mtcars study

Stéphane Lassalvy

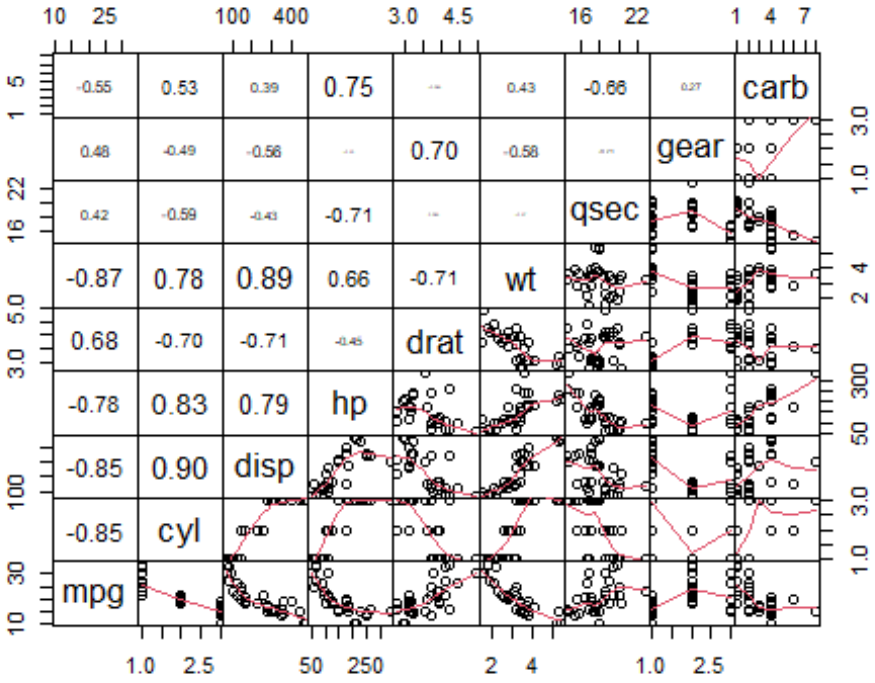
2024-02-02

```
# mtcars study
# title:  mtcars study
# author: Stéphane Lassalvy
# date:   2024-02-02"
# Licence : GPL-3

# install.packages("car")
# install.packages("FactoMineR")
library(car)

## Le chargement a nécessité le package : carData

library(FactoMineR)
```

Abstract

This is a study based on the mtcars data from the R software, trying to determine the influence of the transmission type (Manual vs Automatic) on the consumption (mpg) of the cars. This is the last assignment graded by peers to pass the certificate of the Coursera course “Regression Models” taught by Dr Brian Caffo, Dr Roger D. Peng and Dr Jeff Leek from John Hopkins University. We here performed a backward elimination of the predictors modelling the consumption “mpg” starting with all the predictors possible in the data set. This way we got a more parsimonious model and we tried to get it better by adding it the interactions terms of transmission with weight (am:wt) and transmission with horsepower (am:hp). This ancova model that we fitted had an adjusted R-squared of 0.87 with residuals reasonably following a normal distribution. Our model showed Manual cars had a +14.9 miles/gallon significant effect and showed an interaction slope “am:wt” of -5.21 for Manual cars : on the whole, Manual cars consume less but as their weight grows up they will comparatively consume more than Automatic ones.

descriptive statistics



Frequency table of “am” VS “vs”

```
# Frequencies am VS vs
table(dat$am, dat$vs)
```

```
##
##           V-shaped Straight
##   Auto          12        7
##   Manual         6        7
```

Frequency table for “cyl” and “gear”

