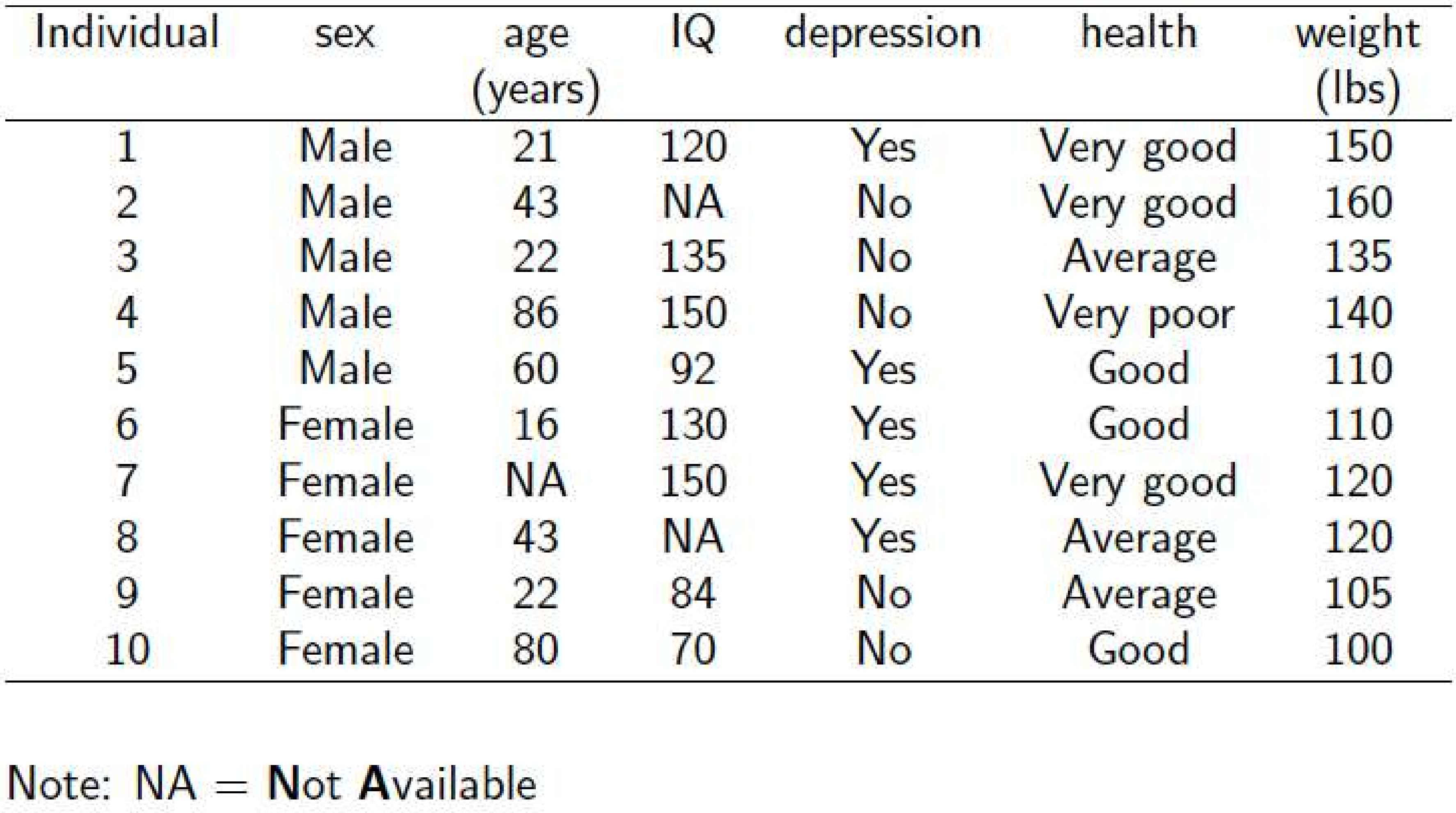
Hypothetical example



Individual	sex	age (years)	IQ	depression	health	weight (lbs)
1	Male	21	120	Yes	Very good	150
2	Male	43	NA	No	Very good	160
3	Male	22	135	No	Average	135
4	Male	86	150	No	Very poor	140
5	Male	60	92	Yes	Good	110
6	Female	16	130	Yes	Good	110
7	Female	NA	150	Yes	Very good	120
8	Female	43	NA	Yes	Average	120
9	Female	22	84	No	Average	105
10	Female	80	70	No	Good	100

