Topic 2:
Data Quality and Profiling

Statistical Modelling and Inference

Master in Data Science

Prof. Lídia Montero & Josep Franquet

lidia.montero@upc.edu *josep.franquet@upc.edu*

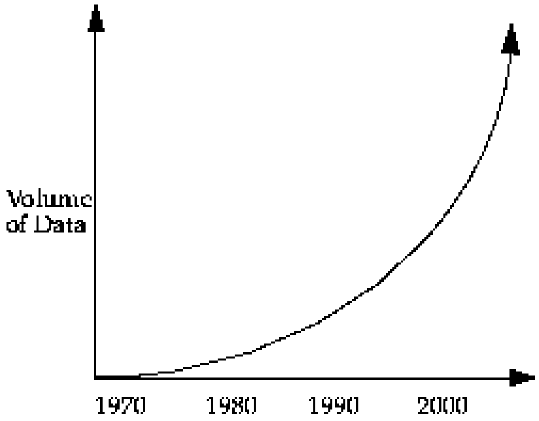
FIB

UPC

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

Data growth rate

The past two decades has seen a dramatic increase in the amount of information or data being stored in electronic format. This accumulation of data has taken place at an explosive rate. It has been estimated that the amount of information in the world doubles every 20 months and the size and number of databases are increasing even faster.



QUALITY of stored data is a fundamental issue

Aspects of data quality

- Problems with data:
  - Redundancy (duplicated information across DDBB)
  - Inconsistencies: changes in names, addresses, telephone numbers, email addresses (perishing validity) …
  - Application-data dependence, lack of flexibility,
  - Inability to share data among applications.
  - Errors, incorrect data
  - Outliers, unusual values for a given data (bias the results)
  - Missing data, non coded data.... (non response: total, partial)
- Effects of low data quality:
  - Loose of accuracy, waste of money, reduction of data size, poor result precision, increment of variability, …

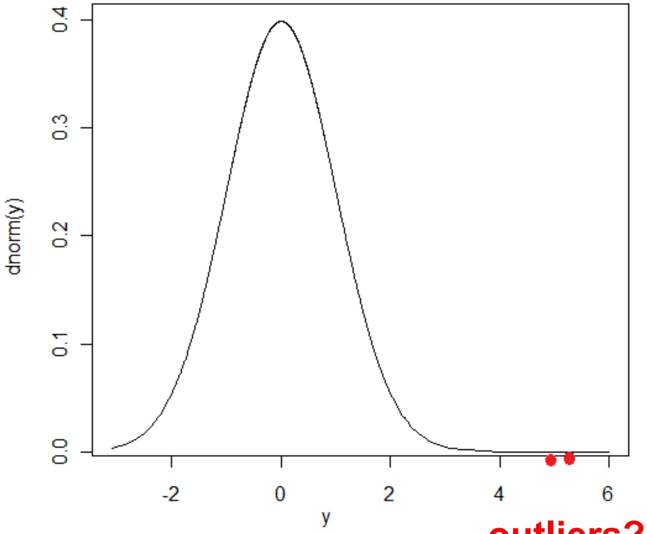
From a statistical point of view, we can only treat outliers and missing data

Outliers

What is an outlier? Definition of Douglas Hawkins: "An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism"

Statistics-based intuition. Normal data follow a "normal generating data mechanism", e.g. some given statistical process. Outlying data may be a:

- very unlikely events for the normal generating mechanism
- data following a different generating mechanism



outliers??

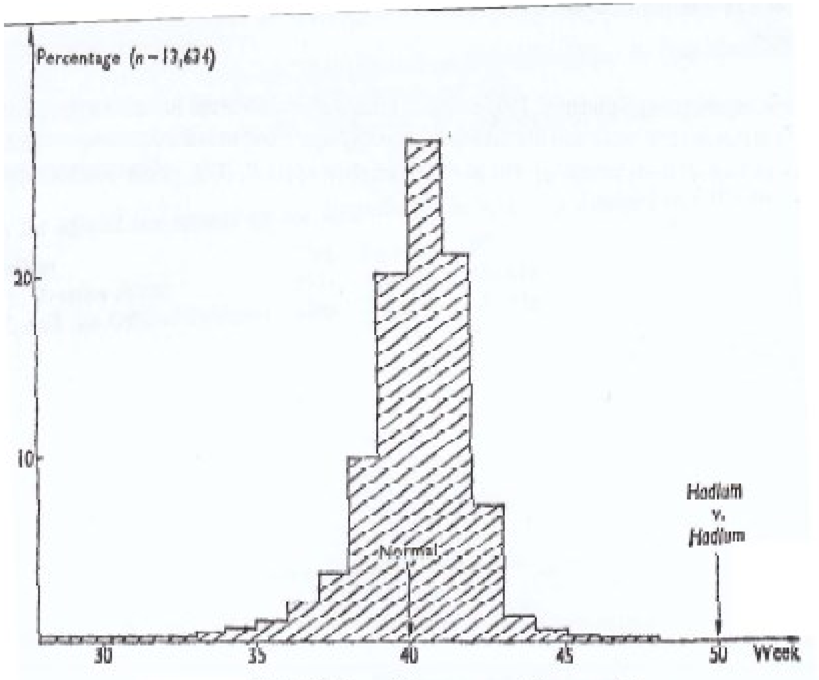
if X~N(0,1)	Prob(x≥X)
1	0.1586553
2	0.02275013
3	0.001349898
4	3.167124e-05
5	2.866516e-07

Example: Hadlum vs. Hadlum (1949) [Barnett 1978]

The birth of a child to Mrs. Hadlum happened 349 days after Mr. Hadlum left for military service.

Average human gestation period is 280 days (40 weeks).

Statistically, 349 days is an outlier.



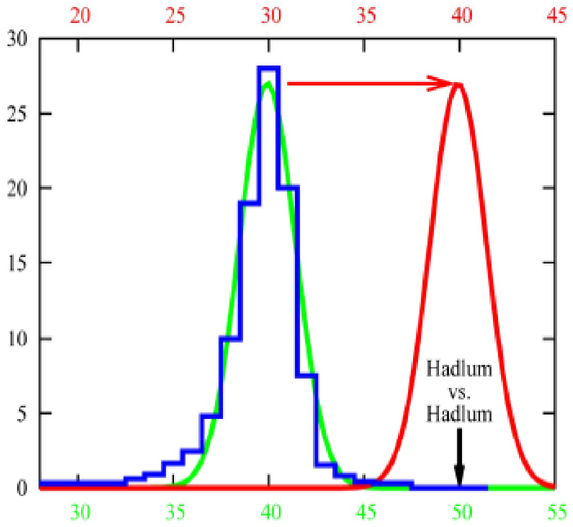
Distribution of human gestation periods.

Example: Hadlum vs. Hadlum (1949) [Barnett 1978]

blue: statistical basis (13634 observations of gestation periods)

green: assumed underlying Gaussian process. Very low probability for the birth of Mrs. Hadlums child for being generated by this process

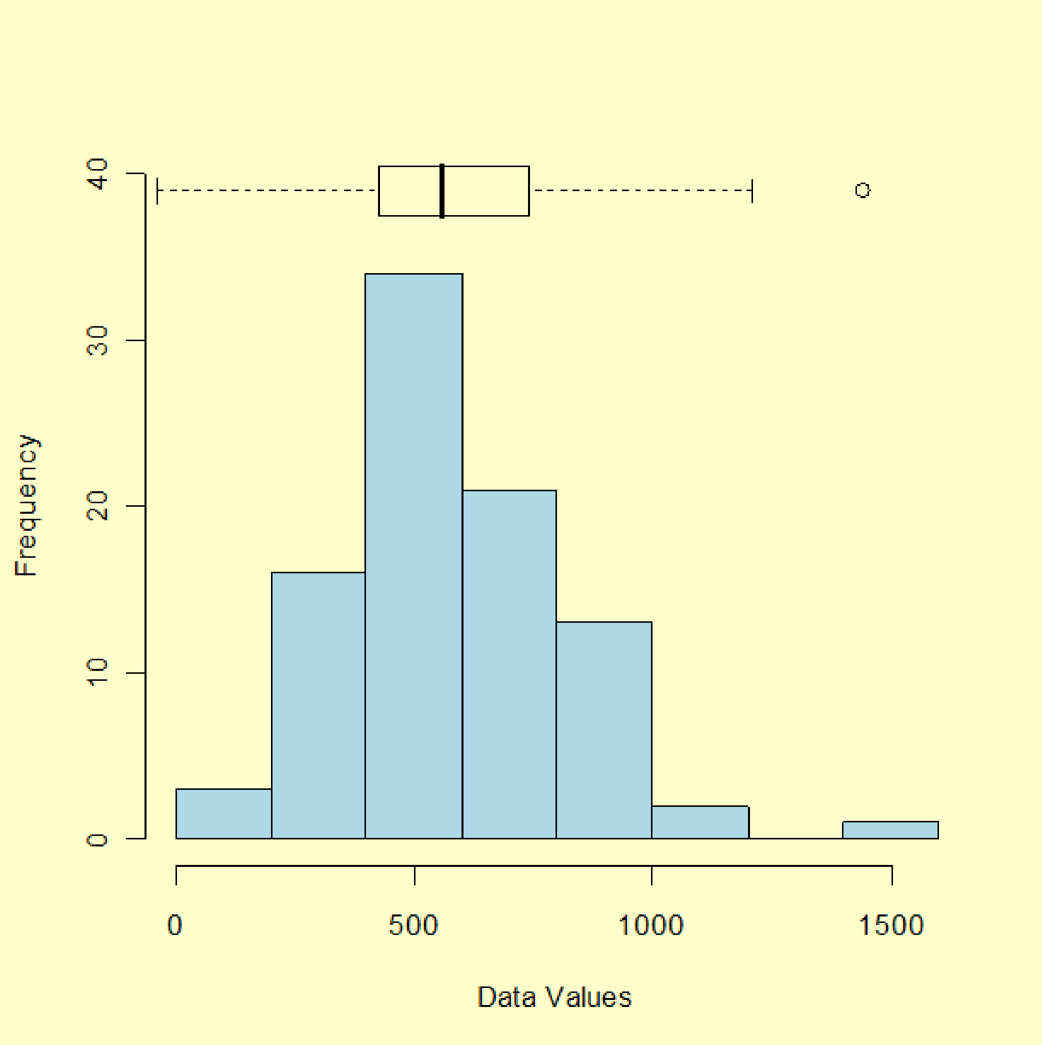
red: assumption of Mr. Hadlum: Another Gaussian process responsible for the observed birth, where the gestation period starts later. Under this assumption the specific birthday has highest-probability.



