Year 1974: Automatic or Manual?

Report compleated by “Motor Trend” US magazine.

As gas prices are climbing up, readers of our magazine are eager to set the final blow in the argument over economic advantages of automatic *vs.* manual transmission. In this report, *Motor Trend* uses a powerful approach of regression models to settle the issue. All conclusions are based on the well-known *mtcars* data set published in 1974. It lists fuel consumption and 10 other aspects of design and performance for 32 automobile models produced in 1973–74. Although the data are old and insufficient, we hope that findings made with this approach will help to choose your dream car.
The data set with 32 observations is too small for a meaningful multivariable analysis. This is reminiscent of the real life problem when car models that we like and can afford at the same time are few.

Although on average cars with manual, rather than automatic, transmission were more economical in 1974, our analysis shows that the type of transmission played little role and that the difference could result from other factors.

The data set is comprised of six continuous (e.g., displacement) and five discrete (e.g., number of carburetors) variables. Histograms of continuous variables are shown in Appendix 1a. Some of them are much skewed. Linear regressions work the best when variables have normal distributions. Hence, we transformed variables with skewed distributions as follows:

$MPG- > GPM = (MPG)^{-1}$;
$DISP- > log(DISP)$;
$HP- > sqrt(HP)$.

Distributions of new variables are shown in Appendix 1b. They look more bell-shaped now. We also consider an additional variable HP/WT. It estimates how overpowered a car is. For the reasons described above, we use $sqrt(HP/WT)$ as a descriptor. Importantly, $sqrt(HP/WT)$ correlates weakly with WT (cor = 0.0724). If the two were to correlate strongly, only one of them would be needed. Such is the case for variables CYL (number of cylinders) and $log(DISP)$, which have cor = 0.1569. Therefore, we consider only $log(DISP)$. In order to avoid over-fitting, we consider only models with five descriptors. Since we have ten descriptors to choose from, the total number of models is C_5^{10} = 252. Model performance is evaluated by the R^2 criteria. The best model (R^2 = 0.8625) is:

(Model A)
$$GPM = 0.0016 + 0.0039 * AM + 0.0133 * WT - 0.0045 * GEAR + 0.001 * CARB + 0.0033 * sqrt(HP/WT)$$.

According to this model, gasoline usage is expected to be higher in cars with automatic, rather than manual, transmission by 0.39 gallons per 100 miles. This estimate is valid if all other descriptors of the car are same.

It is important, however, to evaluate significance of each descriptor in the model. For this, we remove each descriptor one by one and test if we can reject the following null hypothesis: the describtor added to the model does not improve it. Thus obtained p-values are:

AM	0.3971
WT	1.6429×10^{-5}
$GEAR$	0.1738
$CARB$	0.4981
$sqrt(HP/WT)$	0.0248

As we see, descriptors $CARB$ and AM are least significant. Thus we test if we can reject the following null hypothesis: addition of descriptors $CARB$ and AM does not improve the model. The p-value for this test is 0.5297. This is much greater than any reasonable significance level. Thus inclusion of $CARB$ and AM causes over-fitting and it is sufficient to keep the three remaining descriptors:

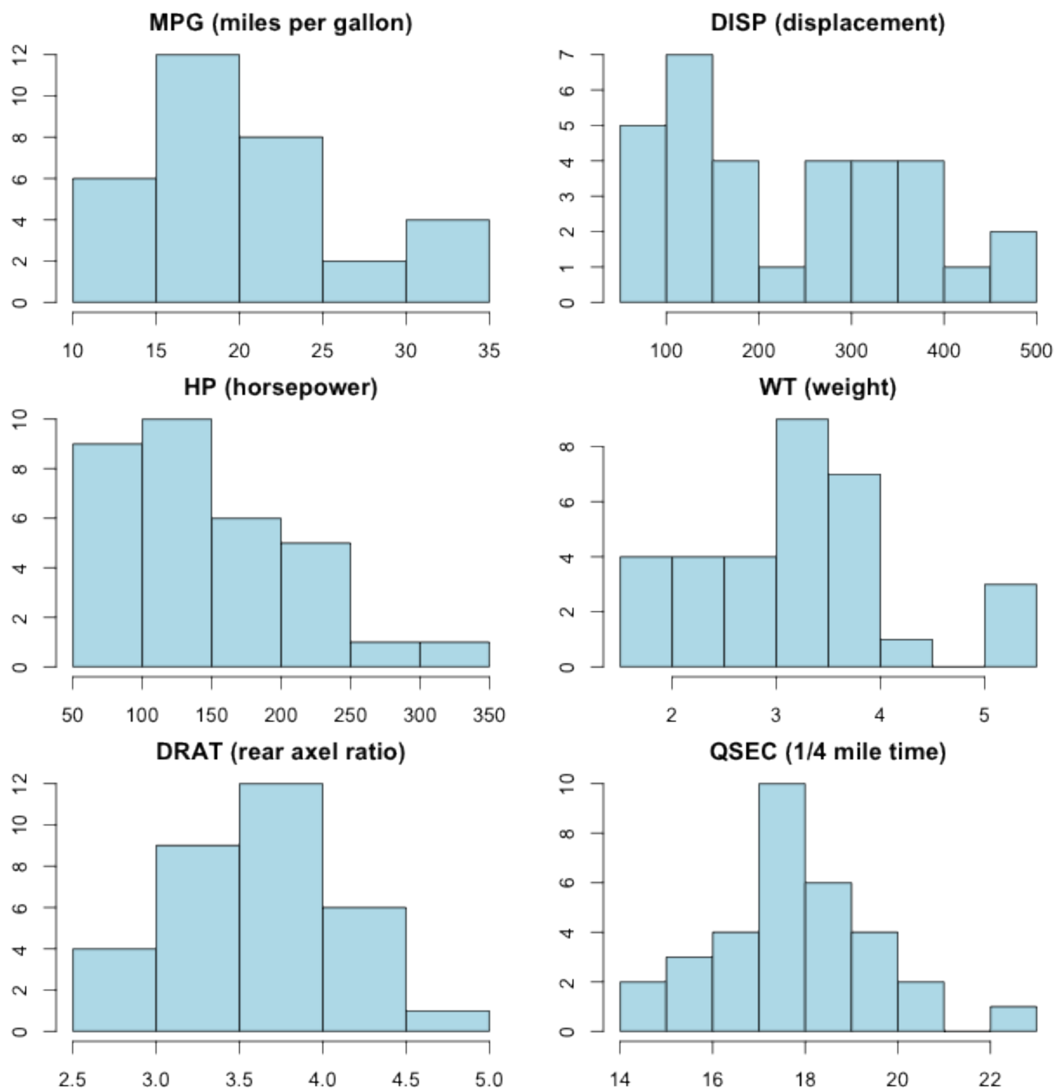
(Model B)
$$GPM = -0.01 + 0.0138 * WT - 0.002 * GEAR + 0.004 * sqrt(HP/WT)$$

Residuals of models A and B are shown in Appendix 2a. Vertical line in the plot separates points corresponding to different transmisson types: automatic transmission is on the left. The graph in Appendix 2b shows similar analysis for $MPG = (GPM)^{-1}$. Residuals for models with three and five variables are similar. Therefore, GPM is well defined by

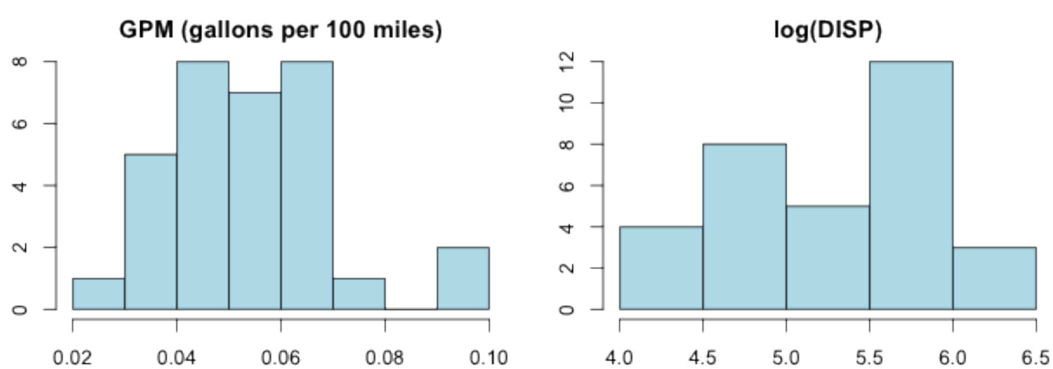
Model B. Graphs in Appendix 3 compare results of fits by Models A and B.

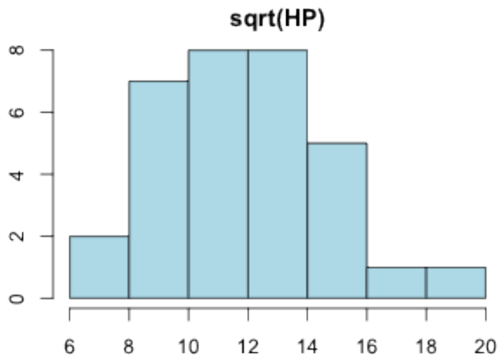
To conclude, cars with manual transmission made in 1974 had smaller on average GPM. However, the type of transmission appears to be an insignificant descriptor of GPM. Other descriptors, such as car's weight, the number of gears, and the power per weight ratio, define the car's fuel consumption.