Note: NA = **N**ot **A**vailable

Understanding the data

- Variables have meaning
  - Normally expressed in the **metadata file**

- **Role of variables**
  - **Response**, output or target. They are the variables we want to study, model, predict, … (Y)
  - **Explanatory**, input or predictors. They are the variables used to predict the former. (X)

- **Data origin**
  - **Primary** source: we collect the data (sampling): surveys, health studies
  - **Secondary** source: existing data (public webs, web scrapping,…)

Two types of analysis

1. **Descriptive Analytics**: Methods to find associations (to explore):
   - **Dimension Reduction: PCA, CA, FA,…**
   - **Clustering**
   - Visualization
   - **Associations rules**
   - Sentiment Analysis
   - …

2. **Predictive Analytics**: Methods to do predictions (by means of a model):
   - Multiple regression
   - Generalized Linear models
   - Partial Least Squares Regression
   - **Discriminant Analysis**
   - **Decision trees**
   - Support Vector Machines
   - Neural networks
   - …

Preprocessing Steps

- **Preparing the data for the analysis**
  - Feature selection: *filtering the uninteresting variables*
  - Feature extraction: *deriving new variables*
  - Transformations
    - Recoding (numeric → categorical)
    - Quantifying a nominal variable (categ. → numeric)
    - Normalizing

$$z = \frac{x - \overline{x}}{s_x}, \quad \frac{x}{\max(x)}, \quad \log(x), \quad \ldots$$

Preprocessing Steps

- **Data Cleaning**
  - Errors: Typos. Detect them and correct them
  - Missing values
  - Outliers

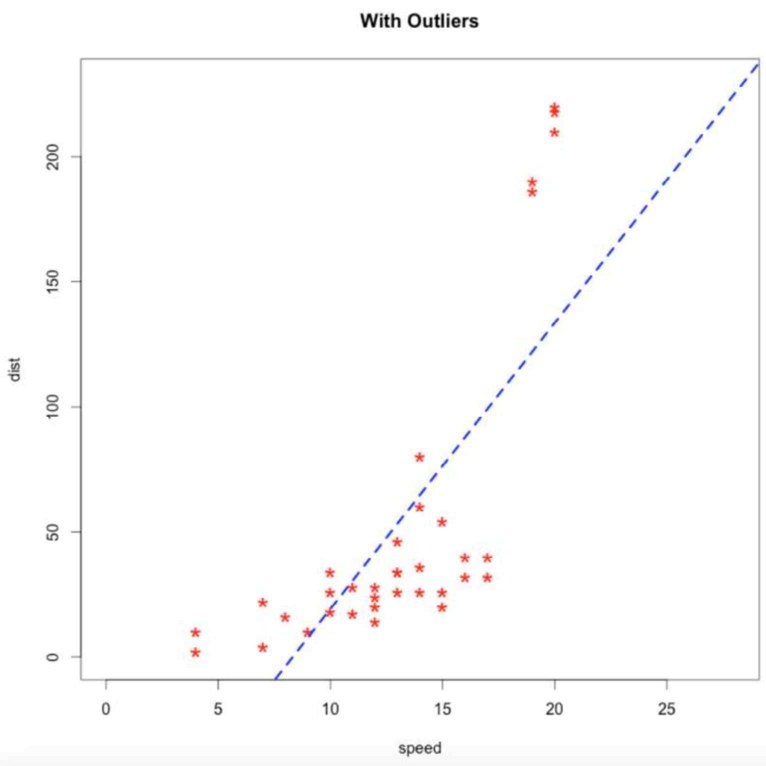
  Missing values can bias the results. In multivariate data, they may arise for several reasons:

  1. Non-response in sample surveys.
  2. Dropouts in longitudinal data.
  3. Refusal to answer particular questions in a questionnaire.