Example of an Outlier in data

- The data set of *N* = 90 ordered observations as shown below is examined for outliers:

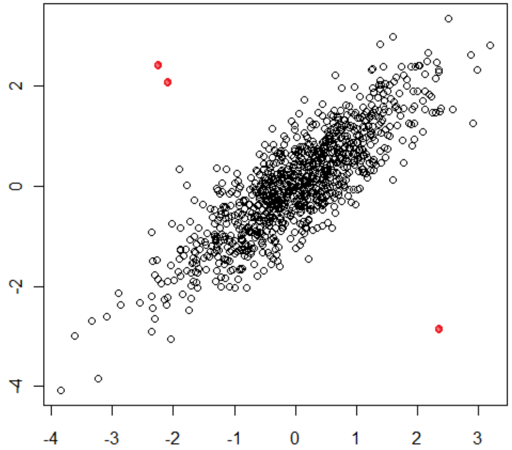
30, 171, 184, 201, 212, 250, 265, 270, 272, 289, 305, 306, 322, 322, 336, 346, 351, 370, 390, 404, 409, 411, 436, 437, 439, 441, 444, 448, 451, 453, 470, 480, 482, 487, 494, 495, 499, 503, 514, 521, 522, 527, 548, 550, 559, 560, 570, 572, 574, 578, 585, 592, 592, 607, 616, 618, 621, 629, 637, 638, 640, 656, 668, 707, 709, 719, 737, 739, 752, 758, 766, 792, 792, 794, 802, 818, 830, 832, 843, 858, 860, 869, 918, 925, 953, 991, 1000, 1005, 1068, 1441



Discussion of the Hadlum vs. Hadlum case

1. Data is usually ***multivariate***, i.e., multi-dimensional, whereas => basic model is assumed to be univariate, i.e., 1-dimensional

2. There is usually ***more than one generating*** mechanism/statistical process underlying the "normal" data; => basic model assumes only one "normal" generating mechanism, where outliers are rare observations. Outliers may represent a different class (generating mechanism) of objects, so there may be a large class of similar objects that are the outliers.



Outliers are multivariate

Univariate detection of outliers doesn't imply multivariate detection

Univariate detection of outliers. The Boxplot

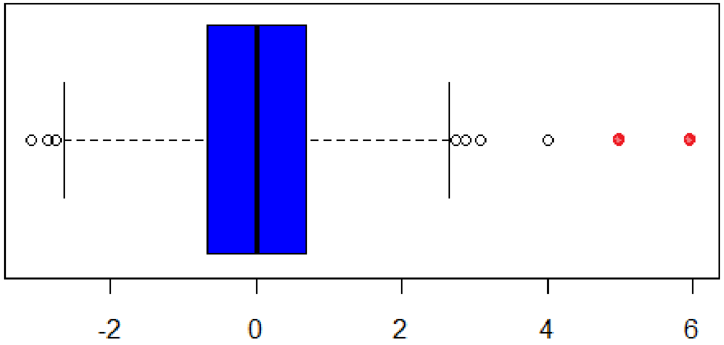
The Boxplot (Tukey, 1977) is a graphical display for exploratory data analysis, where the outliers appear tagged. Two types of outliers are distinguished: *mild* outliers and *extreme* outliers.

An observation x is declared an extreme outlier, if it lies outside of the interval **$(Q1\text{-}3 \times IQR,\ Q3+3 \times IQR)$**, where $IQR=Q3\text{-}Q1$ is called the *Interquartile Range*.
An observation x is declared a mild outlier if it lies outside of the interval **$(Q1\text{-}1.5 \times IQR,\ Q3+1.5 \times IQR)$**.

The numbers 1.5 and 3 are chosen by comparison with a normal distribution.
If $x \sim Normal$:

$Prob(X \geq Q3+1.5 \times IQR)= 0.003488302$

$Prob(X \geq Q3+3 \times IQR)= 1.170971e\text{-}06$



Practice of detecting outliers

- To obtain unbiased results in any statistical/learning algorithm. Including outliers in the training data may invalidate the results.

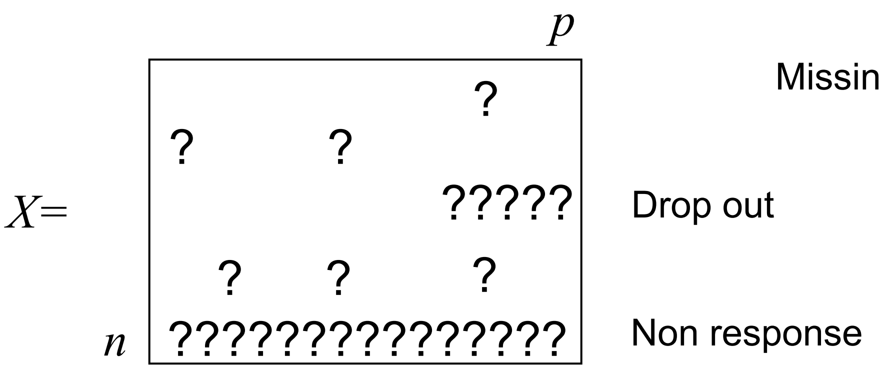
- Once we have detected outliers, what we should do?
  - Eliminate them (but we loose information of the eliminated individuals) and deleting outliers is not the best solution, since outliers are recursive.
  - Weight the individuals inversely to outlying degree of individuals, to diminish its importance (but statistical/learning methods would need to had implemented a weighing option of individuals).
  - Make robust estimation of the parameters of the “normal generating mechanism”, for instance with a given percentage of the “central” individuals.
  - Declare outliers as “missing values” and treat them as missing data.

- Detecting “rare” events:
  - Fraud detection,
  - Detecting network intrusion
  - Detecting changes in the behavior (sales, claims, connections, waiting time, …)

The missing data problem

Typical data set:

Some information is missing for some variables and for some cases.



p

Missing values

?

? ?

$X=$????? Drop out

? ? ?

n ???????????????? Non response

Analysis is just designed for complete data sets (standard methods will fail)

Missing data

- **Databases:**
  - Databases are used for secondary purposes, only information which is currently used is maintained. (i.e. in land registries, addresses are the best up to date field, the characteristics of the premises much less).
  - Not compulsory fields.
  - Errors and outliers as missing values …
- **Surveys:**
  - Outright refusals: unit nonresponse → (reweighing the sample)
  - Non response to some items : item nonresponse → (dealing with missings) (it depends on the data collection method: internet, telephone, mail, face to face)
  - Inapplicable questions to some respondents
  - Dropouts in panel studies

Serious drawback of the data quality (values not recorded, not consistent, …)

Missingness is a nuisance

Is missing data a problem

1. Ignoring missing data can seriously bias the results

2. Missing data represents a loss of information (waste of resources)

3. The impact of missing data depends on its generating mechanism (why some values are missing?)

The best policy to deal with missing data is to avoid them with careful planning of data collection, with proper intelligent interfaces.

Exploring the missingness

Before to start. Identify the missing data

Usual convention:

Assign a missing code to continuous variables (NA, -1, 999999, …)

Assign a new category (missing) to a categorical variable.

Check the quality of the information

Count the number of missing per variable and rank them accordingly.

The more the missings the less reliable is the *information* provided by the variable

Characterize the missingness mechanism

Create a new variable counting the number of missing per individual.

Describe this variable (association analysis).

