

# 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

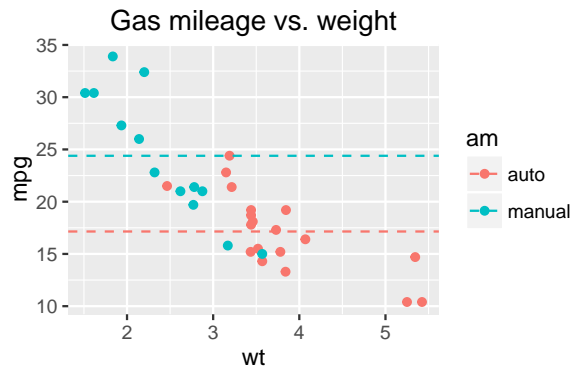
We investigate the gas mileage of the cars in the dataset `mtcars` provided by the R `datasets` package.

The aim of the analysis is to 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<sup>1</sup> (measured in miles per gallon) against the vehicle weight, with transmission type being differentiated by color.



## 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. As a result, we obtain 5 “telescoping” models. Many other models are fitted by the algorithm, but they are not saved as they do not result in the “best” possible fit for the given number of variables.

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
<code>ammanual</code>	2.94	1.41	2.08	0.047
<code>wt</code>	-3.92	0.71	-5.51	0.000
<code>qsec</code>	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 residuals are at least roughly speaking normally distributed.

## Sources

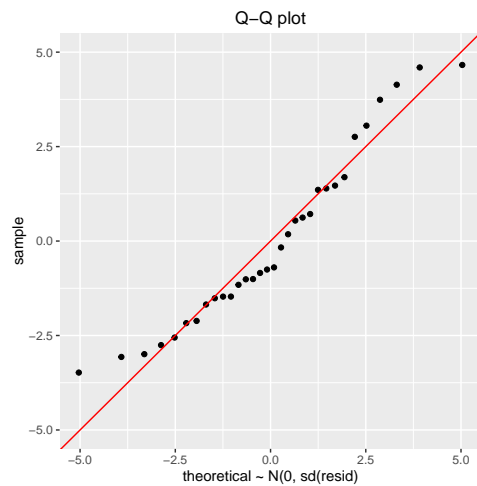
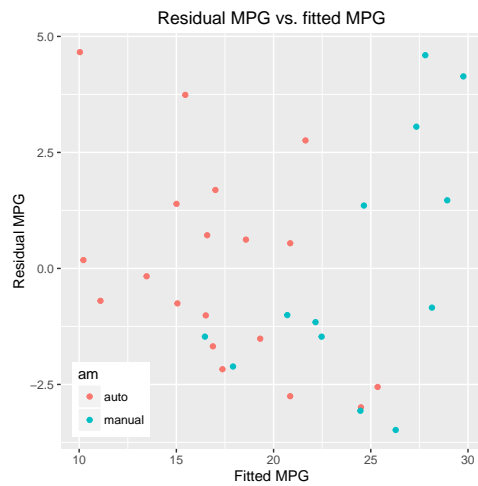
- Ronald R. Hocking, “The Analysis and Selection of Variables in Linear Regression,” *Biometrics* **32** (1976), pp. 1–49. <http://www.jstor.org/stable/2529336>
- Harold V. Henderson and Paul F. Velleman, “Building Multiple Regression Models Interactively,” *Biometrics* **37** (1981), pp. 391–411. <http://www.jstor.org/stable/2530428>

## Source code

Complete R Markdown source is available in GitHub: [https://github.com/tvoipio/JHU\\_RM\\_Project](https://github.com/tvoipio/JHU_RM_Project)

## Appendix

Residual plot and Q-Q plot:



Pairs plot of selected variables::