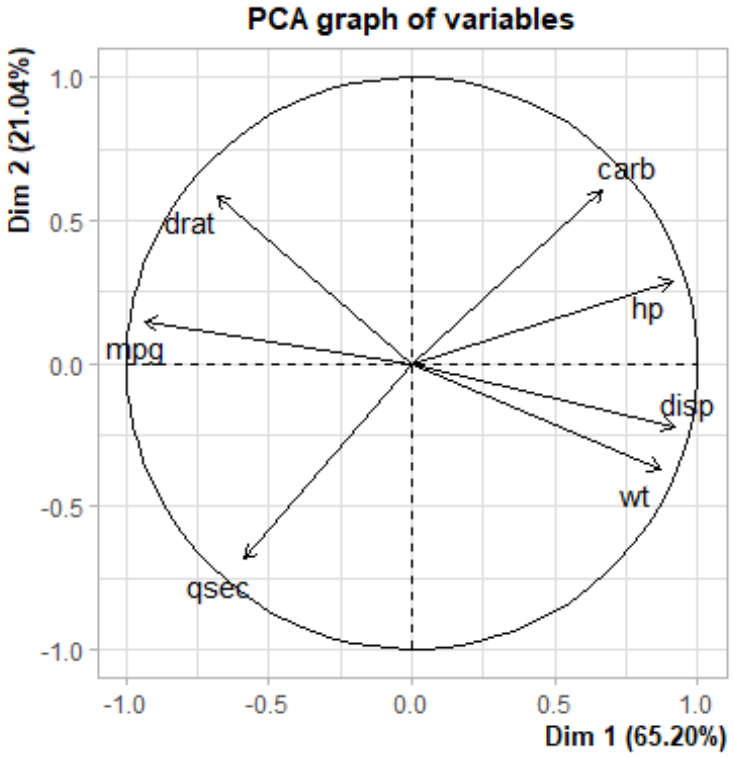
```
table(dat$cyl)
```

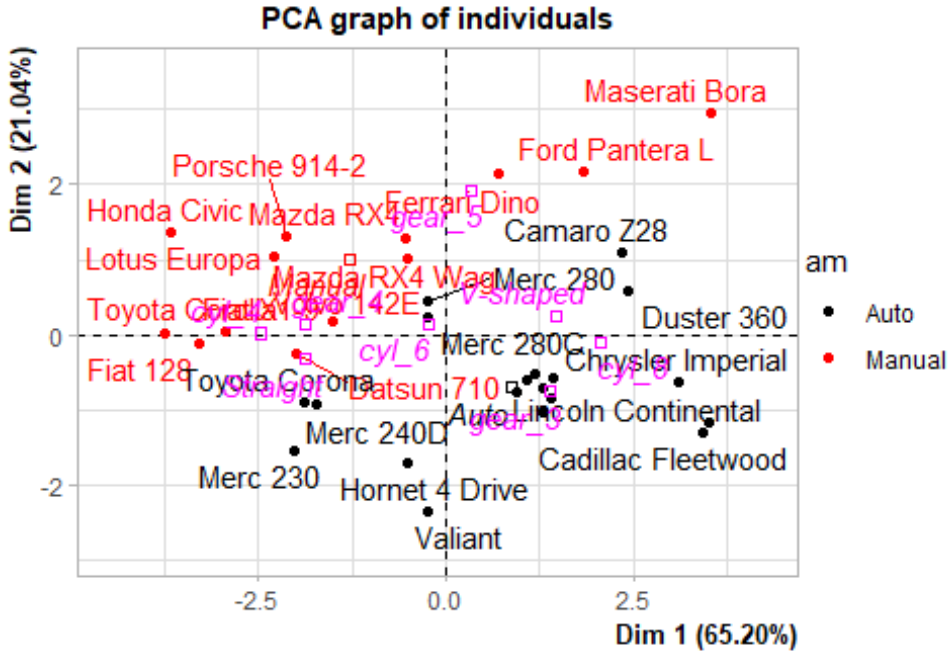
```
##
##  4  6  8
## 11  7 14
```

```
table(dat$gear)
```

```
##
##  3  4  5
## 15 12  5
```