Describe the missing categories by multidimensional methods (missing values form a specific category)

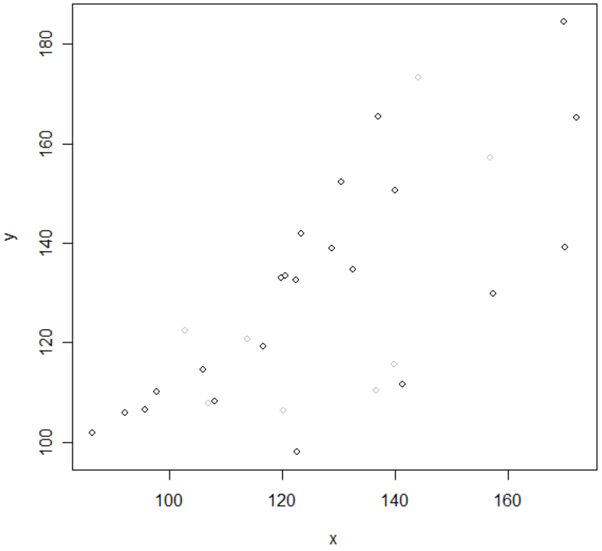
Missingness mechanisms

- MCAR - Completely at random: missing values appear without any pattern. This is the most favorable situation, missing values just implies a reduction of the size.

- MAR - At random: missing values appear related to third observed variables. This is the most usual case, i.e. asking the income of individuals, income is missing but can be imputed from the educational level.

- MNAR - Not at random: missing values depend on the missing variable itself. This is the most difficult case. In the previous example it would be that high incomes tend to not declare it.

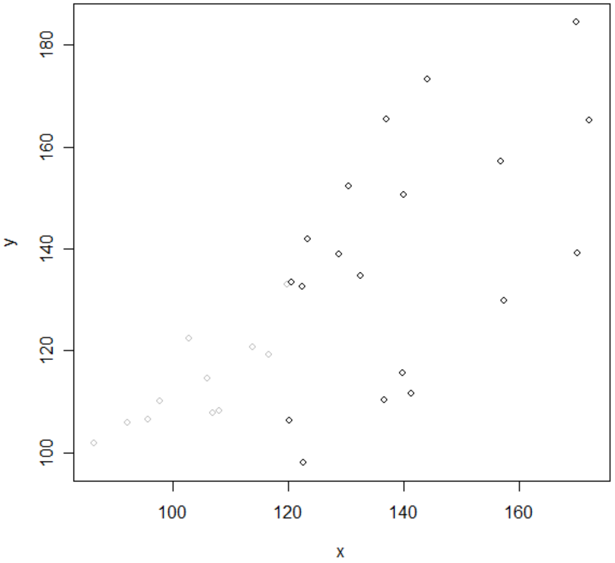
Treatment of missing values



Blood presure month *i+1*

Blood presure month *i*

Complete data



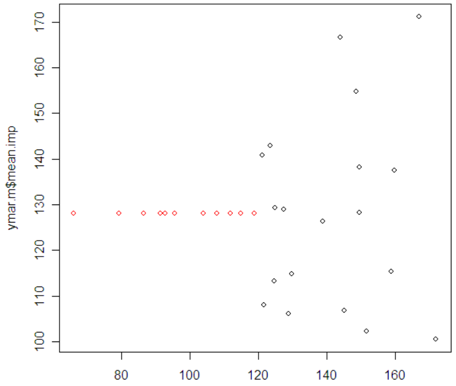
Data with missing values

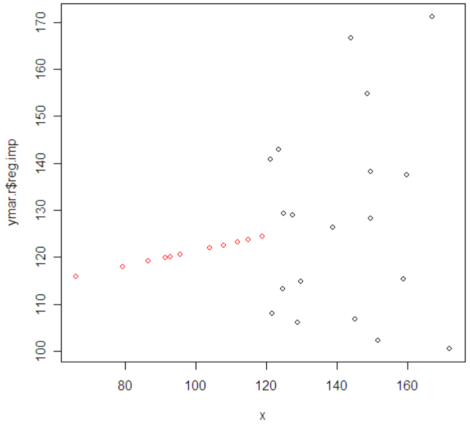
Treatment of missing values

Traditional methods

- **Listwise deletion**. Every individual with a missing value is deleted (loose of information, biasing the results (except in MCAR))

- **Unconditional mean imputation**. Every missing value is substituted by the corresponding global mean of the variable





- **Regression imputation**. Every missing value is substituted by the predicted value from a multiple regression.

Treatment of missing values

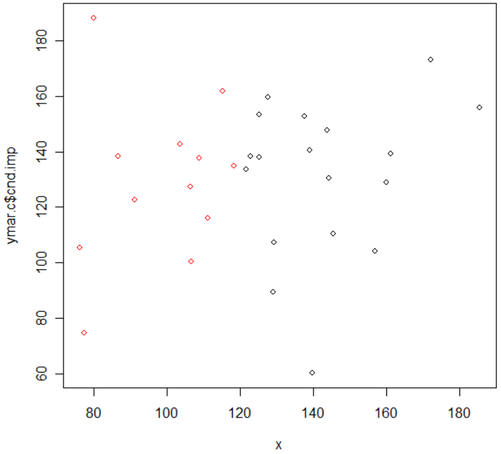
- Stochastic imputation (imputare = to fill in)

Simulate actual data

$$y_{imputed} = f(y/X) + \varepsilon$$

Stochastic regression imputation

$$y_{imputed} = \hat{y} + random\_draw\ N(0, s^2_{iresid})$$

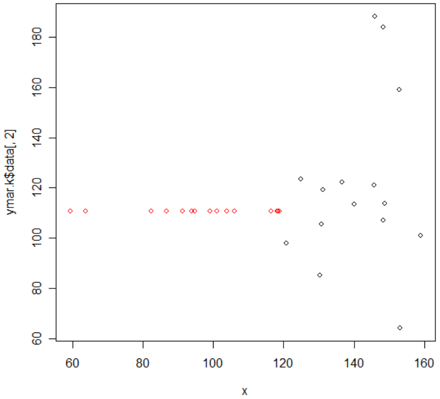


Treatment of missing values

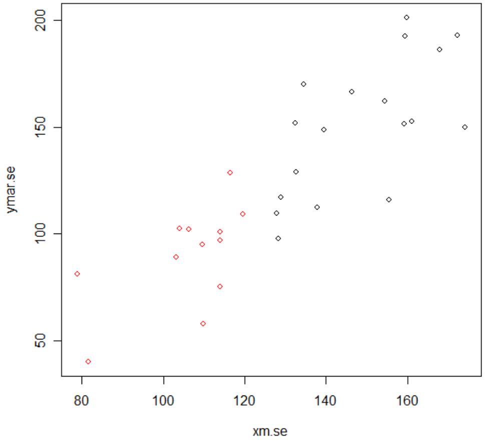
Knn – K nearest neighbor imputation (easy to implement)

- For every individual containing a missing value in a specific variable, we find another individual with minimal distance to the previous one and with complete information.
- Then transfer (copy) the value of the specific variable, of the second individual to the first one.

knn function in R



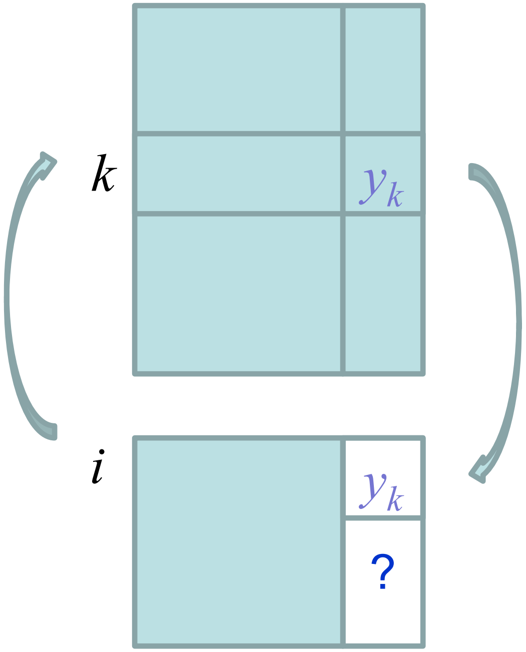
with only x as covariate



with x and many other covariates (age, BMI, sex, …)

Knn imputation

Complete data



Find the closest individual to i, according all variables except y

Copy the y_k value in the i individual

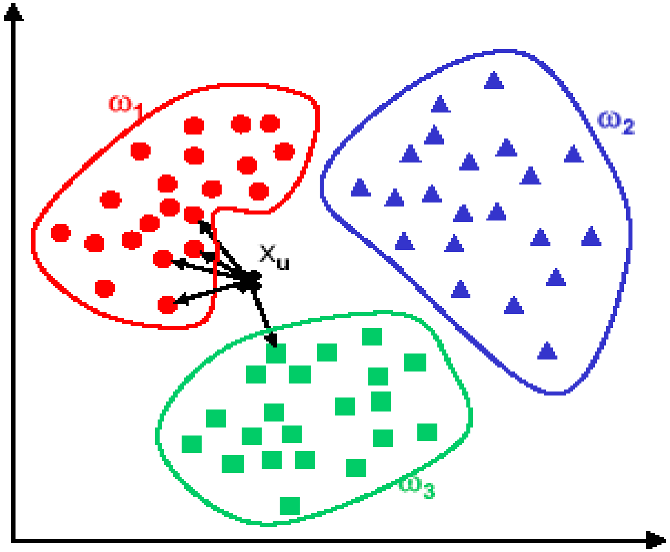
cases with y missing

MissMDA package in R:

- imputePCA(X) for numeric variables only.
- imputeMCA(X) for qualitative variables only.

