

Motor Trend Car Road Tests. Regression Models

Course Project

Michael Berger

21 January 2018

Abstract

The purpose of this study is to evaluate whether gear type of an automobile is a significant predictor for a vehicle efficiency expressed in miles per gallon. Moreover it is important to check whether other parameters appear to be significant predictors. We will perform the study on a standard mtcars dataset included in R: The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models). R help

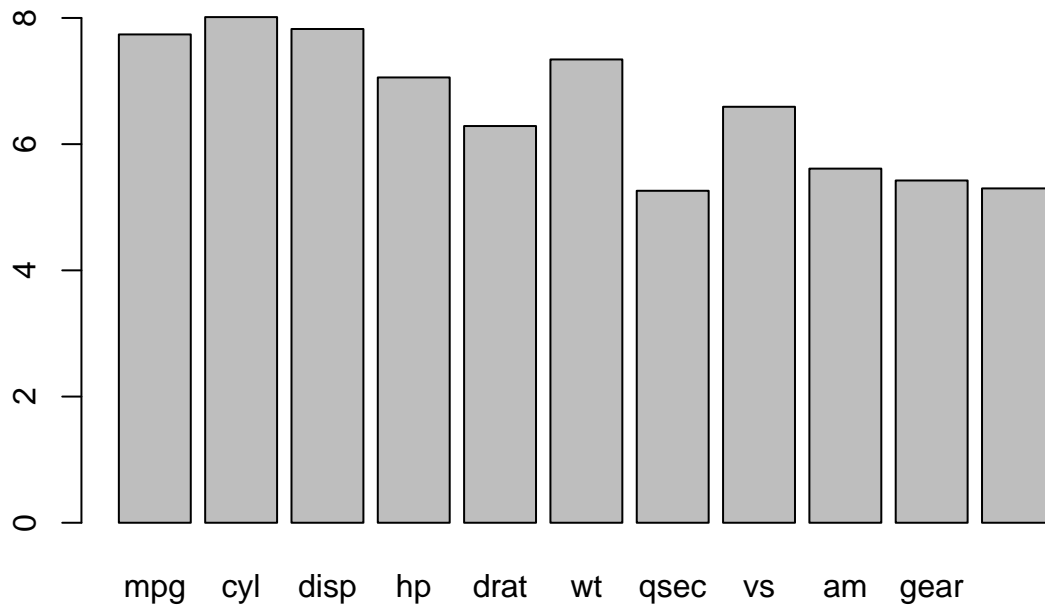
Exploratory analysis

For the sake of visual simplicity We will rearrange the dataset such that each variable will be ordered according with total amount of correlation with other variables. IN this way a more visually clear correlation plot can be built:

