Consumption vs Transmission (mtcars)

geotsa

October 4, 2019

Introduction

For the needs of Motor Trend (magazine about the automobile industry) we're looking at a data set of a collection of cars [mtcars: (Package datasets version 3.6.1 Index)], in order to explore the relationship between a set of variables and miles per gallon (MPG) (outcome). We are particularly interested in the following two questions:

- Is an automatic or manual transmission better for MPG?
- Quantify the MPG difference between automatic and manual transmissions?

Executive Summary

If we consider the transmission type as the only regressor of mpg, then the difference between AT and MT is 7.24 mpg in favour of the latter. Having fitted the simplest linear model, this difference is in reality the average mpg difference for the two cases.

However, this model has major weaknesses. In addition to explaining only one third of the variation, it is characterized by extremely high bias since, except for consumption and speed type, significant correlation exists between all the pairs of variables parameters, and of course both mpg and am, including of course both mpg and am. The mpg regression against am thus conceals and conceals all other significant correlations. Thus, the regression of mpg against am covers/hides all these other significant correlations.

Among all these, most important ultimately emerge the ones between the fuel consumption and the car's weight and acceleration (secondarily between the fuel consumption and car's displacement and horsepower). But including also these regressors, the conclusions about fuel consumption and transmission mode are changing. In the case of weight alone, the transmission mode does not seem to differentiate fuel consumption. The same happens in the case of the prediction model that includs all other variables as regressors of mpg. Finally in the model that we include as regressor the acceleration -in addition to am and weight- we observe that manual transmission results in (for constant weight and acceleration) a statistically significant reduction of 2.5 mpg (increase in consumption). Hence we cannot conclude that an automatic transmission is better for MPG than a manual one.

Hence, we cannot conclude that an automatic transmission is better for MPG than a manual one.

Loading and preprocessing the data and the necessary libraries

```
# Loading necessary libraries
library(dplyr); library(car); library(ggplot2); library(knitr); library(kableExtra)
```

```
# Loading data
data("mtcars")
# Display the structure of the data (data type, variables classes)
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

The object mtcars is already a data frame of 32 observations and 11 variables. We factorize the transmission (am) variable changing also its levels from 0/1 to AM/MT

```
mtcars$am <- as.factor(mtcars$am)
levels(mtcars$am)=c("AT","MT")
```

Exploratory Data Analysis

Display how many of the obs are AM and MT

```
table(mtcars$am)
```

```
##
## AT MT
## 19 13
```

We're using dplyr piping, group_by and summarize to calculate the mean consumption for the AM and the MT cases

```
mtcars %>% group_by(am) %>% summarise(mean = mean(mpg))
```

```
## # A tibble: 2 x 2
##   am     mean
##   <fct> <dbl>
## 1 AT     17.1
## 2 MT     24.4
```

The difference in the mean mpg values for AT and for MT seems alreay important. We 're going a little beyond to calculate the 6 number summaries of the mpg values for AT and for MT

```
summary(mtcars[which(mtcars$am=="AT"),]$mpg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   14.95   17.30   17.15   19.20   24.40
```

```
summary(mtcars[which(mtcars$am=="MT"),]$mpg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   21.00   22.80   24.39   30.40   33.90
```

The new findings also indicate a strong relationhip between the type of transmission (am) and the fuel consumtion (mpg) [A boxplot of these summaries can be found in Appendix A].

```
t.test(mtcars[mtcars$am=="AT",]$mpg, mtcars[mtcars$am=="MT",]$mpg)
```

```
##
##  Welch Two Sample t-test
##
## data:  mtcars[mtcars$am == "AT", ]$mpg and mtcars[mtcars$am == "MT", ]$mpg
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean of x mean of y
##  17.14737  24.39231
```

So, for a confidence interval of 95% (a=0.05), the p-value shows that indeed the mpg means (for AT and MT) cannot be equal.

But before we draw conclusions, we should check the correlation of all the variables to each other, and especially against mpg. Hence, if we see the pairs of plots of Appendix B and specially its first row, we remark a clear correlation, stronger or milder, negative or positive, between everey pair of our variables (and specially between mpg and the other variables).

Regression analysis will help us to clarify and quantify the relation between fuel consumption and transmission type.

Regression Analysis

mpg against am

First we are fitting a model mpg~am

