Session 2:
Data Quality

Anàlisi de Dades i Explotació de la Informació

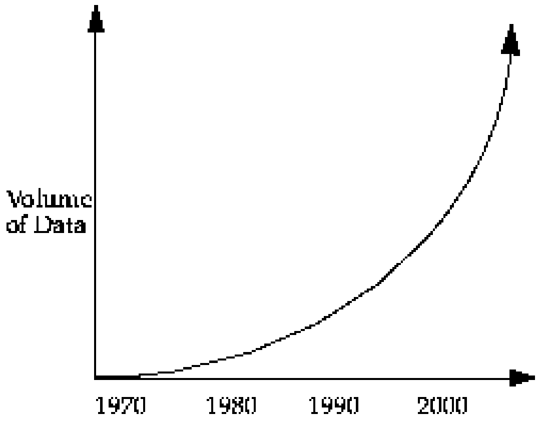
Grau d'Enginyeria Informatica.
Information System track

Prof. Lídia Montero
lidia.montero@upc.edu

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, …

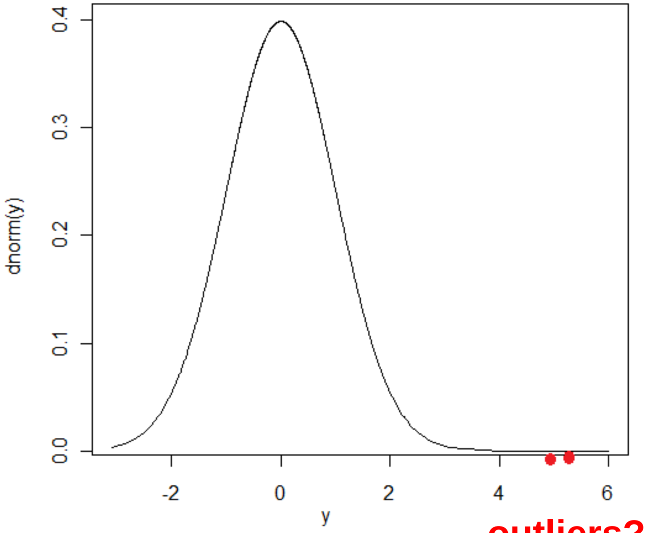
From a statistical point of view, we can only treat outliers and missing data

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??

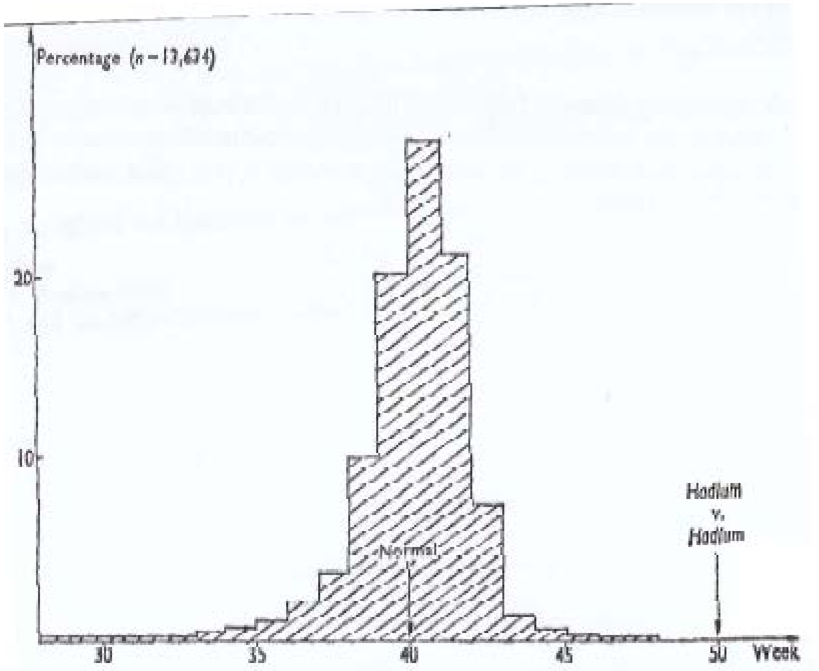
if X~N(0,1)	Prob(x≥X)
1	0.1586553
2	0.02275013
3	0.001349898
4	3.167124e-05
5	2.866516e-07

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.