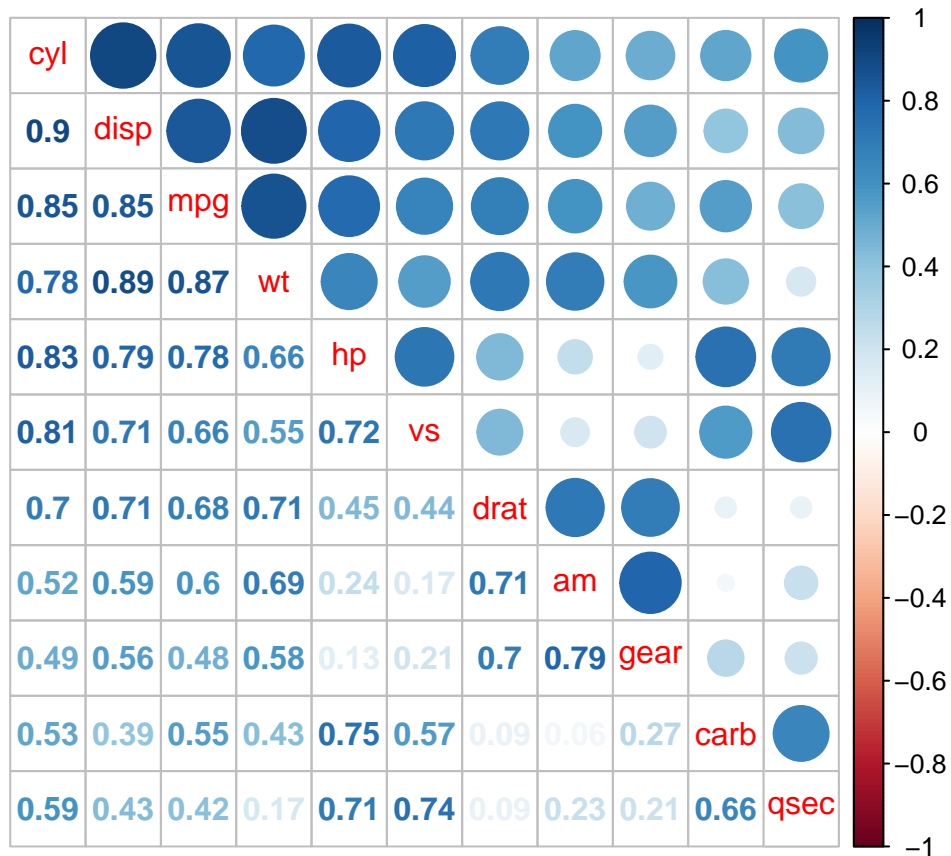
```
data("mtcars")
head(mtcars)
```

##	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
## Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
## Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
## Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
## Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
## Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
## Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3	1

```
co <-abs(cor(mtcars))
barplot(colSums(co))
```



```
mc <- mtcars[,order(colSums(co),decreasing = TRUE)]
#barplot(colSums(abs(cor(mc)))) # just check that the parameters really got sorted by total amount of c
co <-abs(cor(mc))
corrplot.mixed(co, upper = "circle")
```

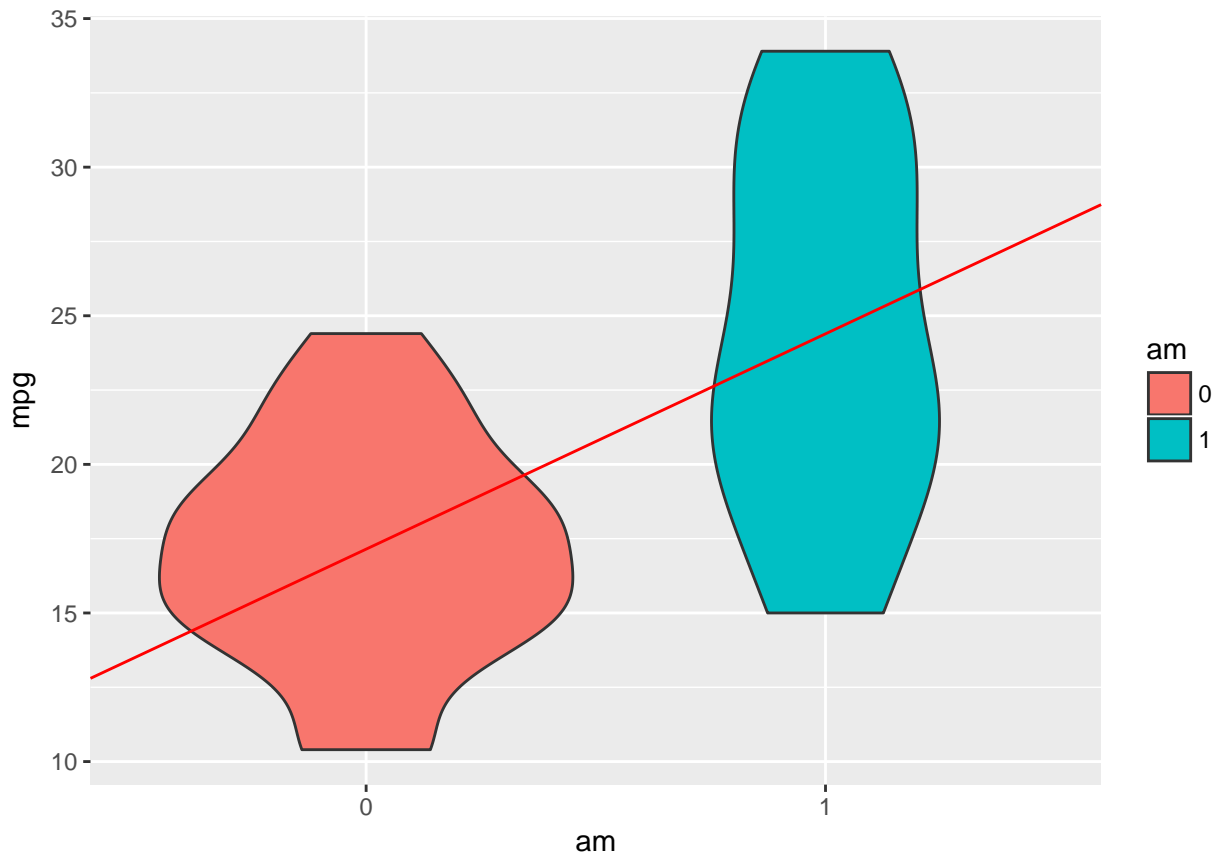


From this plot one can obtain a general understanding of which variables correlate with the most other variables.

