# On effect of transmission type on mpg

*Jan Herman*

*November 22, 2015*

## Executive summary

The goal of this short analysis is to determine wheter the transmission type has the effect the car fuel efficiency. If so, we aim to quantify this effect by means of mpg difference.

We showed that the effect of transmission type on fuel efficiency is not statistically significant based on the given data. However the data suggests that cars with manual transmission might have about 2 mpg better fuel efficiency compared to cars with automatic one. We strongly recommend to collect more data in order to accept this hypothesis.

## Explaining the dataset

In this analysis we work with the `mtcars` dataset shipped with the R language. It contains 11 variables describing a total of 32 different cars. The following table summarizes the meaning of these variables:

| Variable | Description | Variable | Description |
|---|---|---|---|
| `mpg` | fuel economy [miles per gallon] | `qsec` | 1/4 mile time [sec] |
| `cyl` | number of cylinders | `vs` | engine type (0 = vee, 1 = straight) |
| `disp` | engine displacement [in$^2$] | `am` | transmission type (0 = automatic, 1 = manual) |
| `hp` | engine gross horsepower [hp] | `gear` | number of forward gears |
| `drat` | rear axle ratio | `carb` | number of carburetors |
| `wt` | car weight [lb/1000] | | |

## Goal and applied method

Our goal is to quantify the effect of the transmission type on the fuel economy. We will try to find the most appropriate multivariable linear model describing the relationship between `mpg` variable and some of the other variables including `am`. The coefficient of the `am` variable in this model will then mean the effect of the transmission type on the fuel economy measured in mpg.

In order to make an analysis as precise as possible, we have to preproccess the dataset a little – we will convert variables that have only a few integer values into factors. This involves variables `cyl`, `gear`, `carb`, `am` and `vs`. For the last two is this transformation unnecessary in means of the linear reggression (these variables have only two different values – 0 and 1, thus turning them into factors does not change the reggression model at all).

## Choosing the model

This is the most tricky part of the entire analysis. We want to have our model as simple as possible and at the same time it should estimate the `mpg` variable well enough. Especially we want to remove a majority of effect of "skewness" of the data. In order to achive this we at first analyze the covariances of all pairs of the variables. The following table present those.

| | mpg | cyl | disp | hp | drat | wt | qsec | vs | am | gear | carb |
|---|---|---|---|---|---|---|---|---|---|---|---|
| mpg | 1.000 | -0.852 | -0.848 | -0.776 | 0.681 | -0.868 | 0.419 | 0.664 | 0.600 | 0.480 | -0.551 |
| cyl | -0.852 | 1.000 | 0.902 | 0.832 | -0.700 | 0.782 | -0.591 | -0.811 | -0.523 | -0.493 | 0.527 |
| disp | -0.848 | 0.902 | 1.000 | 0.791 | -0.710 | 0.888 | -0.434 | -0.710 | -0.591 | -0.556 | 0.395 |
| hp | -0.776 | 0.832 | 0.791 | 1.000 | -0.449 | 0.659 | -0.708 | -0.723 | -0.243 | -0.126 | 0.750 |
| drat | 0.681 | -0.700 | -0.710 | -0.449 | 1.000 | -0.712 | 0.091 | 0.440 | 0.713 | 0.700 | -0.091 |
| wt | -0.868 | 0.782 | 0.888 | 0.659 | -0.712 | 1.000 | -0.175 | -0.555 | -0.692 | -0.583 | 0.428 |
| qsec | 0.419 | -0.591 | -0.434 | -0.708 | 0.091 | -0.175 | 1.000 | 0.745 | -0.230 | -0.213 | -0.656 |
| vs | 0.664 | -0.811 | -0.710 | -0.723 | 0.440 | -0.555 | 0.745 | 1.000 | 0.168 | 0.206 | -0.570 |
| am | 0.600 | -0.523 | -0.591 | -0.243 | 0.713 | -0.692 | -0.230 | 0.168 | 1.000 | 0.794 | 0.058 |
| gear | 0.480 | -0.493 | -0.556 | -0.126 | 0.700 | -0.583 | -0.213 | 0.206 | 0.794 | 1.000 | 0.274 |
| carb | -0.551 | 0.527 | 0.395 | 0.750 | -0.091 | 0.428 | -0.656 | -0.570 | 0.058 | 0.274 | 1.000 |

We will build a sequence of nested linear models having `mpg` variable as outcome. The first model will be just the constant model, the second one the simple linear regression on `am`. Now we order the remaining variables in descending order with respect to the absolute value of covariance with the outcome varible `mpg` and add them successively to the model. The residual plots of all of the models are presented in the Appendix [since all the models have different predictors we decided to plot on x-axis the row number in `mtcars` dataset corresponding to the residual].