Knn imputation: y qualitative variable

Find the closest individual to i, according all available variables except factor y



Find for each missing case, the most frequent category in the complete data set for closest neighbours.

Y_u ?

Easy to calculate in R

y_u category for u individual would be the red one – category 1

Data Quality report

- Per variable, count:
  - Number of missing values
  - Number of errors (including inconsistencies)
  - Number of outliers
  - Rank variables according the sum of missing values (and errors).
- Per individuals, count:
  - number of missing values
  - number of errors,
  - number of outliers
  - Create a new variable adding the total number missing values (and errors).
  - Describe this variable, *to which other variables exist higher associations*.
    - Compute the correlation with all other variables. Rank these variables according the correlation
    - Compute for every group of individuals (group of age, size of town, singles, married, …) the mean of missing values. Rank the groups according the computed mean.

Example: SwissLabor data in AER library

Usage

data("SwissLabor")

Format

A data frame containing 872 observations on 7 variables.

```
levels(SwissLabor$participation)<-
       paste("Parti.",sep="",levels(SwissLabor$participation))
levels(SwissLabor$foreign)<-
       paste("Foreign.",sep="",levels(SwissLabor$foreign))
```

participation	Factor. Did the individual participate in the labor force?
income	Logarithm of nonlabor income.
age	Age in decades (years divided by 10).
education	Years of formal education.
youngkids	Number of young children (under 7 years of age).
oldkids	Number of older children (over 7 years of age).
foreign	Factor. Is the individual a foreigner (i.e., not Swiss)?

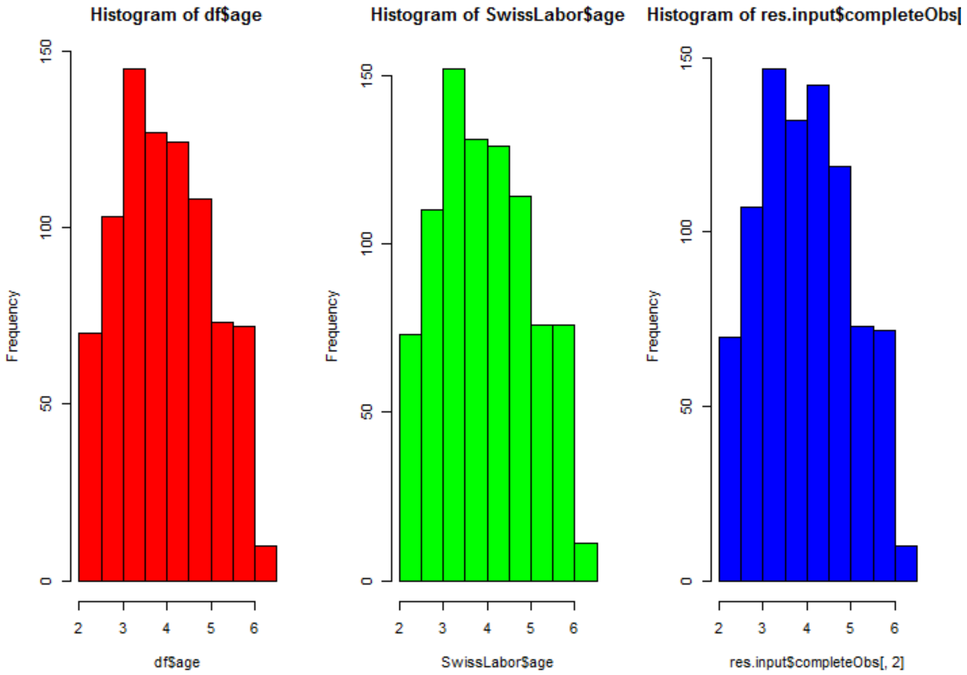
Example: SwissLabor in AER library - Imputation

```
> llista<-sample(1:nrow(SwissLabor),40);llista
> df<-SwissLabor
> df[llista,"age"]<-NA

> library(missMDA)
# Numeric imputation
> vars_con<-names(df)[2:6]
> summary(df[,vars_con])
> res.input<-imputePCA(df[,vars_con],ncp=4)
> summary(res.input$completeObs)

> par(mfrow=c(1,3))
> hist(df$age,col="red")
> hist(SwissLabor$age,col="green")
> hist(res.input$completeObs[,2],col="blue")

> quantile(df$age,seq(0,1,0.1),na.rm=T)
> quantile(SwissLabor$age,seq(0,1,0.1),na.rm=T)
> round(quantile(res.input$completeObs[,2],seq(0,1,0.1),na.rm=T),dig=1)
```

Example: SwissLabor data in AER library - Imputation

```
quantile(df$age,seq(0,1,0.1),na.rm=T)
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
2.0 2.6 3.0 3.3 3.6 3.9 4.3 4.6 5.0 5.5 6.2
quantile(SwissLabor$age,seq(0,1,0.1),na.rm=T)
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
2.0 2.6 3.0 3.3 3.6 3.9 4.3 4.6 5.0 5.5 6.2
round(quantile(res.input$completeObs[,2],seq(0,1,0.1),na.rm=T),dig=1)
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
2.0 2.6 3.0 3.3 3.6 4.0 4.3 4.6 4.9 5.5 6.2
```

```
>
```



Example: SwissLabor data in AER library - Imputation

R Code imputeMCA()

```
> llista<-
sample(1:nrow(SwissLabor),40);llista
> df<-SwissLabor
> df[llista,“participation"]<-NA

> library(missMDA)
# Categorical imputation
> vars_dis<-names(df)[c(1,7)]
> summary(df[,vars_dis])

> nb <- estim_ncpMCA(df[,
vars_dis],ncp.max=25)
> res.input<-imputeMCA(df[,vars_dis],ncp=10)
> summary(res.input$completeObs)
```

Results

- Check category frequences
- For the given example, with such a few factors, the example code does not work

Example: SwissLabor in AER library – mice imputation

R Code mice()

```
> llista<-sample(1:nrow(SwissLabor),
> df<-SwissLabor
> df[llista, c("foreign","age")] <- NA
> library(mice)
> # Imputation
> res.imp <- mice( df )
---
```

Results

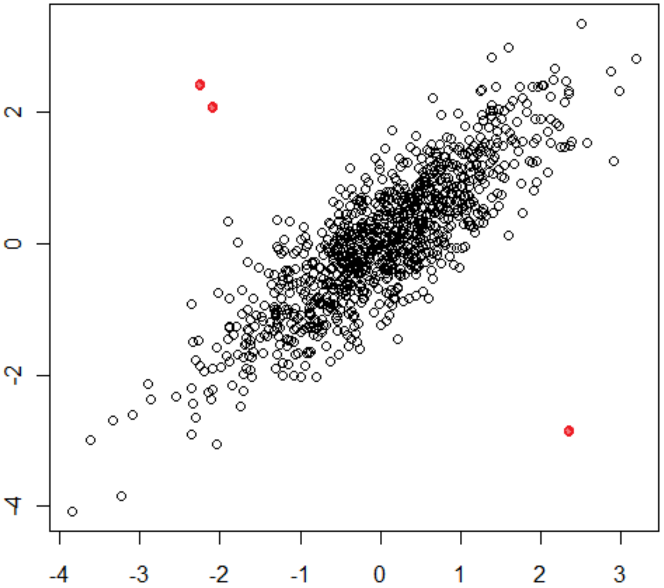
- Validate consistency of numeric values
- Validate imputed categories

```
> summary(complete(res.imp))
 participation     income            age          education        youngkids         oldkids       foreign        mout        
 no :471       Min.   : 7.187   Min.   :2.000   Min.   : 1.000   Min.   :0.0000   Min.   :0.0000   no :651   Min.   :0.00000  
 yes:401       1st Qu.:10.472   1st Qu.:3.200   1st Qu.: 8.000   1st Qu.:0.0000   1st Qu.:0.0000   yes:221   1st Qu.:0.00000  
               Median :10.643   Median :4.000   Median : 9.000   Median :0.0000   Median :1.0000             Median :0.00000  
               Mean   :10.686   Mean   :4.003   Mean   : 9.307   Mean   :0.3119   Mean   :0.9828             Mean   :0.01491  
               3rd Qu.:10.887   3rd Qu.:4.800   3rd Qu.:12.000   3rd Qu.:0.0000   3rd Qu.:2.0000             3rd Qu.:0.00000  
               Max.   :12.376   Max.   :6.200   Max.   :21.000   Max.   :3.0000   Max.   :6.0000             Max.   :1.00000  
```

