Introduction and Data pre-processing

Jan Graffelman[1]

[1]Department of Statistics and Operations Research
Universitat Politècnica de Catalunya
Barcelona, Spain

jan.graffelman@upc.edu

February 8, 2020

Contents

Data Analysis

In this course we study two types of data

- Multivariate data sets (many observations, many variables)
- Time series (variables that are observed repeatedly over time (daily, weekly, monthly, etc.)

The subdisciplines of statistics dedicated to the analysis of such data sets are

- Multivariate analysis
- Time series analysis

This course is an introduction to both multivariate analysis and to time series analysis.

Multivariate analysis

Topics in multivariate analysis we study in this course:

- Matrix algebra
- Numerical and graphical summaries of data matrices; Biplots
- Principal component analysis (PCA)
- Distances
- Multidimensional scaling (MDS; metric & non-metric)
- Simple and multiple correspondence analysis (CA and MCA)
- Multivariate normal distribution
- Multivariate inference: comparison of two or more groups
- Group detection: Cluster analysis
- Classification: Linear and quadratic Discriminant analysis (LDA and QDA)

There are many additional topics in MVA we do not deal with in this course

- Factor analysis (latent variable models)
- Canonical correlation analysis (CCO)
- Procrust analysis
- ...

Organizing methods

Methods are sometimes organized into groups, considering different criteria:

- Interdependence versus dependence methods
- Exploratory methods and inferential methods
- One-table methods versus multiple table methods
- ...

Bibliography

- Manly, B.F.J. (1989) Multivariate statistical methods: a primer. 3rd edition. Chapman and Hall, London.
- Johnson & Wichern, (2002) Applied Multivariate Statistical Analysis, 5th edition, Prentice Hall.
- Peña, D. (2002) Análisis de datos multivariantes. McGraw-Hill, Madrid.

Preprocessing

- It is, in general, not recommended to fit statistical models directly to raw data sets.
- There are many data cleaning or preprocessing steps that help to ensure data quality and that pave the way for a more sensible analysis.
- Some aspects of preprocessing
  - Conversion of the data to a convenient file format
  - Does the data come from a single source or from multiple sources?
  - Avoiding duplications
  - Checking for the existence of missing data
  - Checking for the existence of outliers
  - Checking for the existence of zeros
  - Look if transformations of the data are needed.
  - ...

Dealing with missing values

- Are there any missing values?
- How are missing values coded?
- What percentage of the data is missing?
- Do missings concentrate in some variables or individuals?

Classification of missingness

- Missing completely at random (MCAR)
- Missing at random (MCAR)
- Missing not at random (MNAR)

MCAR: Missing Completely At Random

1. There are missing observations, but one can envision a (hypothetical) data set of completely observed individuals.
2. If the observed items are a random sample of this ideal data set, then data is MCAR.
3. The missing observations are also a random sample of this ideal data set.
4. Discarding missings is not too problematic, if there are not too many.

MAR: Missing At Random

1. The probability that a result is missing for a particular variable may depend on the observed data (e.g. other variables registered)
2. Conditional on the observed data, this probability may not depend on the values of the variable itself.

MNAR: Missing Not At Random

1. The probability of a missing result does depend on the values of the variable under consideration
2. Even so after controlling for the relationships of this variable with other relevant variables

Approaches

- Delete the missings.
- Impute the missings with a "reasonable value" and behave as if the dataset would be complete (single imputation).
- Impute the missings many times and do the analysis for each imputed data set (multiple imputation).
- ...

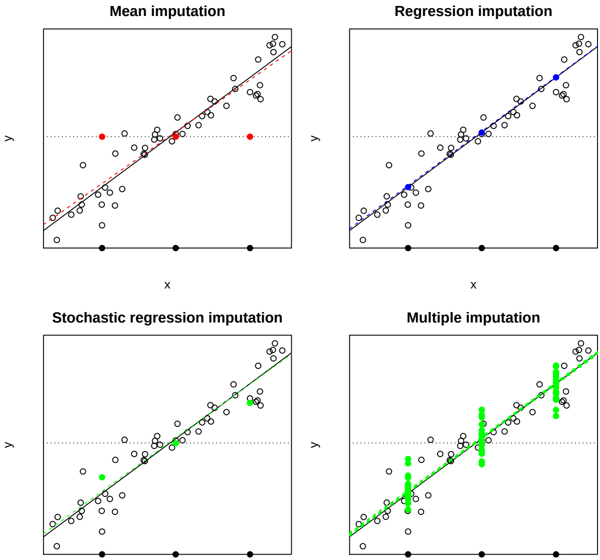
Deleting missing observations

Why not just delete the missings?