We end with 11 different models – the first one is the constant model, then in each step we add one variable to the predictors in the following order: `am, wt, cyl, disp, hp, drat, vs, carb, gear, qsec`.

Now we make an analysis of variance on these nested models in order to determine predictors that we want to keep in the model and those we want to exclude.

| Added predictor | Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
|---|---|---|---|---|---|---|
| NA | 31 | 1126.047 | NA | NA | NA | NA |
| am | 30 | 720.897 | 1 | 405.151 | 50.474 | 0.000 |
| wt | 29 | 278.320 | 1 | 442.577 | 55.137 | 0.000 |
| cyl | 27 | 182.968 | 2 | 95.351 | 5.940 | 0.013 |
| disp | 26 | 182.869 | 1 | 0.099 | 0.012 | 0.913 |
| hp | 25 | 150.409 | 1 | 32.461 | 4.044 | 0.063 |
| drat | 24 | 150.101 | 1 | 0.308 | 0.038 | 0.847 |
| vs | 23 | 142.655 | 1 | 7.445 | 0.928 | 0.351 |
| carb | 18 | 125.618 | 5 | 17.038 | 0.425 | 0.825 |
| gear | 16 | 121.644 | 2 | 3.974 | 0.248 | 0.784 |
| qsec | 15 | 120.403 | 1 | 1.241 | 0.155 | 0.700 |

If we keep only predictors with $F$-statistic greater than one (there is a large gap) and exclude the others, we get the model described by the formula `mpg ~ am + wt + cyl + hp`:

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 33.7083239 | 2.6048862 | 12.940421 | 0.0000000 |
| ammanual | 1.8092114 | 1.3963045 | 1.295714 | 0.2064597 |
| wt | -2.4968294 | 0.8855878 | -2.819404 | 0.0090814 |
| cyl6 | -3.0313445 | 1.4072835 | -2.154040 | 0.0406827 |
| cyl8 | -2.1636753 | 2.2842517 | -0.947214 | 0.3522509 |
| hp | -0.0321094 | 0.0136926 | -2.345025 | 0.0269346 |

Finally we exclude the `cyl` predictor since its effect is not monotonous and the $p$-value for the 8 cylinder is too high. The final model thus has the formula `mpg ~ am + wt + hp` and its summary is as follows:

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 34.0028751 | 2.6426593 | 12.866916 | 0.0000000 |
| ammanual | 2.0837101 | 1.3764202 | 1.513862 | 0.1412682 |
| wt | -2.8785754 | 0.9049705 | -3.180850 | 0.0035740 |
| hp | -0.0374787 | 0.0096054 | -3.901830 | 0.0005464 |

