Automobile transmission type and fuel mileage

Timo Voipio

26 Aug 2016

Executive summary

We analyze the fuel consumption of various automobiles from early 1970s in order to determine whether the transmission type, manual or automatic, significantly affects fuel consumption. Based on a linear model, our results indicate that a manual transmission is connected with better gas mileage ($\alpha < 0.05$). The other covariates present in the linear model are the vehicle weight and the quarter-mile (dragstrip) time.

Introduction

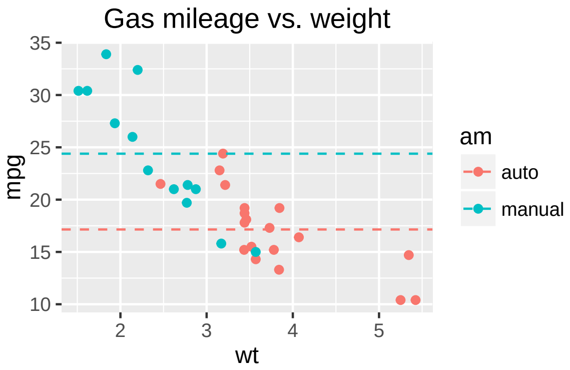
We investigate the gas mileage of the cars in the dataset `mtcars` provided by the R `dataset` package.

The aim of the analysis is two answer two questions: is an automatic or manual transmission better for gas mileage (measured in mpg, miles traveled per one US gallon of fuel consumed), and how the gas mileage is quantitatively affected by the transmission type.

Data description and exploratory analysis

The data consists of design and performance data of 32 cars, as published in the *Motor Trends* magazine printed in the United States in 1974. For a description of the dataset, see the help page of `mtcars` in R. As suggested by Henderson and Velleman (1981), a new variable `pwr` (hp/1000 lbs) is created for the power-to-weight ratio and used instead of the engine power in order to reduce multicollinearity of the covariates.

As a first glance to the data, we plot the gas mileage[1] (measured in miles per gallon) against the vehicle weight, with transmission type being differentiated by color.



The plot shows that the mean gas mileage of cars with manual transmission is significantly higher (i.e., better) compared to cars with automatic transmission. However, the plot similarly illustrates that the gas mileage seems to have a clear negative correlation with the weight of the vehicle. We therefore conclude that analyzing at the effect of transmission type on the gas mileage cannot meaningfully be done without adjusting for the effect of other factors affecting the fuel efficiency.

[1] The research question stated in the project assignment exclusively asks for analysis of the gas mileage (MPG), so for consistency reasons the data is not converted into SI units, even though the conversion would make the data more accessible to most parts of the world. Additionally, analysing the consumption via the amount of fuel consumed per fixed distance (e.g., litres/100 km) would be both physically and statistically more reasonable choice, as suggested also by Henderson and Velleman, than gas mileage. However, as the assignment explicitly asks for effect of transmission type on gas mileage, the data will not be converted for analysis.

Linear model for gas mileage

Next we try to isolate the effect of the transmission type by fitting a linear model to the data. This entails also model selection, i.e., which variables best explain the variability of MPG, without overfitting. The dataset contains variables which we would expect to be correlated to at least some degree, such as engine displacement and power, and power-to-weight ratio and quarter-mile time (see the Appendix for a pairs plot). In order to keep the model interpretable, we consider the following variables whose connection to gas mileage has a logical explanation: weight, engine displacement, power-to-weight ratio, quarter-mile time, and transmission type.

We approach the model fitting by starting from a model with transmission type as the only regressor. Additional regressors are then added, one at a time, always choosing the one which results in the smallest deviance for the resulting model.

Our analysis shows that the model with weight, quarter mile time, and transmission type as the regressors gives the best fit (in the sense that residual sum of squares is minimized), with the constraint that the last added regressor is still significant. The significance is measured using the F test, and the null hypothesis that the regressor is not significant ($\alpha < 0.05$). For three regressors, the largest P value is much less than 0.05. Adding the power-to-weight ratio `pwr` would result in a "better" fit in the sense of smaller residual sum of squares, but the P value of 0.29 indicates that this could constitute overfitting, and thus a 3-regressor model is used.