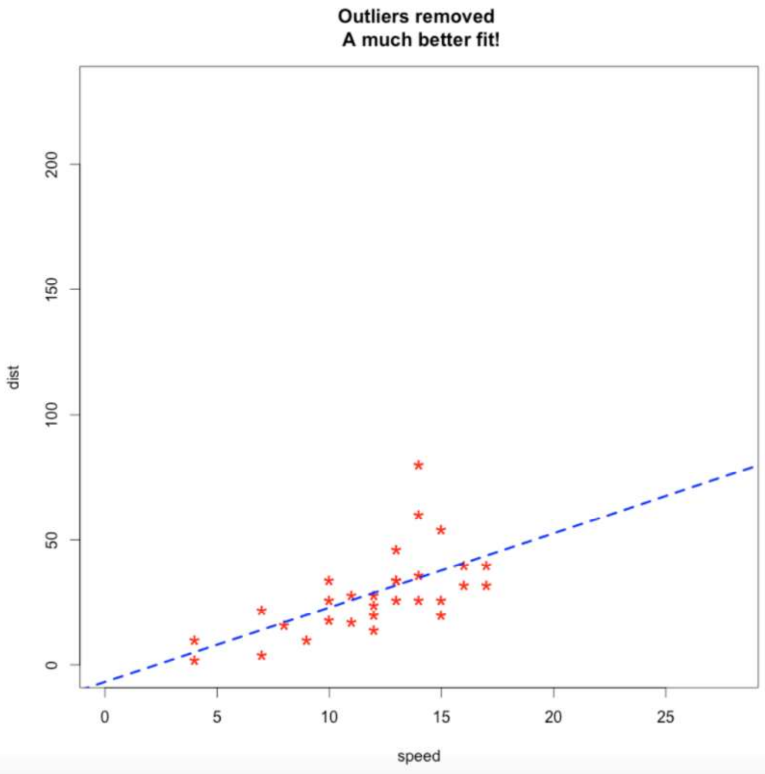
- **Complete-case analysis**: omit any case with a missing value on any of the variables.
- **Available-case analysis**: use all the cases available to estimate quantities of interest.
- **Imputation**: the practice of “filling in” missing data with plausible values.

Preprocessing Steps

- **Data Cleaning**
  - Errors: typos. Detect them and correct them
  - Missing values
  - Outliers: **They can bias the results**. Remove them or treat them as NA.



With Outliers



Outliers removed
A much better fit!

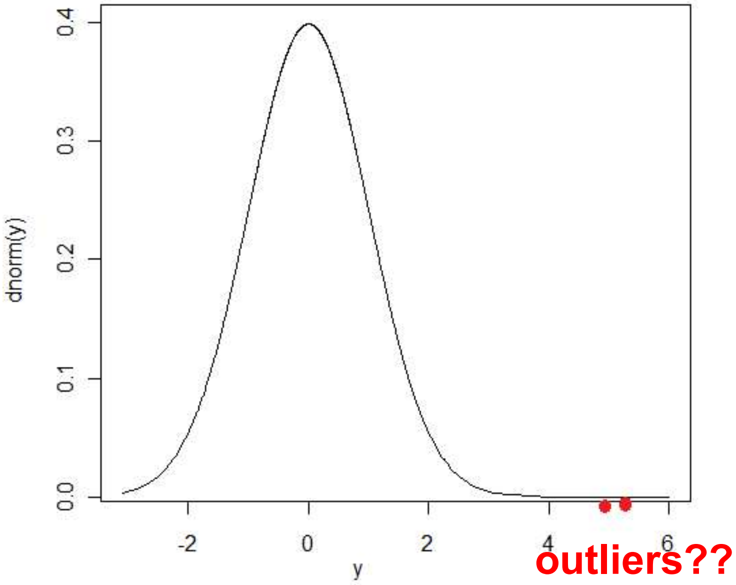
Outlier detection

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”

https://link.springer.com/book/10.1007%2F978-94-015-3994-4

Statistics-based intuition. Data is always generated by a mechanism that bestow a specific probability distribution (i.e., normal data follow a “normal generating data mechanism”). Outlying data may be a:

- very unlikely events for the current 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

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 <u>outside</u> of the interval

(Q1-3×IQR, Q3+3×IQR),

where IQR=Q3-Q1 is called the Interquartile Range.

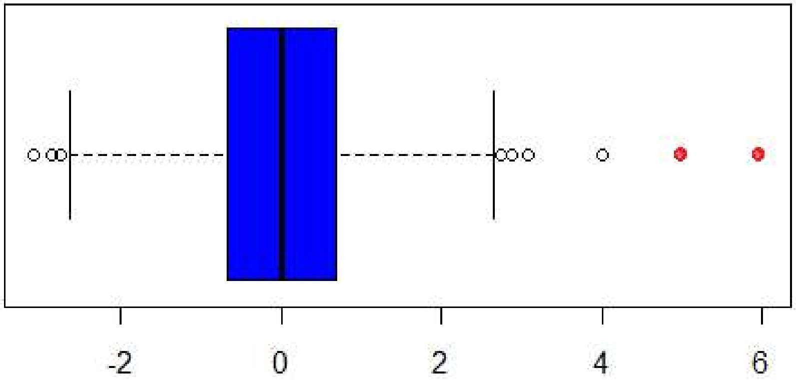
➢ An observation x is declared a **mild outlier** if it lies <u>outside</u> of the interval

(Q1-1.5×IQR, Q3+1.5×IQR).