- You reduce your sample size and loose power to detect effects.
- Statistical inference may be biased if the missing observations are not MCAR.

A better idea is to impute the missings somehow.

The imputation of missing values

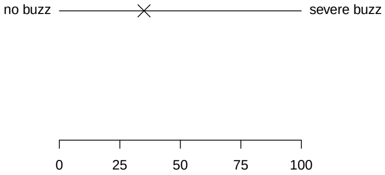


About zeros

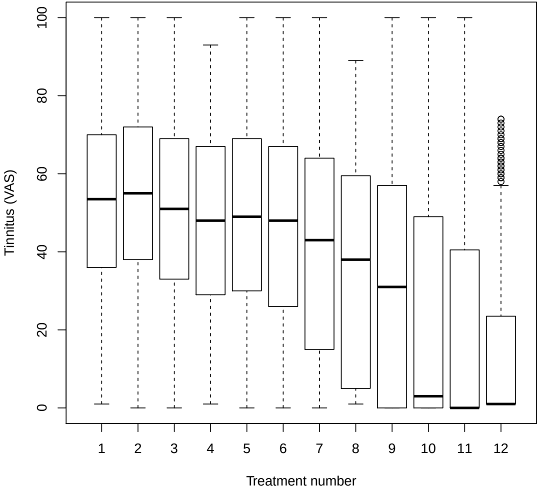
- Dealing with zeros is a very delicate matter
- Zeros are sometimes, in fact missing values
- Complications arise if some zeros are real, and others represent missings
- The coding of the data is very important, and it is imperative to use a special code for a missing value.
- NA represents a missing value in the R environment.

Example

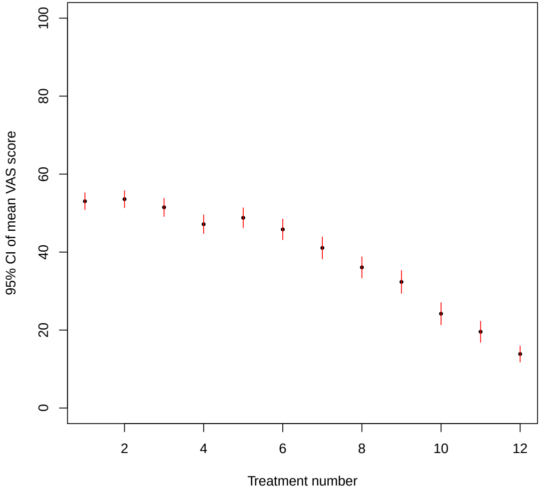
- A medical doctor applies a treatment to relieve tinnitus (ear buzzing) to 400 patients on 12 successive occasions
- On each occassion, patients score their complaint on a visual analog scale (VAS)



Plotting the data

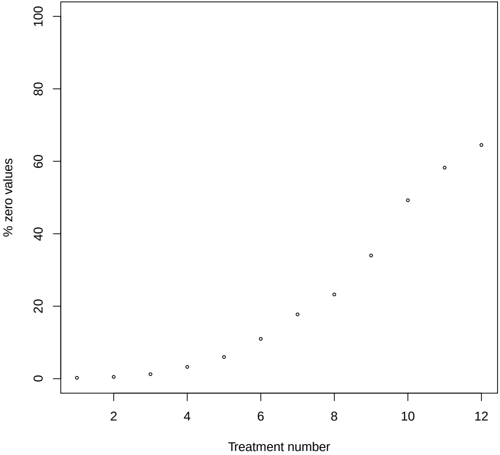


Is it significant?

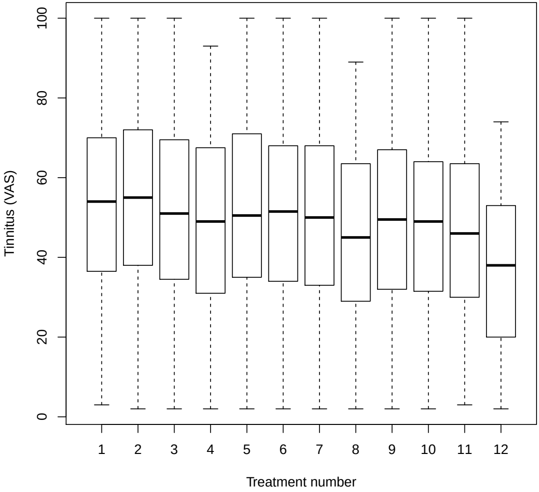


How about zeros?

