Automobile transmission type and fuel mileage

Timo Voipio

21 Aug 2016

Executive summary

Introduction

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. The dataset is `mtcars` provided by the R `dataset` package, and it consists of mechanical and performance data on 32 different cars from the model years 1973 and 1974.

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

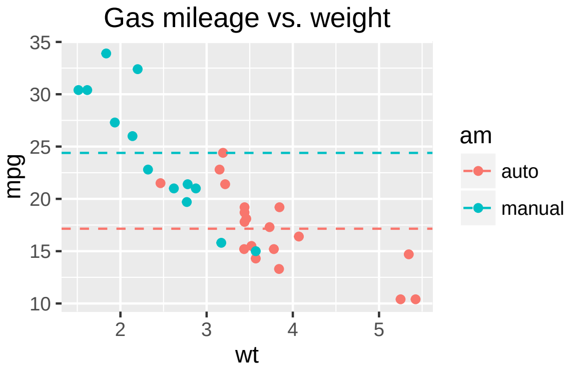
The data consists of design and performance data of 32, as published in the *Motor Trends* magazine published in the United States in 1974. According to the dataset documentation, the variables in the following table are included. Text in *italics* gives explanatory notes not present in the original data.

Column name	Variable description and units
`mpg`	Miles/(US) gallon (*gas mileage; 235 l/100 km = 1/(1 mpg)*)
`cyl`	Number of cylinders
`disp`	Displacement (cu.in.) (*1 litre = 61.0 cu in*)
`hp`	Gross horsepower (*1 kW = 1.34 hp*)
`drat`	Rear axle ratio (*driveshaft rpm/axle rpm*[1])
`wt`	Weight (1000 lbs = *454 kg*)
`qsec`	1/4 mile time
`vs`	V/S [*V engine (0) or straight (inline) engine (1)*]
`am`	Transmission (0 = automatic, 1 = manual)
`gear`	Number of forward gears
`carb`	Number of carburetors

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[2] 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. As suggested by Henderson and Velleman, a new variable `pwr` (hp/1000 lbs) is created for the power-to-weight ratio.

[1] "the drive-axle ratio is a comparison of the number of gear teeth on the ring gear of the rear axle and the pinion gear on the driveshaft. - - For example, a 4.11:1 ratio means there are 4.11 teeth on the axle's ring gear for each tooth on the driveshaft's pinion gear. Or, put another way, the driveshaft must turn 4.11 times to turn the rear wheels one full revolution. - - typical rule of thumb: The higher the numerical ratio, the slower the gear will be. This higher ratio gives a truck greater pulling power, but since the engine must work harder to spin the driveshaft more times for each turn of the rear wheels, top-end speed and fuel economy are sacrificed." From [worktruckonline.com][drat]

[2] As suggested also by Henderson and Velleman



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.

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: `wt`, `disp`, `pwr`, `qsec`, and `am`.

We approach the model fitting by starting from a model with `am` 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. The ANOVA table resulting from this process is presented below.

Model 3, with weight, quarter mile time, and transmission type as the regressors, gives the best fit (in the sense that deviance is minimized) where the attained significance leads us to reject the null hypothesis that the added regressor is not significant. 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 would lead to overfitting.

The coefficients of the 3-regressor model are

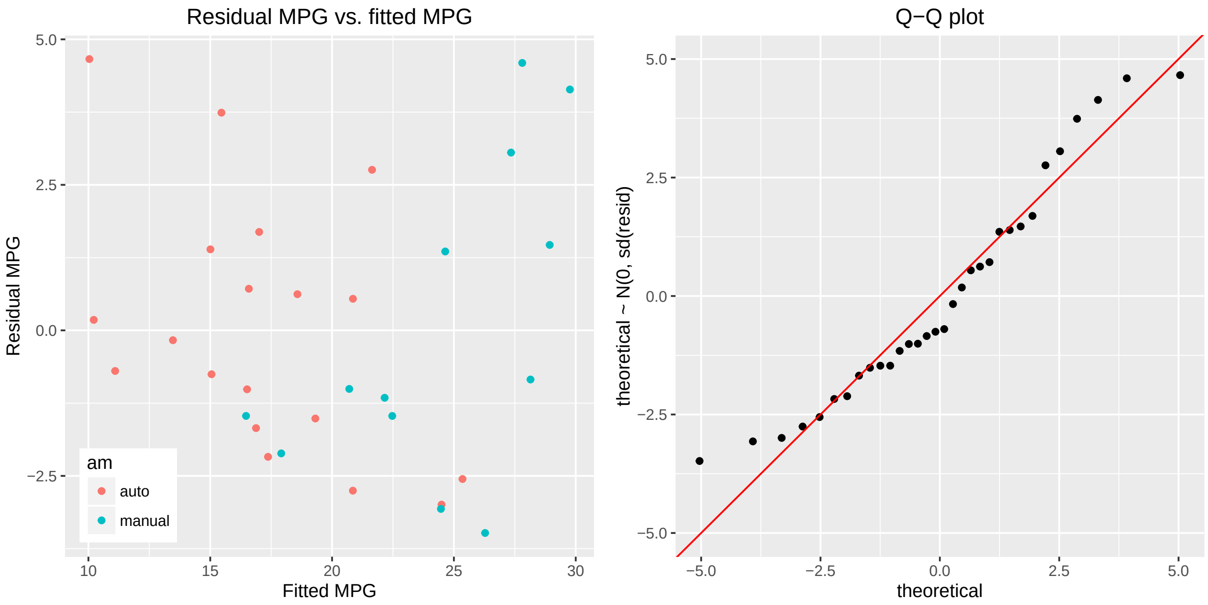
```
mpgmodel <- fits[[3]]
summary(mpgmodel)$coefficients
```