## Conclusion

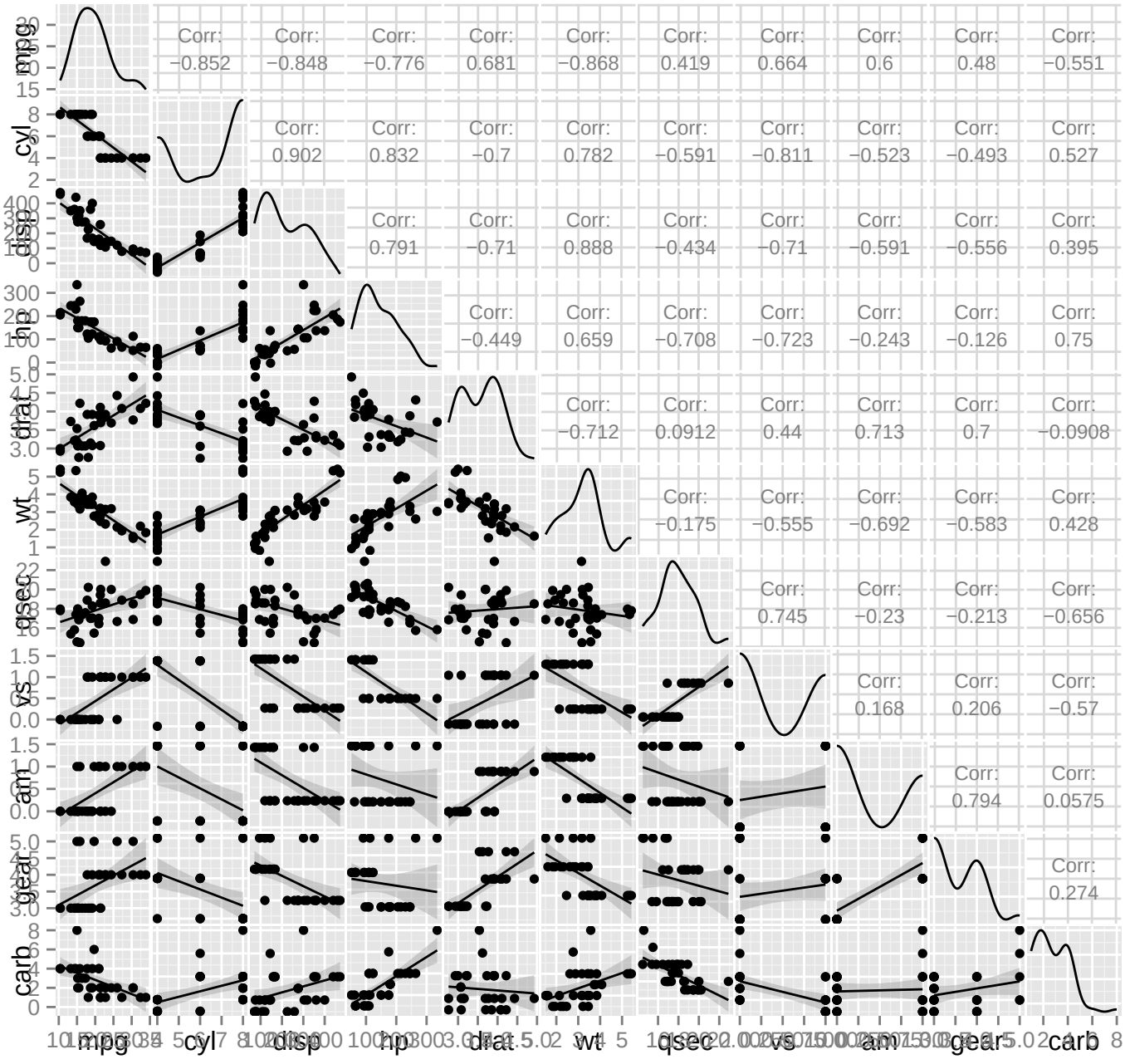
From the table above we see that the $p$-value associated with the `am` regressor is greater than 0.05, and so we failed to reject the null hypotheses that the type of trasmission has no impact on car fuel efficiency at 95% confidence level. However the data suggests that the hypothesis might be valid, so we higly recommend to collect more data to confirm it definitely.

The second proposed question about the quantification of the effect of the transmission type is thus meaningless. We hope that the further analysis on larger data will prove our hypotheses that manual transmissions are better for fuel efficiency by approximately 2 mpg.

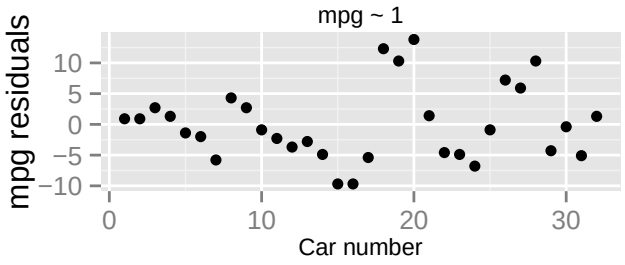
Also the residual plots in the appendix suggests that we do not take into accounts some of the car properties having an effect on the fuel efficiency. We thus recommend to try to find those.