- The data matrix was complete (400×12) and had no missing values
- However, 22.4% of the entries in the data matrix were zeros



What if zeros are missings?



Outliers

You may have:

- univariate outliers
- bivariate outliers
- multivariate outliers

Issues:

- How to identify outliers?
- What to do with outliers?

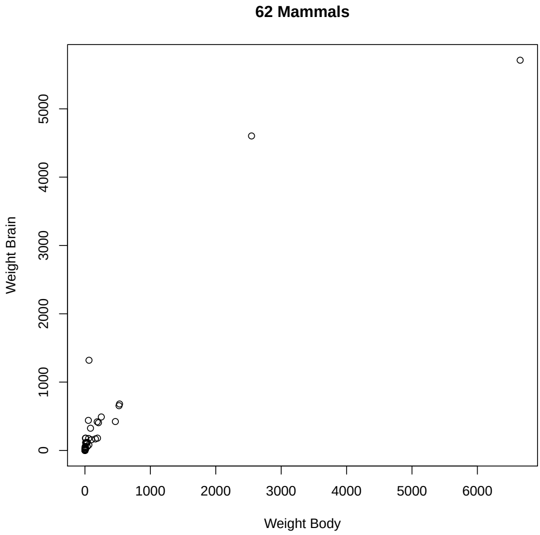
How to deal with outliers?

- First check if the outlier corresponds to a correct measurement or a clearly erroneous or impossible outcome
- Consult the scientists who generated the data
- Depending on the statistical technique used, an outlier may be problematic or not
- If problematic, consider a transformation to reduce the effect of the outlier
- Perform the analysis with and without outlier, and compare results
- ...

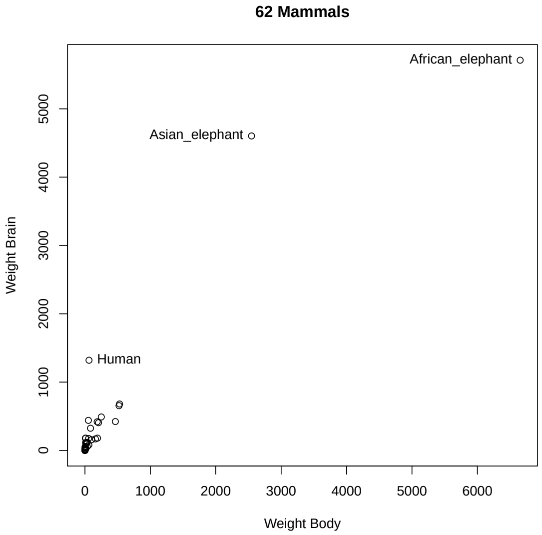
Transformations

- Transformations are often used in statistics
- Why are transformations used?
  - Reduce the effect of outliers
  - Make a distribution more symmetric
  - To produce homocedasticity
  - Achieve approximate normality of a variable
  - Remove a constraint that operates on the data
  - ....
- Some common transformations
  - $y = \ln(x)$ (zeros not allowed)
  - $y = \sqrt{x}$ (zeros allowed)
  - Logit transformation for probabilities $y = \ln\left(\frac{p}{1-p}\right)$
  - replacing observations by their rank
  - power transformation $y = x^a$
  - Box-Cox transformation
  - ...

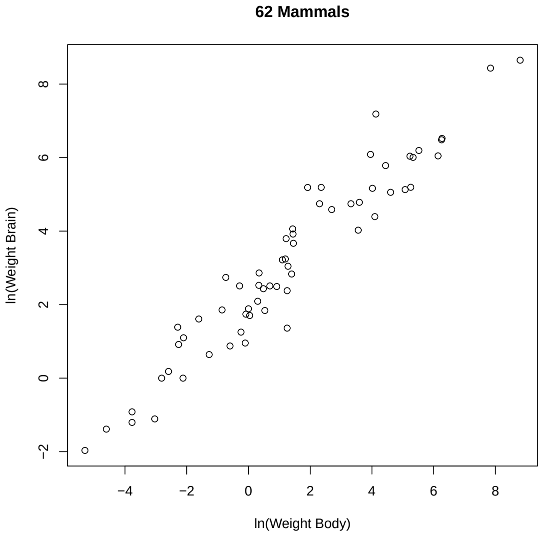
Example logaritmic transformation



Identifying outliers



Logaritmically transformed data



The Box-Cox transformation

- The Box-Cox transformation can be employed to achieve approximate normality
- The Box-Cox transformation is given by