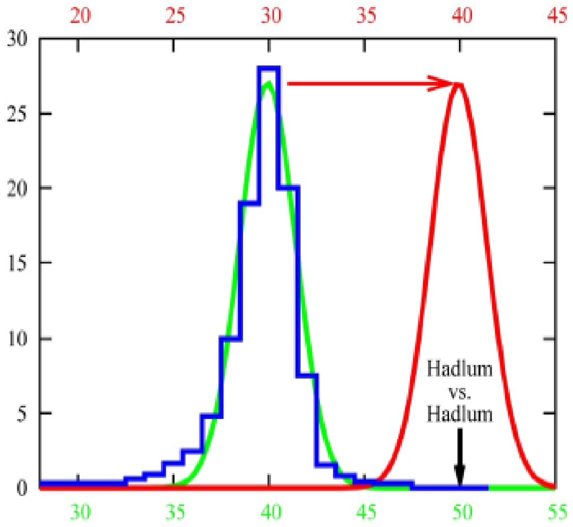
Distribution of human gestation periods.

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

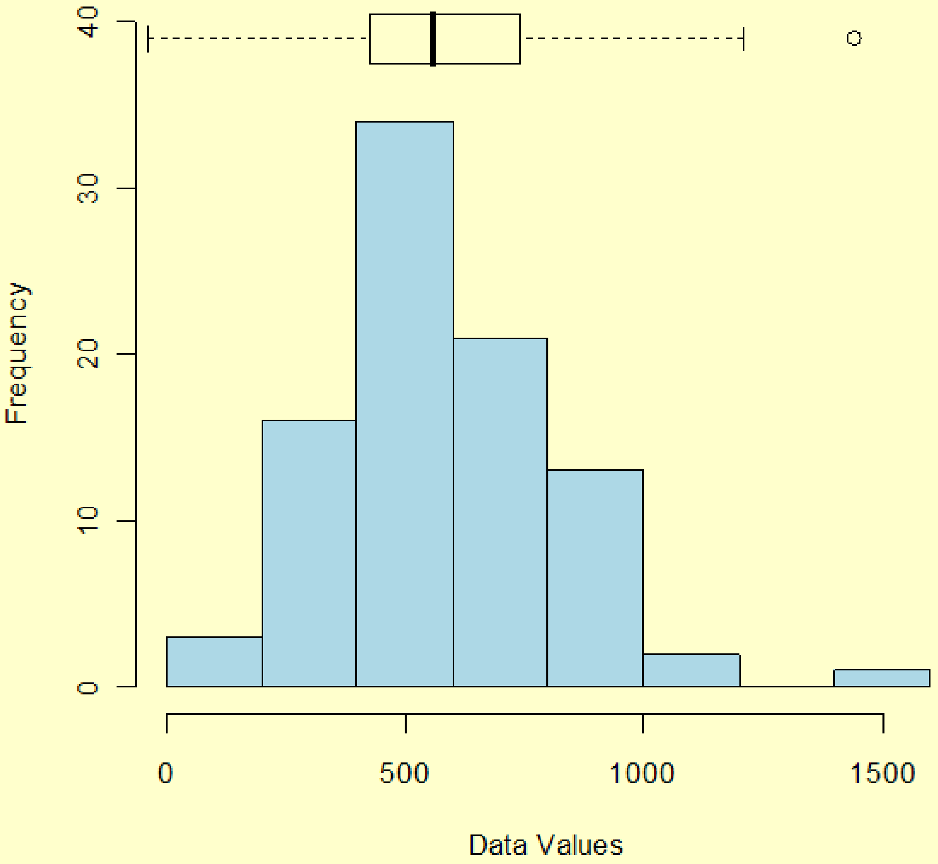
red: assumption of Mr. Hadlum: Another Gaussian process responsible for the observed birth, where the gestation period starts later. Under this assumption the specific birthday has highest-probability.



Example of an Outlier in data

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

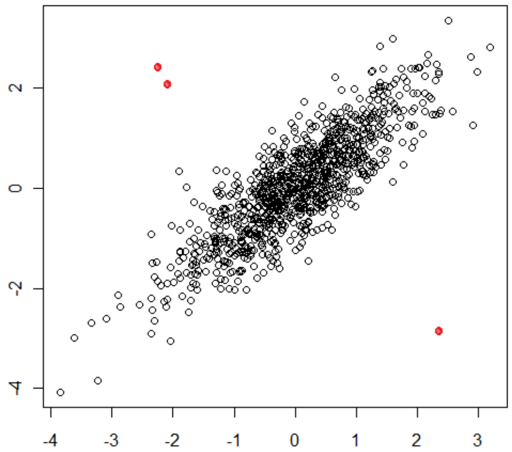
30, 171, 184, 201, 212, 250, 265, 270, 272, 289, 305, 306, 322, 322, 336, 346, 351, 370, 390, 404, 409, 411, 436, 437, 439, 441, 444, 448, 451, 453, 470, 480, 482, 487, 494, 495, 499, 503, 514, 521, 522, 527, 548, 550, 559, 560, 570, 572, 574, 578, 585, 592, 592, 607, 616, 618, 621, 629, 637, 638, 640, 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

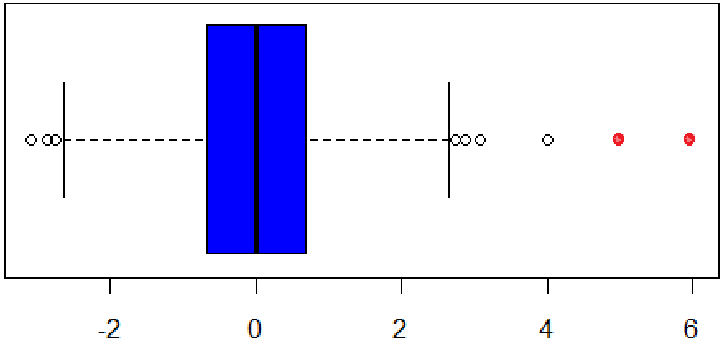
Univariate detection of outliers doesn't imply multivariate detection

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

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.

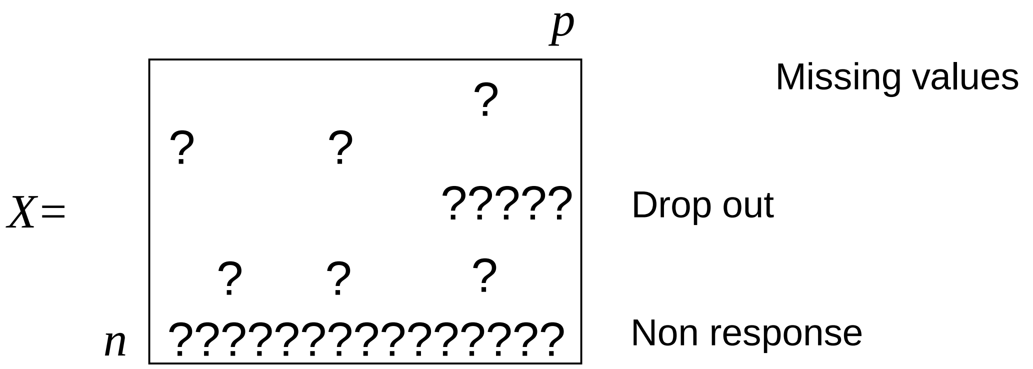
- Once we have detected outliers, what we should do?
  - Eliminate them (but we loose information of the eliminated individuals) and deleting outliers is not the best solution, since outliers are recursive.
  - Weight the individuals inversely to outlying degree of individuals, to diminish its importance (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.



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)

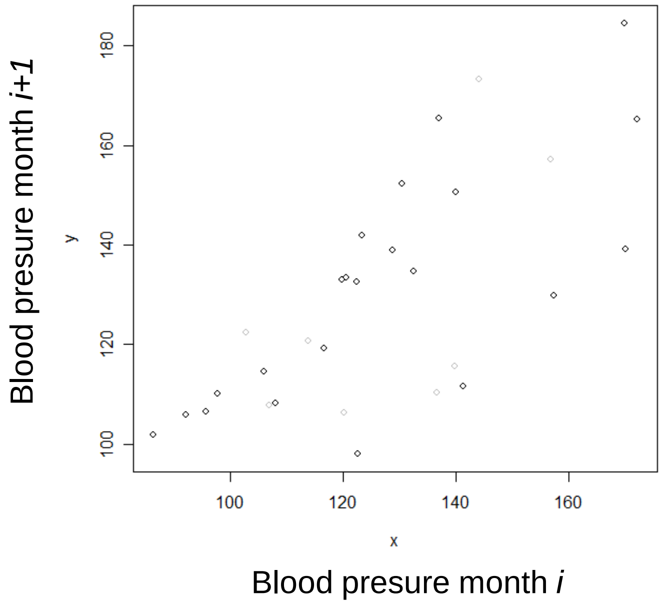
Missingness mechanisms

- 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.

