

# ADVANCED STATISTICS

*DR. MEIKE MORREN*

Lecture 2

# contents

---

- Data cleaning
- Missing data
- Outliers
  
- M(C)AR assumption
- Imputation
- Sensitivity analyses

# literature

---

Very clear paper on missing data by Paul D. Allison,  
who is *the* authority on missing data

# DATA CLEANING



# Data cleaning

---

- Remove unrealistic values
- Correct mistakes in creating the file

# Recode missing values

- Recode 99 to missing values

- ▣ `df$first[df$first==99] <- NA`

- Recode 99 to missing values for entire df

- ▣ `df[df==99] <- NA`

- Recode negative to missing values

- ▣ `df$first[df$first<0] <- NA`

- Recode multiple values to missing values

- ▣ Use loop (next lecture)

# MISSING VALUES



# Listwise

	Var1	Var2	Var3	Var4
Individual 1	2	4	2	2
Individual 2	NA	2	2	3
Individual 3	1	1	2	3
Individual 4	2	2	2	NA
Individual 5	2	3	NA	2
Individual 6	3	NA	NA	2
Individual 7	1	1	NA	1
...	1	1	2	4
Individual N	3	2	1	2



# Listwise

	Var1	Var2	Var3	Var4
Individual 1	2	4	2	2
Individual 2	NA	2	2	3
Individual 3	1	1	2	3
Individual 4	2	2	2	NA
Individual 5	2	3	NA	2
Individual 6	3	NA	NA	2
Individual 7	1	1	NA	1
...	1	1	2	4
Individual N	3	2	1	2

# Listwise

- Advantage

- Easy

- Disadvantage

- Massive losses of data

- Increase of type II errors

- i.e. decrease of power: do you find what is really there in the population?

# Pairwise

	Var1	Var2	Var3	Var4
Individual 1	2	4	2	2
Individual 2	NA	2	2	3
Individual 3	1	1	2	3
Individual 4	2	2	2	NA
Individual 5	2	3	NA	2
Individual 6	3	NA	NA	2
Individual 7	1	1	NA	1
...	1	1	2	4
Individual N	3	2	1	2

# Pairwise (1)

In pairwise deletion, each of these 'moments' is estimated using all available data for each variable or each pair of variables.

- Advantage

- ▣ you do not have to throw away data

# Pairwise (2)

## □ Disadvantage

- Can **only** be used when parameters can be expressed as functions of means, variances and covariances (correlations) (i.e. factor analysis)
- Each covariance (correlation) is based on different sample size
- This leads to inaccurate standard error estimates

# complete.cases

- Only look at rows with missing values
  - ▣ `df[!complete.cases(df), ]`
- Only look at cases that are complete
  - ▣ `df[complete.cases(df), ]`
- Only look at cases that are complete in first var
  - ▣ `df[complete.cases(df[, "first"]), ]`

# is.na

Returns TRUE when value is missing

## □ One value

- ▣ `x <- 9`
- ▣ `is.na(x)`

## □ Vector

- ▣ `x <- c(1, 4, 5, 6, NA)`
- ▣ `is.na(x)`

## □ Select nonmissing values

- ▣ `x[!is.na(x)]`

# na.rm

- Exclude missing values from analyses
  - ▣ `mean(x)`
  - ▣ `mean(x, na.rm=TRUE)`



# na.omit

Removes all rows with missing data

- Create new dataset without missing data
  - ▣ `newdf <- na.omit(df)`
  - ▣ Creates list:
    - df without missing data
    - number of observations with missing values
- `na.exclude` does not remove the rows but excludes them from analysis/ print

# Exercise 2\_1.r

- Find & correct mistakes in recoded variables
  - ▣ Age
  - ▣ Education
- Explore missing data
  - ▣ PC
- Explore which variables are completely missing

# MISSING VALUES ASSUMPTIONS



# Assumptions

- Missing at random
  - ▣  $P(Y \text{ is missing} \mid X, Y) = P(Y \text{ is missing})$
- chance that y is missing is unrelated to y or x
- X needs to be variables in the model of interest

# MCAR, MAR, NMAR

## Missing Completely at Random (MCAR)

- Missing value ( $y$ ) neither depends on  $x$  nor  $y$
- Example: some survey questions asked of a simple random sample of original sample

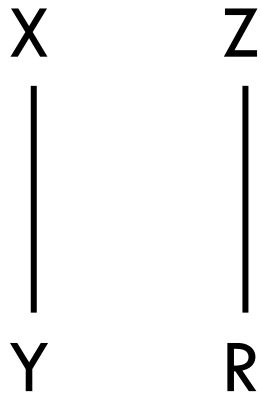
## Missing at Random (MAR)

- Missing value ( $y$ ) depends on  $x$ , but not  $y$
- Example: Respondents in service occupations less likely to report income

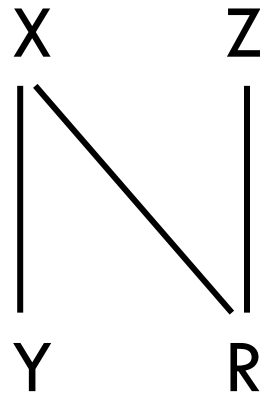
## Missing not at Random (NMAR)

- The probability of a missing value depends on the variable that is missing
- Example: Respondents with high income less likely to report income