- The numbers 1.5 and 3 are chosen by comparison with a normal distribution.
- If x ~ Normal :

Prob(X≥ Q3+1.5×IQR)= 0.003488302

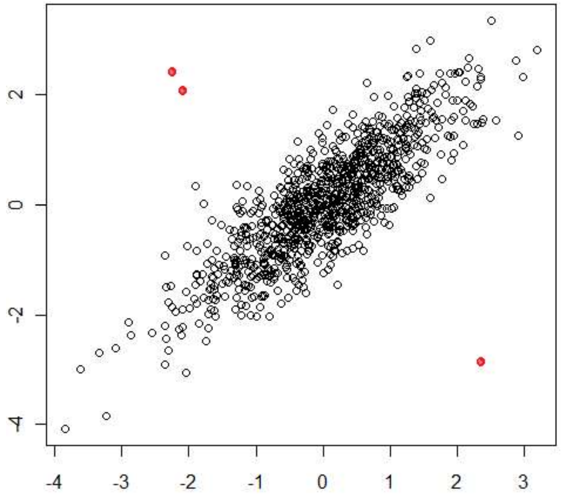
Prob(X≥ Q3+3×IQR)= 1.170971e-06



Multivariate outliers

Outliers are multivariate

Univariate detection of outliers **doesn't imply multivariate detection**



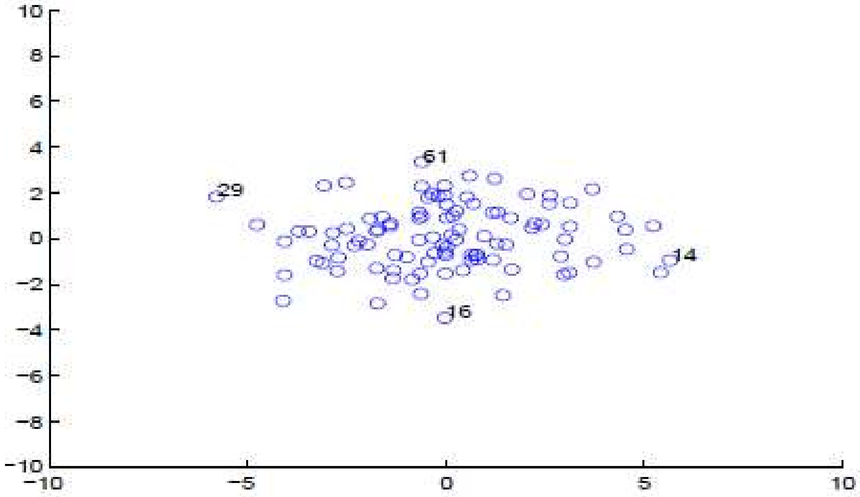
Then, the detection of outliers is based on computing the Mahalanobis distances to the central point of data.

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

Mahalanobis distance

The **Mahalanobis distance** is a measure of the distance between a point *i* and a distribution *G*. It is a multi-dimensional generalization of the idea of measuring how many standard deviations away *i* is from the mean of *G*.

Mahalanobis Distance vs. Euclidean distance



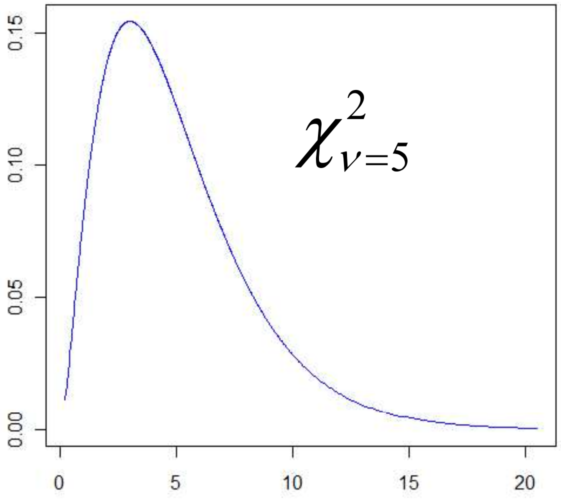
Point Pairs	Mahalanobis	Euclidean
(14,29)	5.07	11.78
(16,61)	4.83	6.84