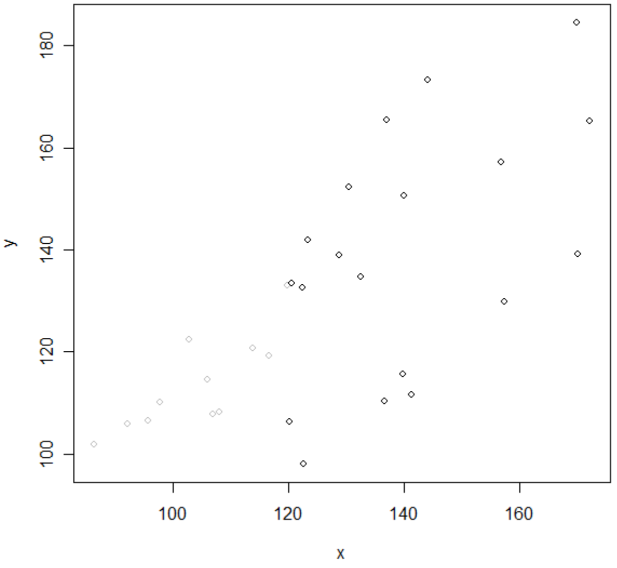
- MAR - At random: missing values appear related to third observed variables. This is the most usual case, i.e. asking the income of individuals, income is missing but can be imputed from the educational level.

- 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



Complete data



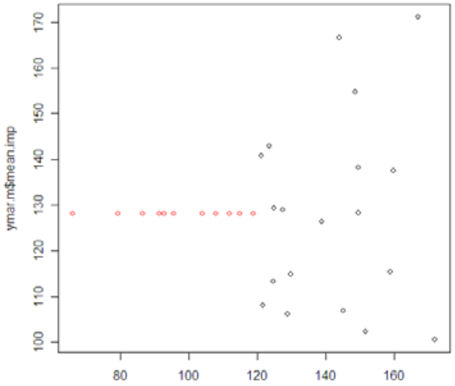
Data with missing values

Treatment of missing values

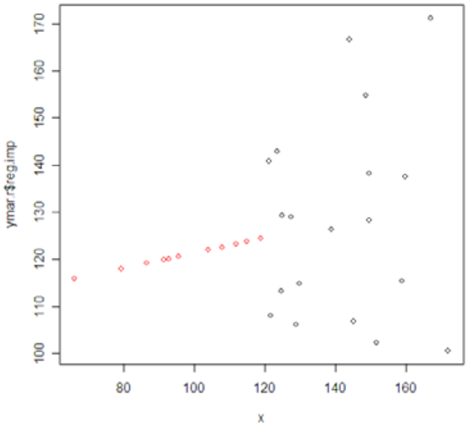
Traditional methods

- **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

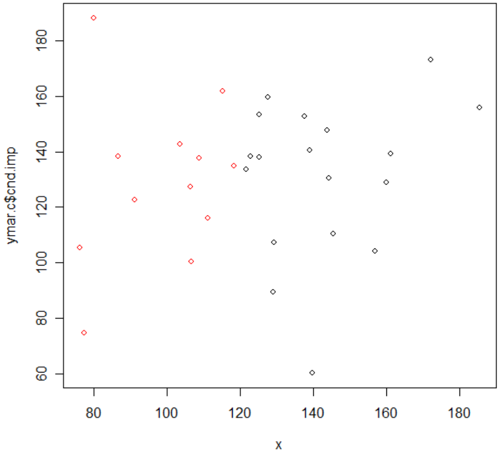
- Stochastic imputation (imputare = to fill in)

Simulate actual data

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

Stochastic regression imputation

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

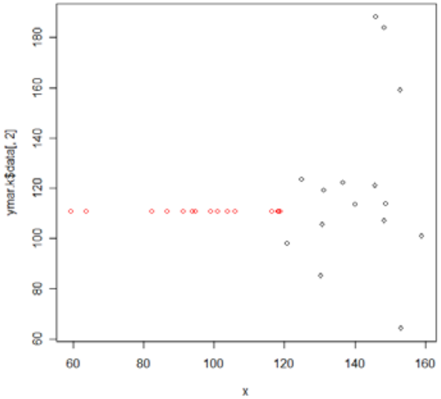


Treatment of missing values

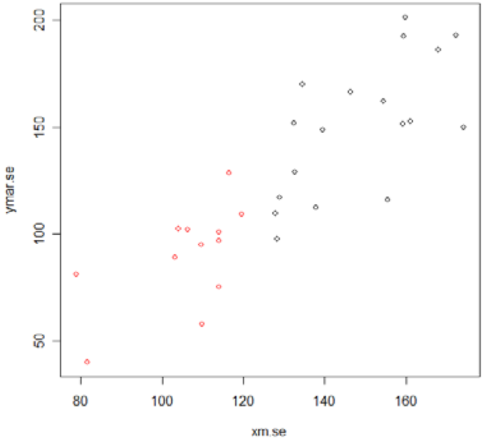
Knn – K nearest neighbor imputation (easy to implement)