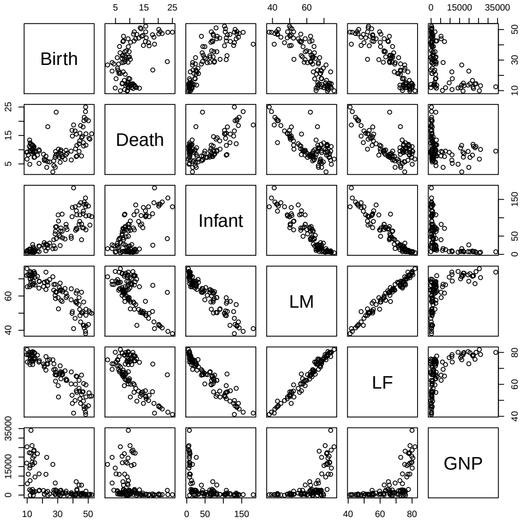
$$y_i^{(\lambda)} = \begin{cases} \frac{y_i^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln\left(y_i\right) & \text{if } \lambda = 0. \end{cases}$$

- An optimal value for the transformation parameter λ is obtained by maximum likelihood.

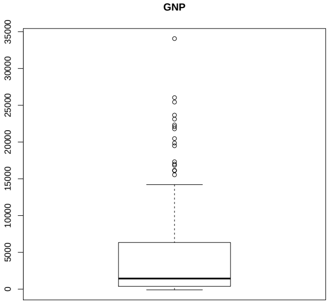
Example: poverty data set

Country	Birth	Death	Infant	LM	LF	GNP
Albania	24.7	5.7	30.8	69.6	75.5	600
Bulgaria	12.5	11.9	14.4	68.3	74.7	2250
Czechoslovakia	13.4	11.7	11.3	71.8	77.7	2980
FormerEastGermany	12.0	12.4	7.6	69.8	75.9	-99
Hungary	11.6	13.4	14.8	65.4	73.8	2780
Poland	14.3	10.2	16.0	67.2	75.7	1690
Romania	13.6	10.7	26.9	66.5	72.4	1640
Yugoslavia	14.0	9.0	20.2	68.6	74.5	-99
USSR	17.7	10.0	23.0	64.6	74.0	2242
Byelorussian_SSR	15.2	9.5	13.1	66.4	75.9	1880
Ukrainian_SSR	13.4	11.6	13.0	66.4	74.8	1320
Argentina	20.7	8.4	25.7	65.5	72.7	2370
Bolivia	46.6	18.0	111.0	51.0	55.4	630
Brazil	28.6	7.9	63.0	62.3	67.6	2680
Chile	23.4	5.8	17.1	68.1	75.1	1940
Columbia	27.4	6.1	40.0	63.4	69.2	1260
Ecuador	32.9	7.4	63.0	63.4	67.6	980
Guyana	28.3	7.3	56.0	60.4	66.1	330
Paraguay	34.8	6.6	42.0	64.4	68.5	1110
Peru	32.9	8.3	109.9	56.8	66.5	1160
Uruguay	18.0	9.6	21.9	68.4	74.9	2560
Venezuela	27.5	4.4	23.3	66.7	72.8	2560
Mexico	29.0	23.2	43.0	62.1	66.0	2490
Belgium	12.0	10.6	7.9	70.0	76.8	15540
Finland	13.2	10.1	5.8	70.7	78.7	26040
Denmark	12.4	11.9	7.5	71.8	77.7	22080
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Exploring relationships



Boxplot of GNP



	N	N*	Mean	Stdev	Med	Q1	Q3	Min	Max
GNP	97	0	5380	7963.25	1440	380	6340	-99	34064

What to do with -99?

```
> sum(X$GNP==-99)
[1] 6
> X$GNP[X$GNP==-99] <- NA
> sum(is.na(X$GNP))/nrow(X)
[1] 0.06185567
```

A quick solution

```
me <- median(X$GNP,na.rm=TRUE)
X$GNP[is.na(X$GNP)] <- me
```

Box-Cox transformation

```
library(MASS)
boxcox(X$GNP~1,lambda = seq(-1, 1,by=0.1))
```