If generating mechanism is Normal distributed:

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

It allows to establish a threshold for outlying points:

$$\chi^2_{\nu=\text{dim space}}(0.99)$$



Short distances occur more often

```
> qchisq(0.99,5)
[1] 15.08627
```

Detection of multivariate outliers

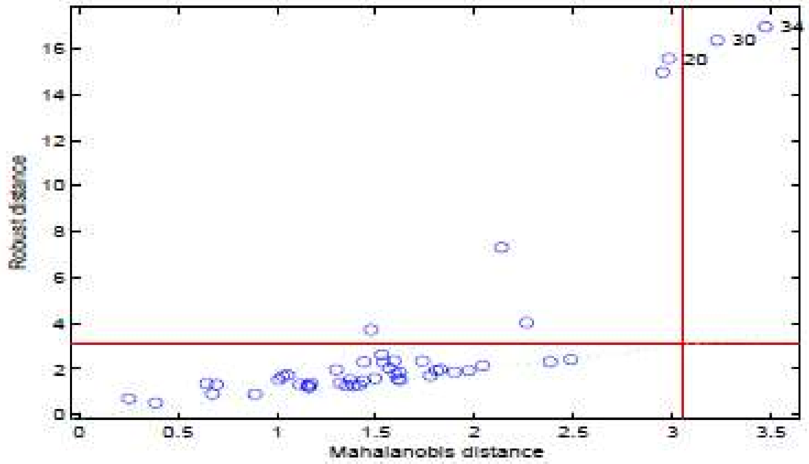
Problem: computation of G and V are contaminated by outliers
(G = mean of variables. V = matrix of variances

Take a value of h (size of data assumed not containing outliers), *h* must be > p (number of variables).

Initialization of an estimation of G and V : Compute the Mahalanobis distances $D^2_M(i,G)$ for all points i.

1. Rank the $D^2_M(i,G)$ and retain the h individuals with lower $D^2_M(i,G)$
2. Update G and V till convergence.

Plot the final “robustified” Mahalanobis distances with the initial Mahalanobis distances to detect the outliers



Detecting outliers in R

Moutlier {chemometrics}

Multivariate outlier detection using the Mahalanobis distance can be used. Plot of the classical and the robust (based on the MCD) Mahalanobis distance is drawn.

```
Moutlier(X, quantile = 0.975, plot = TRUE, ...)
```

Arguments

X	numeric data frame or matrix
quantile	cut-off value (quantile) for the Mahalanobis distance
plot	if TRUE a plot is generated