- 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

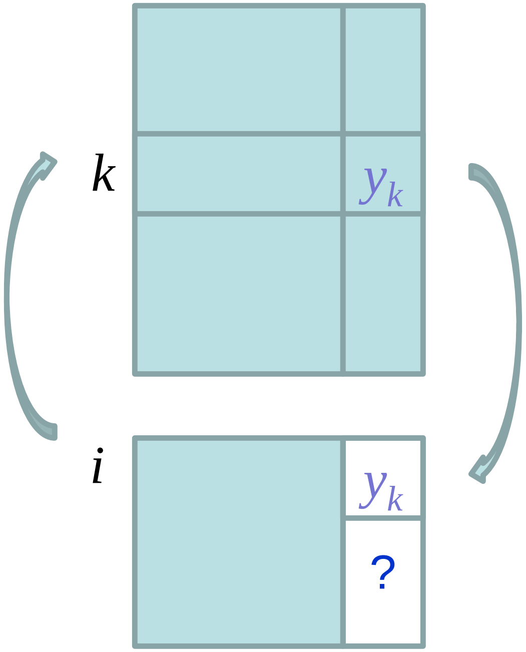


with x and many other covariates (age, BMI, sex, …)

Knn imputation

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

Complete data

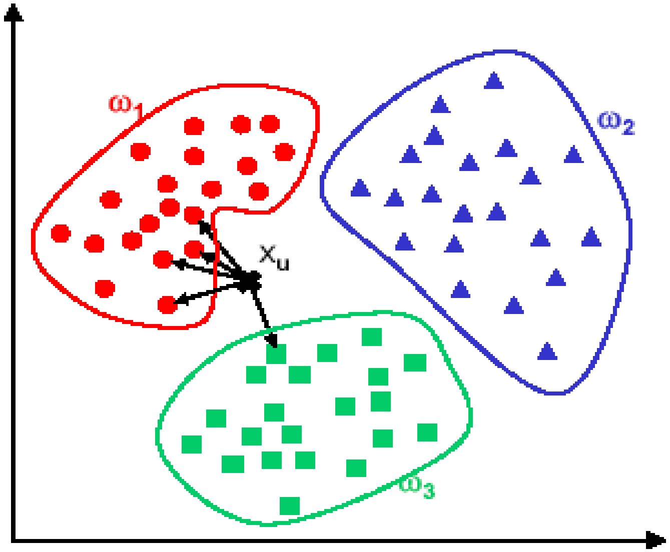


cases with y missing

Copy the y_k value in the i individual

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

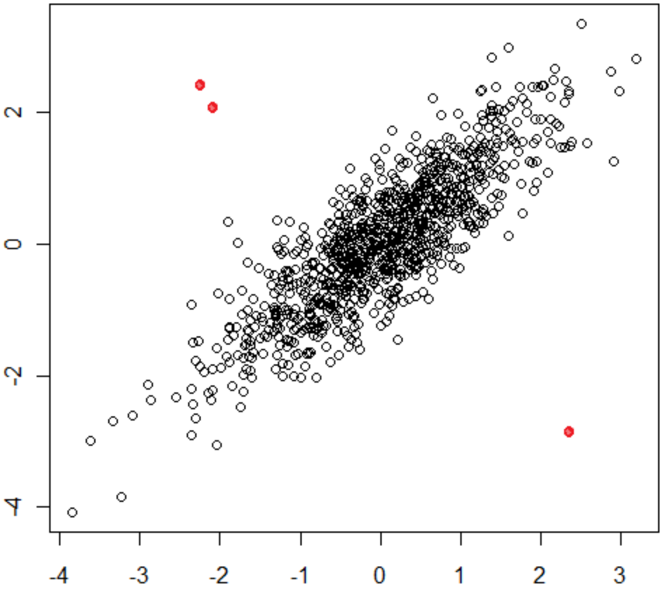
Data Quality report

- 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.

