Automatic transmission effect on miles per gallon

Julius Neviera

2015/11/21

Summary

This report investigates how automatic transmissions in cars influence their miles per gallon. Data on car fuel consumption was analysed and linear regression model to explain then data was built. Influence of automatic transmission on miles per gallon was confirmed and it is positive - cars with automatic transmission on average consume 5.3 gallons of fuel more than cars with manual transmission.
This report and R markdown file used to create report are available in my public github repository.

Background and data

Data is taken from 1974 *Motor Trend* US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models). Fuel consumption is expressed in miles per (US) gallon - mpg.

There are 2 questions that needs answer:
1. Is an automatic or manual transmission better for MPG?
2. Quantify the MPG difference between automatic and manual transmissions.

To answer these questions multivariate linear model will be constructed.
mpg is dependent. It is variable in question. It will be analyzed in context of other automobile design and performance variables. For full list of data set variables see Table 1 in Appendix.
There are 32 data points in data and there are no missing values, thus all data points can be used in data analysis.

Technical reasoning

Figure 1 shows correlations matrix for data set variables.
mpg is correlated to every other variable in data set, but correlation sign and levels differ. Also many of the predictors are correlated, thus having all 10 variables as predictors is not recommended as model will suffer from multicollinearity.
From point of technical view what is car, data set has 2 types of variables - structural car features and performance features that depend on car structure.
Structural car features are: number of cylinders *cyl*, engine displacement *disp*, rear axle ratio *drat*, engine type *vs*, transmission type *am*, number of forward gear *gear* and number of carburetors *carb*. Car weight *wt*, gross horsepower *hp* and quarter mile time *qsec* are automobile performance features that are directly dependent on car's structural design, thus highly correlated to structural variables (though not all of them). Performance variables can be used as way to reduce model dimensions, as they will have combined structural feature effect in them (known and unknown).

Predictors, model selection and diagnostics

Main question is if transmission type has effect on *mpg*, thus first predictor is *am* and simplest model is $\mathbf{mpg{=}beta_0{+}beta_1 as.factor(am)}$. This model explains 36% percent of *mpg* variation. This level is not satisfactory.
In following steps new predictors will be added to model. When all probable predictor will be added, nested

models approach will be used to compare all incrementally created models.
Next variable **hp** is included: $\mathbf{mpg=beta_0+beta_1as.factor(am)+beta_2hp}$.
Gross horsepower is highly correlated (fig. 1) to number of cylinders *cyl*, displacement *disp*, number of carburetors *carb*, engine type *vs* and weight *wt*. By automotive design logic those 5 car features directly influence gross horsepower, thus these variables will be last to add (if used). Using gross horsepower is viable also because it can contain effect of other unknown features.
At this point there are 3 more candidates to add to model - *drat*,*qsec*,*gear*.
qsec is result of gross horsepower effect and other car features (incl. *drat* and *gear*), also it has lowest correlation to *mpg*, therefore we do not add *qsec* as it has great potential to introduce bias. *drat* and *gear* are highly correlated to transmission type, but we add them to model as causality of relation is unknown. Last expanded model is $\mathbf{mpg=beta_0+beta_1as.factor(am)+beta_2hp+beta_3drat+beta_4gear}$.
Three nested models have been created and anova analysis on the nested models show results pictured in figure 2 (appendix). Anova shows great improvement is RSS from model 1 to model 2 and H_{null}: $RS^2{}_1 = RS^2{}_2$ is rejected, thus H_{alt}: $RS^2{}_1 < RS^2{}_2$ is accepted. Model 2 has better fit than model 1. Also we fail to reject H_{null}: $RS^2{}_2 = RS^2{}_3$, model 2 has better fit than model 3.
Model 2 explains 78% of dependent variation.
Model 2 is following: **mpg=26.584914+5.277085*as.factor(am)-0.058888*hp** where $\mathbf{beta_0=26.584914}$, $\mathbf{beta_1=5.277085}$, $\mathbf{beta_2=-0.058888}$.
Figure 3 shows diagnostic plots of residuals versus fitted values. Visually examining residuals vs fitted plot confirms that there are no visible pattern in scatter plot that would suggest hidden and unaccounted factors that introduce significant bias. From leverage (fig. 3) plot it can be seen that one data point has high leverage and potentially significantly influences model. Doing model refit without that data point results in only marginal reduction in residual standard error (2.909 -> 2.878), thus it will not be removed. This concludes model diagnostics.

