Regression Models Assignment: Motor Trend Case Study

Francisco Marco-Serrano

Executive Summary

Eleven variables are analised for thirty-two cars in search of differences in consumption for cars with automatic and manual transmission. A difference of 2.94 miles per gallon is found in manual transmission car, a difference significant at a 5% level (95% confidence).

Dataset and Exploratory Analysis

The analysed dataset is the `mtcars` data from the `datasets` package. It contains information on 11 variables for 32 observations 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). See variables description in the Appendix.

Analysis

Different nested models are tested starting with all the variables as confounders. Significance tests and comparison of adjusted coeficient of determination are used to assess the suitability of the model. See Appendix for detailed results.

Finally, to the best fitted model is shown below.

```
##
## Call:
## lm(formula = mtcars$mpg ~ mtcars$wt + mtcars$qsec + factor(mtcars$am),
##     data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           9.6178     6.9596   1.382 0.177915
## mtcars$wt            -3.9165     0.7112  -5.507 6.95e-06 ***
## mtcars$qsec           1.2259     0.2887   4.247 0.000216 ***
## factor(mtcars$am)1    2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

The variable `am` is significant at any 5% significance level (p-value = 0.046716, which suggests that at 1% we would not be rejecting the null hypothesis). As the variable is defined for taking unitary values for manual transmision, the positive coefficient points out towards a higher `mpg` for the manual transmision cars versus the consumption of the automatic transmision ones. The mean difference, after accounting for the weight and mile time effects, is of 2.9358 miles/(US) gallon.

All the diagnostics pointed out to the validity of the model (see Appendix for details on normality, homocedasticity, and multicollinearity).

Appendix

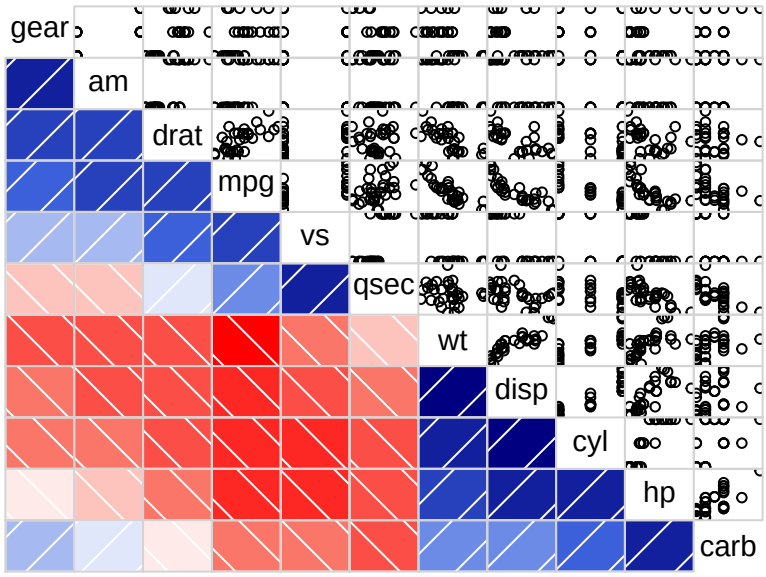
Variables

Variable	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
am	Transmission (0 = automatic, 1 = manual)
gear	Number of forward gears
carb	Number of carburettors

Correlation Structure

The variables are correlated in the following manner:

```
## Warning: package 'corrgram' was built under R version 3.1.3
```



As we can see, all five numerical variables seem to have either an indirect relationship with the miles per gallon (i.e. disp, hp, wt) or a direct one (i.e. drat, qsec).

Model Selection

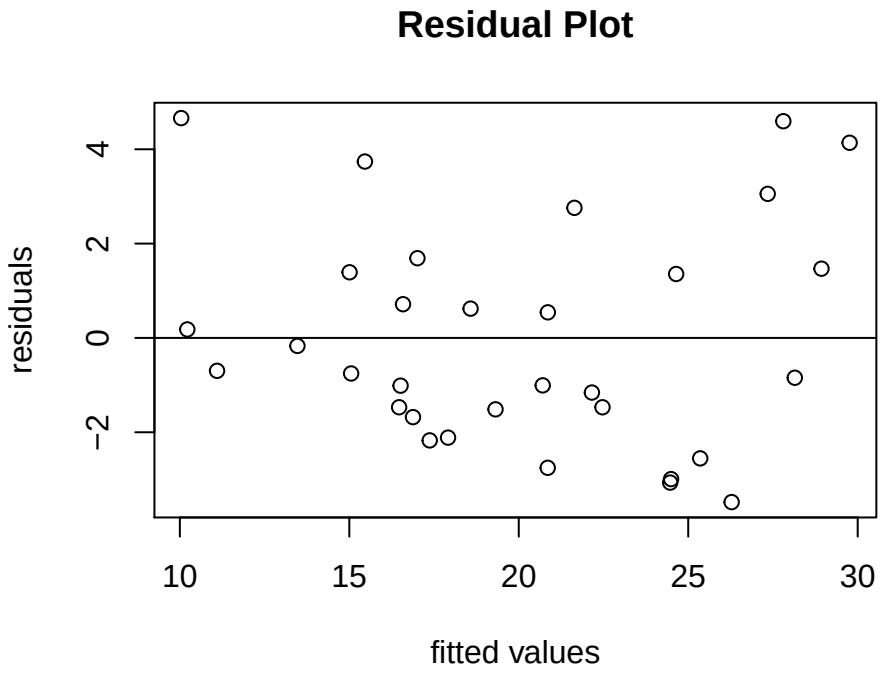
```
##
## Call:
## lm(formula = x$mpg ~ ., data = x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5399 -1.7398 -0.3196  1.1676  4.5534
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.36190    9.74079   1.474  0.15238
## disp         0.01124    0.01060   1.060  0.29897
## hp          -0.02117    0.01450  -1.460  0.15639
## wt          -4.08433    1.19410  -3.420  0.00208 **
## qsec         1.00690    0.47543   2.118  0.04391 *
## am           3.47045    1.48578   2.336  0.02749 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.429 on 26 degrees of freedom
## Multiple R-squared:  0.8637, Adjusted R-squared:  0.8375
## F-statistic: 32.96 on 5 and 26 DF,  p-value: 1.844e-10

##
```

```
## Call:
## lm(formula = x$mpg ~ ., data = x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## am            2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

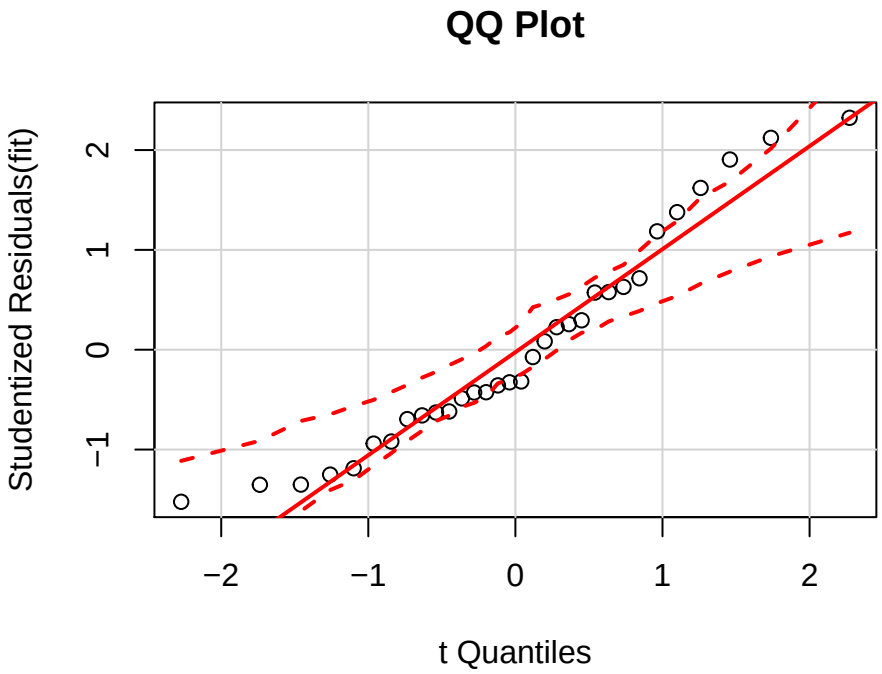
Selected model is `fit3`, which explains mpg by wt, qsec and am. All three variables significant at 5% level, and an adjusted coeficient of determination of 83.36% (unadjusted is 84.97%).

Diagnostics



Diagnose normality by using the QQ plot: the diagram suggests the error term is normally distributed.

```
## Loading required package: car
```



Diagnose homocedasticity by using the Breusch-Pagan test: the p-value is 0.1029 so if we consider a significance level of 5%, we could not reject the hypothesis of homocedasticity (at a 10%, we would be in a border line situation).

```
## Loading required package: lmtest

## Warning: package 'lmtest' was built under R version 3.1.3

## Loading required package: zoo
##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

##
##  studentized Breusch-Pagan test
##
## data:  fit
## BP = 6.1871, df = 3, p-value = 0.1029
```

Diagnose multicollinearity using the Variance Inflation Factors (VIF): none of the values exceeds the rule of thumb of 2. So, the presence of multicollinearity is not considered.

```
##         mtcars$wt       mtcars$qsec factor(mtcars$am)
##          1.575738          1.168049          1.594189
```