```
## Multiple regression lab

## Multiple regression analysis with a dataset on automobiles

## Goal: analyse what is the best model to predict "accelerazione"
(accel, sec -from 0 to 100km/h) when you have two potential important
covariates: "potenza del motore" (potenzacv, cavalli vapore) and "massa"
(kg)

library(dplyr)
library(PerformanceAnalytics)

## import data (in automobili.Rdata) and form your current dataset of
interest

???

# check what you imported

automobili %>% ???
# in Rstudio
View(automobili)

# Define row names by merging the first three columns (use e.g. paste())

row.names(automobili) <- ???(automobili$marca, automobili$modello,
automobili$allestimento)

# create dataset by selecting three variables (choose between subset() or
the “piping” syntax)

auto <- ???(automobili, select=c("accel","potenzacv","massa"))
# or
auto <- auto %>% ???

# export dataset
write.table(auto, file="data/auto2.csv", ...)

# let's start!

d <- auto

## explore data graphically

pairs(d)

## we can start by fitting a simple regression of accel on "potenza"
# first plot the raw (untransformed) data

???

# examine whether transforming "potenza" into ??? or ??? or etc. make the
relationship linear (or (approximately linear)

???()
???()
???()
```

```r
## fit a linear model of accel on the properly transformed covariate

fitrecsqrt <- lm(accel ~ ???, data=d)

sfit <- summary(fitrecsqrt)

# look at coefs and error variance estimates, moreover, at R-squared

???

# plot the regression line

plot(d$accel ~ ???)
abline(fitrecsqrt)

## Build and plot the confidence and prediction interval along the entire
range of the regressor.

new_data <- ???
A <- predict(???)
B <- predict(???)
???
???


## multiple regression for data in auto2.csv (the same as the current d)

d<-read.table(file="auto2.csv", sep=",", dec=".", ???, ???)
head(d)

# explore
pairs(d)

chart.Correlation(d)

# comment on the relationships between variables
# Does a model with both covariates make sense? (problems: underfitting?,
overfitting?)

# fit on the original (two) covariates
fit0 <- lm(accel ~ potenzacv + massa, data=d)
fit0
summary(fit0)

# comment on the results, in particular the significance and
interpretation of the coefs

# now a brief look at residual analysis

???

# do we stop here?

## fit a second model after somehow transforming variables

fit1 <- lm( ??? ~ ??? + ???, data=d)
```

```
summary(fit1)

## compare the two models

anova(???, ???)

# can you compare fitrecsqrt with fit0? Why?
# Is mass an unnecessary variable?


# A look at the variance inflation factor
library(car)
vif(???)


## Model representation  (with multiple covariates)

plot(d$accel, fitted(fit1))
abline(0, 1)

# now represent y as function of the covariates

fitYX=lm(accel~I(1/sqrt(potenzacv)),data=d)
fitVX=lm(massa~I(1/sqrt(potenzacv)),data=d)

# fitYX is an old frienf of us ...

plot(resid(fitYX)~resid(fitVX))
abline(h=0,lty=2)
abline(v=0,lty=2)
abline(lm(resid(fitYX)~resid(fitVX)))
summary(lm(resid(fitYX)~resid(fitVX)))

# What coincidence do you note?

#####################################
#####################################

## A polynomial model for a nonlinear reationship

fitpolin2 <- lm(accel ~ potenzacv + I(potenzacv^2), data=d)
summary(fitpolin2)

# Check the model


# Are we satisfied with it?

# Update the model adding the cubic power

fitpolin3 <- update(fitpolin2, .~.+???)
summary(fitpolin3)

# was it really worth the effort if we compare it to the simple regr
model?
# compare the fitted model of the simple regr to the polynomial regr.
```

???
???
???