Results

Model 2 summary (figure 4) infers that we do reject hypotheses that model coefficients are equal to zero, thus alternatives are accepted. Now answers to questions can be stated.
Is an automatic or manual transmission better for MPG? Manual transmission is better for MPG as model shows that, while accounting for gross horsepower influence, on average automatic transmission increases MPG -> $beta_1>0$.
Quantify the MPG difference between automatic and manual transmissions. This difference, while accounting for gross horsepower influence, on average is extra **5.277085** MPG for cars with automatic transmission.

APPENDIX

id	Description
mpg	Miles/(US) gallon
cyl	Number of cylinders
disp	Displacement (cu.in.)
hp	Gross horsepower
drat	Rear axle ratio
wt	Weight (lb/1000)
qsec	1/4 mile time
vs	V/S - engine type “Vee”/“Straight”
am	Transmission (0 = automatic, 1 = manual)
gear	Number of forward gears
carb	Number of carburetors

Table 1. List of dataset variables.

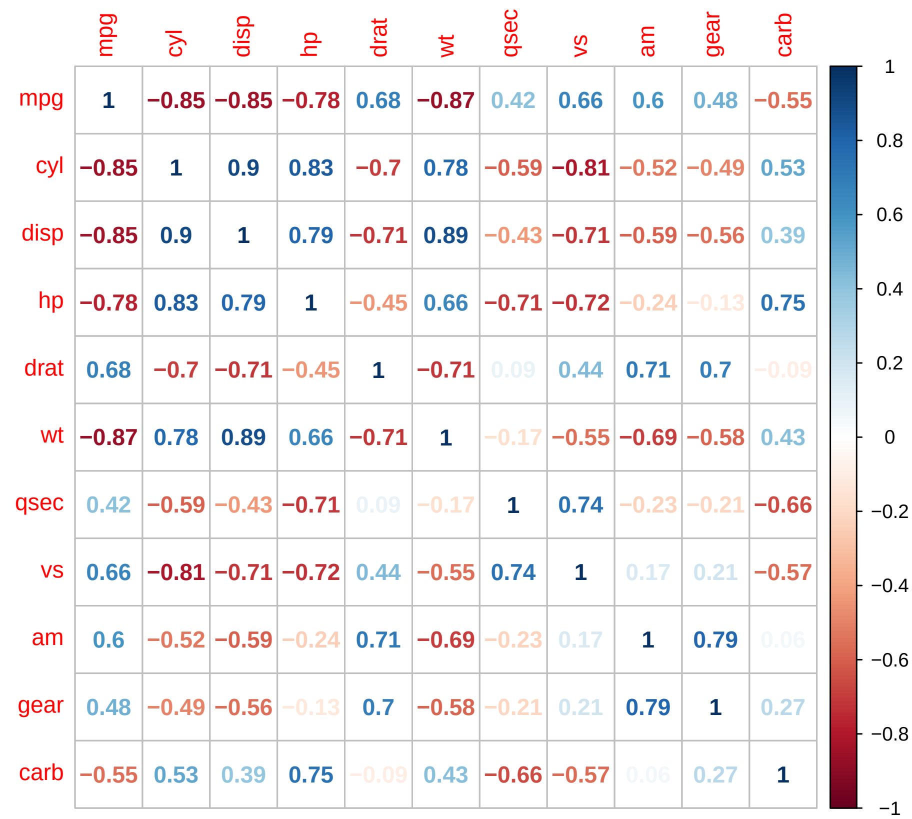

Figure 1. Variable correlations matrix.