Multivariate outliers

But outliers are multivariate

Univariate detection of outliers doesn't imply multivariate detection

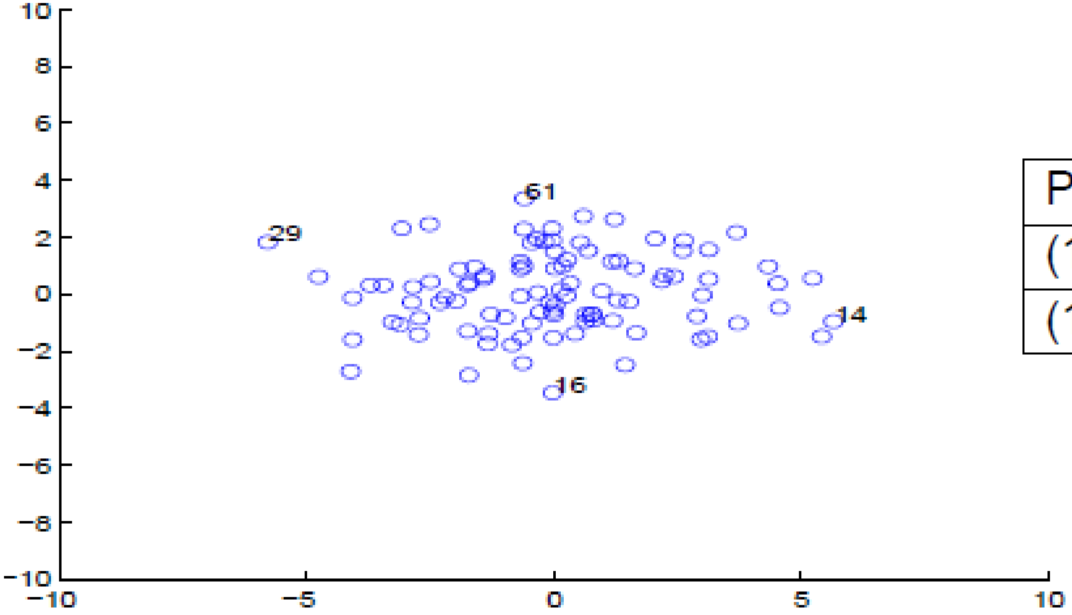


Then, detection of outliers is based in computing distances to the central point of data, by means an Iterative algorithm

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

Mahalanobis distance

Mahalanobis Distance vs. Euclidean distance

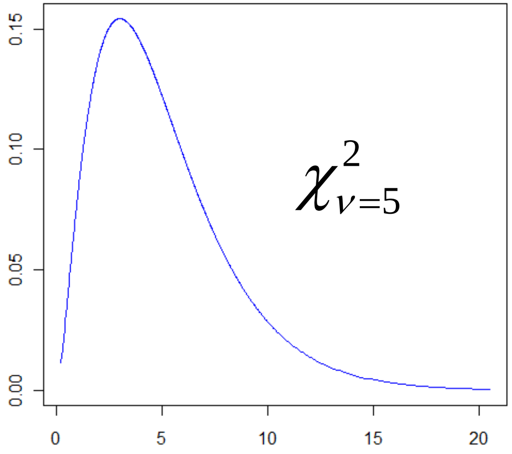


Point Pairs	Mahalanobis	Euclidean
(14,29)	5.07	11.78
(16,61)	4.83	6.84

If generating mechanism is Normal:

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

Short distances occur more often



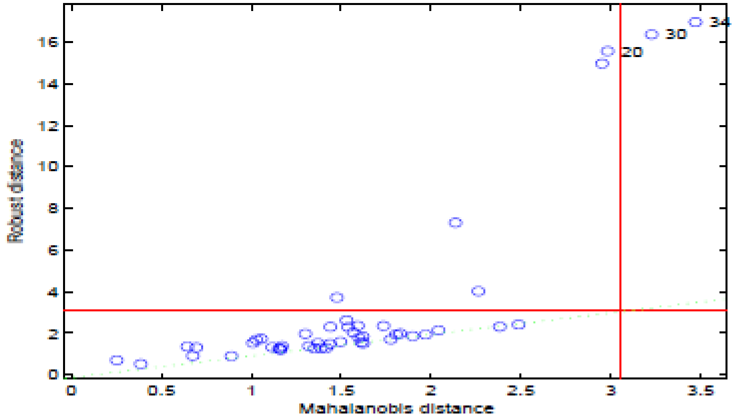
Detection of multivariate outliers

Take a value of h (size of data assumed not containing outliers), h must be $> p$ (number of variables). Usual values = $0.75n$ (at most 25% 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