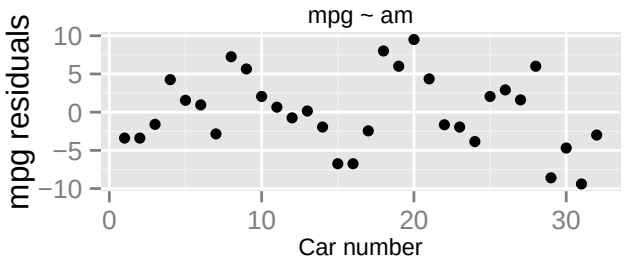
## Apendix

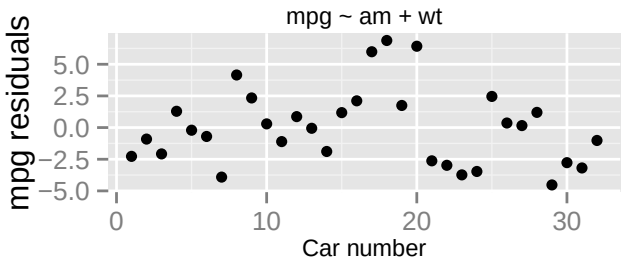
At first present one figure – the matrix of one variable linear regressions together with covariances of each pair of variables. In the main part it was replaced by covariance table as it covers less space.

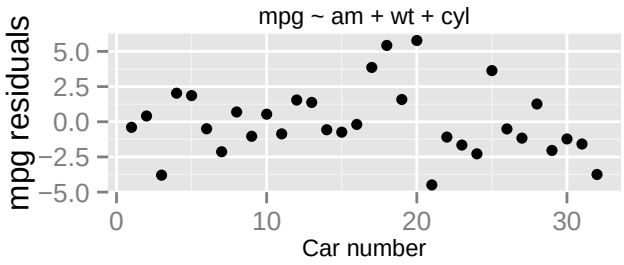


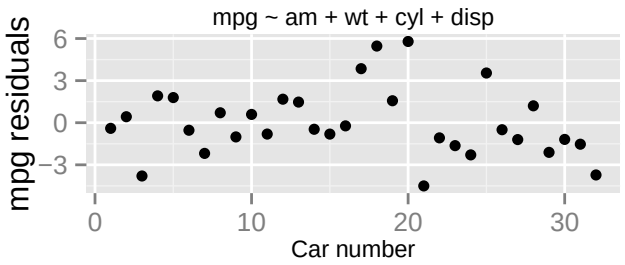
Next come the residual plots of all the nested models:

