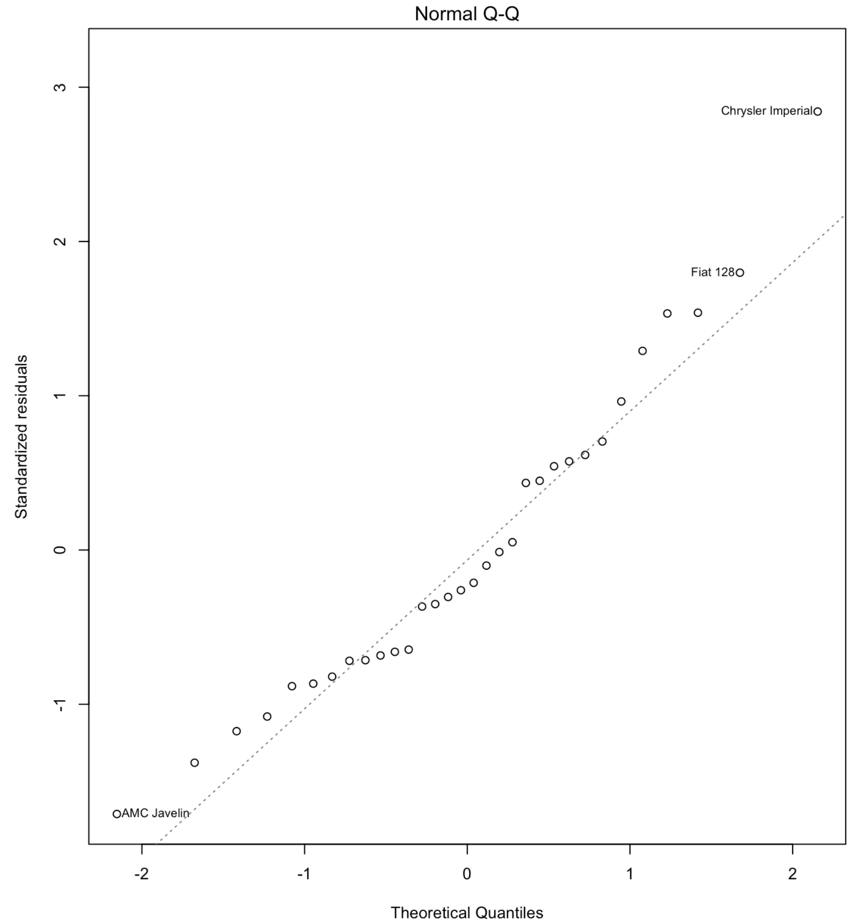
Adjustments

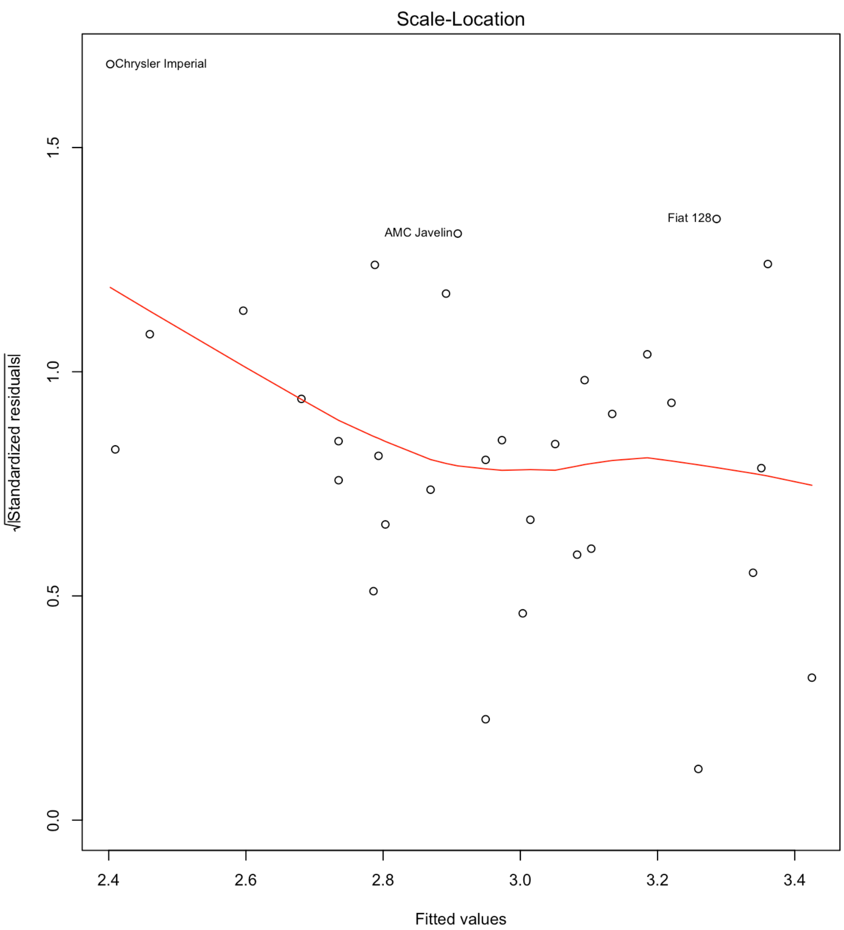


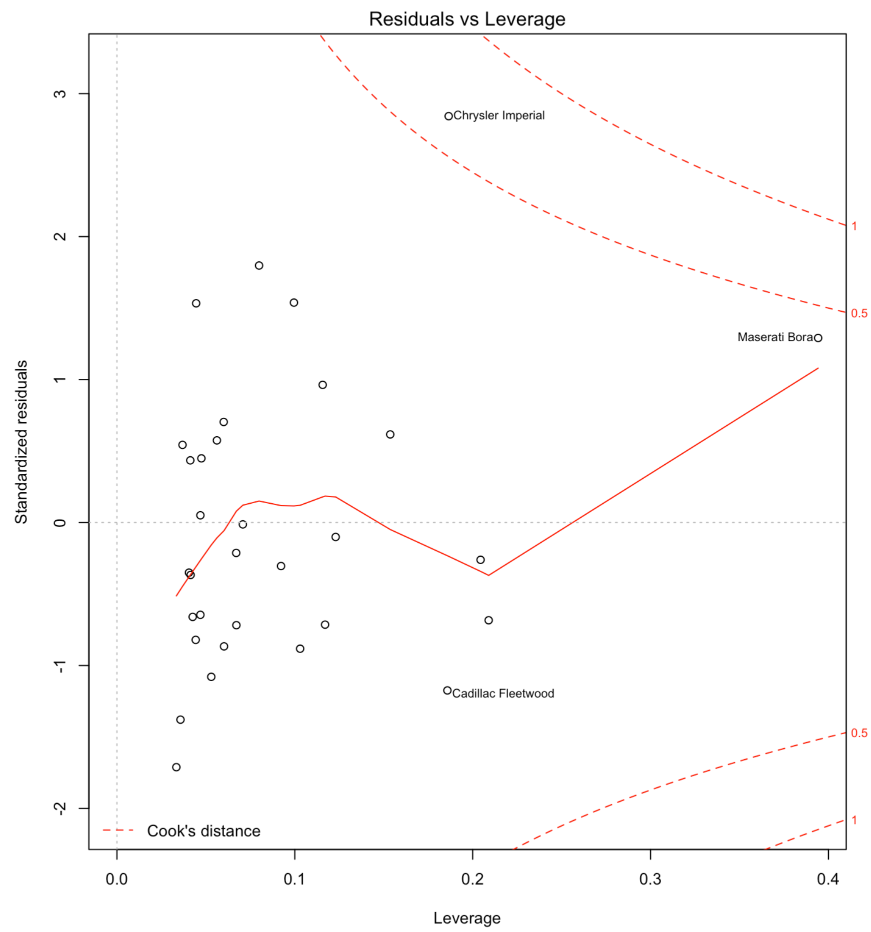


Diagnostic plot


Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Fiat 128
AMC Javelin


Normal Q-Q
Standardized residuals
Theoretical Quantiles
Chrysler Imperial
Fiat 128
AMC Javelin


Scale-Location
√|Standardized residuals|
Fitted values
Chrysler Imperial
AMC Javelin
Fiat 128


Residuals vs Leverage
Standardized residuals
Leverage
Chrysler Imperial
Maserati Bora
Cadillac Fleetwood
Cook's distance