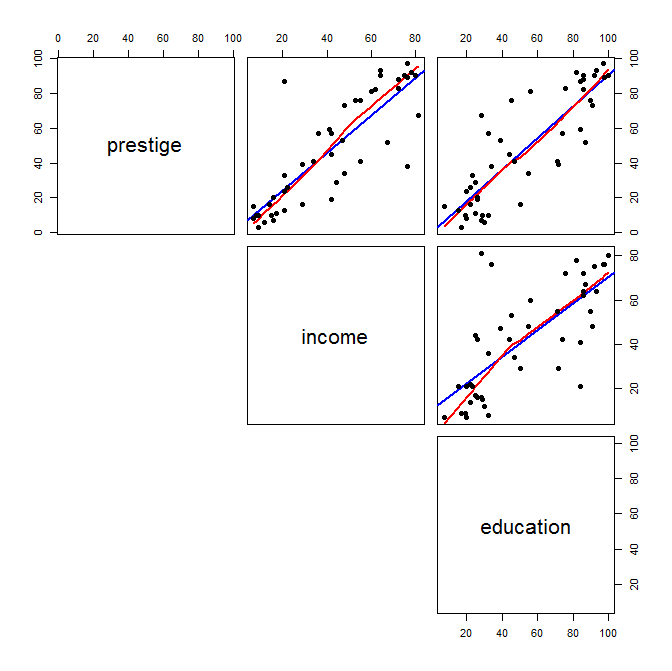
# Multiple regression:

****

## Example: Duncan data on prestige of professions or weight vs height in Davis

Use data(Prestige) in library(car).

Study correlations between numeric variables appearing in the work space. Explicative variables are income and education. Response variables is prestige and we have to propose a multiple regression model to explain the prestige of jobs.

****

### Suggested steps

* Correlation matrix in R, cor(duncan1, use="pairwise.complete.obs" )

1. Matrix of 2 by 2 scatterplots.
2. Forward regression from the nul model with a direction forward option in method step().

> duncan1.lm0 <- lm( prestige ~1, data=duncan1)

> summary(duncan1.lm0)

> step(duncan1.lm0, ~income+education, direction=”forward”, data=duncan1)

1. Backward regression from the model with INCOME+EDUCATION in backward direction option in method step().

> duncan1.lm2 <- lm( prestige ~ income+education, data=duncan1)

> summary(duncan1.lm2)

> step(duncan1.lm2, direction=”backward”,data=duncan1)

1. Use method step(.) in R from the nul model to the maximal model with direction specification “both” (it is the default)

> duncan1.lml <- lm( prestige ~income+education, data=duncan1)

> summary(duncan1.lm1)

> duncan1.lm<- step(duncan1.lm1, ~income+education, data=duncan1)

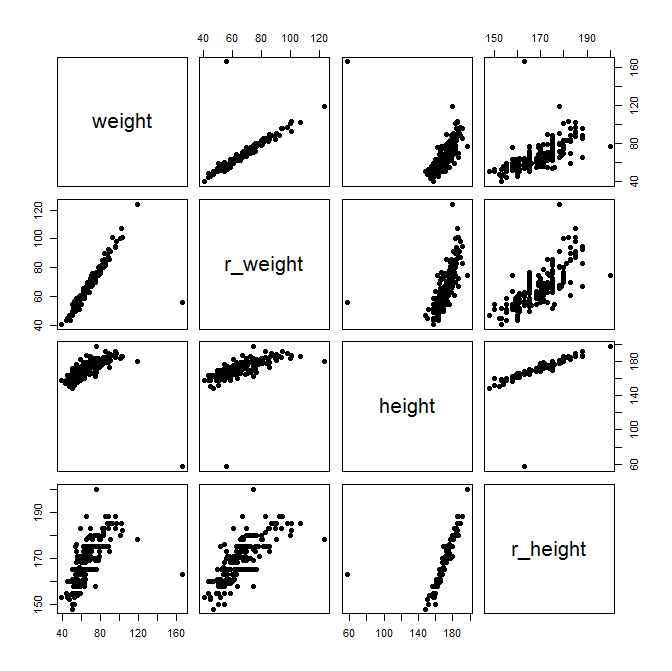
1. Linear correlation between a response variable and explicative variables might not be significative once some of the explicative variables are already included in the model.
2. ***A touch on diagnostics:*** Check outliers in residuals and influent data in the selected model. Compute histogram of studentized residuals (rstudent(model)), leverage (hatvalues(model)) and Cook’s distance (cooks.distance(model)).
3. R2 and global regression test .
4. Residual analysis:

* Detection of o*utliers*.
* Scatterplot of studentitzed residual *vs.* .
* Scatterplot of studentitzed residual *vs.* *vs.* .
* Detection of *a priori* and *a posterior influent data*.
* Scatterplot of studentitzed residual *vs.* *leverage*.
* Scatterplot of studentitzed residual *vs.* Cook’s distance.

## Example: weight vs height in Davis

The Davis data frame has 200 rows and 5 columns. The subjects were men and women engaged in regular exercise. There are some missing data. This data frame contains the following columns:

* sex: A factor with levels: F, female; M, male.
* weight:Measured weight in kg.
* height: Measured height in cm.
* r\_weight : Reported weight in kg.
* r\_height : Reported height in cm.



Firstly, we examine the relationship between the reported weight and the actual weight in order to assess how data behaves. Pay attention to outliers.

Secondly, we focus on the classical relationship between weight (Y) and height (X): does a quadratic fit hold? Why?

### Suggested steps

* Correlation matrix in R, cor(Davis, use="pairwise.complete.obs" )

1. Matrix of 2 by 2 scatterplots.
2. Multiple regression weight (Y) vs r\_weight (Y). Interpret the regression equation and quality of the fit
3. Multiple regression weight (Y) vs height (X). Interpret the regression equation and quality of the fit
4. Multiple regression weight (Y) vs poly(height,2) (X). Can you Interpret the regression equation and quality of the fit?