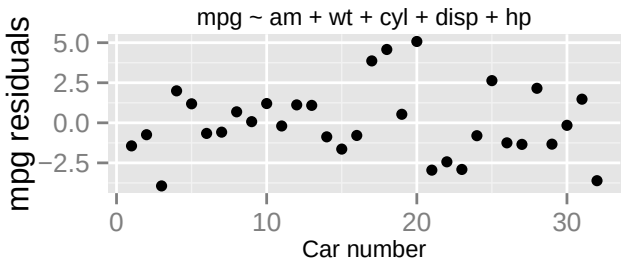


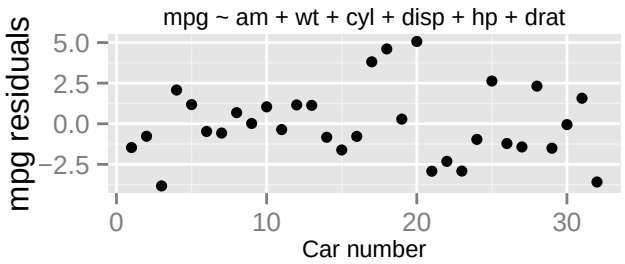



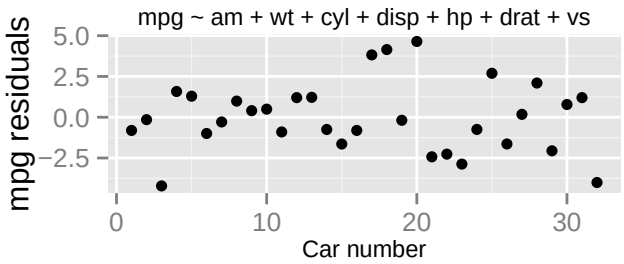



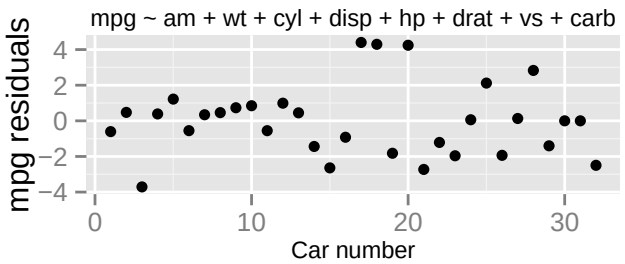



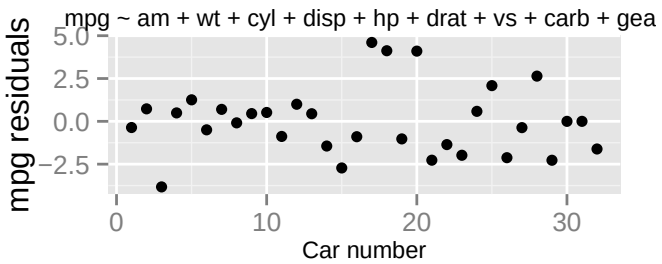



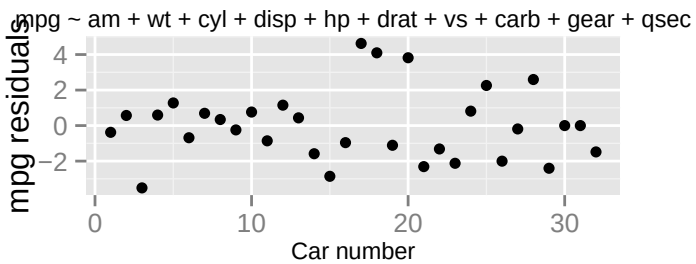



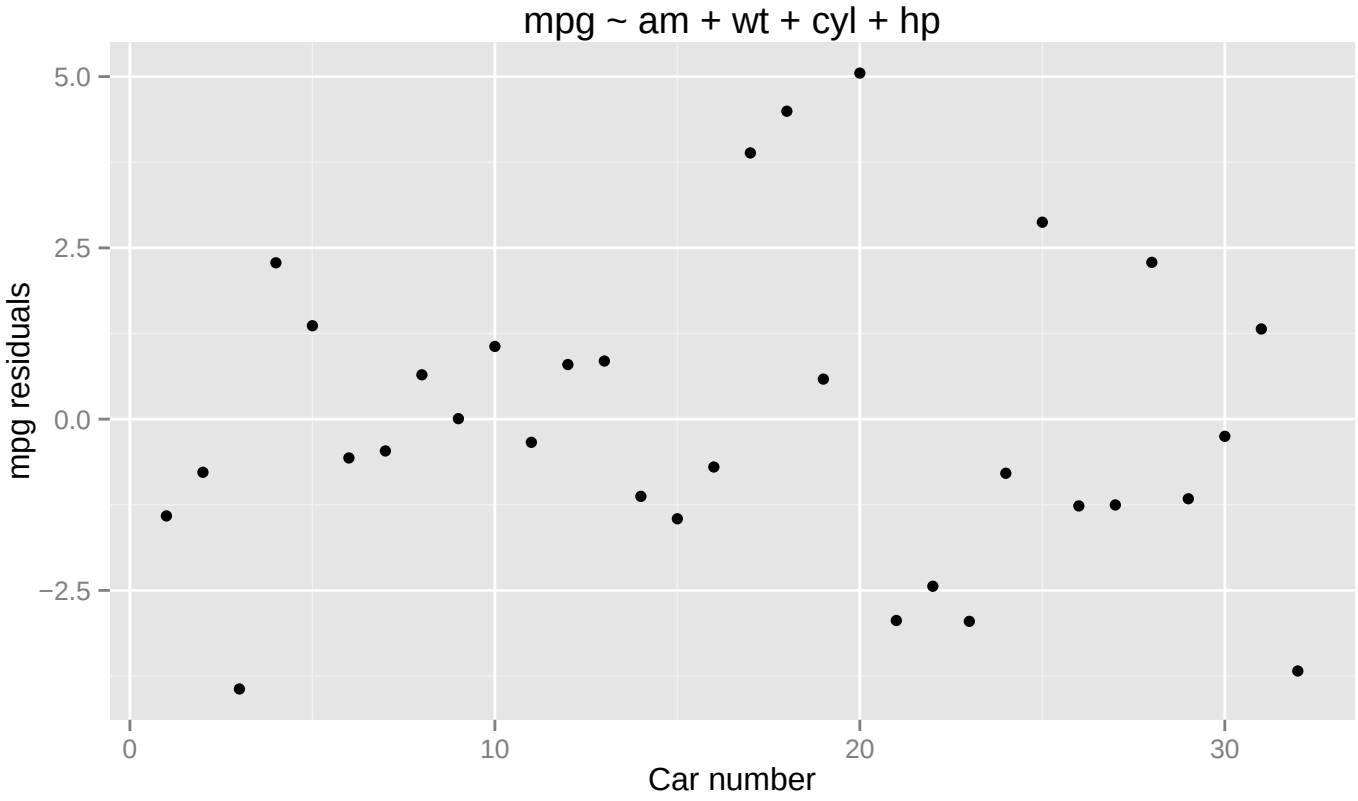



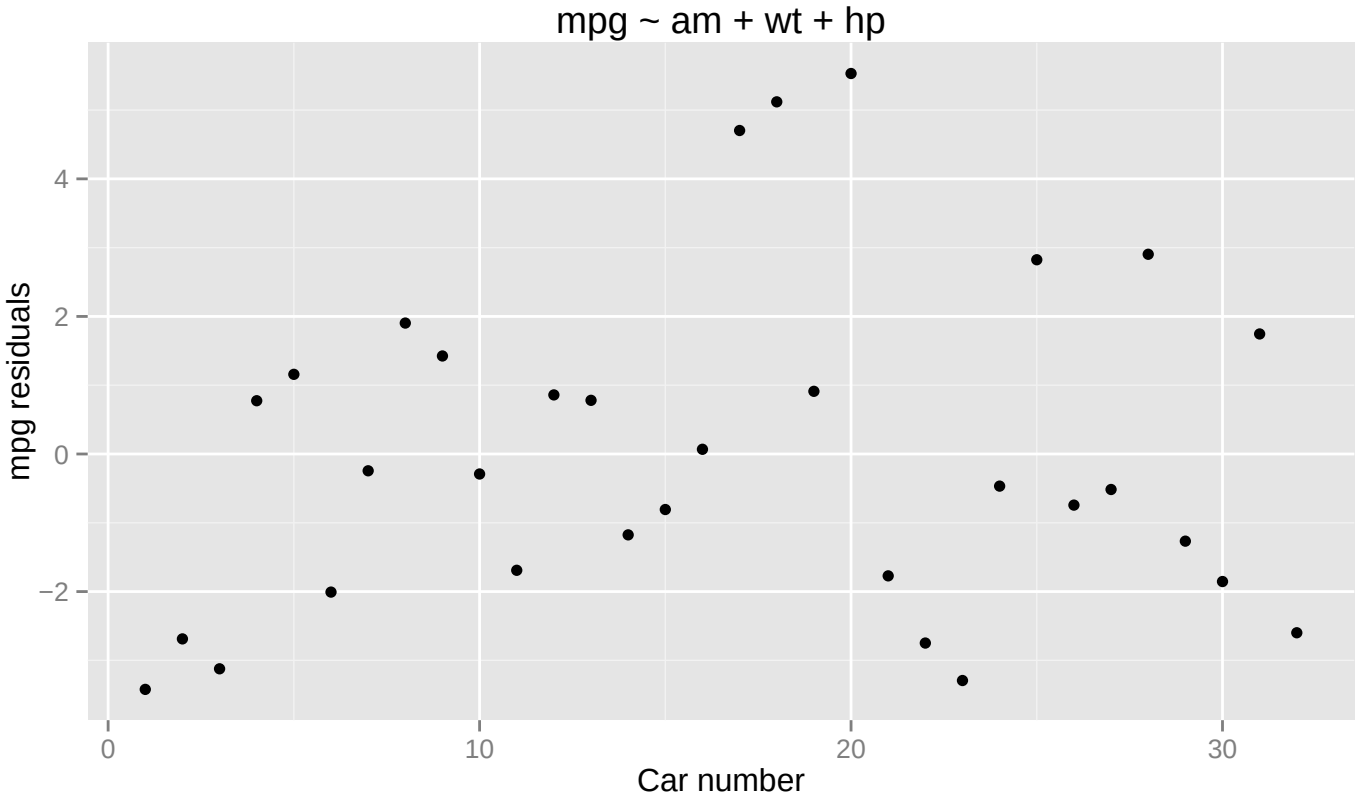



And finally the residual plots for the reduced and final model:




## Notice

The knitr source code of this analysis including dirty tricks to fit the main part on 2 pages is available at github.