Simple linear modeling

Since the outcome, mpg, is an interval variable, it is reasonable to use simple linear model for the whole study. The first obvious step is to create a simple model with one predictor, am

```
##  
## Call:  
## lm(formula = mpg ~ am, data = mtcars)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max   
## -9.3923 -3.0923 -0.2974  3.2439  9.5077   
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)      
## (Intercept)   17.147      1.125   15.247 1.13e-15 ***  
## am1           7.245      1.764    4.106 0.000285 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 4.902 on 30 degrees of freedom  
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385   
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```



The line looks fine and the p-values are highly significant. However, Adjusted R-squared points out that this model accounts for only 34% of variance in the data. Hence, we should expand the model to include more variables. Not all the available variables are necessary in the model, and we know that excess variables cause variance inflation and overfitting. As can be seen below, even though the Adjusted R-squared is rather high,

none of the ceafficent are even remotely significant.

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.5087 -1.3584 -0.0948  0.7745  4.6251
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  23.87913    20.06582   1.190  0.2525
## cyl6         -2.64870     3.04089  -0.871  0.3975
## cyl8         -0.33616     7.15954  -0.047  0.9632
## disp         0.03555     0.03190   1.114  0.2827
## hp          -0.07051     0.03943  -1.788  0.0939 .
## drat         1.18283     2.48348   0.476  0.6407
## wt          -4.52978     2.53875  -1.784  0.0946 .
## qsec         0.36784     0.93540   0.393  0.6997
## vs1          1.93085     2.87126   0.672  0.5115
## am1          1.21212     3.21355   0.377  0.7113
## gear4        1.11435     3.79952   0.293  0.7733
## gear5        2.52840     3.73636   0.677  0.5089
## carb2       -0.97935     2.31797  -0.423  0.6787
## carb3        2.99964     4.29355   0.699  0.4955
## carb4        1.09142     4.44962   0.245  0.8096
## carb6        4.47757     6.38406   0.701  0.4938
## carb8        7.25041     8.36057   0.867  0.3995
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.833 on 15 degrees of freedom
## Multiple R-squared:  0.8931, Adjusted R-squared:  0.779
## F-statistic:  7.83 on 16 and 15 DF,  p-value: 0.000124
```

We will use stepwise model selection procedure to sort out variables which do not contribute to the interpretability.

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.70832    2.60489  12.940 7.73e-13 ***
## cyl6         -3.03134     1.40728  -2.154  0.04068 *
## cyl8         -2.16368     2.28425  -0.947  0.35225
## hp          -0.03211     0.01369  -2.345  0.02693 *
## wt          -2.49683     0.88559  -2.819  0.00908 **
## am1          1.80921     1.39630   1.296  0.20646
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10

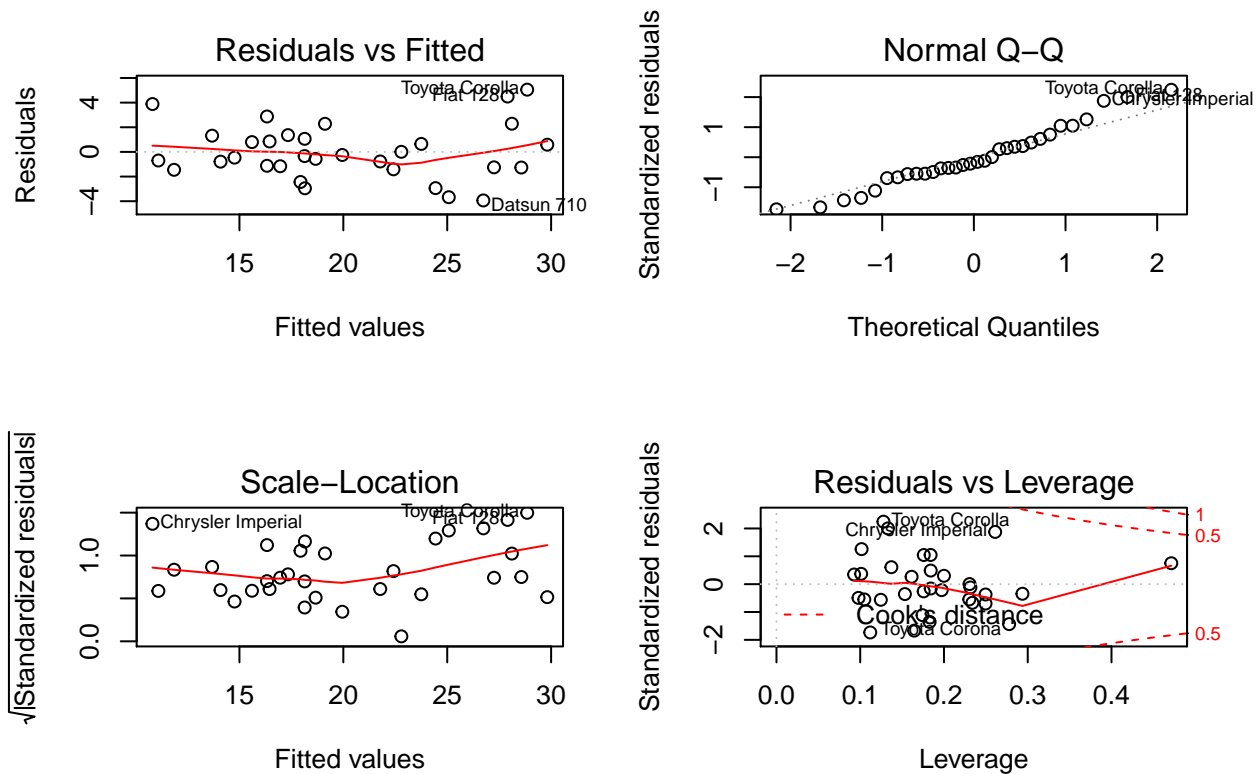
##          GVIF Df GVIF^(1/(2*Df))
## cyl 5.824545  2      1.553515
## hp  4.703625  1      2.168784
## wt  4.007113  1      2.001778
## am  2.590777  1      1.609589
```

Stepwise Algorithm yeilds a model with 4 variables: cyl, hp, wt, am. Adjusted R-squared of 84% is significantly larger than that of the simplest model (with am as only predictor). This points out that 84 of variability in the dat can be explained by the improved model. Variance inflation factor analysis does not reveal abnormally large (more than 10) values, which implies that model does not contain variables with high corelation with each other.

p-values for coefficents are significant except for the cyl8 and am1. The first one implies that there is no significant difference between 6 and 8 cylinder engine cars. The second one says there is no difference between gear types.

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + hp + wt + am
## Model 3: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1      30 720.90
## 2      26 151.03  4    569.87 17.7489 1.476e-05 ***
## 3      15 120.40 11     30.62  0.3468  0.9588
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Anova points out that the mdlstep's variables are a significant (p-value less than 0.001) improvement over the simplest model, while full model does not give significant improvement over the mdlstep



Residuals plots do not express any non-random pattern which implies that what little variance is left in the model is normally random and unpredictable.

Standardised residuals are situated close to the normal line on Q-Q plot, another indication that model leftover variance is random normal.

Cook'd distance plot does not reveal any extreme data that may influence the model in a game-changing way.

Conclusion

In this study we have proposed a model which tackles an amount of influence which different cars parameters exert on fuel consumption. A stepwise selection algorithm gave us a model with 4 main influencing parameters: number of cylinders (cyl), horsepower (hp), weight (wt) and gearbox type (am). While all the parameters are significant, the gearbox type is not. Hence we can conclude that the answer to the study's question, whether am is a significant predictor for the mpg, is negative. As for the assignment questions, It is not reasonable to answer them since am significance level is not sufficient to be able to answer those questions. Nevertheless, suppose the coefficient was sufficiently significant, what would have been the answers to the questions? 1. "Is an automatic or manual transmission better for MPG" am1 coefficient is positive, meaning that the manual transmission vehicles travel more miles on one gallon of fuel, so the manual transmission is better than automatic. 2. "Quantify the MPG difference between automatic and manual transmissions" am1 coefficient is 1.8, so the manual transmission vehicles travel 1.8 miles more than automatic transmission vehicles on one gallon of fuel.

Appendix

=====