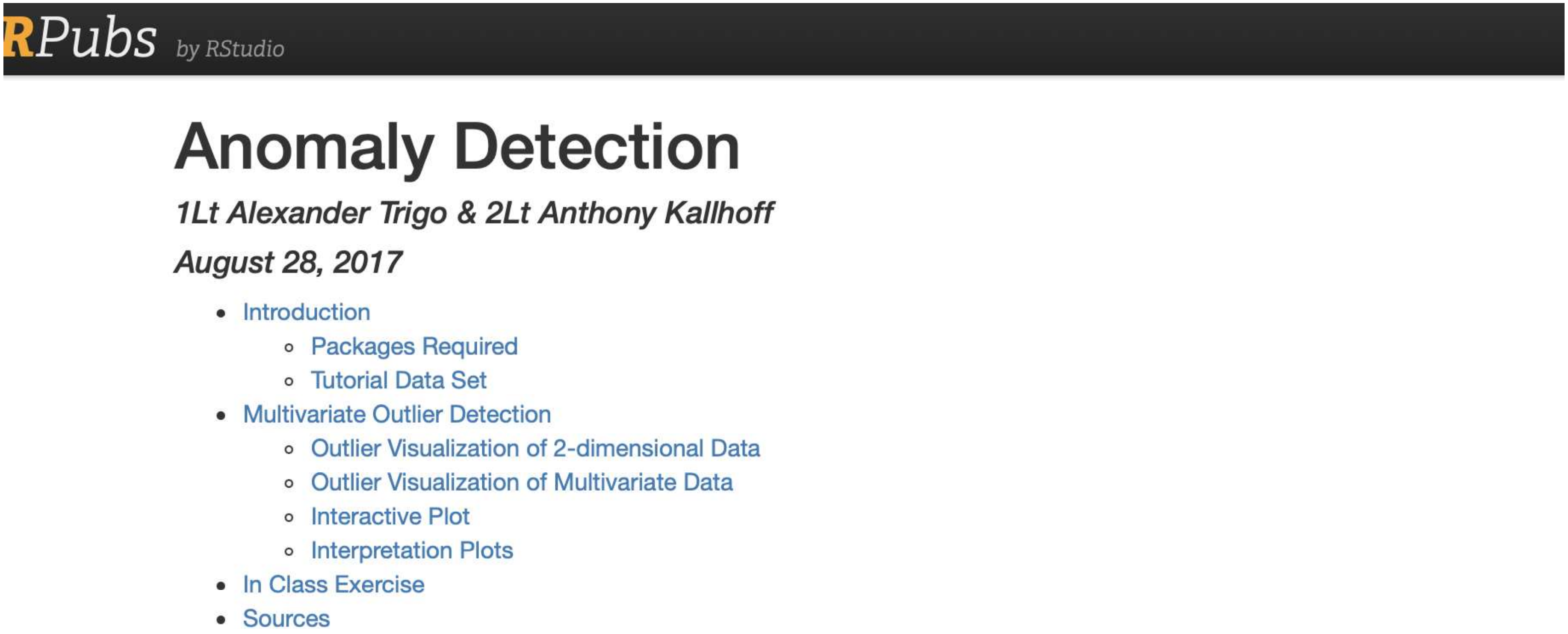
For multivariate normally distributed data, a fraction of 1-quantile of data can be declared as potential multivariate outliers. These would be identified with the Mahalanobis distance based on classical mean and covariance. For deviations from multivariate normality center and covariance have to be estimated in a robust way, e.g. by the MCD estimator. The resulting robust Mahalanobis distance is suitable for outlier detection. Two plots are generated, showing classical and robust Mahalanobis distance versus the observation numbers.

Values

md	Values of the classical Mahalanobis distance
rd	Values of the robust Mahalanobis distance
cutoff	Value with the outlier cut-off

Detecting outliers in R (2)

https://rpubs.com/Treegonaut/301942



Anomaly Detection

1Lt Alexander Trigo & 2Lt Anthony Kallhoff

August 28, 2017

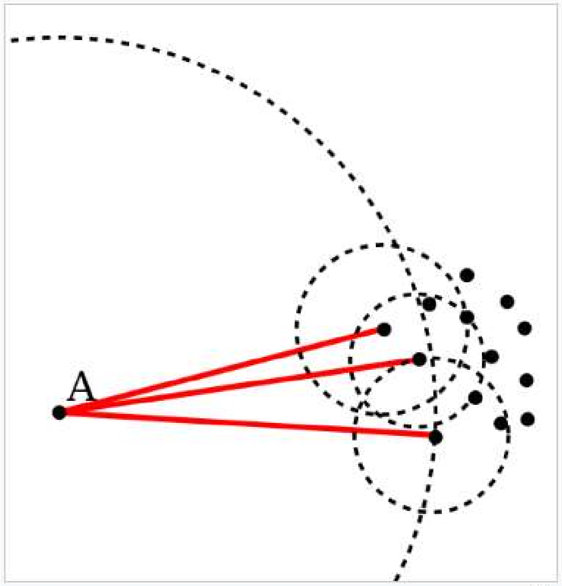
Non parametric detection of outliers

The LOF (Local Outlier Factor)

LOF is an algorithm for identifying density-based local outliers [Breunig et al., 2000]
https://dl.acm.org/doi/abs/10.1145/342009.335388

$$LOF_k(x) = \frac{\sum_{nei_x} \frac{\max dist_k(x)}{\max dist_k(nei_x^k)}}{k}$$

Comparison of the maxdist of the neighborhood of a point, respect to the maxdist of the neighborhood of the neighbors of the point (detection of outliers based on local density).



Basic idea of LOF: comparing the local density of a point with the densities of its neighbors. A has a much lower density than its neighbors.

The outcome is an outlying value per individual. Values greater than 1 suggest outliers.

```
library(DMwR)

outlier.scores <- lofactor(data, k=5)
plot(density(outlier.scores))

# pick top 5 outliers
outliers <- order(outlier.scores, decreasing=T)[1:5]
# who are outliers
print(outliers)
```

Practice of detecting outliers

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

- 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?

  1. **Eliminate them** (but we lose the information of the eliminated individuals) and deleting outliers is not the best solution, since outliers are recursive.