Multivariate outliers

But outliers are multivariate

Univariate detection of outliers doesn't imply multivariate detection

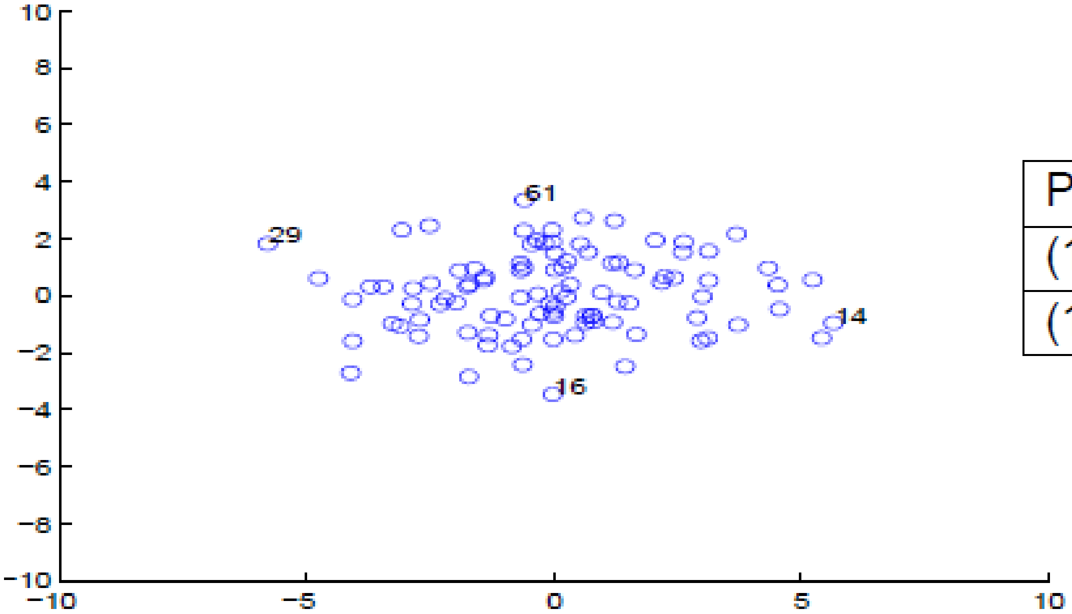


Then, detection of outliers is based in computing distances to the central point of data, by means an Iterative algorithm

$$D_M^2(i,G) = (x_i - G)' V^{-1} (x_i - G)$$

Mahalanobis distance

Mahalanobis Distance vs. Euclidean distance

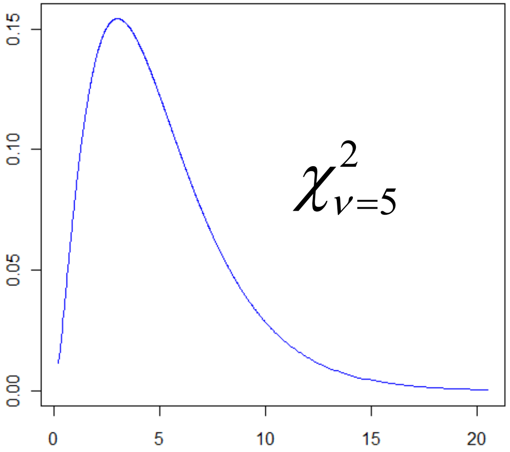


Point Pairs	Mahalanobis	Euclidean
(14,29)	5.07	11.78
(16,61)	4.83	6.84

If generating mechanism is Normal:

$$D_M^2(i,G) \sim \chi^2_{\nu=\text{dim space}}$$

Short distances occur more often



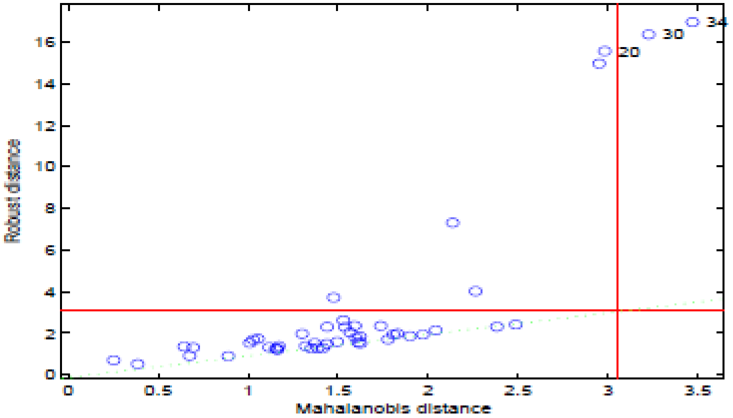
Detection of multivariate outliers

Take a value of h (size of data assumed not containing outliers), h must be $> p$ (number of variables). Usual values = $0.95n$ (at most 5% of outliers)

Initialization of an estimation of G and V : G = mean of variables. V = matrix of variances

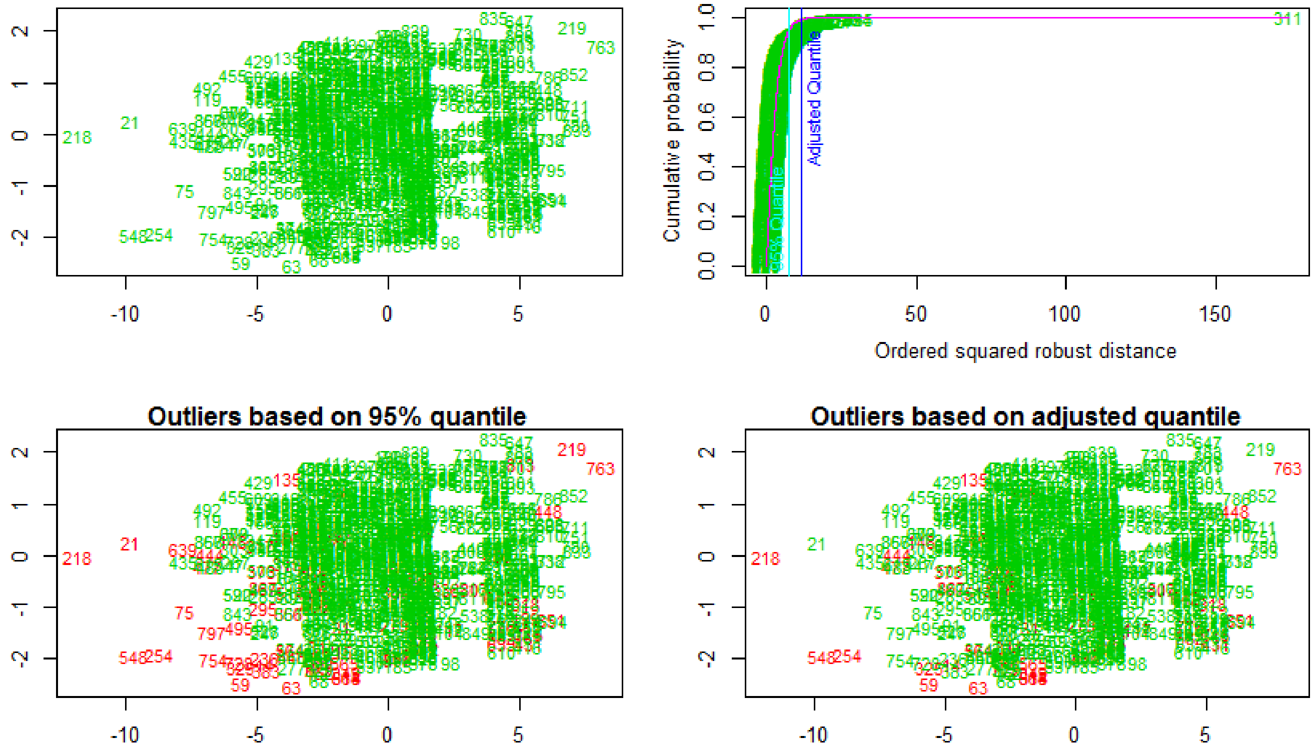
1. Compute the Mahalanobis distances $D^2_M(i,G)$ for each point i.
2. Rank the $D^2_M(i,G)$ and retain the h individuals with lower $D^2_M(i,G)$
3. Update G and V till convergence.