```
fit1 <- lm(mpg~am, mtcars)
coef(fit1)
```

```
## (Intercept)        amMT
##   17.147368    7.244939
```

The model fit1 indicates that (for all the other variables fixed) the MT provokes a 7.24 mpg increase comparing to the AT.

mpg against all the variables

But the important correlations of the pairs of all the other variables leads us to examine the coefficients of a model with all of them as regressors

```
fit10 <- lm(mpg~., mtcars)
coef(fit10)
```

```
## (Intercept)         cyl        disp          hp        drat          wt
## 12.30337416 -0.11144048  0.01333524 -0.02148212  0.78711097 -3.71530393
##        qsec          vs        amMT        gear        carb
##  0.82104075  0.31776281  2.52022689  0.65541302 -0.19941925
```

This time, the model fit10 indicates that (for all the other variables fixed) the MT provokes a 2.52 mpg increase comparing to the AT. From the same coefficients (and from the corresponding variables' means: Appendix C) we guess that except transimission type it would be the displacement (disp), the rear axle ratio (drat), the horsepower (hp), the weight (wt), number of f/w gears (gear) and the acceleration (qsec)

Analysis of Variance

In order to simplify our model and keep the least possible regressors, we are proceeding to a nested analysis of variance (ANOVA). In other words we are analysing if there is a (statistically) significant difference in the mpg mean between each model and its previous one.

```
# We fit the rest 8 models
fit2 <- lm(mpg~am+disp, mtcars)
fit3 <- lm(mpg~am+disp+hp, mtcars)
fit4 <- lm(mpg~am+disp+hp+drat, mtcars)
fit5 <- lm(mpg~am+disp+hp+drat+wt, mtcars)
fit6 <- lm(mpg~am+disp+hp+drat+wt+qsec, mtcars)
fit7 <- lm(mpg~am+disp+hp+drat+wt+qsec+vs, mtcars)
fit8 <- lm(mpg~am+disp+hp+drat+wt+qsec+vs+cyl, mtcars)
fit9 <- lm(mpg~am+disp+hp+drat+wt+qsec+vs+cyl+gear, mtcars)
summary(fit9)
```

```
##
## Call:
## lm(formula = mpg ~ am + disp + hp + drat + wt + qsec + vs + cyl +
##     gear, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3038 -1.6964 -0.1796  1.1802  4.7245
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.83084   18.18671   0.706  0.48790
## amMT         2.55093    2.00826   1.270  0.21728
## disp         0.01623    0.01290   1.259  0.22137
## hp          -0.02424    0.01811  -1.339  0.19428
## drat         0.70590    1.56553   0.451  0.65647
## wt          -4.03214    1.33252  -3.026  0.00621 **
## qsec         0.86829    0.68874   1.261  0.22063
## vs           0.36470    2.05009   0.178  0.86043
## cyl         -0.16881    0.99544  -0.170  0.86689
## gear         0.50294    1.32287   0.380  0.70745
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.593 on 22 degrees of freedom
## Multiple R-squared:  0.8687, Adjusted R-squared:  0.8149
## F-statistic: 16.17 on 9 and 22 DF,  p-value: 9.244e-08
```

```
# and we proceed to the analysis of variance of the ten total nested models
anova(fit1, fit2, fit3, fit4, fit5, fit6, fit7, fit8, fit9, fit10)
```

```
## Analysis of Variance Table
##
## Model  1: mpg ~ am
## Model  2: mpg ~ am + disp
## Model  3: mpg ~ am + disp + hp
## Model  4: mpg ~ am + disp + hp + drat
## Model  5: mpg ~ am + disp + hp + drat + wt
## Model  6: mpg ~ am + disp + hp + drat + wt + qsec
## Model  7: mpg ~ am + disp + hp + drat + wt + qsec + vs
## Model  8: mpg ~ am + disp + hp + drat + wt + qsec + vs + cyl
## Model  9: mpg ~ am + disp + hp + drat + wt + qsec + vs + cyl + gear
## Model 10: 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      29 300.28  1    420.62 59.8865 1.397e-07 ***
## 3      28 226.10  1     74.18 10.5613  0.003833 **
## 4      27 221.04  1      5.06  0.7207  0.405503
## 5      26 175.67  1     45.37  6.4603  0.018983 *
## 6      25 150.09  1     25.57  3.6412  0.070130 .
## 7      24 149.45  1      0.65  0.0919  0.764754
## 8      23 148.87  1      0.57  0.0819  0.777594
## 9      22 147.90  1      0.97  0.1384  0.713653
## 10     21 147.49  1      0.41  0.0579  0.812179
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This analysis shows that the most significant regressors (except am that interests us) are the disp, the hp, the wt and the qsec (in some accordance with our previous remarks). The addition of these regressors gives statistically significant differences in the mean of the prediction, comparing to their omission.

Appendix D shows that, among them, mainly the wt (weight), but also, the qsec (accelaration) are very important to prevent the bias that comes into the model mpg~am.

Models Selection

mpg ~ am (mdl1)

```
summary(mdl1 <- lm(mpg ~ am, mtcars))
```

```
##
## 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 ***
## amMT           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
```

According to the mdl1 model, which explains 36% of the variation, in MT cars the mpg price increases by 7.245 units (consumption reduction). This increase is considered statistically significant (Pr<|t|). Looking at the residual plots (Appendix E.1.), we observe that there are few ouliers available on the dataset but nothing significantly skews the data (residuals vs. leverage). No pattern is seen across the residuals, which is good for our model.

mpg ~ am + wt (mdl2)

```
summary(mdl2 <- lm(mpg ~ am + wt, mtcars))
```

```
##
## Call:
## lm(formula = mpg ~ am + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5295 -2.3619 -0.1317  1.4025  6.8782
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37.32155    3.05464  12.218 5.84e-13 ***
## amMT        -0.02362    1.54565  -0.015    0.988
## wt          -5.35281    0.78824  -6.791 1.87e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.098 on 29 degrees of freedom
```

```
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7358
## F-statistic: 44.17 on 2 and 29 DF,  p-value: 1.579e-09
```

For a constant weight, the mdl2 model —which explains this time 75% of the variation of mpg— predicts for MT cars a slight reduction in the mpg value by 0.024 mpg (slight increase in consumption). However, this increase is considered not statistically significant (Pr>|t|). Looking at the residual plots (Appendix E.2.), we observe that there are few ouliers available on the dataset but nothing significantly skews the data (residuals vs. leverage). No pattern is seen across the residuals, which is good for our model.

mpg ~ am + wt + qsec (mdl3)

```
summary(mdl3 <- lm(mpg ~ am + wt + qsec, mtcars))
```

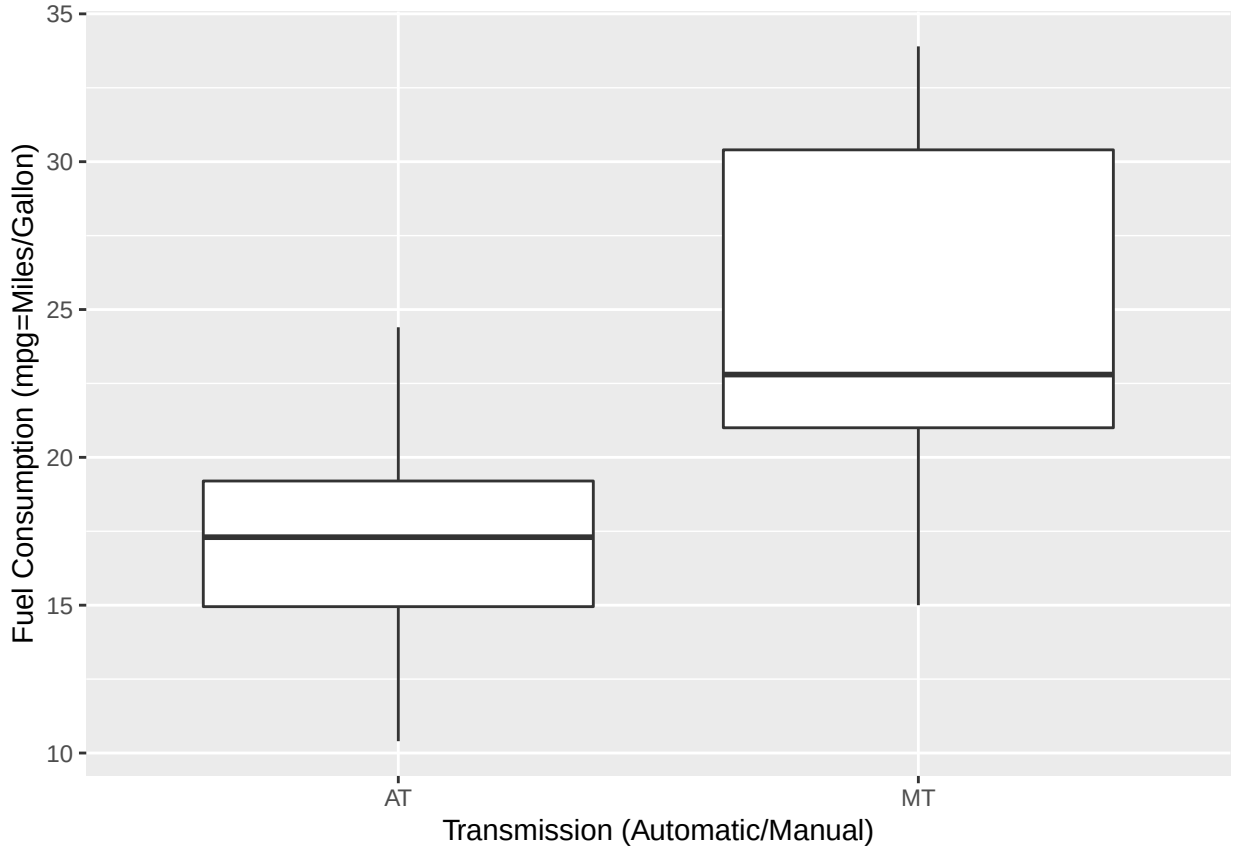
```
##
## Call:
## lm(formula = mpg ~ am + wt + qsec, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## amMT          2.9358     1.4109   2.081 0.046716 *
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## ---
## 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
```

Finally, the mdl3 model, that explains 85% of the variation of mpg, keeping weight and acceleration constant, indicates that MT cars have a mpg value decrease of 2.936 mpg (increase in consumption) in comparison with the AT cars. This increase is considered statistically significant (Pr<|t|). Looking at the residual plots (Appendix E.3.), we observe that there are few ouliers available on the dataset but nothing significantly skews the data (residuals vs. leverage). No pattern is seen across the residuals, which is good for our model.

Appendix

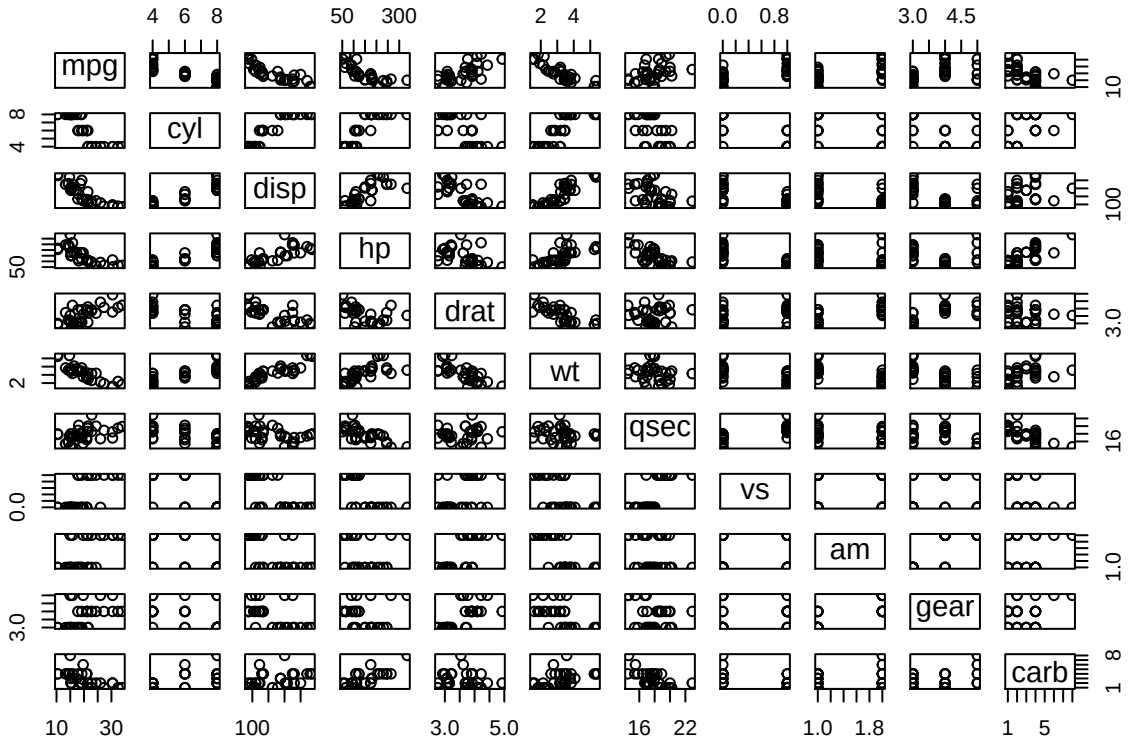
A. MPG: AT vs MT

```
ggplot(mtcars, aes(am,mpg)) + geom_boxplot() + labs(x= "Transmission (Automatic/Manual)",y = "Fuel Consu
```



B. CORRELATIONS

```
corr <- select(mtcars, mpg,cyl,disp,hp, drat,wt,qsec,vs,am,gear,carb)
pairs(corr)
```



C. COEFFICIENTS x MEAN

```
Mean <- sapply(select(mtcars,-c(mpg,am)),mean)
Coeff <- c(coef(fit10)[2:8],coef(fit10)[10:11])
.x. <- Mean*Coeff
round(rbind(Mean, Coeff, .x.),3)
```

```
##          cyl    disp      hp  drat      wt   qsec    vs  gear   carb
## Mean   6.188 230.722 146.688 3.597   3.217 17.849 0.438 3.688  2.812
## Coeff -0.111   0.013  -0.021 0.787  -3.715  0.821 0.318 0.655 -0.199
## .x.   -0.690   3.077  -3.151 2.831 -11.953 14.655 0.139 2.417 -0.561
```

D. ANOVA - Comparison of all the possible nested models - Average Pr(>F)

We are creating nested models so as each one of the four regressors to be added/included in all the four possible steps and we take the average of their Pr.

```
anova(fit1, fit2, fit3, fit5, fit6)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + disp
## Model 3: mpg ~ am + disp + hp
```

```
## Model 4: mpg ~ am + disp + hp + drat + wt
## Model 5: mpg ~ am + disp + hp + drat + wt + qsec
##   Res.Df    RSS Df Sum of Sq       F   Pr(>F)
## 1     30 720.90
## 2     29 300.28  1    420.62 70.0590 1.02e-08 ***
## 3     28 226.10  1     74.18 12.3553  0.00170 **
## 4     26 175.67  2     50.44  4.2004  0.02675 *
## 5     25 150.09  1     25.57  4.2598  0.04955 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fi2 <- lm(mpg~am+hp, mtcars)
fi3 <- lm(mpg~am+hp+wt, mtcars)
fi5 <- lm(mpg~am+hp+wt+qsec, mtcars)
fi6 <- lm(mpg~am+hp+wt+qsec+disp, mtcars)
anova(fit1, fi2, fi3, fi5, fi6)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + hp
## Model 3: mpg ~ am + hp + wt
## Model 4: mpg ~ am + hp + wt + qsec
## Model 5: mpg ~ am + hp + wt + qsec + disp
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 245.44  1    475.46 80.5661 1.913e-09 ***
## 3     28 180.29  1     65.15 11.0393  0.002654 **
## 4     27 160.07  1     20.22  3.4271  0.075528 .
## 5     26 153.44  1      6.63  1.1232  0.298972
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fii2 <- lm(mpg~am+wt, mtcars)
fii3 <- lm(mpg~am+wt+qsec, mtcars)
fii5 <- lm(mpg~am+wt+qsec+disp, mtcars)
fii6 <- lm(mpg~am+wt+qsec+disp+hp, mtcars)
anova(fit1, fii2, fii3, fii5, fii6)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt
## Model 3: mpg ~ am + wt + qsec
## Model 4: mpg ~ am + wt + qsec + disp
## Model 5: mpg ~ am + wt + qsec + disp + hp
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 74.9946 3.877e-09 ***
## 3     28 169.29  1    109.03 18.4757  0.000214 ***
## 4     27 166.01  1      3.28  0.5551  0.462912
## 5     26 153.44  1     12.57  2.1303  0.156387
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fiii2 <- lm(mpg~am+qsec, mtcars)
fiii3 <- lm(mpg~am+qsec+disp, mtcars)
fiii5 <- lm(mpg~am+qsec+disp+hp, mtcars)
fiii6 <- lm(mpg~am+qsec+disp+hp+wt, mtcars)
anova(fit1, fiii2, fiii3, fiii5, fiii6)
```

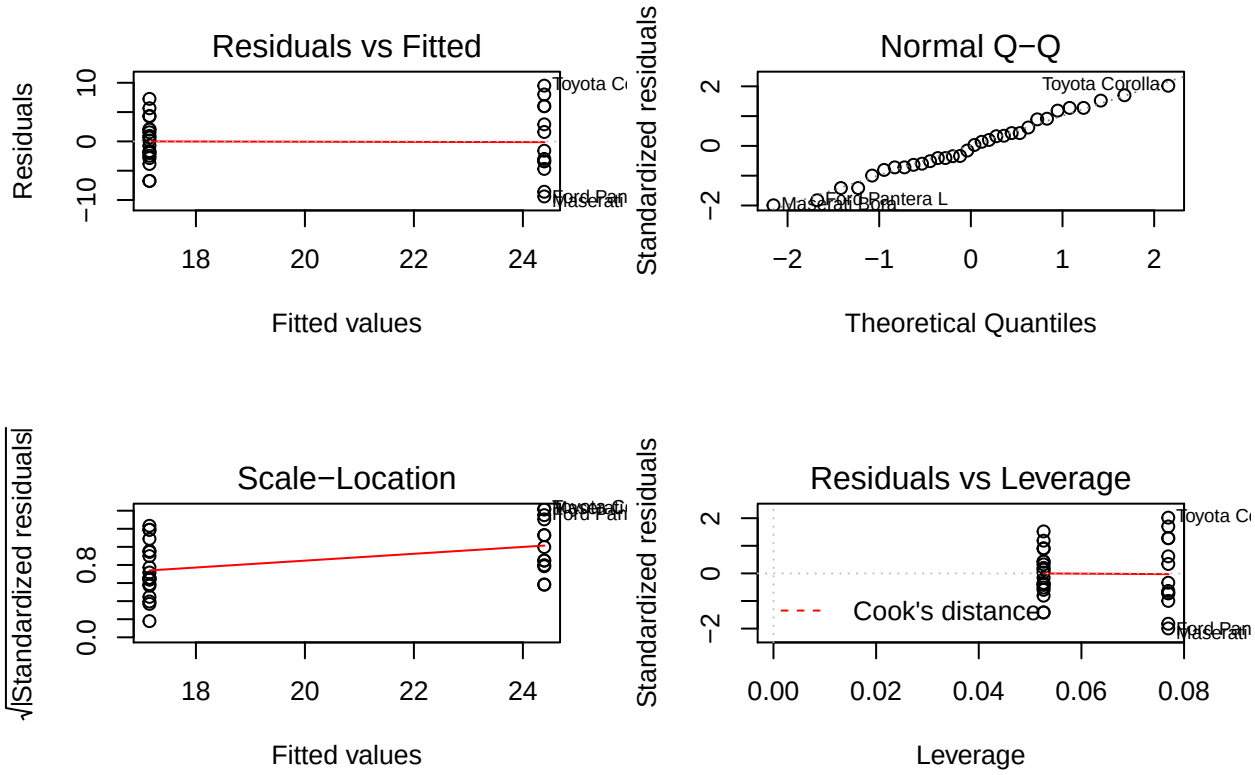
```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + qsec
## Model 3: mpg ~ am + qsec + disp
## Model 4: mpg ~ am + qsec + disp + hp
## Model 5: mpg ~ am + qsec + disp + hp + wt
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 352.63  1    368.26 62.4021  2.24e-08 ***
## 3     28 261.09  1     91.54 15.5117 0.0005487 ***
## 4     27 222.48  1     38.61  6.5426 0.0167086 *
## 5     26 153.44  1     69.04 11.6993 0.0020750 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
disp <- round((anova(fit1, fit2, fit3, fit5, fit6)$Pr[2]+
anova(fit1, fi2, fi3, fi5, fi6)$Pr[5]+
anova(fit1, fii2, fii3, fii5, fii6)$Pr[4]+
anova(fit1, fiii2, fiii3, fiii5, fiii6)$Pr[3])/4,4)
hp <- round((anova(fit1, fit2, fit3, fit5, fit6)$Pr[3]+
anova(fit1, fi2, fi3, fi5, fi6)$Pr[2]+
anova(fit1, fii2, fii3, fii5, fii6)$Pr[5]+
anova(fit1, fiii2, fiii3, fiii5, fiii6)$Pr[4])/4,4)
wt <- round((anova(fit1, fit2, fit3, fit5, fit6)$Pr[4]+
anova(fit1, fi2, fi3, fi5, fi6)$Pr[3]+
anova(fit1, fii2, fii3, fii5, fii6)$Pr[2]+
anova(fit1, fiii2, fiii3, fiii5, fiii6)$Pr[5])/4,4)
qsec <- round((anova(fit1, fit2, fit3, fit5, fit6)$Pr[5]+
anova(fit1, fi2, fi3, fi5, fi6)$Pr[4]+
anova(fit1, fii2, fii3, fii5, fii6)$Pr[3]+
anova(fit1, fiii2, fiii3, fiii5, fiii6)$Pr[2])/4,4)
#kable_styling(kable(rbind(c("","disp","hp","wt","qsec"),c("mean(Pr(>F))",disp,hp,wt,qsec)),
#                    format = "latex"))
```

E. MODEL RESIDUALS

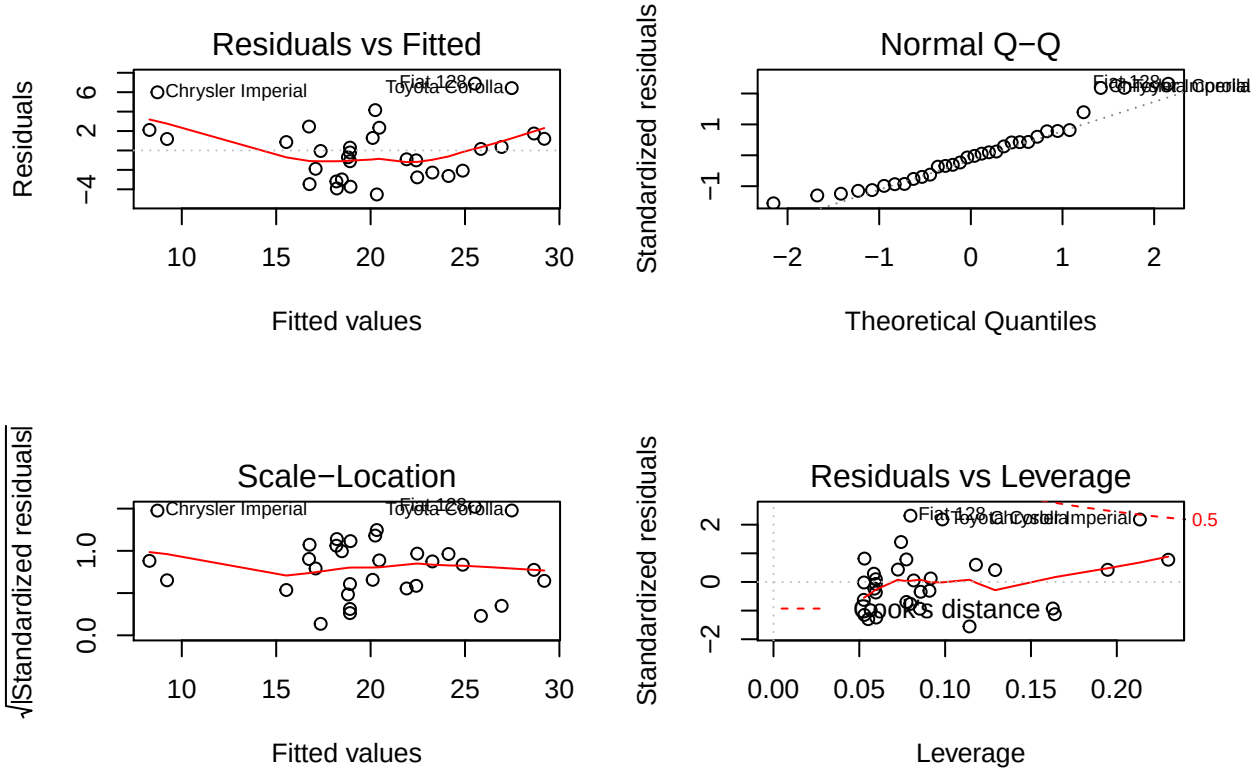
1. mpg ~ am

```
par(mfrow = c(2,2)); plot(md11)
```



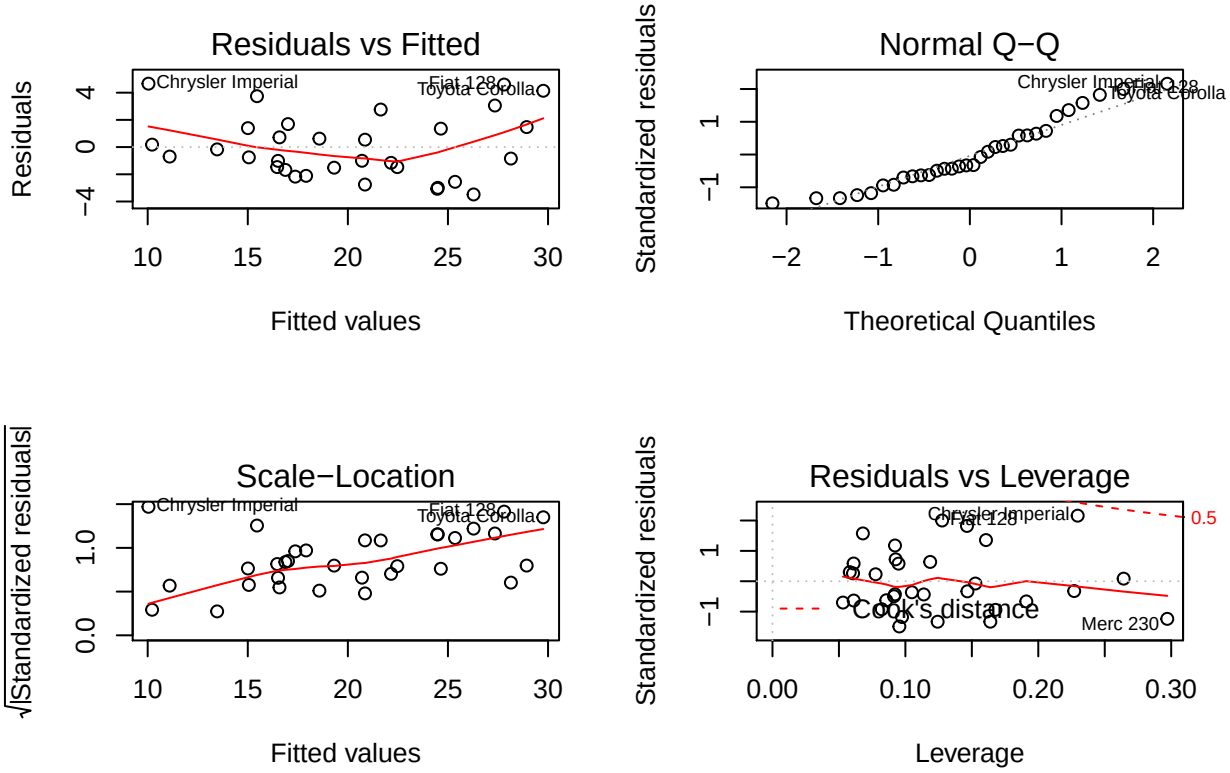
2. mpg ~ am + wt

```
par(mfrow = c(2,2)); plot(md12)
```



3. mpg ~ am + wt + qsec

```
par(mfrow = c(2,2)); plot(md13)
```


Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Residuals vs Leverage
Cook's distance
Leverage
Merc 230
0.5