  2. **Weight the individuals inversely to the outlying degree of individuals**, to diminish its importance (but statistical/learning methods would need to had implemented a weighing option of individuals).

  3. Make a robust estimation of the parameters of the “normal generating mechanism”, for instance **with a given percentage of the “central” individuals.**

  4. Declare outliers as “**missing values**” and treat them as missing data.

Missing data

- **Databases:**
  - Databases are used for secondary purposes, only information that is currently used is maintained. (i.e., inland registries, addresses are the best up to date field, the characteristics of the property much less).
  - Not compulsory fields.
  - Errors and outliers may be taken as missing values …
- **Surveys:**
  - Outright refusals: *unit nonresponse* → reweighing the sample
  - Nonresponse 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 → this is not missing data
  - Dropouts in panel studies → this is not missing data. Deal as censored data

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 it with careful planning of data collection, with proper intelligent interfaces.

Dealing with missing data

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 missings per variable and rank them accordingly.
The more the missing the less reliable is the information provided by the variable

Characterize the missingness mechanism

Create a new variable counting the number of missings per individual.
Describe this variable (association analysis). Profiling of global missingness analysis
Create a new variable per variable, indicating the (missing/non-missing) and compare both groups of cases. Profiling of variable missingness analysis

Dealing with missing values

Ignore records with missing values
If categorical, treat missing value as a separate level (or impute them).
If continuous, impute (fill in) with mean or median values, 1nn, EM algorithm, DA algorithm, …

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.

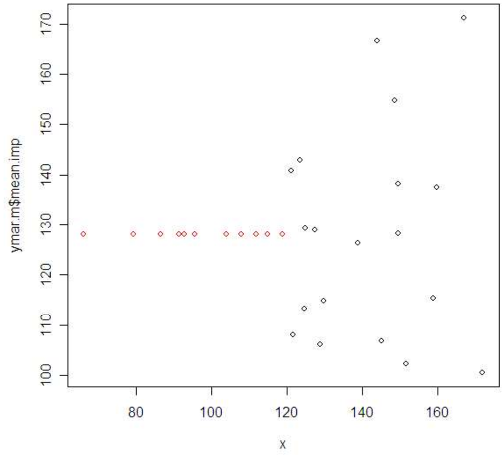
- MAR - At random: missing values appear related to third observed variables. This is the most usual case, i.e., asking the income of individuals, income is missing but can be imputed from the educational level.

- 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

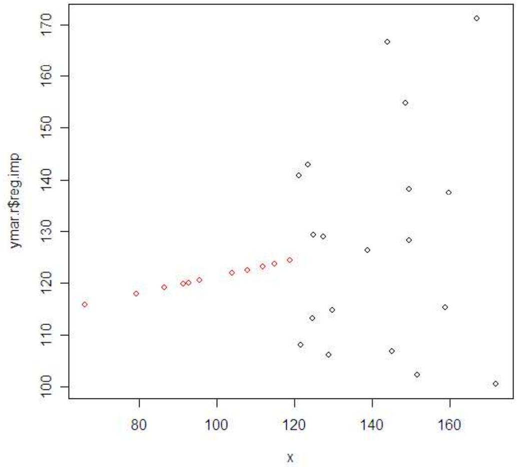
Traditional methods

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



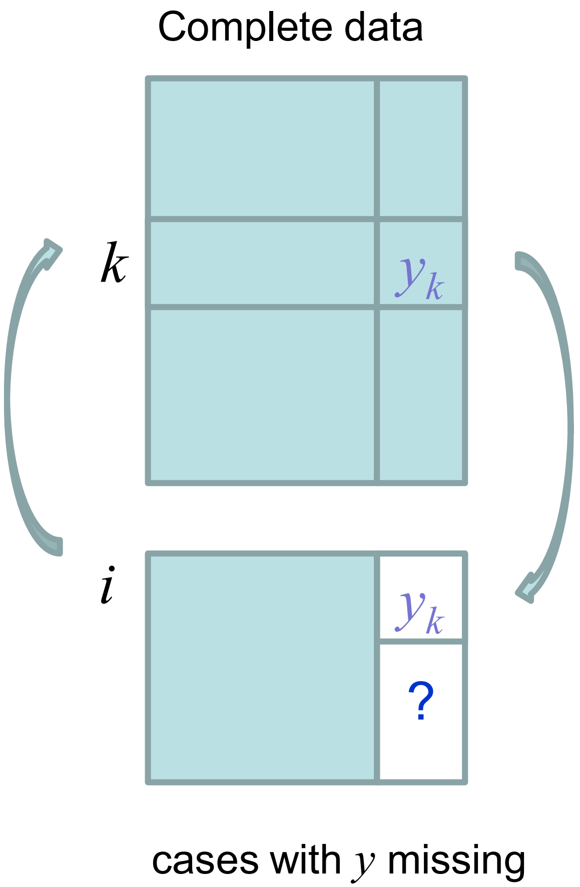
- Unconditional mean imputation. Every missing value is substituted by the corresponding global mean of the variable