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. According to our model, using a manual transmission as opposed to automatic leads to a 2.89 mpg increase in the gas mileage. The respective confidence interval is $[0.05, 5.83]$.

The coefficients `wt` and `qsec` indicate that each 1000 lbs increase in weight leads to a gas mileage decrease of -4.01 and each additional second in a quarter-mile run is connected with a 0.86 increase in gas mileage.

```
resplot <- qplot(x = fitted(mpgmodel), y = resid(mpgmodel), data = mtcars,
                 mapping = aes(color = am),
                 xlab = "Fitted MPG", ylab = "Residual MPG",
                 main = "Residual MPG vs. fitted MPG") +
    theme(legend.justification=c(0,0), legend.position=c(0,0))

qqplot <- ggplot(data = data.frame(y = resid(mpgmodel)), aes(sample = y)) +
    stat_qq(dparams = list(sd = sd(resid(mpgmodel)))) +
    geom_abline(slope = 1, intercept = 0, color = "red") +
    ylim(-5, 5) +
    ggtitle("Q-Q plot") +
    ylab("theoretical ~ N(0, sd(resid)")
```



There is no obvious pattern present in the residuals (left panel), which supports our model selection. The Q–Q plot shows that the distribution of the residual quantiles is roughly approximated by a normal distribution (red line).