Usage

data("SwissLabor")

Format

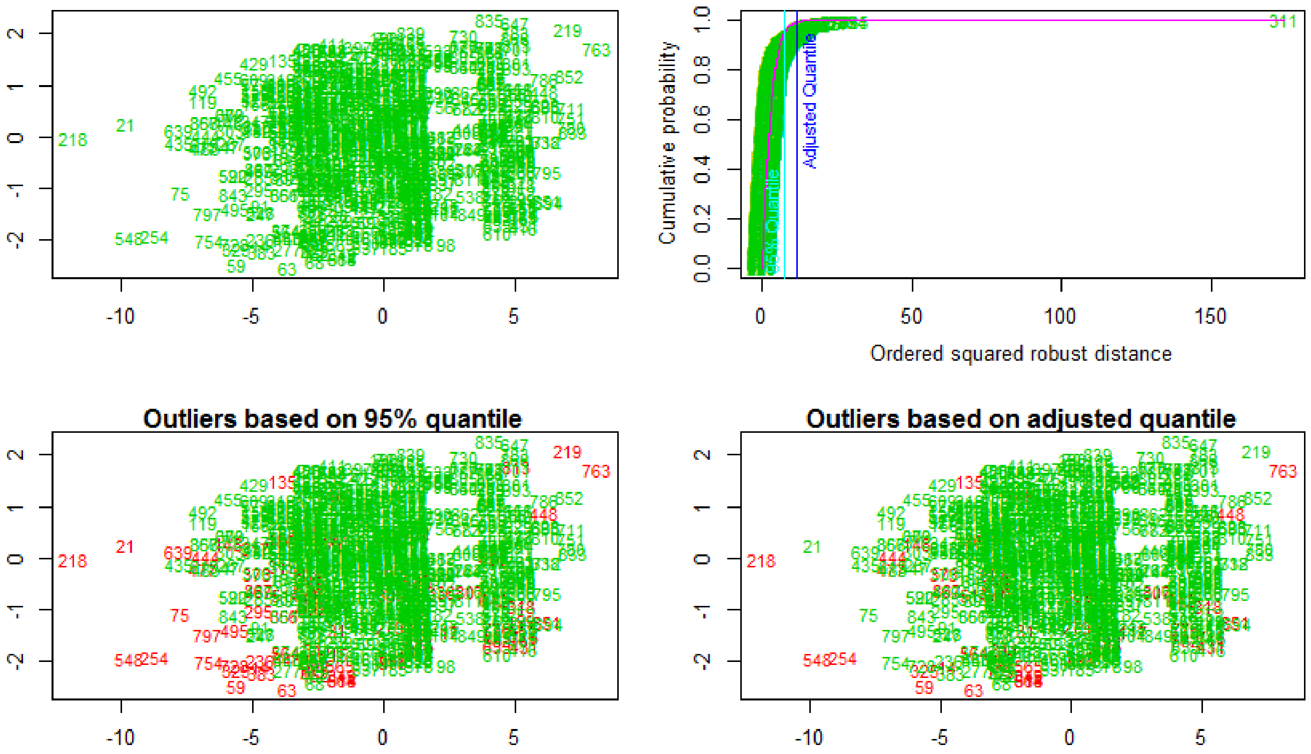
A data frame containing 872 observations on 7 variables.

```
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 data in AER library

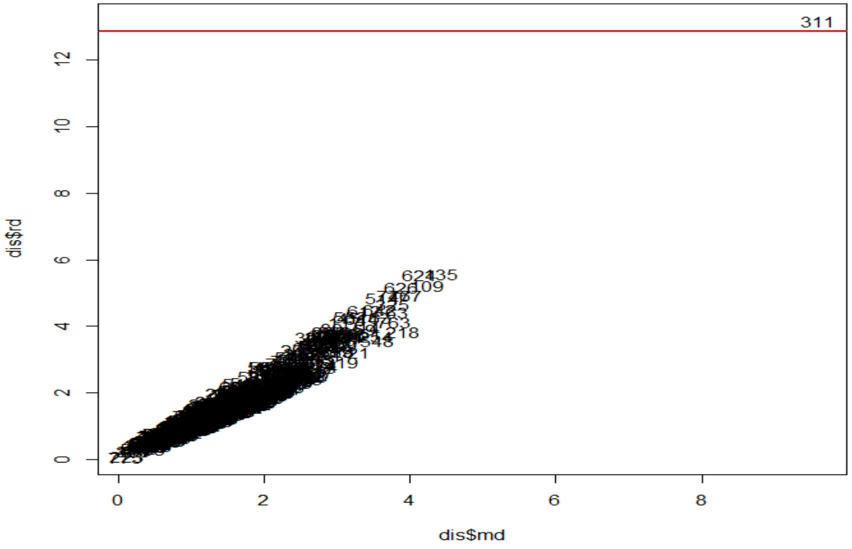
```
library(mvoutlier)
vout<-aq.plot(SwissLabor[,2:4], delta=qchisq(0.95,
df=ncol(x)),alpha=0.05)
```



Example: SwissLabor data in AER library