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



Knn imputation

For every observation to be imputed, it identifies ‘k’ closest observations based on the Euclidean distance and computes the weighted average (weighted based on distance) of these ‘k’ obs.



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

Copy the y_k value in the i individual

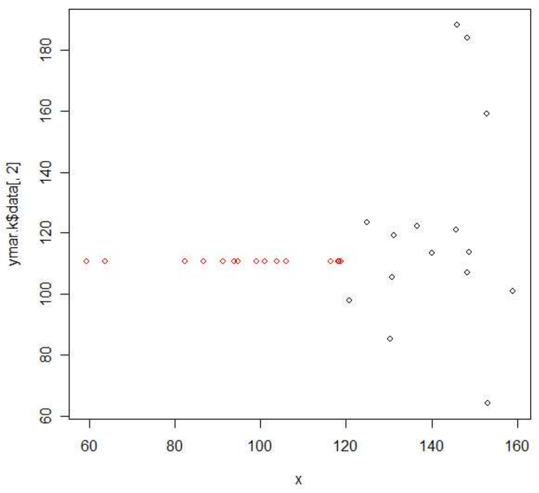
https://www.r-bloggers.com/2016/04/missing-value-treatment/

Knn imputation (2)

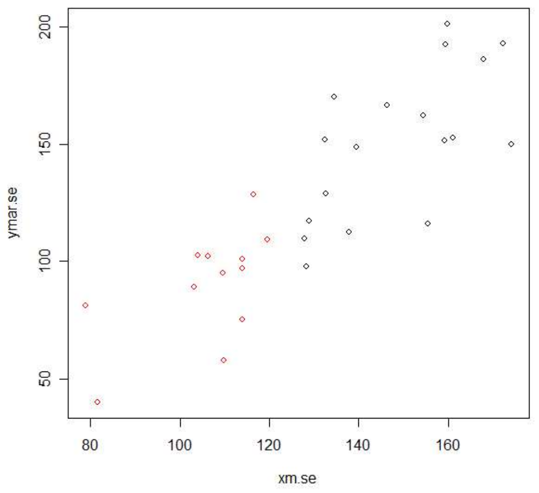
Knn – K nearest neighbor imputation (easy to implement)

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

https://www.rdocumentation.org/packages/DMwR/versions/0.4.1/topics/knnImputation



with only x as covariate



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

Imputation by chained equations

Let X be a data set with missing observations

Order the data set trying to follow a monotone increasing missingness pattern

1.Start filling in the missing data with values at random
2.For every variable with missing values

    a. Impute the missing values of the variable from the predicted values of the regression of the current variable with the remaining ones.

Iterate the above procedure till the convergence

Apply 1nn to impute every missing value from the closest individual to obtain the final realistic imputed values

```
library mice;    imp <- mice(data, m = 1);    data_imp <- complete(imp)
```

Imputed data is REAL data, but it is just a plausible value of the missing data

https://www.r-bloggers.com/2016/06/handling-missing-data-with-mice-package-a-simple-approach/

Imputation by random Forests

Non-parametric method of imputation

Let X be a data set with missing observations of any type (continuous and categorical)

Order the data set trying to follow a monotone increasing missingness pattern

1. Start filling in the missing data with values at random
2. For every variable with missing values

   Impute the missing values of the variable from the predicted values from the random forest of the individuals with the current variable as response using the remaining ones as predictors.

Iterate the above procedure till imputed values converge

$$\frac{\sum_i (x_{new}^{imp} - x_{old}^{imp})^2}{\sum_i (x_{new}^{imp})^2}$$

In convergence, output the OOB error

```
library(missForest);  mf_imp <- missForest(data); data_imp <- mf_imp$ximp
```

Imputed data is not REAL data, it is just an estimate of possible data

https://rpubs.com/david-deming-tung/missing_dat