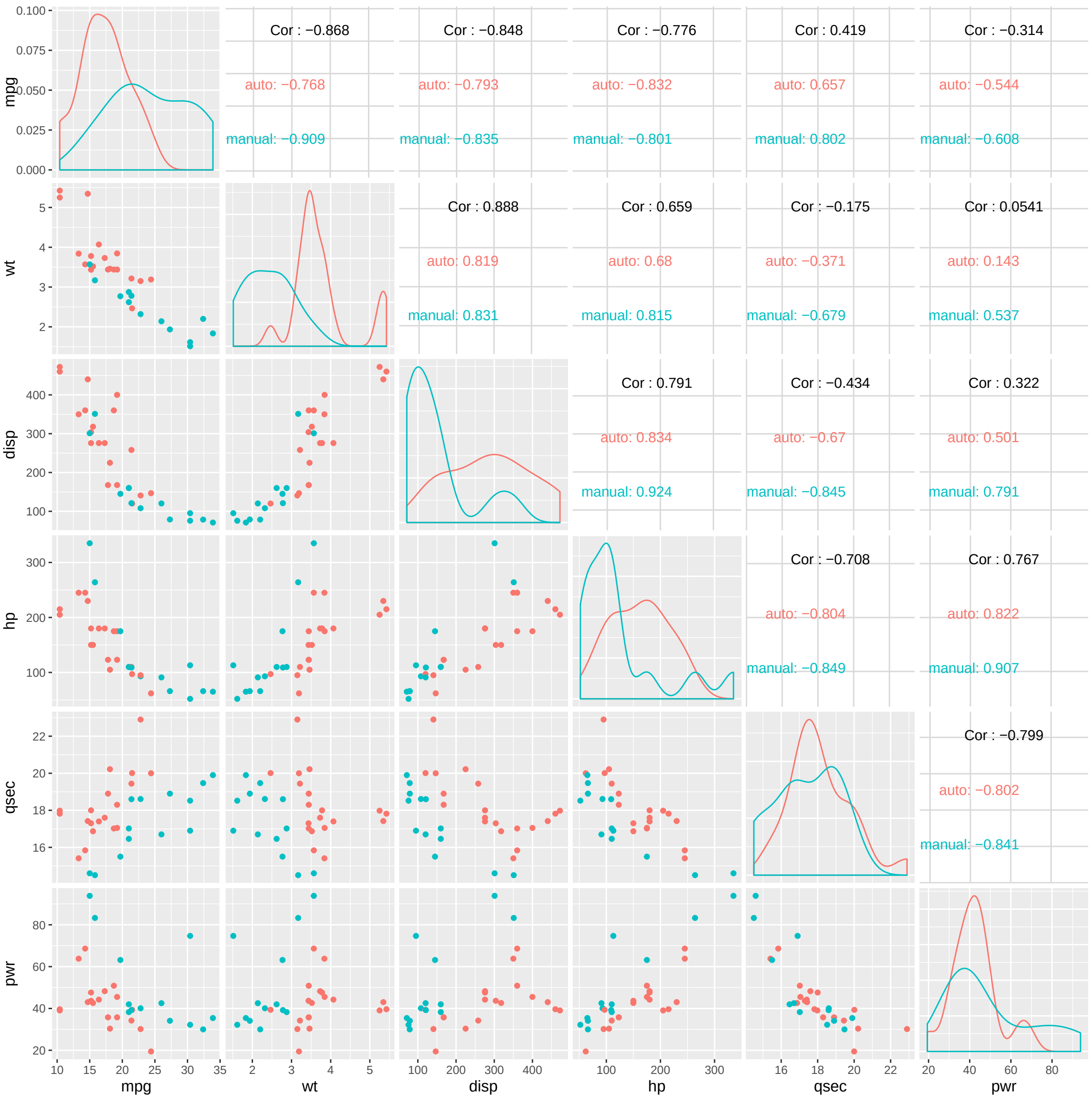
Sources

- 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

Appendix

Pairs plot of selected varibles

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



Source code for init, exploratory plot, and model selection

Initialization:

```r
library(datasets)
data("mtcars")

library(ggplot2)
library(GGally)
```

Exploratory plot:

```r
# Convert transmission type, (cylinder count, carburetor count), and engine type
# to factor variables
mtcars$am <- factor(mtcars$am, levels = c(0, 1),
                    labels = c("auto", "manual"))
#mtcars$cyl <- factor(mtcars$cyl)
#mtcars$carb <- factor(mtcars$carb)
mtcars$vs <- factor(mtcars$vs, levels = c(0, 1),
                    labels = c("v", "straight"))

# Create a new variable for gross power/weight ratio
# (hp/1000 lb; 1 hp/1000 lb   1.64 W/kg)
mtcars$pwr <- mtcars$hp/mtcars$wt

g <- ggplot(mtcars, aes(x = wt, y = mpg))
g <- g + geom_point(aes(color = am))
g <- g + geom_hline(aes(yintercept = mpg, color = am),
                    data = aggregate(mpg ~ am, data = mtcars, mean),
                    linetype = "dashed")
g <- g + ggtitle("Gas mileage vs. weight")

print(g)
```

Model selection:

```
# Function for constructing a formula containing the desired variables
fitformula <- function(fitvar, modelvars = NULL, keepvar = NULL) {
    modelstr <- paste("mpg ~", paste(c(keepvar, modelvars, fitvar),
                                     collapse = " + "))
    as.formula(modelstr)
}

# Calculate the deviance resulting from fitting MPG using the listed
# variables as regressors
dev.next <- function(fitvar, modelvars = NULL)
{
    fit <- lm(fitformula(fitvar, modelvars), mtcars)
    deviance(fit)
}

fitvars <- c("wt", "disp", "pwr", "am", "qsec")
resids <- dev.next("am")
modelvars <- "am"
fits <- vector("list", length(fitvars))
fits[[1]] <- lm(fitformula("am"), mtcars)

# Create a sequence of fits, starting from mpg ~ am.
# In each step, choose as the next regressor the one which
# results in the fit having the smallest deviance (square sum
# of residuals)
while (is.null(fits[[length(fits)]]))
{
    # Determine the deviances of the linear models when each of the
    # as of yet unused variables is added
    devs <- sapply(setdiff(fitvars, modelvars),
                   function(fitvar) dev.next(fitvar, modelvars))

    # Choose the variable which results in least deviance
    resids <- c(resids, devs[which.min(devs)])
    modelvars <- c(modelvars, names(which.min(devs)))

    # Create a fit involving the chosen variables
    fits[[length(modelvars)]] <- lm(fitformula(modelvars), mtcars)
}
```

```r
fits <- fits[1:length(modelvars)]

# Call the anova function with the incremental fits
do.call(anova, fits)
```

Diagnostic plot:

```r
# Thanks to http://zevross.com/blog/2014/08/04/beautiful-plotting-in-r-a-ggplot2-cheatsheet-3/
# for the hint on gridExtra
library(gridExtra)

grid.arrange(resplot, qqplot, ncol=2)
```