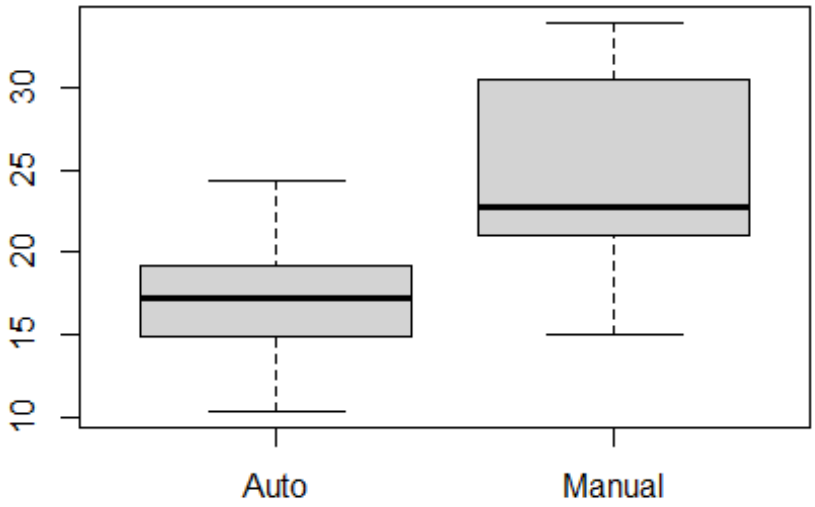
PCA of data


PCA graph of variables
Dim 2 (21.04%)
Dim 1 (65.20%)
1.0
0.5
0.0
-0.5
-1.0
carb
drat
hp
mpg
disp
wt
qsec


PCA graph of individuals
Dim 2 (21.04%)
Dim 1 (65.20%)
2
0
-2
-2.5
0.0
2.5
Maserati Bora
Ford Pantera L
Porsche 914-2
Honda Civic
Mazda RX4
Ferrari Dino
gear_5
Camaro Z28
Lotus Europa
Merc 280
V-shaped
Duster 360
Merc 280C
cyl_6
Chrysler Imperial
Fiat 128
Toyota Corona
Datsun 710
Auto
Lincoln Continental
Merc 240D
Cadillac Fleetwood
Merc 230
Hornet 4 Drive
Valiant
am
Auto
Manual

Axis 1 of the PCA opposes “mpg” with the engine power variable like “hp” or “carb”, but also with the weight variable “wt”. Variable “drat” may contribute positively to “mpg”. The “am” levels are not very well discriminated 1st axis of the PCA which correspond also to “mpg”.

Boxplot of “mpg” VS “am” levels



Two sample T-test for “mpg” VS “am”

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means between group Auto and
group Manual is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
##   mean in group Auto mean in group Manual
##             17.14737             24.39231
```

Trying backward elimination

```
# First model
mdl <- lm(mpg ~ ., data = dat)
```

```r
mdl.sel <- invisible(step(mdl, direction = "backward"))
```

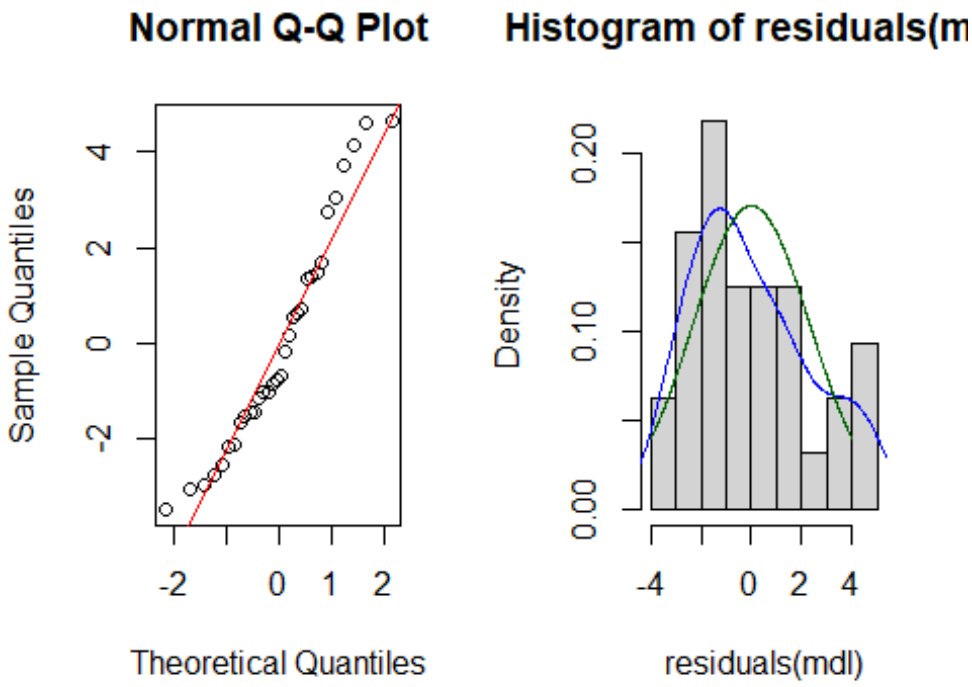
```
## Start:  AIC=70.87
## Stop:   AIC=61.31
## mpg ~ wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## <none>              169.29 61.307
## - am    1    26.178 195.46 63.908
## - qsec  1   109.034 278.32 75.217
## - wt    1   183.347 352.63 82.790
```

```r
summary(mdl.sel)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = dat)
##
## 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 ***
## amManual      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
```

```r
plot.res(mdl.sel)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(mdl)
## W = 0.9411, p-value = 0.08043
```



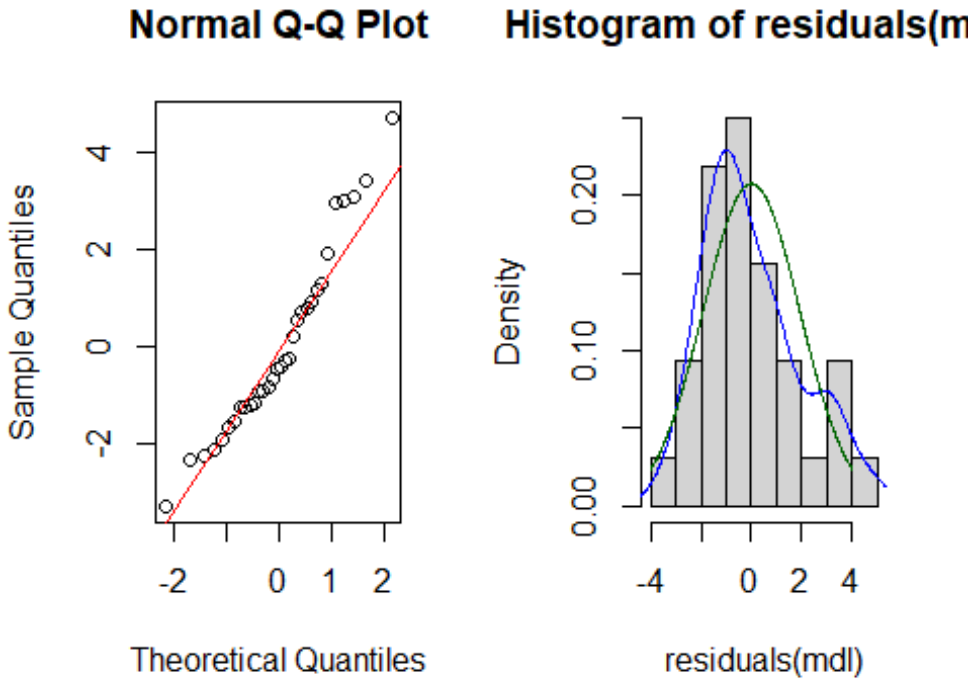
Trying to add interaction terms am:wt and am:hp

```
mdl.ancova <- update(mdl.sel, mpg ~ . + am:wt + am:hp)

summary(mdl.ancova)

##
## Call:
## lm(formula = mpg ~ wt + qsec + am + wt:am + am:hp, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2875 -1.2387 -0.4519  0.9905  4.7161
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.225233   8.894399   1.374 0.181487
## wt          -2.511527   0.934863  -2.687 0.012646 *
## qsec         0.890492   0.411294   2.165 0.040126 *
## amManual    14.876348   3.987068   3.731 0.000985 ***
## wt:amManual -5.211613   1.956068  -2.664 0.013311 *
## amAuto:hp   -0.011257   0.018727  -0.601 0.553159
## amManual:hp  0.003566   0.014833   0.240 0.811992
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

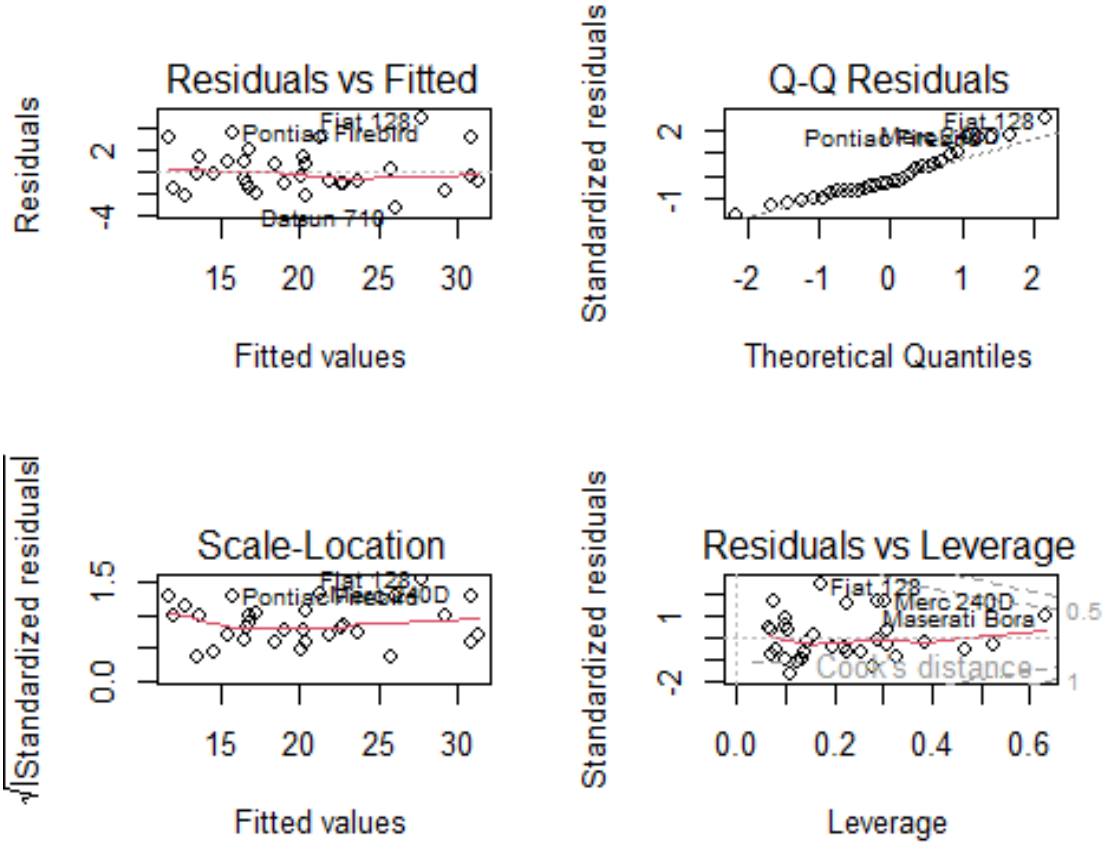
```
## Residual standard error: 2.14 on 25 degrees of freedom
## Multiple R-squared:  0.8984, Adjusted R-squared:  0.874
## F-statistic: 36.83 on 6 and 25 DF,  p-value: 3.11e-11
```

```
plot.res(mdl.ancova)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(mdl)
## W = 0.9477, p-value = 0.1239
```



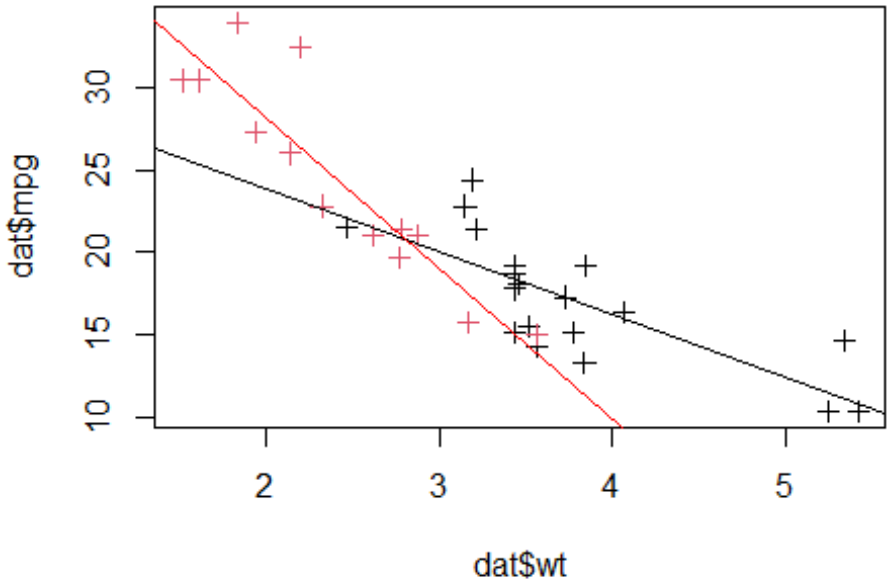

```
par(mfrow = c(2,2))
plot(mdl.ancova)
```



Residuals of the model mdl.ancova are normal and its adjusted R^2 of 0.86 is quite good. It appears now that "am" and "am:wt" are significant.

Description of the relationship of mpg ~ wt by am to see if Auto and Manual groups have different slopes

```
# see if wt splope differs following am level
plot(dat$wt, dat$mpg, col = dat$am, pch = 3)
reg0 <- lm(mpg ~ wt, dat[dat$am == "Auto",])
coef(reg0)

## (Intercept)          wt
##   31.416055   -3.785908

reg1 <- lm(mpg ~ wt, dat[dat$am == "Manual",])
coef(reg1)

## (Intercept)          wt
##   46.294478   -9.084268

abline(reg0, col = "black")
abline(reg1, col = "red")
```



The slope of "wt" appears to be different for Automatic and Manual cars. The slope for Manual cars is about 5 miles/gallon/1000lbs lesser than for Automatic ones.

Conclusion

The model mdl.ancova fit well the data with an adjusted R^2 equals to 0.87, but we can't assure it has a good predictive value for new data to come. This model allows us to test that the "am" effect is significant being 14.9 miles/gallon higher for Manual cars. The "am:wt" interaction is also significant showing a slope of -5.21 miles/gallon/1000lbs Manual cars : on the whole, Manual cars consume less, but as their weight grows up they will comparatively consume more than Automatic ones.

References

Caffo B *et al* (2024) : Regression Models - Online course on Coursera from John Hopkins University

Fox J, Weisberg S (2019). *An R Companion to Applied Regression*, Third edition. Sage, Thousand Oaks CA. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Le S, Josse J, Husson F (2008). FactoMineR: An R Package for Multivariate Analysis. Journal of Statistical Software, 25(1), 1-18. 10.18637/jss.v025.i01

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.