Plot the final "robustified" Mahalanobis distances with the initial Mahalanobis distances to detect the outliers **mvoutlier** library and method aq.plot()



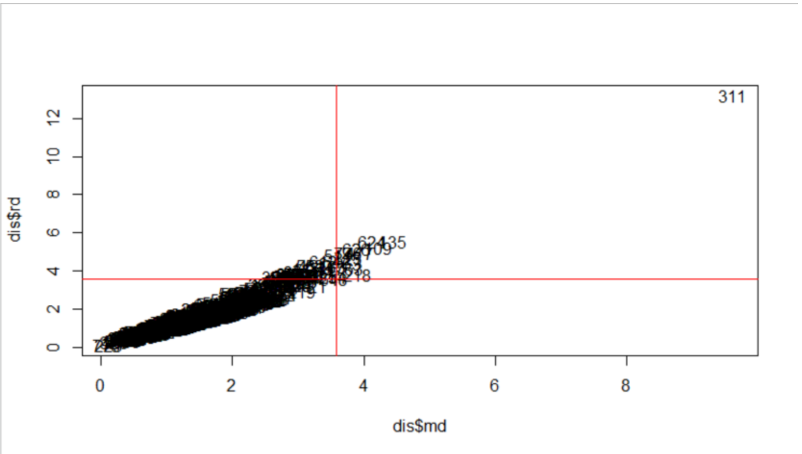
Example: SwissLabor data in AER library

```
library(mvoutlier)
vout<-aq.plot(SwissLabor[,2:4], delta=qchisq(0.95,
df=ncol(x)),alpha=0.05)
```



Example: SwissLabor data in AER library

```
> library(chemometrics)
> dis <- Moutlier(SwissLabor[,2:4], quantile = 0.995)
> plot(dis$md,dis$rd, type="n")
> text(dis$md,dis$rd,labels=rownames(SwissLabor[,2:4]))
> abline(h=qchisq(0.995, col(SwissLabor[,2:4])),col="red",lwd=2)
> str(dis) # List of 3
 $ md    : Named num [1:872] 1.20
  ..- attr(*, "names")= chr [1:87
 $ rd    : Named num [1:872] 1.20
  ..- attr(*, "names")= chr [1:87
 $ cutoff: num 3.58
> SwissLabor$mout<-0
> sel<-which((dis$rd>dis$cutoff)&(dis$md>dis$cutoff))
> SwissLabor[sel,"mout"]<-1
```



Role of variables

- **Response**

  Variables that we want to study, by building a model, finding associations, … (number of products bought, passing or failing a course, income, …)

  It can be either continuous or categorical

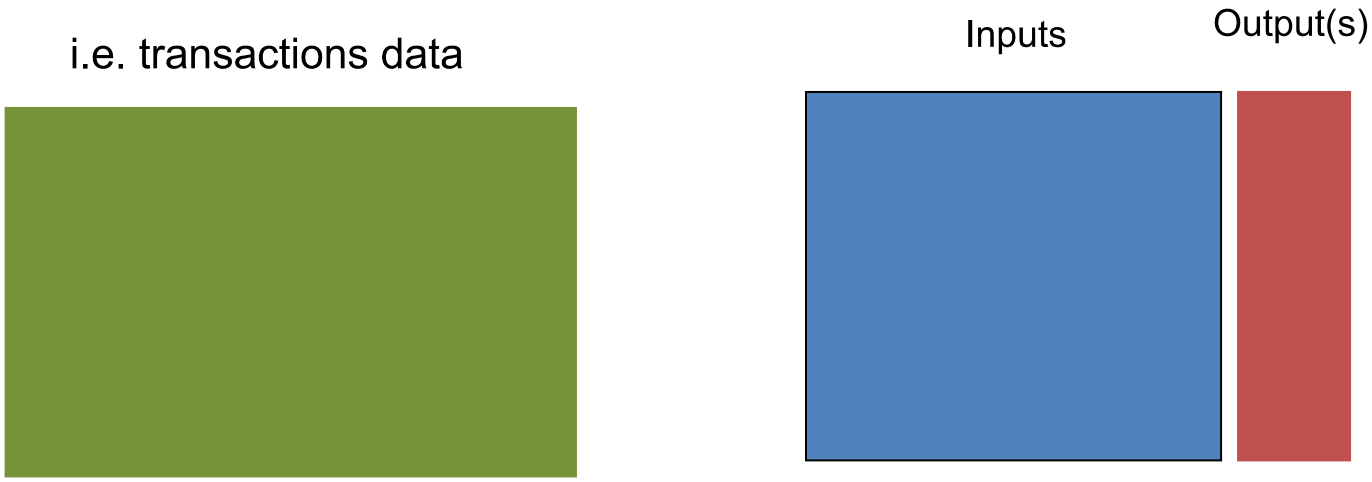
- **Explanatory**

  Variables which serve to explain the behaviour of the response variables (all the variables present in the data matrix except the response)

  They can be either continuous or categorical

Types or data matrix

With or without response(s) variable



Data to explore, to describe, to find associations (i.e. itemsets), …

Idem, but we want to **find a model to predict the response**

Paradigm

Any stored data from any process always contain information about the generating phenomenon (**statistical regularity**).

Goal: **To reveal the information** (model, patterns, associations, trends, clusters, … hidden in the data

Data are routinely stored (and most will never be analyzed)

Data is a treasure for organizations (be aware of the data quality)

Any transactional process con be enhanced by analysis of its collected data

How? *Selecting and reporting what is interesting*

SQL queries are NOT ENOUGH. How many A products sold last month?.

Profiling. What is the profile of A buyers? *Automatic detection of significant deviations*

Automatic profiling of groups of individuals

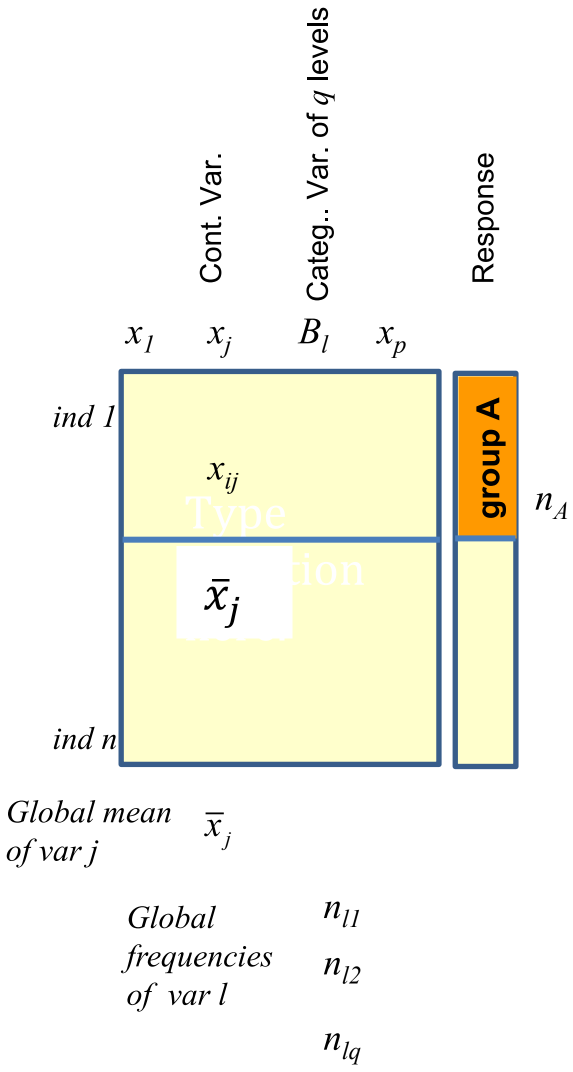
We have a group of individuals defined **by a level of a categorical variable (target).**

Problem: For every group of individuals detect which other groups of individuals (identified by the levels of the explanatory variables) or what continues variables, deviate significantly from what were expected.

- We take as response variable the variable identifying the groups that we want to find their profile.
- The explanatory variables are either categorical or continuous.

Tool: Hypothesis test

- For each group to profile, rank the modalities of the categorical explanatory variables according their p-value (ascending). Likewise, rank the continuous variables according their p-value
- Select the most significant by a threshold (0.05, 0.01, ..) defined a priori. (what matters is the ordering, actual significance depends on the number of individuals)