```
library(chemometrics)
dis <- Moutlier(SwissLabor[,2:4], quantile = 0.995)
dis$cutoff
par(mfrow=c(1,1))
plot(dis$md,dis$rd, type="n")
text(dis$md,dis$rd,labels=rownames(SwissLabor[,2:4]))
abline(h=qchisq(0.995, ncol(SwissLabor[,2:4])),col="red",lwd=2)
```

```
SwissLabor$mout<-0
sel<-which(dis$rd>12)
SwissLabor[sel,"mout"]<-1
```



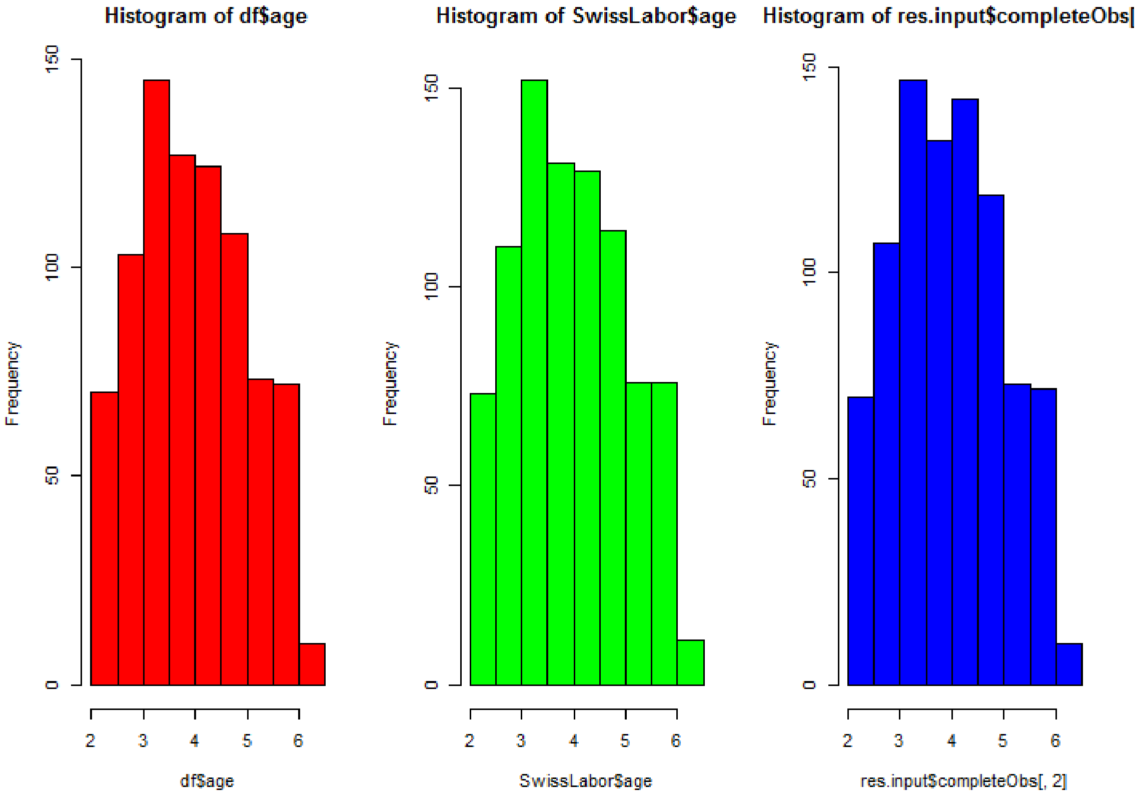
Example: SwissLabor data in AER library - Imputation

```
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