The coefficients of the 3-regressor model are

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.62	6.96	1.38	0.178
ammanual	2.94	1.41	2.08	0.047
wt	-3.92	0.71	-5.51	0.000
qsec	1.23	0.29	4.25	0.000

The coefficient `ammanual` quantifies the effect of the transmission type when weight and 1/4 mile time has been adjusted for. The P value is less than 0.05, so we conclude that, with a 95 % probability, `ammanual` differs from zero and thus the gas mileage difference between manual and automatic transmissions is statistically significant. According to our model, using a manual transmission as opposed to automatic leads to a 2.94 mpg *increase* in the gas mileage, i.e., a manual transmission is connected with better mileage. The respective confidence interval is $[0.05, 5.83]$.

The coefficients `wt` (weight) and `qsec` (quarter-mile time) indicate that each 1000 lbs increase in weight leads to a gas mileage decrease of -3.92 and each additional second in a quarter-mile run is connected with a 1.23 increase in gas mileage.

The quality of the fit is assessed via a residual plot and a quantile-quantile plot (see the Appendix). There is no easily discernible pattern and no heteroskedasticity in the residuals when plotted as a function of the fitted values, and the Q–Q plot shows that the residurals are at least roughly speaking normally distributed.

Sources

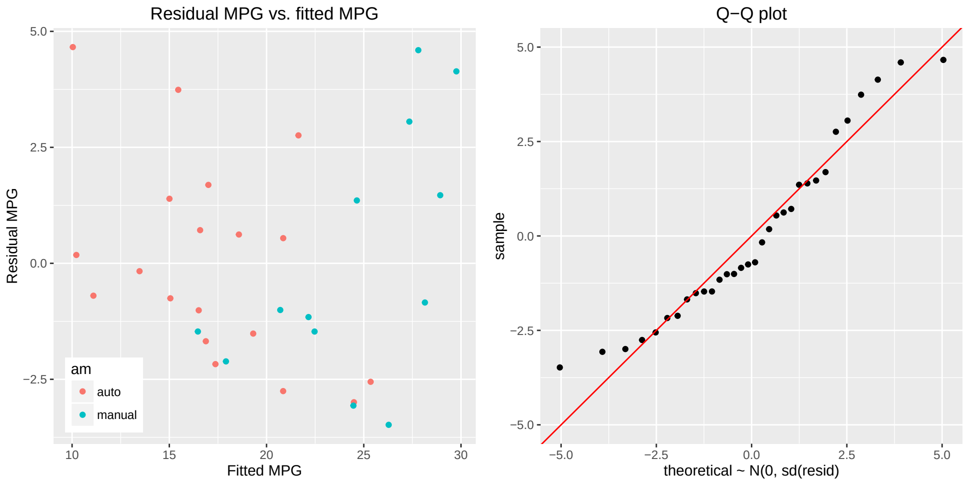
- Ronald R. Hocking, "The Analysis and Selection of Variables in Linear Regression," *Biometrics* **32** (1976), pp. 1–49. http://www.jstor.org/stable/2529336
- Harold V. Henderson and Paul F. Velleman, "Building Multiple Regression Models Interactively," *Biometrics* **37** (1981), pp. 391–411. http://www.jstor.org/stable/2530428

Source code

Complete R Markdown source is available in GitHub: https://github.com/tvoipio/JHU_RM_Project