```
## Analysis of Variance Table
##
## Model 1: mpg ~ as.factor(am)
## Model 2: mpg ~ as.factor(am) + hp
## Model 3: mpg ~ as.factor(am) + hp + drat + gear
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 245.44  1    475.46 55.7293 4.966e-08 ***
## 3     27 230.35  2     15.09  0.8842    0.4247
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 2. Anova results on 3 models.

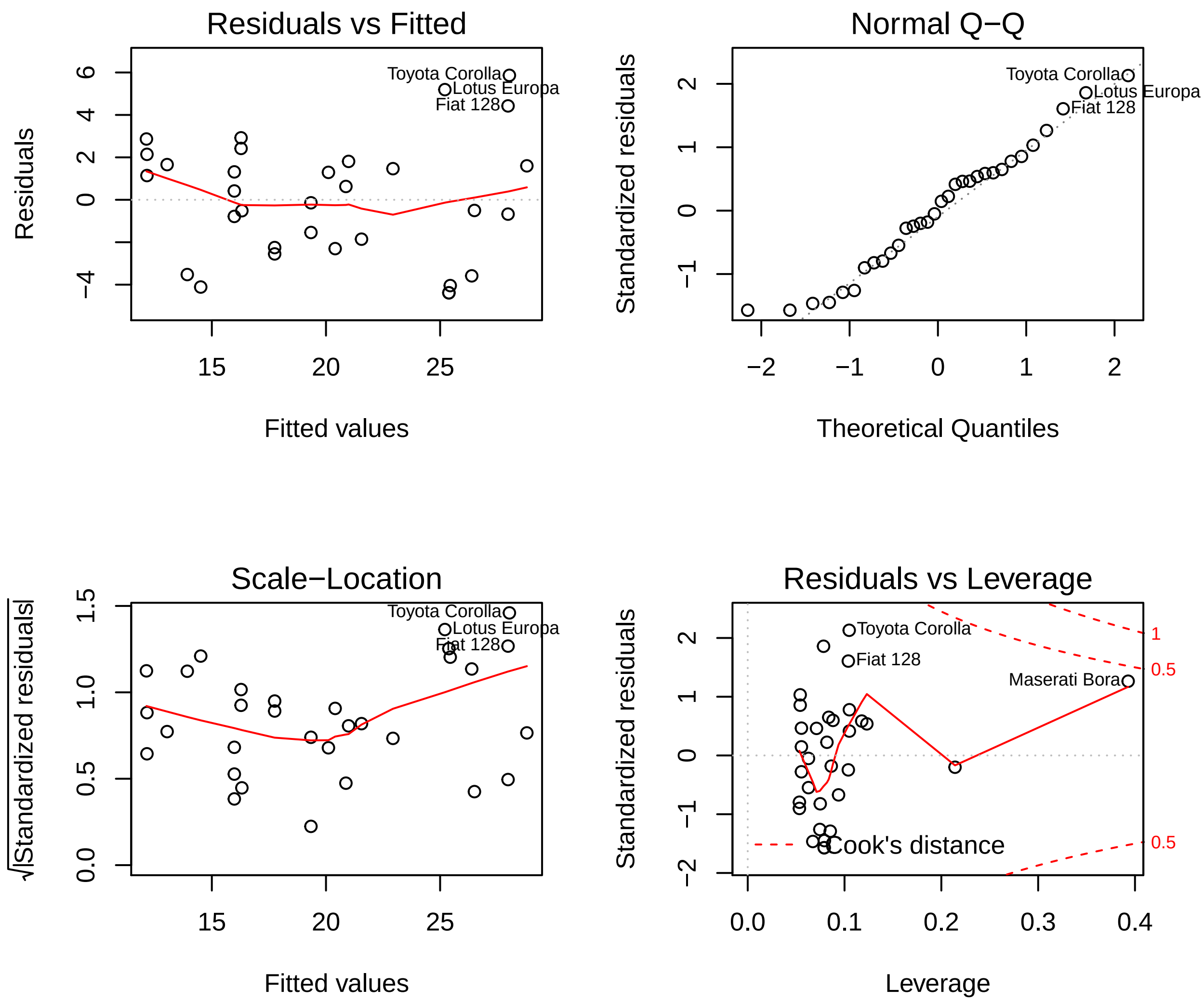

Figure 3. Model 2 plot of residual vs fitted.

```
##
## Call:
## lm(formula = mpg ~ as.factor(am) + hp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3843 -2.2642  0.1366  1.6968  5.8657
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    26.584914   1.425094  18.655  < 2e-16 ***
## as.factor(am)1  5.277085   1.079541   4.888 3.46e-05 ***
## hp             -0.058888   0.007857  -7.495 2.92e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2.909 on 29 degrees of freedom
## Multiple R-squared:  0.782,  Adjusted R-squared:  0.767
## F-statistic: 52.02 on 2 and 29 DF,  p-value: 2.55e-10
```

Figure 4. Model 2 summary.