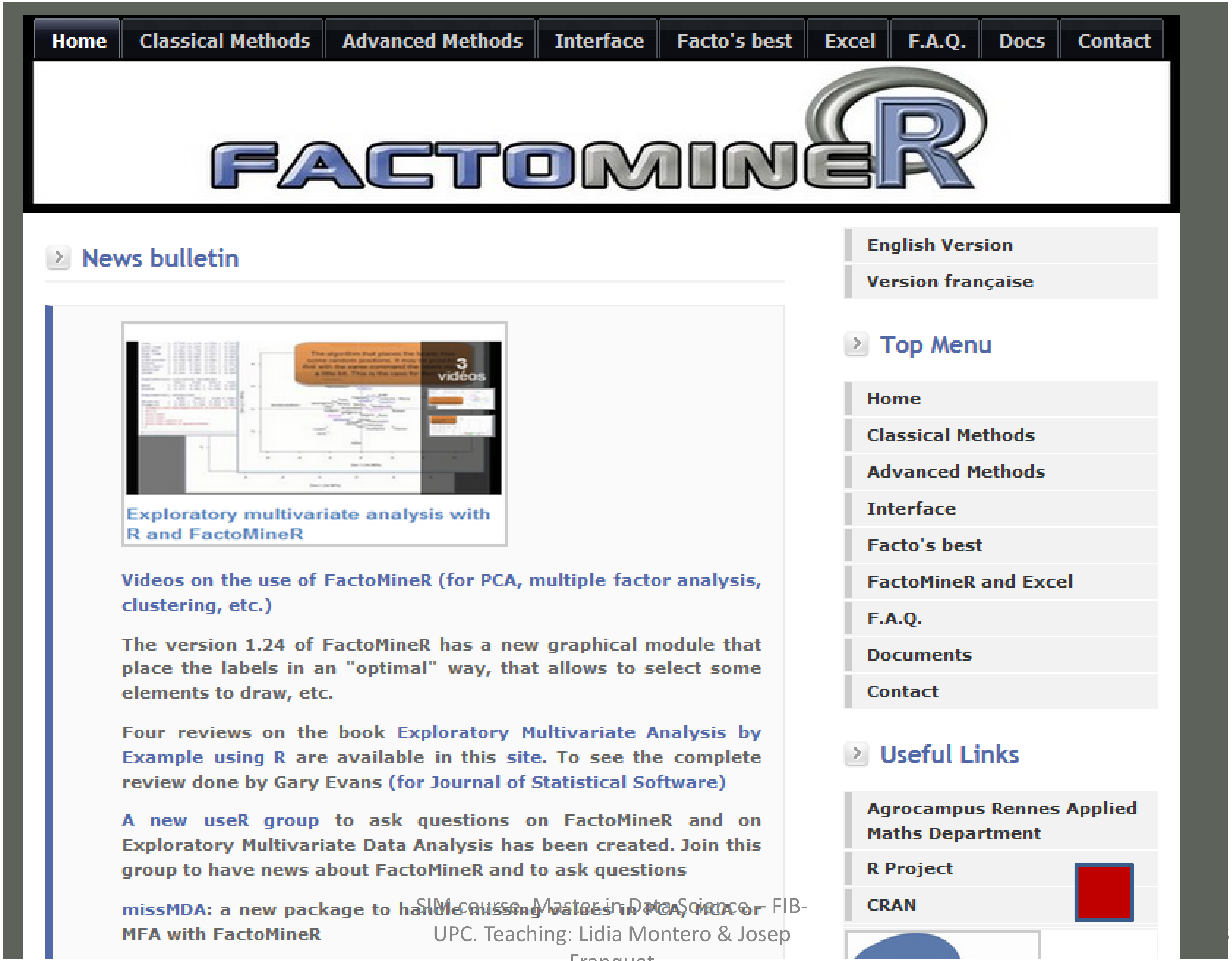
R function available in FactoMineR

We will use FactoMineR Package (cran R)

You can also consult (and download this R function from) http://factominer.free.fr/ where a large documentation is provided, with theoretical background, examples, tutorials and so on.

The functions of this package corresponding to this sessions are:

- **Catdes:** description of the categories of a categorical variable by quantitative variables, categorical variables and categories
- **Condes:** description of a quantitative variable by quantitative and categorical variables



News bulletin

Exploratory multivariate analysis with R and FactoMineR

Videos on the use of FactoMineR (for PCA, multiple factor analysis, clustering, etc.)

The version 1.24 of FactoMineR has a new graphical module that place the labels in an "optimal" way, that allows to select some elements to draw, etc.

Four reviews on the book Exploratory Multivariate Analysis by Example using R are available in this site. To see the complete review done by Gary Evans (for Journal of Statistical Software)

A new useR group to ask questions on FactoMineR and on Exploratory Multivariate Data Analysis has been created. Join this group to have news about FactoMineR and to ask questions

missMDA: a new package to handle missing values in PCA, MCA or MFA with FactoMineR

Response target: factor (B)
Explanatory variable: numerical (X)

B ~ X

Consider background for X ~ B

Profiling a categorical target from a continuous variable

$groups$	$means$	$counts$
1	$\bar{x}_1$	n_1
$\vdots$	$\vdots$	$\vdots$
p	$\bar{x}_p$	n_p
$Global$	$\bar{x}$	n

$$H_0\colon \mu_1 = \cdots = \mu_p = \mu$$

Null Hypothesis: All group means are equal to the global mean

Ronald Fisher 1890, 1962

In R:

- Assuming normal distribution on X: oneway.test(X~B).
- Without normality assumption (non –parametric test): Kruskal-Wallis test
  kruskal.test(X~B)
- Global association: Tested using a F-Fisher based-test

Profiling target categorical variables from continuous variables

$$H_0 : \mu_k = \mu \quad k = 1, \ldots, p$$

$groups$	$means$	$counts$
1	$\overline{x}_1$	n_1
$\vdots$	$\vdots$	$\vdots$
p	$\overline{x}_p$	n_p
$Global$	$\overline{x}$	n

Test statistic: Difference between the mean in group *k* and the global mean. T-Student based-test

Highlight groups with a significant different mean: level specific association tests

$$t = \frac{\overline{x}_k - \overline{x}}{\sqrt{(1 - \frac{n_k}{n})\frac{s^2}{n_k}}} \square\ t_{n-1}$$

Student's *t*

William Gosset "Student", English, 1876, 1937

Rank the continuous variables by p.value (ascending)

Function to compute p-values for profiling a categorical target from continuous variables – Globally and Specific Level

To Rank variables and groups according to pvalues:

```
p.xk <- function(vec,fac)  {
       nk <- as.vector(table(fac));
       n <- sum(nk);
       xk <- tapply(vec,fac,mean);
       txk <- (xk-mean(vec))/(sd(vec)*sqrt((n-
nk)/(n*nk)));
       pxk <- pt(txk,n-1,lower.tail=F)}
```

Rank the continuous variables by p.value (ascending)

FactoMineR solution:

- catdes(data.frame,num.var): sections
  - ➢ Link between the cluster variable and the quantitative variables
  - ➢ Description of each cluster by quantitative variables

Response target: factor (B)
Explanatory variable: factor (A)

B ~ A

Profiling categories from categorical variables

$$\begin{array}{c|ccc|c}
 & 1\ldots & j & \ldots q & \\ \hline
1 & & \vdots & & \\
k & \cdots & n_{kj} & \cdots & n_k \\
p & & \vdots & & \\ \hline
 & & n_j & &
\end{array}$$

– **Global Relationship between each category** of the target variable and other categorical variables: **a chi-square-test is performed**

– **Relationship between each category** of the variable target and each category of another categorical variable: comparison of two proportions, taking into account an hypergeometric model and normal approximations

–Descriptive tools: contingency tables (numeric) and mosaic plot (graphical)

Rank the levels of the categorical explanatory variables/ the categories by p-value (ascending)

(Global) Characterization of a target categorical variable by the other categorical variables

Test

Null hypothesis H_0: conservative hypothesis. Both variables are independent

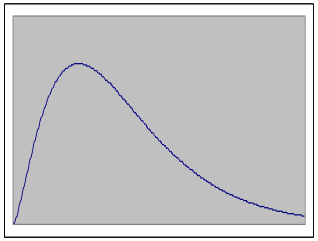
Alternative hypothesis H_1: Both variables are not independent

Test statistic:

$$\chi^2_{obs} = \sum_i \sum_j \frac{\left(n_{ij} - {n_{i.} n_{.j}}/{n}\right)^2}{{n_{i.} n_{.j}}/{n}} = \\ = \sum_i \sum_j \frac{\left(n_{ij} - n p_{i.} p_{.j}\right)^2}{n p_{i.} p_{.j}}$$

Reference distribution: Distribution of the *test statistic* under H_0 (that is, if H_0 is true).

Chi-2 distribution, with the convenient degrees of freedom



Significance threshold: Risk of rejecting H_0 although H_0 being true

(significance depends on the number of individuals) P-value

Characterization of a <u>target categorical variable</u> by the levels of other categorical variables

$$\begin{array}{c|ccc|c} & 1\ldots & j & \ldots q & \\ \hline 1 & & \vdots & & \\ k & \cdots & n_{kj} & \cdots & n_k \\ p & & \vdots & & \\ \hline & & n_j & & \end{array}$$

$$\boxed{H_0 : p_{j/k} = p_j \quad k = 1,\ldots,p;\, j = 1,\ldots,q}$$

Assumption of normality of proportions:

$$\frac{n_{kj}}{n_k} \sim N\left(p_j = \frac{n_j}{n}, \left(1 - \frac{n_k}{n}\right)\frac{p_j(1-p_j)}{n_k} \right)$$

Test statistic: Difference between proportion of modality *j* in group *k* and proportion of modality *j* in whole data

$$z = \frac{\dfrac{n_{kj}}{n_k} - \dfrac{n_j}{n}}{\sqrt{\left(1 - \dfrac{n_k}{n}\right)\left(\dfrac{p_j\left(1-p_j\right)}{n_k}\right)}} \sim N(0,1)$$

Rank the levels of the categorical explanatory variables by p.value (ascending)

R function to compute the table of p-values of a categorical variable

```
p.zkj <- function(res,expl){
   taula <- table(res,expl)
   n <- sum(taula);
   pk <- apply(taula,1,sum)/n;
   pj <- apply(taula,2,sum)/n;
   pf <- taula/(n*pk);
   pjm <- matrix(data=pj,nrow=nrow(pf),ncol=ncol(pf), byrow=T);
   dpf <- pf - pjm;
   dvt <- sqrt(((1-pk)/(n*pk))%*%t(pj*(1-pj)));
   zkj <- dpf/dvt;
   pzkj <- pnorm(zkj,lower.tail=F);
list(rowpf=pf,vtest=zkj,pval=pzkj)}
```

FactoMineR solution:

- catdes(data.frame,num.var)
  - ➢ Link between the cluster variable and the categorical variables (chi-square test)
  - ➢ Description of each cluster by categories

Example: SwissLabor data in AER library

Usage

data("SwissLabor")

Format

A data frame containing 872 observations on 7 variables.

```
levels(SwissLabor$participation)<-
       paste("Parti.",sep="",levels(SwissLabor$participation))
levels(SwissLabor$foreign)<-
       paste("Foreign.",sep="",levels(SwissLabor$foreign))
```

participation	Factor. Did the individual participate in the labor force?
income	Logarithm of nonlabor income.
age	Age in decades (years divided by 10).
education	Years of formal education.
youngkids	Number of young children (under 7 years of age).
oldkids	Number of older children (over 7 years of age).
foreign	Factor. Is the individual a foreigner (i.e., not Swiss)?

Profiling a categorical target by the categories of the other categorical variables

In SwissLabor dataset in library(AER): Participation –Yes (target) vs Foreign

H_0: The category “foreign=NO” is neither infra nor supra represented

H_1: The category “foreign=NO” is infra (versus supra) represented

```
>table(SwissLabor$foreign,SwissLabor$participation)
               Target-no Target-yes
Foreign-no     402      254
Foreign-yes    69       147
➢prop.table(table(SwissLabor$foreign,SwissLabor$par
ticipation),1)
               Target-no Target-yes
Foreign-no     0.6128049 0.3871951
Foreign-yes    0.3194444 0.6805556
>
prop.table(table(SwissLabor$foreign,SwissLabor$part
icipation),2)
               Target-no Target-yes
Foreign-no     0.8535032 0.6334165
Foreign-yes    0.1464968 0.3665835
```

```
round(prop.table
(table(SwissLabor$foreign)),dig=
2)
Foreign.no Foreign.yes
    0.75        0.25
```

Profiling a categorical target by other categorical variables

In SwissLabor dataset in library(AER): Participation –Yes (target) vs Foreign

H_0: Global Association is not present among Target and Explanatory Factor

H_1: Global Association is present

```
>table(df$foreign,df$participation)
               Target-no Target-yes
Foreign-no     402       254
Foreign-yes    69        147
```

```
Under H0
       Target-no Target-yes
 f.For-no   354.3303 301.66972
 f.For-yes 116.6697   99.33028
```

```
> chisq.test(table(df$foreign,df$participation))
Pearson's Chi-squared test with Yates' continuity correction
data:  table(df$foreign, df$participation)
X-squared = 55.126, df = 1, p-value = 1.131e-13

> res.cat$test.chi2 # Global association target is factor and
explanatory factors - catdes()
             p.value df
foreign 6.220116e-14  1
```

Characterization of a categorical variable by the categories of the other categorical variables

In SwissLabor dataset in library(AER): Participation –Yes (target) vs Foreign

H_0: The category “foreign=NO” is neither infra nor supra represented

H_1: The category “foreign=NO” is infra (versus supra) represented

68% of class Foreign-Yes belongs to Category Target-Yes P([B-TargetYes]/[A-Foreign-Yes])

36.7% of Category Target-Yes belongs to class Foreign-Yes : P ([A-Foreign-Yes] /[B-TargetYes])

```
$category$`Target-yes`
                       Cla/Mod  Mod/Cla   Global      p.value   v.test
foreign=Foreign-yes   68.05556 36.65835 24.77064 5.591005e-14  7.517321
foreign=Foreign-no    38.71951 63.34165 75.22936 5.591005e-14 -7.517321
```

```
prop.table(table(SwissLabor$foreign))
Foreign-no Foreign-yes
0.7522936  0.2477064
```

Foreign women represent 25% of the sample, but 36.7% in the target class Target-Yes

Characterization of a categorical variable by a quantitative variable

Characterization of the categorical variable "*target*" by the quantitative variables

$$Y_{ki} = \mu + \alpha_k + \varepsilon_{ki}$$

target (level *k*)= grand mean + effect for *k* level + error

H_0 (no category effect): $\alpha_1 = \ldots = \alpha_k = \ldots = \alpha_K = 0$
H_1: There are at least two "target" levels *k* and *k'* such as: $\alpha_k \neq \alpha_{k'}$

```
> res.cat$quanti.var  # Global association target is factor
and explanatory variables numeric
                 Eta2      P-value
youngkids 0.029968826 2.695567e-07
income    0.029891180 2.794460e-07
education 0.010516854 2.429641e-03
age       0.008521288 6.375401e-03
oldkids   0.006445786 1.772877e-02
```

Profiling categories from quantitative variables

Characterization of the categories of "*target*" by the quantitative variables

H_0 mean of the variable in the category= grand mean

H_1: mean in the category≠global mean

```
> res.cat$quanti # Especific association: target factor and numeric variables
$`f.Par-no`
             v.test Mean in category Overall mean sd in category Overall sd      p.value
youngkids  5.109095        0.4097665    0.3119266      0.6770660  0.6125185 3.237063e-07
income     5.102472       10.7513327   10.6855675      0.4351131  0.4122522 3.352458e-07
education  3.026579        9.5944798    9.3073394      2.8484531  3.0345172 2.473382e-03
age        2.724342        4.0853503    3.9955275      1.1599921  1.0545623 6.442967e-03
oldkids   -2.369447        0.9023355    0.9827982      1.0622927  1.0861630 1.781471e-02

$`f.Par-yes`
             v.test Mean in category Overall mean sd in category Overall sd      p.value
oldkids    2.369447        1.0773067    0.9827982      1.1060968  1.0861630 1.781471e-02
age       -2.724342        3.8900249    3.9955275      0.9040227  1.0545623 6.442967e-03
education -3.026579        8.9700748    9.3073394      3.2067731  3.0345172 2.473382e-03
income    -5.102472       10.6083220   10.6855675      0.3689875  0.4122522 3.352458e-07
youngkids -5.109095        0.1970075    0.3119266      0.5029499  0.6125185 3.237063e-07
```

Response target: Numeric (Y)
Explanatory variable: Numeric (X)
Explanatory variable: factor (A)

Y ~X

Y ~ A

- Target type: numeric or factor
- Type of explanatory variates
  - Global association (target, explanatory variate)
  - Specific association (target, explanatory variate)
- FactoMineR:
  - Target is a factor: catdes()
  - Target is numeric: condes()

Profiling a quantitative target from quantitative or categorical variables

– **Description by quantitative variables (condes) : global association**

Correlation (Pearson)-> cor(*data.frame*)

– **Description by categorical variables and categories**

ANOVA: test F (global association) and t-tests (level specific associations)

Response target: Numeric (Y)
Explanatory variable: Numeric (X)

Y ~ X

Relationship between a quantitative target and the other quantitative variables

H_0) no relationship (correlation is null ρ=0)
H_1) relationship (correlation non null $\rho \neq 0$)

Statistics:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Pearson linear correlation

```
> condes(SwissLabor,2) #Numeric target income
$quanti
            correlation      p.value
education     0.3273458 3.166132e-23
oldkids       0.1391036 3.758541e-05
```

Response target: Numeric (Y)
Explanatory variable: Factor (A)

Y ~ A

Global Relationship between the quantitative target "income" and the categorical variables

income (foreign group *k*)= mean + effect of foreign group *k* + error

$$Y_{ki} = \mu + \alpha_k + \varepsilon_{ki}$$

H_0 (no category effect): $\alpha_1=\ldots=\alpha_k=\ldots=\alpha_K=0$
H_1: There are at least two "factor" levels *k* and *k'* such as: $\alpha_k \neq \alpha_{k'}$

Global association Fisher F-Based

```
> condes(SwissLabor,2) #Numeric target income
…

$quali
                       R2      p.value
foreign       0.04389655 4.170824e-10
participation 0.02989118 2.794460e-07
```

Relationship between the quantitative variable "income" and the levels of categorical variables

H_0 The coefficient of category *k* is null $\alpha_k = 0$

H_1: The coefficient of category *k* is non-null $\alpha_k \neq 0$

Level specific association t-Student based tests – Only significant levels included in the output

```
> condes(SwissLabor,2) #Numeric target income

$category
               Estimate      p.value
Foreign.no   0.10004281 4.170824e-10
Parti.no     0.07150532 2.794460e-07
Parti.yes   -0.07150532 2.794460e-07
Foreign.yes -0.10004281 4.170824e-10
```

Example: SwissLabor data in AER library

Usage

```
data("SwissLabor")
```

Format

A data frame containing 872 observations on 7 variables.

participation	Factor. Did the individual participate in the labor force?
income	Logarithm of nonlabor income.
age	Age in decades (years divided by 10).
education	Years of formal education.
youngkids	Number of young children (under 7 years of age).
oldkids	Number of older children (over 7 years of age).
foreign	Factor. Is the individual a foreigner (i.e., not Swiss)?

Example: SwissLabor data in AER library

```
> condes(SwissLabor,2) #Numeric target income
$quanti
            correlation      p.value
education     0.3273458 3.166132e-23
oldkids       0.1391036 3.758541e-05

$quali
                       R2      p.value
foreign        0.04389655 4.170824e-10
participation 0.02989118 2.794460e-07

$category
                 Estimate      p.value
Foreign.no     0.10004281 4.170824e-10
Parti.no       0.07150532 2.794460e-07
Parti.yes     -0.07150532 2.794460e-07
Foreign.yes   -0.10004281 4.170824e-10
```

Global association

Global association

Profiling on categories