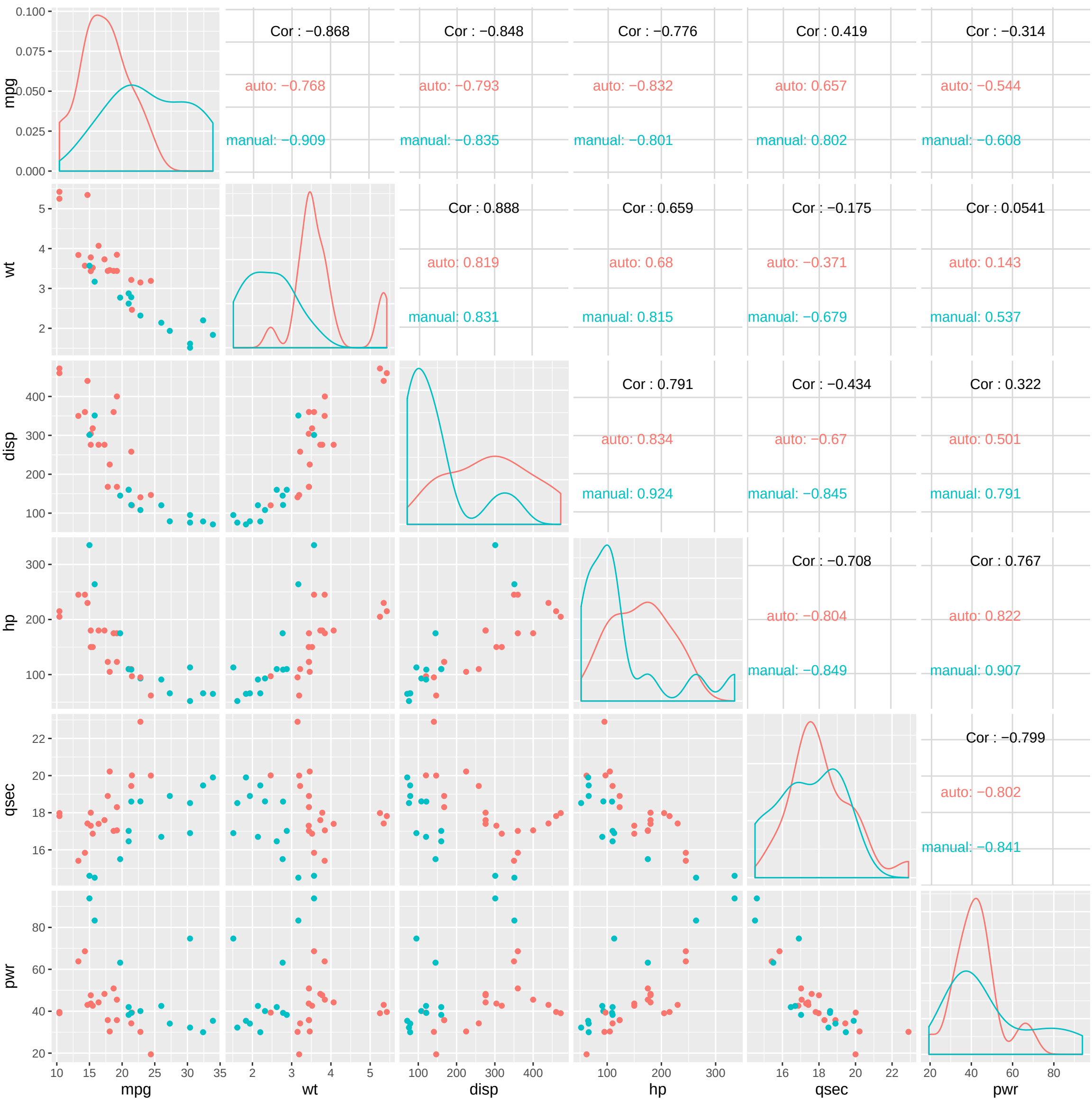
Appendix

Residual plot and Q–Q plot:



Pairs plot of selected variables::

```
paircols <- c("mpg", "wt", "disp", "hp" ,"qsec", "pwr")
ggpairs(mtcars, aes(color = am),
        columns = match(paircols, names(mtcars)))
```


mpg
wt
disp
hp
qsec
pwr
Cor : −0.868
auto: −0.768
manual: −0.909
Cor : −0.848
auto: −0.793
manual: −0.835
Cor : −0.776
auto: −0.832
manual: −0.801
Cor : 0.419
auto: 0.657
manual: 0.802
Cor : −0.314
auto: −0.544
manual: −0.608
Cor : 0.888
auto: 0.819
manual: 0.831
Cor : 0.659
auto: 0.68
manual: 0.815
Cor : −0.175
auto: −0.371
manual: −0.679
Cor : 0.0541
auto: 0.143
manual: 0.537
Cor : 0.791
auto: 0.834
manual: 0.924
Cor : −0.434
auto: −0.67
manual: −0.845
Cor : 0.322
auto: 0.501
manual: 0.791
Cor : −0.708
auto: −0.804
manual: −0.849
Cor : 0.767
auto: 0.822
manual: 0.907
Cor : −0.799
auto: −0.802
manual: −0.841