Appendix 1a. Histograms of continuos variables:

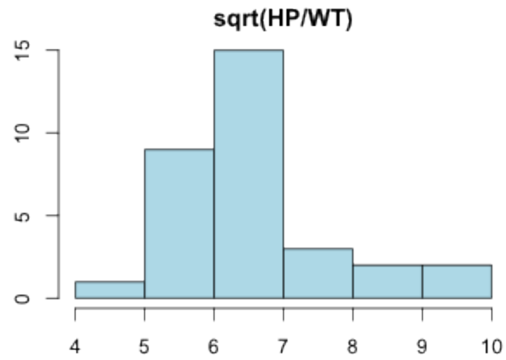


Appendix 1b. Histograms of transformed variables:



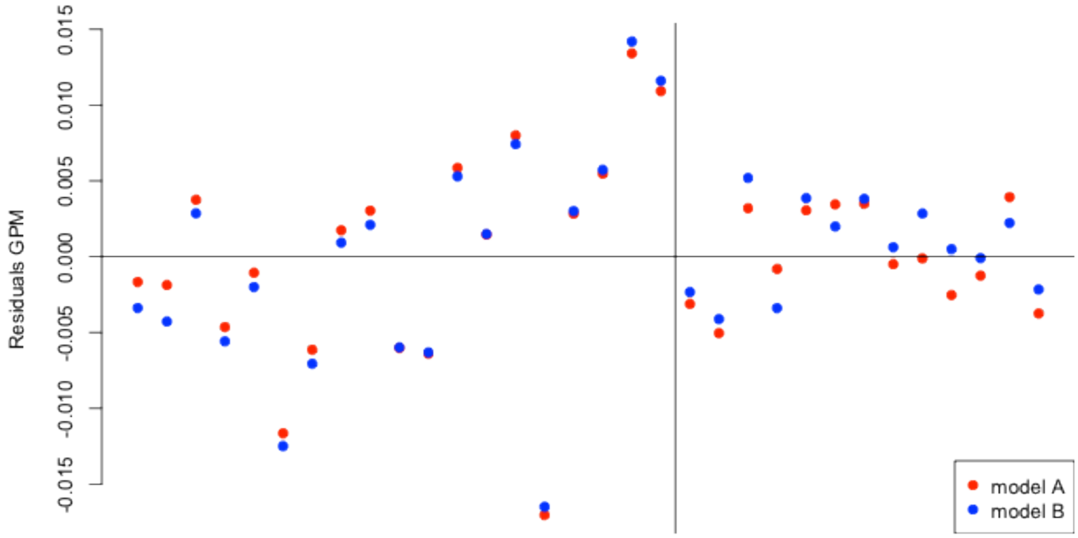


sqrt(HP)
0
2
4
6
8
6
8
10
12
14
16
18
20

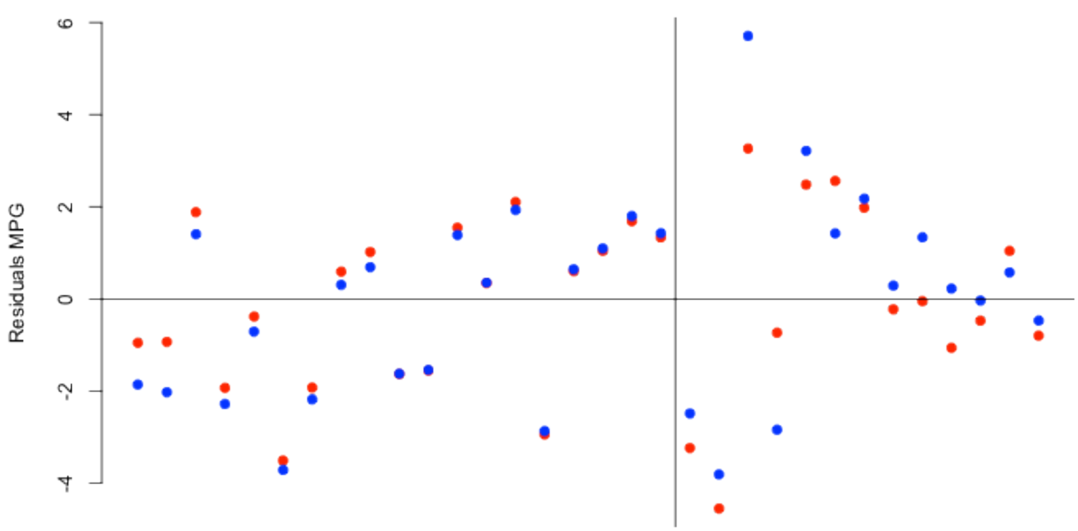


sqrt(HP/WT)
0
5
10
15
4
5
6
7
8
9
10

Appendix 2a. Residuals for GPM models with 5 and 3 variables:



Appendix 2b. Residuals for MPG models with 5 and 3 variables:



Appendix 3. Models fit 5 and 3 variables: