Effect of type of transmission on fuel consumption

Peter Knapen

10/25/2015

Summary

This document describes the analysis of data extracted from 1974 Motor Trend US magazine, which comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-1974 models). Investigated are the variables cyl: nr. of cylinders, disp: stroke displacement in cu, hp: gross power in hp, drat: rear axle ratio, wt: weight in pounds/1000, qsec: the 1/4 mile time in seconds, vs (0=V-build/1=Straight in line engine), Transmission(0=automatic, 1=manual), gear: number of forward gears, carb: number of carburetors.
Questions to be answered are:
- Is an automatic or manual transmission better for fuel consumption.
- Quantify the difference between automatic and manual transmission in relation with fuel consumption.
Conclusions are only valid for cars mentioned in this survey, as today due to down sizing and applying other modern techniques (e.g. fuel injection, turbo charging), fuel efficiency is increased a lot, 78 mpg is possible. The cars mentioned in the survey are not a reflection of the cars which were bought at that time (number of sport/muscle cars)except for the Mazda's (Wankel engine) and the Mercedes240D (diesel engine) all models are equipped with a reciprocating petrol engine.
Based on a ggpairs plot a first choice can be made on appropriate candidates. Also the correlations between the input parameters can be evaluated.
Next an attempt is made by taking all 1 to all 10 factors in consideration and looking which factors are important, using the p-values in the anova table.
Next a physical model is derived, which did not lead to a valid model.
At last a crude modeling approach was followed by calculating all possible models with 2, 3 and 4 variables, besides am, and investigating the best model based on the standard error of the fit. For each series the model is chosen with the smallest standard error for the residuals. At last the best model is selected on the Shapiro test for the residuals.
As the derivation of the physical models takes almost 1 page, but necessary to understand the model, the requirement for 2 pages main text will not be met. It means also that most models are described, but not printed. The .Rmd file can be found on: https://github.com/pk28831/regression_model_fuelconsumption

Introduction to first models.

In the following of this text am is always taken into consideration as an input (as factor).
First the ggpairs plot is made of mtcars6, which is the same as mtcars, however the columns are reshuffled, see appendix, page 5. From this it can be seen that disp, wt and cyl with correlation coefficients of -0.848,-0.868 and -0.852 respectively , are good candidates. It is also noticed that there is a strong correlations between the inputs disp-wt (0.888), disp-cyl (0.902), hp-cyl (0.832) and cyl-vs (-0.811). From this it is also obvious that there is no strong correlation between the fuel consumption and whether the car has an automatic or manual transmission (correlation coefficient 0.6), very much overlap in the graph mpg/am.
Next the models are built with 1 to all 10 factors, using the anova table to find the important input variables.

Deriving a physical model

For a petrol engine a stoichiometric mixture means that the weight ratio air to fuel is about 15. A gallon of fuel therefore needs a certain volume of air, say b cubic inch. When nc is the number of cylinders, there are nc

ignitions per two cycles of the crank shaft. If the swept volume is $disp$ cubic inch, then per revolution of the crank shaft $disp/2$ cubic inch air is used. So in order to burn one gallon of fuel we need $b/(disp/2) = 2b/disp$ number of revolutions of the crank shaft.
When i represents the overall transmission ratio $i = gearratio * endratio$ and r represents the dynamic wheel radius in miles, then we can travel with one gallon: $2b/disp * i * r$ miles.
So the miles per gallon is proportional to $b/disp * i$.
In this formula nothing is said about acceleration or speed.
The power P is proportional to $p_i * disp$, here p_i stands for the induced pressure in the cylinder.
So mpg is proportional to $1/hp$.
Another way of writing is $P = F * v = (m * a + c_d * 1/2\rho v^2 + m * g * sin(\alpha)) * v$, in this F is the sum of the force needed for acceleration, constant driving at speed v and driving up a hill with inclination α.
So mpg is proportional to $1/wt$.
How are the other factors related to mpg?
drat is a part of i, so mpg is proportional to drat.
The other variables are linked indirectly to mpg:
- an automatic transmission has usually more losses than a manual gearbox, so in general the manual gearbox will have a better fuel consumption.
- gear: the more gears the bigger portion of time, the engine can work at its optimum, or an " overdrive" could be created, so in general the more gears, the better fuel consumption.
- qsec: this is giving a ratio between power, acceleration, weight, gear ratio and the first 3 or 4 gears. The most dominant factors are hp and wt. So the expectation is that mpg is proportional to $1/qsec$
- cyl: the number of cylinders is mostly coupled to power, so mpg is proportional to $1/cyl$
- vs: v-built is usually connected to the 8 cylinder , so mpg is proportional to $1/vs$
- carb: the higher number of carburetors is, the more fuel can be mixed with the air, so mpg is proportional to $1/carb$

A new data frame mtcars3 is generated using mutate of dplyr to generate it with the following variables: $displ = 1/disp$, $weight = 1/wt$, $qseq = 1/qsec$, $hpo = 1/hp$, $cyli = 1/cyl$, $vsi = 1/vs$ and $carbu = 1/carb$.
With this data frame a new ggpairs plot is generated, see appendix, page . From this it is shown that the most appropriate factors are : displ, weight, hpo and cyli, with correlation coefficients to mpg of 0.927, 0.892, 0,859 and 0.824 respectively. The inputs which are correlated according this plot are displ-weight (0.898), displ-cyli (0.893), displ-hpo(0.870) and hpo-cyl (0.849).

Deriving models with number crunching

mtcars6 is used for deriving models with combinations of am (as factor) in combination with 1, 2,3 and 4 variables, which leads to $\binom{n}{k}$, with n 9 and k 1, 2, 3 and 4 respectively 9, 36, 84 and 252 models. The most appropriate models are selected based on the standard error of the residuals.

Results.

```
fit0<-lm( mpg~as.factor(am),data=mtcars)
fit0$coefficients
```

```
##    (Intercept) as.factor(am)1
##      17.147368       7.244939
```

From the coefficients it is shown that the average fuel consumption for automatics is 17.1 mpg, while for manual transmissions it is the sum of the 2 factors: 24.3 mpg. This is under the assumption that other factors don't have an influence on the fuel consumption. In fact is is the mean of the fuel consumption for each factor.

```
View(mtcars)
fit1<-lm(mpg~as.factor(am),data=mtcars)
fit2<-lm(mpg~as.factor(am)+hp,data=mtcars)
fit3<-(lm(mpg~as.factor(am)+hp+gear,data=mtcars))
fit4<-lm(mpg~as.factor(am)+hp+gear+wt,data=mtcars)
fit5<-lm(mpg~as.factor(am)+hp+gear+wt+cyl,data=mtcars)
fit6<-lm(mpg~as.factor(am)+hp+gear+wt+cyl+disp,data=mtcars)
fit7<-lm(mpg~as.factor(am)+hp+gear+wt+cyl+disp+drat,data=mtcars)
fit8<-lm(mpg~as.factor(am)+hp+gear+wt+cyl+disp+drat+vs,data=mtcars)
fit9<-lm(mpg~as.factor(am)+hp+gear+wt+cyl+disp+drat+vs+carb,data=mtcars)
fit10<-lm(mpg~as.factor(am)+hp+gear+wt+cyl+disp+drat+vs+carb+qsec,data=mtcars)
anova(fit1,fit2,fit3,fit4,fit5,fit6,fit7,fit8,fit9,fit10)
```

Based on the gpairs plot we choose the following model

```
fit5a<-lm(mpg~as.factor(am)+disp+wt+cyl, data=mtcars)
fit5a$coefficients
```

```
##    (Intercept) as.factor(am)1           disp             wt            cyl
##   40.898313414    0.129065571    0.007403833   -3.583425472   -1.784173258
```

Now we can't state anymore what the average fuel consumption will be. From this model we can deduct that the manual transmission has a .12 mpg better fuel consumption, under the assumptions that disp, wt and cyl are kept constant. The intercept 40.90 mpg for the automatic transmission is for the car with 0 displacement, 0 weight and 0 cylinders. It is obvious that such an engine or car does not exist. So we don't bother about the intercept.
If we derive 10 models, each time adding a variable, we can choose the best variant based on the anova table. The largest model is printed below.

```
fit10<-lm(mpg~as.factor(am)+hp+gear+wt+cyl+disp+drat+vs+carb+qsec,data=mtcars)
anova(fit1,fit2,fit3,fit4,fit5,fit6,fit7,fit8,fit9,fit10)
```

From the anavo table it is clear that hp and wt are dominant factors, resulting in the following model:

```
fit3a<-lm(mpg~as.factor(am)+hp+wt,data=mtcars)
fit3a$coefficients
```

From this model we can deduct that the manual transmission has a 2.08 [mpg] better fuel consumption than the automatic transmission under the assumption that output and weight are kept constant.

```
require(datasets)
require(ggplot2)
require(GGally)
require(dplyr)
require(printr)
data(mtcars)
summary(mtcars)
#g=ggpairs(mtcars,lower=list(continuous="smooth"), params=c(method="loess"))
#g
gg<-ggplot(mtcars,aes(x=hp, y=mpg), colour=am)
gg<-gg+                    geom_line(aes(color="am"))+
                  geom_line(linetype=2,mapping=aes(x=wt,y=mpg,color="am"),data=mtcars)
##gg
```

When fitting 10 models based on the physical model, it is shown from the anavo table that the most important factors are drat, displ and weight.From this model we must conduct that the manual transmission has a 1.02 mpg worse fuel consumption than an automatic transmission. Or the physical model is wrong or the data has such an overlap, that this is possible when the other factors are kept constant. It is believed the latter is true.

```
mtcars2<-mutate(mtcars2,displ=1/disp,weight=1/wt,qseq=1/qsec,hpo=1/hp,cyli=1/cyl, vsi=2*(1/(vs+1)),carbu
mtcars3<-select(mtcars2,mpg,am,drat,displ,weight,qseq,hpo,cyli,vsi,gear,carbu)
fit34<- lm(mpg~as.factor(am)+drat+displ+weight+qseq+hpo+cyli+as.factor(vsi)+gear+carbu, data=mtcars3)
anova(fit25,fit26,fit27,fit28,fit29,fit30,fit31,fit32,fit33,fit34)
```

```
fit28<- lm(mpg~as.factor(am)+drat+displ+weight, data=mtcars3)
fit28$coefficients
fit28$sigma
```

After model crunching we can summarize all described models in a table:

```
stdError2<-100000
index2<-0
count2<-0
for (i in c(2:8,10:11)){
        fit2param<- lm(mtcars[,1]~as.factor(mtcars[,9])+mtcars[,i])
        stdErrorfit2<-summary(fit2param)$sigma
        print(stdErrorfit2)
        count2<-count2+1
        if (stdErrorfit2 < stdError2){
                stdError2<-stdErrorfit2
                index2<-i
        }
}
print(index2)
```

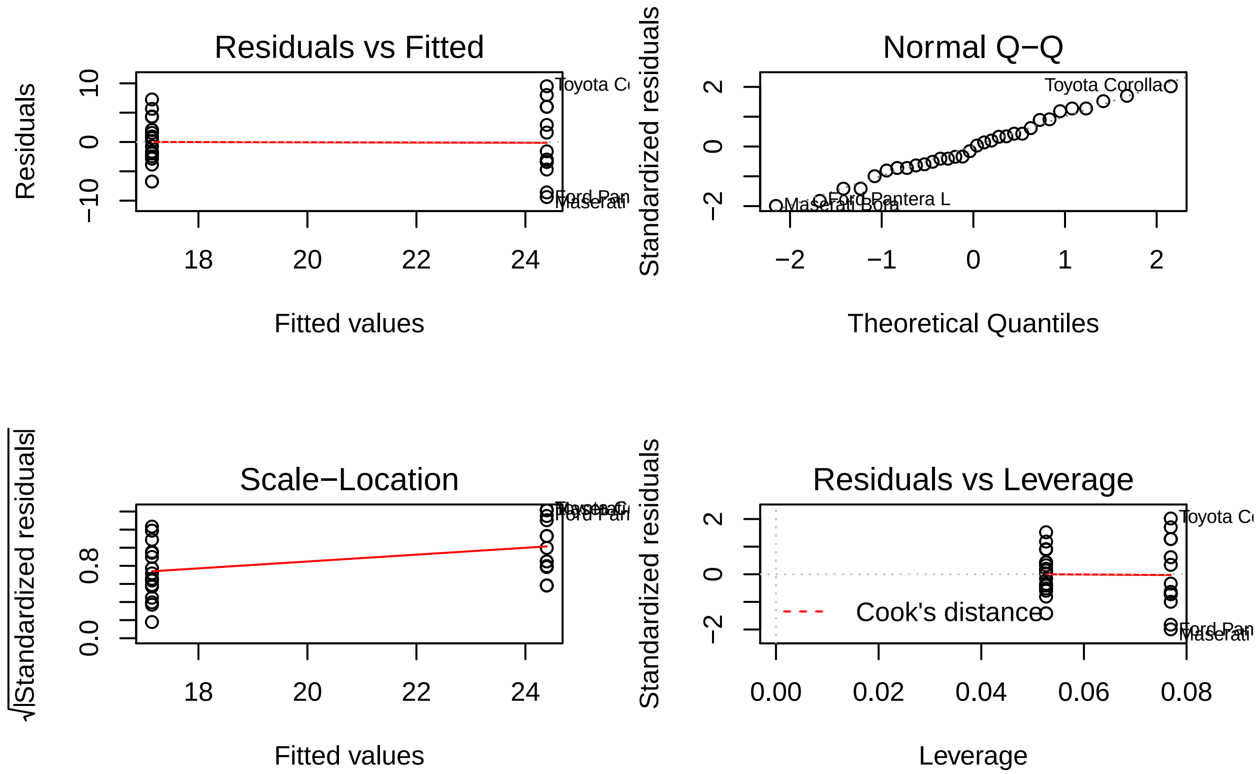
```
library(dplyr)
mtcars5<-as.tbl(mtcars)
mtcars4<-select(mtcars5,mpg,am,drat,disp,wt,qsec,hp,cyl,vs,gear,carb)
mtcars6<-data.frame(mtcars4)
```

```
index3<-c(0,0)
stdError3<-100000
count3<-0
for (i in c(3:10)){
for (j in c((i+1):11)) {
        fit3param<-lm(mtcars6[,1]~as.factor(mtcars6[,2])+mtcars6[,i]+mtcars6[,j],data=mtcars6)
        stdErrorfit3<-summary(fit3param)$sigma
        count3<-count3+1
        print(stdErrorfit3)
        if (stdErrorfit3 < stdError3){
                stdError3<-stdErrorfit3
                index3<-c(i,j)
        }
        }
}
```

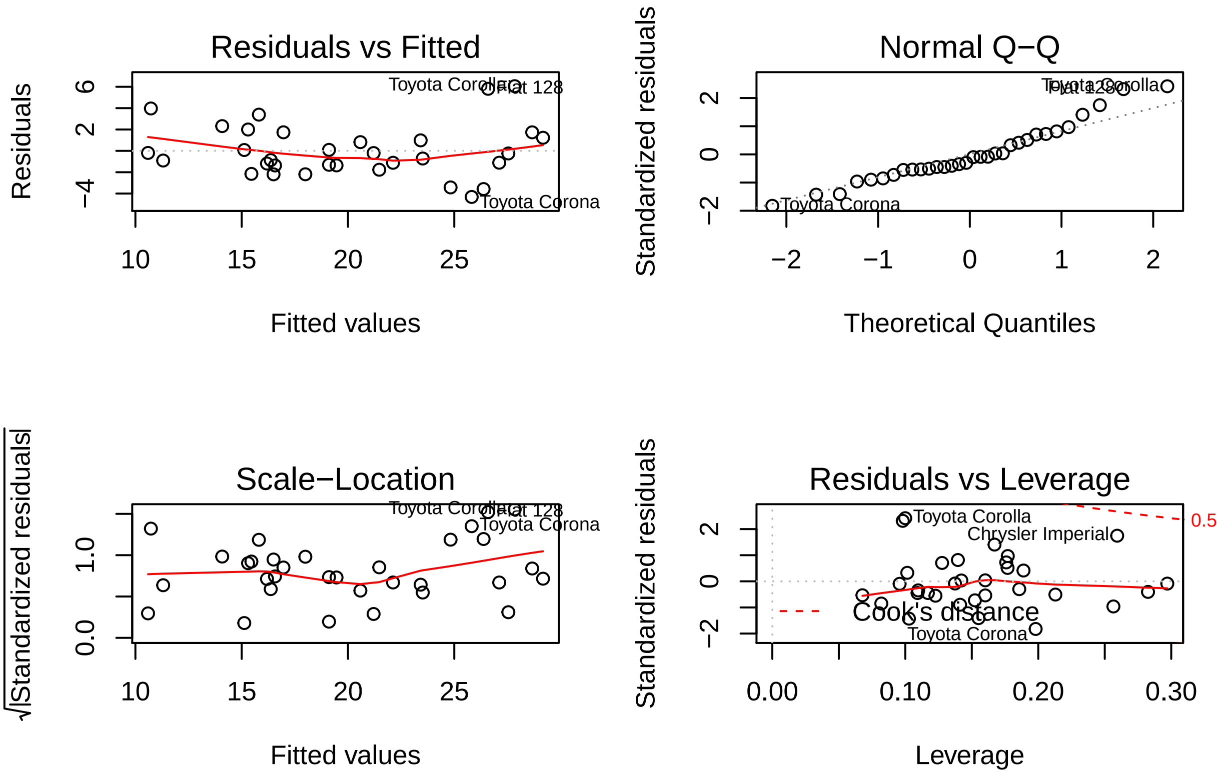
```r
index4<-c(0,0,0)
stdError4<-100000
count4<-0
for (i in c(3:9)){
        for (j in c((i+1):10)) {
                for(k in c((j+1):11)) {
                        fit4param<-lm(mtcars6[,1]~as.factor(mtcars6[,2])+mtcars6[,i]+mtcars6[,j]+mtcars6
                        stdErrorfit4<-summary(fit4param)$sigma
                        count4<-count4+1
                        print(stdErrorfit4)
                        if (stdErrorfit4 < stdError4){
                                stdError4<-stdErrorfit4
                                index4<-c(i,j,k)}}}}
```

```r
index5<-c(0,0,0,0)
stdError5<-100000
count5<-0
for (i in c(3:8)){
        for (j in c((i+1):9)) {
                for(k in c((j+1):10)) {
                        for (l in c((k+1):11)){
                               fit5param<-lm(mtcars6[,1]~as.factor(mtcars6[,2])+mtcars6[,i]+mtcars6[,j]+
                                stdErrorfit5<-summary(fit5param)$sigma
                                count5<-count5+1
                                print(stdErrorfit5)
                                if (stdErrorfit5 < stdError5){
                                stdError5<-stdErrorfit5
                                index5<-c(i,j,k,l)
                                }
                        }
                }
        }
}
```

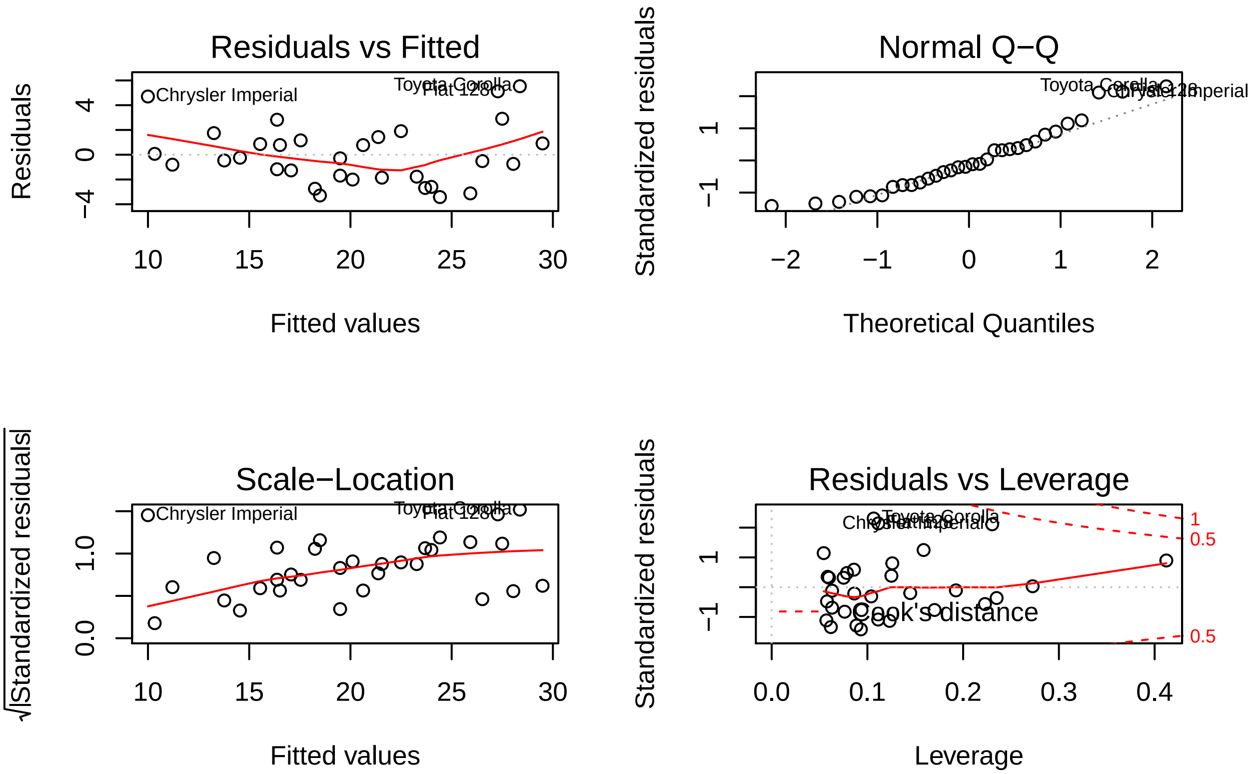
```r
# fisrt approach
# fit0<-lm( mpg~as.factor(am),data=mtcars)
fit0$coefficients
summary(fit0)$sigma
shapiro.test(fit0$residuals)
par(mfrow=c(2,2))
plot(fit0)
```



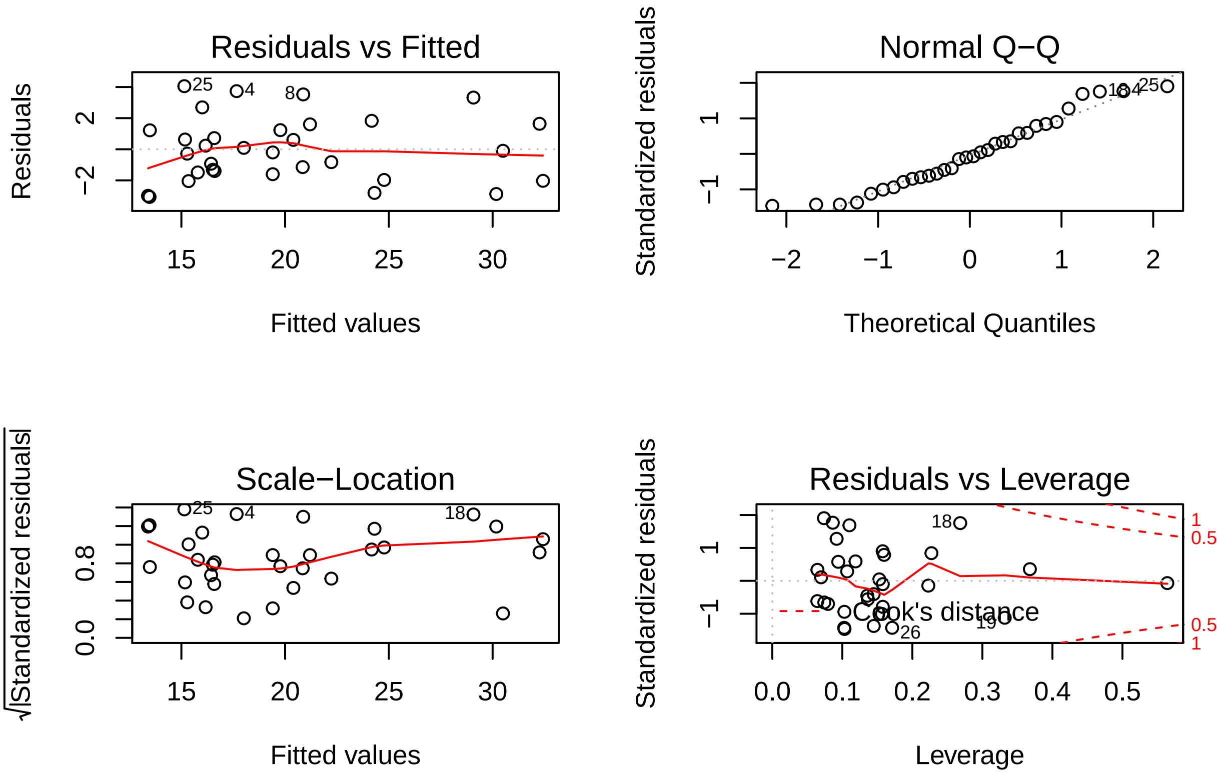

```
# based on gpairs plot of mtcars6
#fit5a<-lm(mpg~as.factor(am)+disp+wt+cyl, data=mtcars)
fit5a$coefficients
summary(fit5a)$sigma
shapiro.test(fit5a$residuals)
par(mfrow=c(2,2))
plot(fit5a)
```



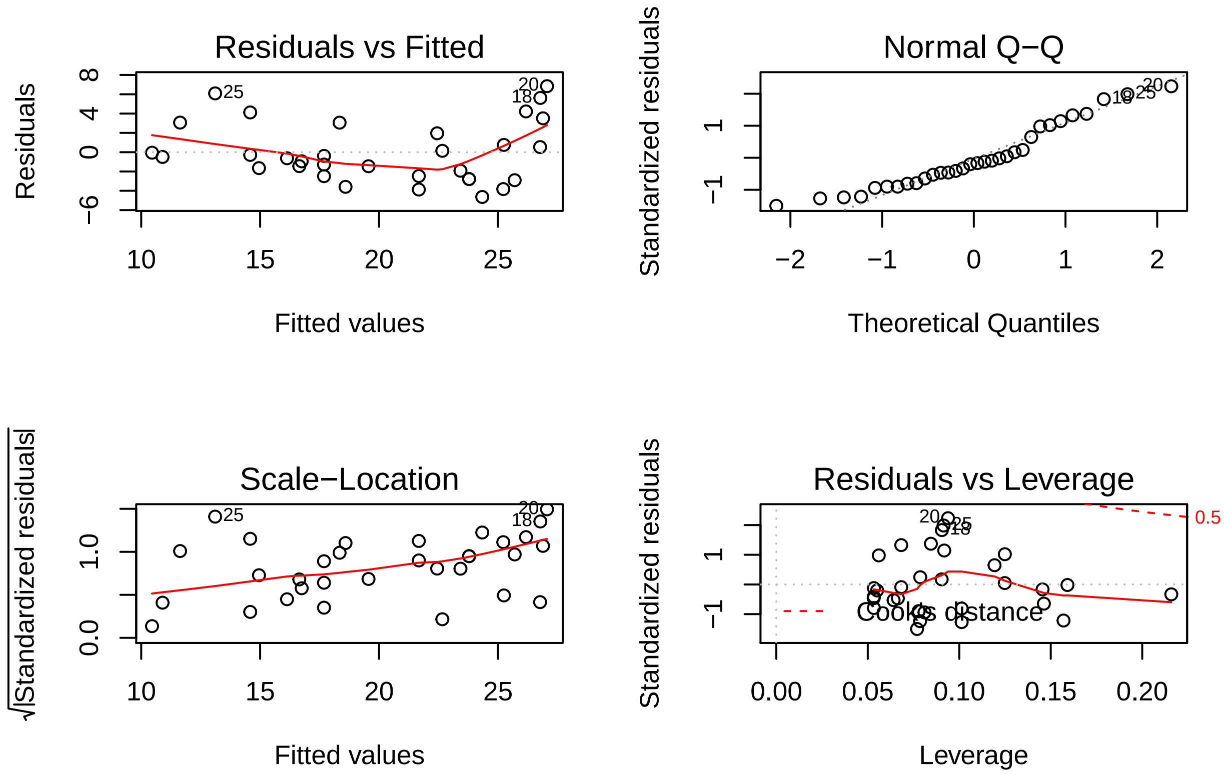

```
# based on the best of 10 models of mtcars6
#fit3a<-lm(mpg~as.factor(am)+hp+wt,data=mtcars)
fit3a$coefficients
summary(fit3a)$sigma
shapiro.test(fit3a$residuals)
par(mfrow=c(2,2))
plot(fit3a)
```



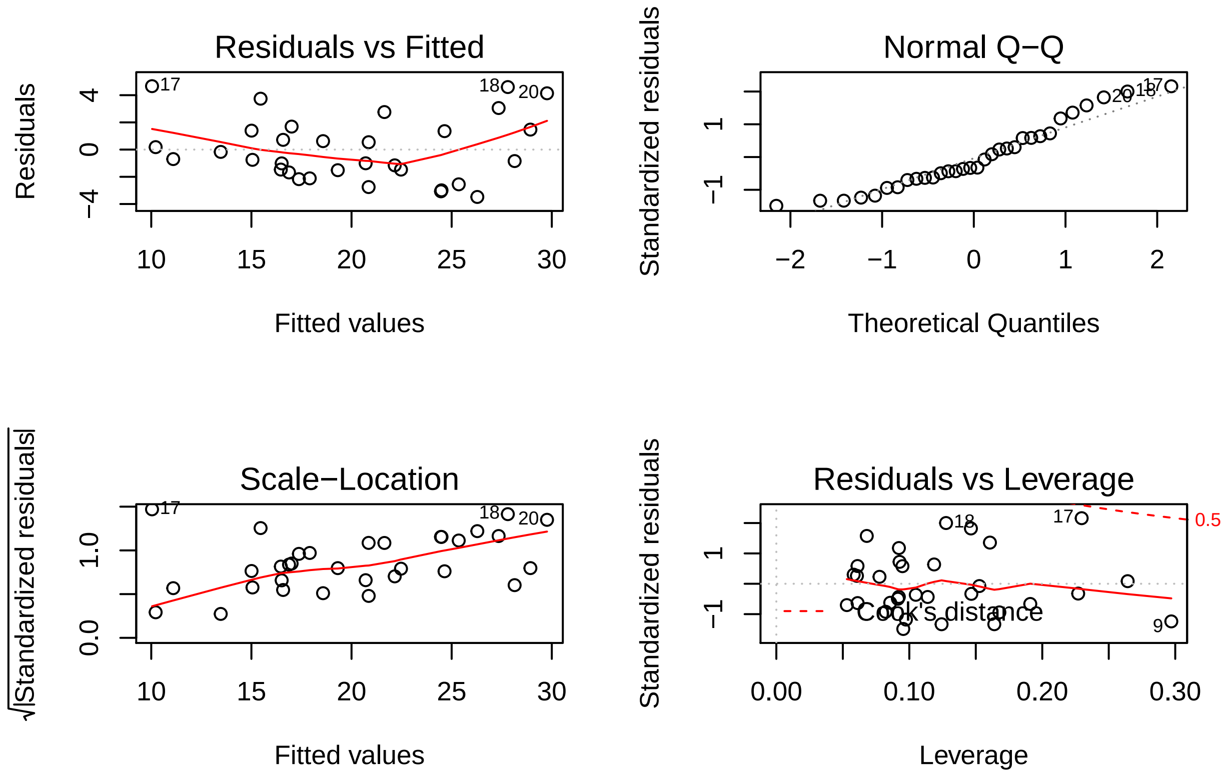

```
# based on the best of 10 models based on the physical model mtcars3
#fit28<- lm(mpg~as.factor(am)+drat+displ+weight, data=mtcars3)
fit28$coefficients
summary(fit28)$sigma
shapiro.test(fit28$residuals)
par(mfrow=c(2,2))
plot(fit28)
```



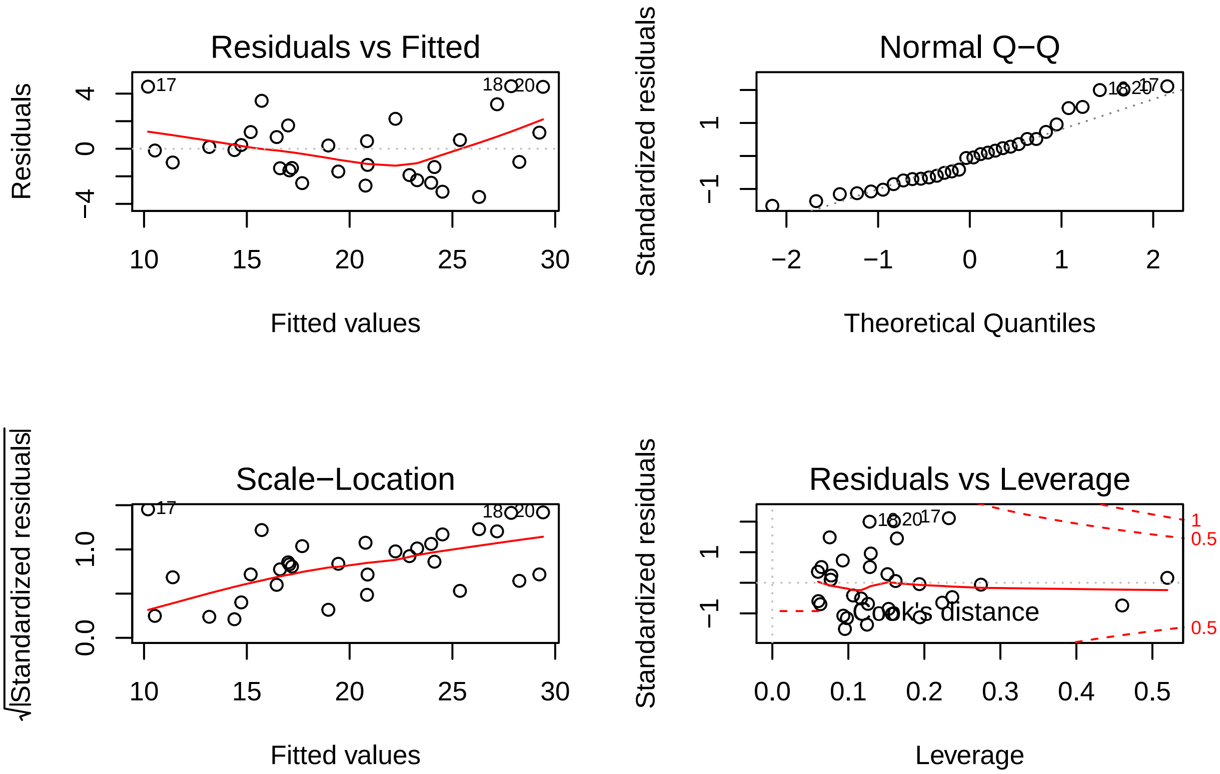

```
# number crunching 2 parameters
index2
fit35<-lm(mpg~as.factor(am)+disp, data=mtcars6)
fit35$coefficients
summary(fit35)$sigma
shapiro.test(fit35$residuals)
par(mfrow=c(2,2))
plot(fit35)
```



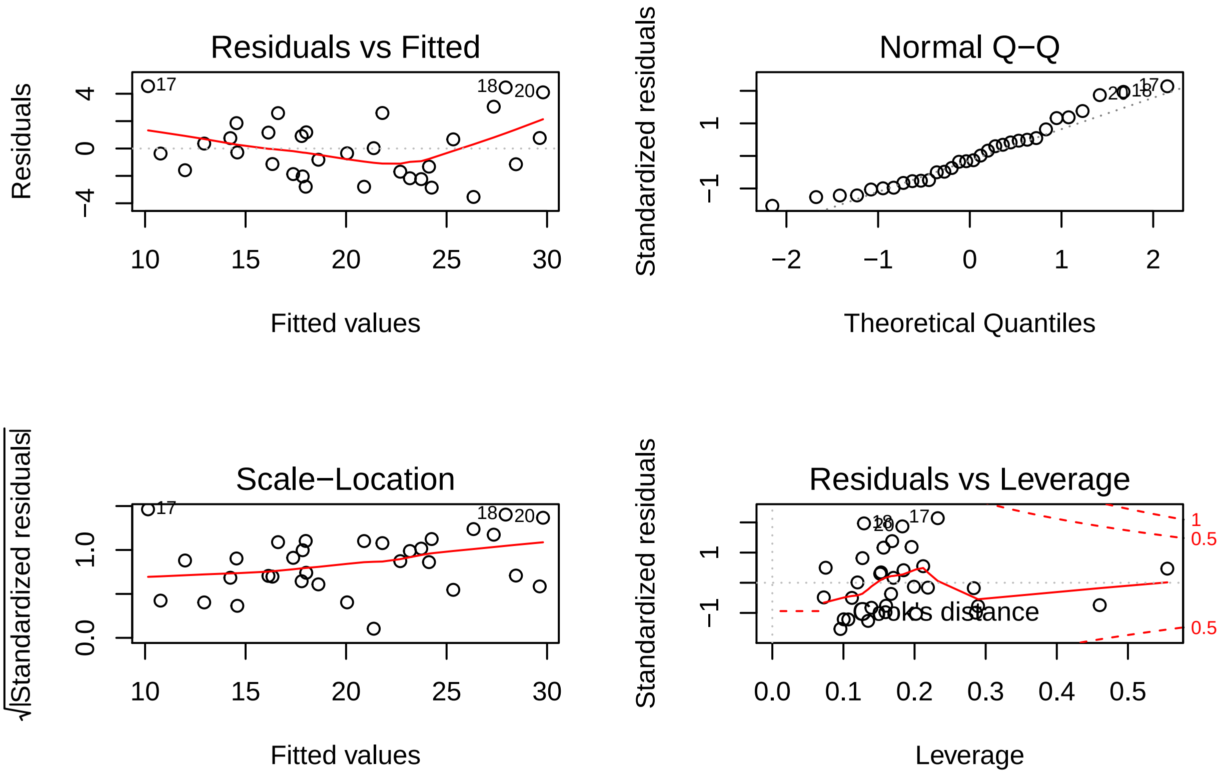

```r
# number crunching based on 3 parameters
index3
fit36<-lm(mpg~as.factor(am)+wt+qsec, data=mtcars6)
fit36$coefficients
summary(fit36)$sigma
shapiro.test(fit36$residuals)
par(mfrow=c(2,2))
plot(fit36)
```




```r
# number crunching based on 4 parameters
index4
fit37<-lm(mpg~as.factor(am)+wt+qsec+hp, data=mtcars6)
fit37$coefficients
summary(fit37)$sigma
shapiro.test(fit37$residuals)
par(mfrow=c(2,2))
plot(fit37)
```




```
# number crunching basedd on 5 parameters
index5
fit38<-lm(mpg~as.factor(am)+disp+wt+qsec+hp, data=mtcars6)
fit38$coefficients
summary(fit38)$sigma
shapiro.test(fit38$residuals)
par(mfrow=c(2,2))
plot(fit38)
```



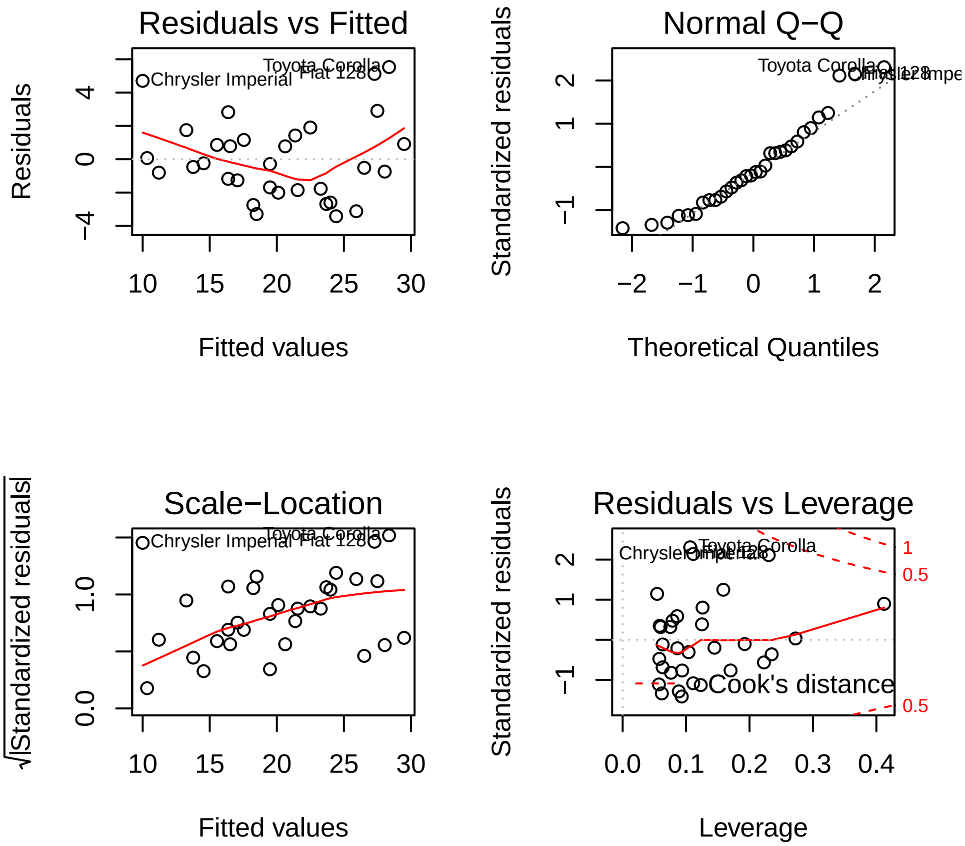
fits	amCoef	stdErrorFit	pValueShapiro	otherFactorsModel
fit0	7.24	4.90	0.85730	none
fit5a	0.13	2.64	0.12030	disp+wt+cyl
fit3a	2.08	2.53	0.10590	hp+wt
fit28	-1.02	2.21	0.20460	drat+displ+weight
fit35	1.83	3.21	0.05251	disp
fit36	2.94	2.46	0.08430	wt+qsec
fit37	2.93	2.43	0.07126	wt+qsec+hp
fit38	3.47	2.42	0.18580	disp+wt+qsec+hp

Conclusion.

Based on the residuals and the pvalues from the Shapiro tests, we would prefer model 37. So the manual transmission is better for fuel consumption compared to the automatic transmission by 2,43 mpg. However due to the the fact that not a real comparison between automatic and manual transmission for each type of car is made, it is very doubtful whether the model would work on another data set.

Appendix

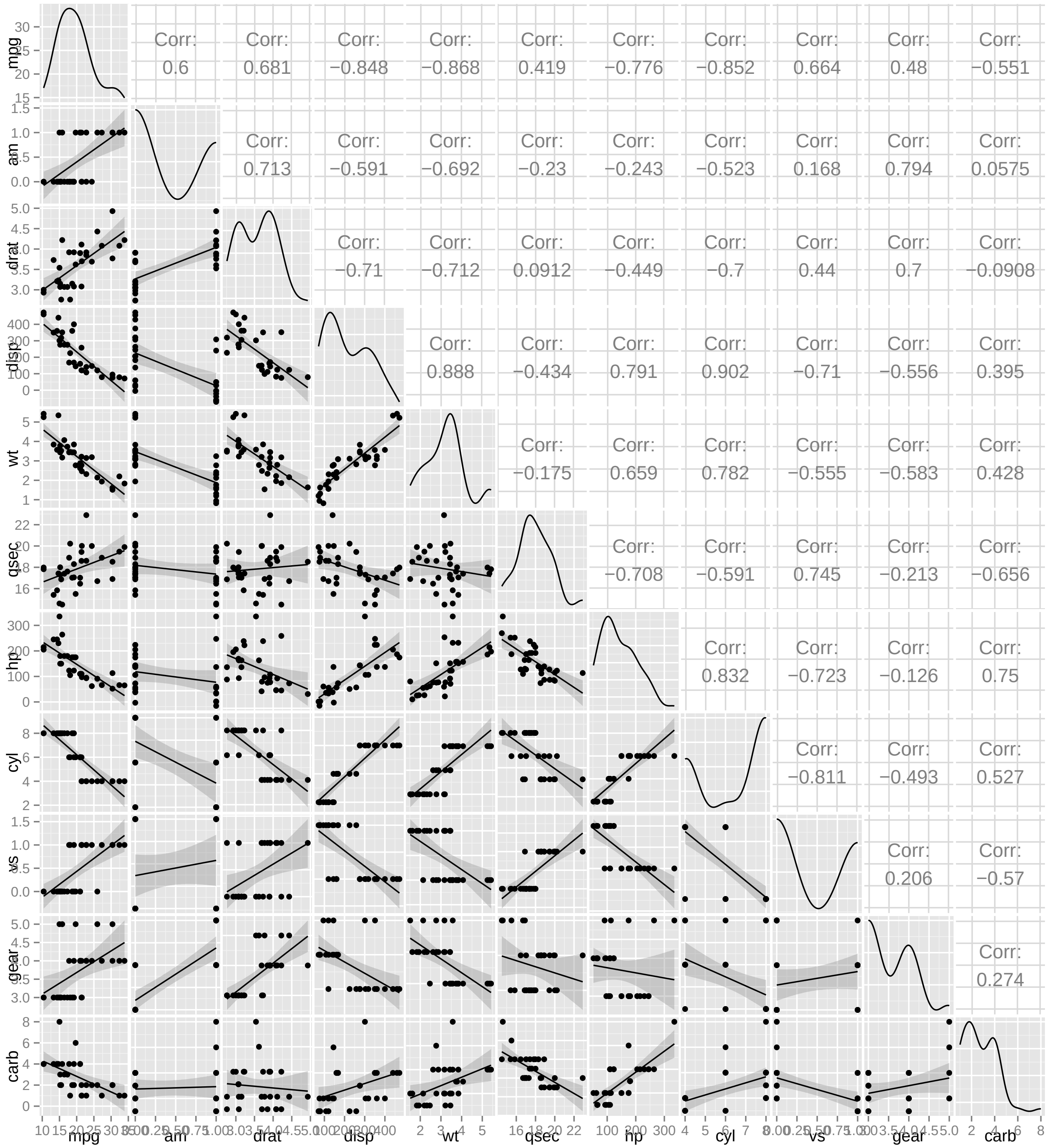
Example of an residuals plot, as used to draw conclusions.



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

The ggpairs plot of mtcars6

```
g3<-ggpairs(mtcars6,lower=list(continuous="smooth"), params=c(method="loess"))
g3
```



The ggpairsplot of mtcars3 (the “physical”" data frame)

```r
g2<-ggpairs(mtcars3,lower=list(continuous="smooth"), params=c(method="loess"))
g2
```