tbs EDUCATION
Toulouse Business School

MSc AIBA

UE 1 - Analytical Theory, Methods and Models

Econometrics and Statistical Models

Class #1 - Data preprocessing

Anne VANHEMS
a.vanhems@tbs-education.fr

EFMD EQUIS ACCREDITED · AACSB ACCREDITED · Accredited by Association of MBAs

www.tbs-education.fr

Motivation

- Check data quality before fitting a model!

  ➔ Data structure

  ➔ Outliers / Normality of quantitative variables

  ➔ Missing values

Data structure

- Check for data structure: are the variables types correctly identified by Rstudio?

➔ Useful to perform appropriate statistical tools!

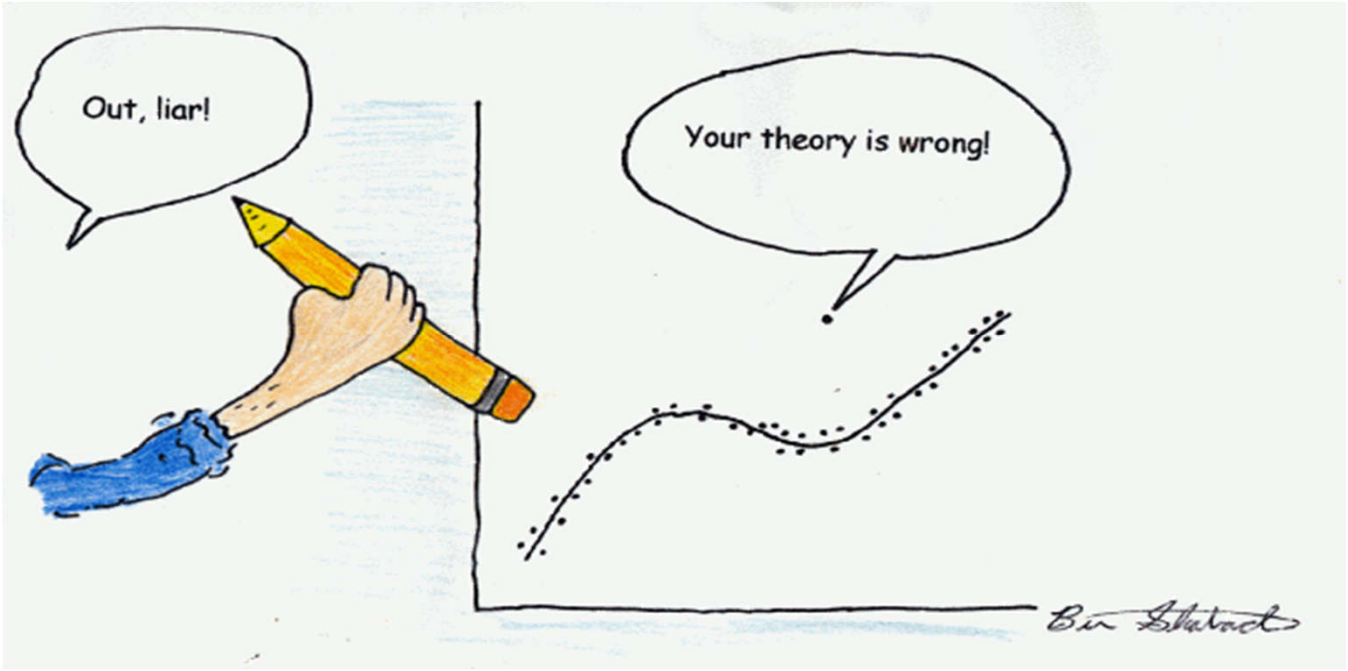
character, factor, numeric

➔ Recode if necessary

as.character, as.factor, as.numeric

Outliers

- Check for outliers: potential aberrant value. May pertubate estimation and fitting.



Outliers

➔ Descriptive statistics on qualitative variables

- Categories with not too few observations
- Remedies: delete observations or merge categories

➔ Descriptive statistics on quantitative variables

- Histograms, boxplots
- Normality tests: Jarque-Bera, Shapiro-Wilks
- Remedies: delete observations or transform data (log transformation for positive values)

Missing values

- When no value is available in one or more variables of an individual.
  - ➔ interviewer mistakes, anonymization purposes, refusal to respond by the participant (*item non response*)
  - ➔ Nonresponse has different causes such as a lack of knowledge about the question, an abortion of the questionnaire, or the unwillingness to respond to sensitive questions.

- Missing values are an issue of essentially every survey
  - ➔ might introduce bias in your estimates
  - ➔ And lead to wrong conclusions of your survey.
  - ➔ ISSUE with R: any calculations on variables that have missing values always return NA as a result.

Missing values

- Reasons for missing values:
  - ➔ Some responses were accidentally deleted (MCAR, missing completely at random)
  - ➔ Some non responses are linked to explanatory variables or the variable of interest (NMAR, non missing at random)
    - Participants with higher age are less likely to respond to their political opinion
    - Participants with higher incomes report their income less often.

Missing values-solutions

- Delete observations or variables
  - ➔ Delete variable: when all missing values are identified in one variable
  - ➔ Delete observations: when they are only a few and the sample size is large
    - Easiest method but that might create biases unless in the MCAR case

Missing values-solutions

- Imputation methods

  ➔ Simple method: replace by the mean or the median (for quantitative variable) or replace by the mode (for qualitative method)

    - Might often lead to biases

  ➔ Prediction methods: linear regression methods (for quantitative variables), multinomial logistic regression (for qualitative variables)

  ➔ More sophisticated iterative methods

Which method choose?

No method is perfect…. Some practical ideas:

- When the missing values are MCAR and less than 5% of the sample size, you can delete them
- When the variable with missing values is homogenous and not dispersed, you can impute the missing values with the mean or the median
- When you suspect the missing values are not missing at random, select more sophisticated predictive methods!

Conclusion

Points	Problems	Detection	Remedies
Representativity of the sample	If you work on a sample that is not representative of the population, your study makes no sense	Be careful when you collect the data	Justify the constitution of your sample
Missing values	Can be misunderstood by R or prevent the use of some R functions	Exploratory data analysis	Cleansing or replace if possible by mean or predictors using more sophisticated methods
Outliers	Can distort statistics (mean, std dev…), graphs or results (regression)	Exploratory data analysis (min, max), graphs (boxplot)	Cleansing (up to 5% of your data)
Variables	Statistical treatments depends on the type of the variables	Check the type of the variables (quantitative, qualitative, textual)	If necessary recode the variable (quantitative in qualitative, qualitative into dummies)