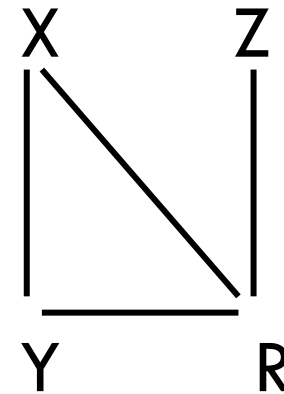
# Mechanism of MCAR, MAR, MNAR



MCAR



MAR



MNAR

$X = Y_{\text{obs}}$ ,  $Y = Y_{\text{mis}}$ ,  $R = \text{response indicator}$

$Z = \text{'explanation' of why participants don't respond}$

# Pairwise / Listwise

- If data are MAR listwise and pairwise yield unbiased estimates
- But NOT when data are MCAR
  - ▣ Simulations show that pairwise might even be less efficient than listwise in these situations

# MISSING VALUES TEST





# tests

---

- Little's test
- T-tests
- Regression with dummy variables (read Allison's chapter)

# Little's MCAR test

- Uses all of the available data
- Assumes multivariate normal distribution
  - ▣ i.e. only includes interval/ratio variables
  - ▣ (often likert scales are assumed to be interval)
- Chisquare test
- Reduces to t-test when data are bivariate with missing data confined to a single variable

# Little's MCAR test in R

- Package 'BaylorEdPsych'
- (also needs package mvnlme)
- Function to conduct test is:
- Where XXX is a dataframe with variables that contain missing values coded as NA
  - ▣ `LittleMCAR(XXX)`

# Install packages into R

First install the package:

- `install.packages("BaylorEdPsych")`

Then load the package into your environment so you can use the commands:

- `library("BaylorEdPsych",  
lib.loc="C:/R-3.2.2/library")`

# Install packages into R (interface)

The screenshot displays the RStudio environment. The main editor window shows a script named 'lecture2\_BIS.r' with R code for reading a dataset and exploring data. The console window at the bottom shows the execution of the script, including the command `install.packages("BayorEdPsych")` and an error message: `Error: could not find function "install.package"`. The 'Packages' pane on the right lists installed and available packages, with a blue arrow pointing to the 'Update' button. The 'Environment' pane shows the 'System Library' and various installed packages.

```
1 #####
2 ##### read datafile #####
3 #####
4 # set directory
5 setwd("...") # note / instead of \ in windows
6 setwd("...") # macs without c:
7 getwd() # print the current working directory
8
9
10 # read dataset as table & create dataframe
11 # first save spss dataset to csv (comma separated file by ;) in the directory
12 # (already done for you)
13 # look up function read.table in help (right lower pane) or use ?read.table
14
15 |
16 # read the csv file
17 # assign a name to the file
18
19
20
21 #####
22 ##### explore data #####
23 #####
24
25 # how does the first line of the file look like?
26
27
28 # select first column
29
30
31 # how many variables are there
32
33
34 # how many observations
35
36
```

Console output:

```
U:\Rcourses_2016\
creation() on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[workspace loaded from U:\Rcourses_2016\RData]

> x<- c(1,4,5,6, NA)
>
> x
[1] 1 4 5 6 NA
> na.omit(x)
[1] 1 4 5 6
attr(,"na.action")
[1] 5
attr(,"class")
[1] "omit"
> install.packages("BayorEdPsych")
Error: could not find function "install.package"
> |
```

Name	Description	Version
abind	Bind Multiple Arrays	1.4-3
acepack	Functions for selecting regression transformations	1.3-3.3
arm	Using Regression and Multilevel/Hierarchical Models	1.8-6
BH	Binding and Handling Files	1.58.0-1
boot	Bootstrap Functions (Originally by Angelo Canty for S)	1.3-17
brew	Tools for Report Generation	1.0-6
car	Comprehensive Regression Diagnostics	2.1-1
class	Functions for Classification	7.3-13
cluster	"Finding Groups in Data: Cluster Analysis Extended Rousseeuw et al.	2.0.3
coda	Output Analysis for Diagnostics for MCMC	0.18-1
codetools	Code Analysis Tools for R	0.2-14
colorspace	Color Space Manipulation	1.2-6
compiler	The R Compiler Package	3.2.2
corpcor	Efficient Estimation of Covariance and (Partial) Correlation	1.6.8
curl	A Modern and Flexible Web Client for R	0.9.3
d3Network	Tools for creating D3 JavaScript network, tree, dendrogram, and Sankey graphs from R	0.5.2.1
datasets	The R Datasets Package	3.2.2
devtools	Tools to Make Developing R Packages Easier	1.9.1
DiagrammeR	Create Graph Diagrams and Flowcharts Using R	0.8.1
dichromat	Color Schemes for Dichromats	2.0-0
digest	Create Cryptographic Hash Digests of R Objects	0.6.8
doParallel	Foreach Parallel Adaptor for the 'parallel' Package	1.0.10
ellipse	Functions for drawing ellipses and ellipse-like confidence regions	0.3-8
evaluate	Parsing and Evaluation Tools that Provide More Details than the Default	0.8
fdttool	Estimation of (Local) False Discovery Rates and Higher Criticism	1.2.15
foreach	Provides Foreach Looping Construct for R	1.4.3
foreign	Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, Weka, dBase, ...	0.8-65
Formula	Extended Model Formulas	1.2-1
ggm	Functions for graphical Markov models	2.3
ggplot2	An Implementation of the Grammar of Graphics	2.0.0
git2r	Provides Access to Git Repositories	0.11.0
glasso	Graphical lasso- estimation of Gaussian graphical models	1.8
graphics	The R Graphics Package	3.2.2
grDevices	The R Graphics Devices and Support for Colours and Fonts	3.2.2
grid	The Grid Graphics Package	3.2.2
gridBase	Integration of base and grid graphics	0.4-7
gridExtra	Miscellaneous Functions for 'Grid' Graphics	2.0.0

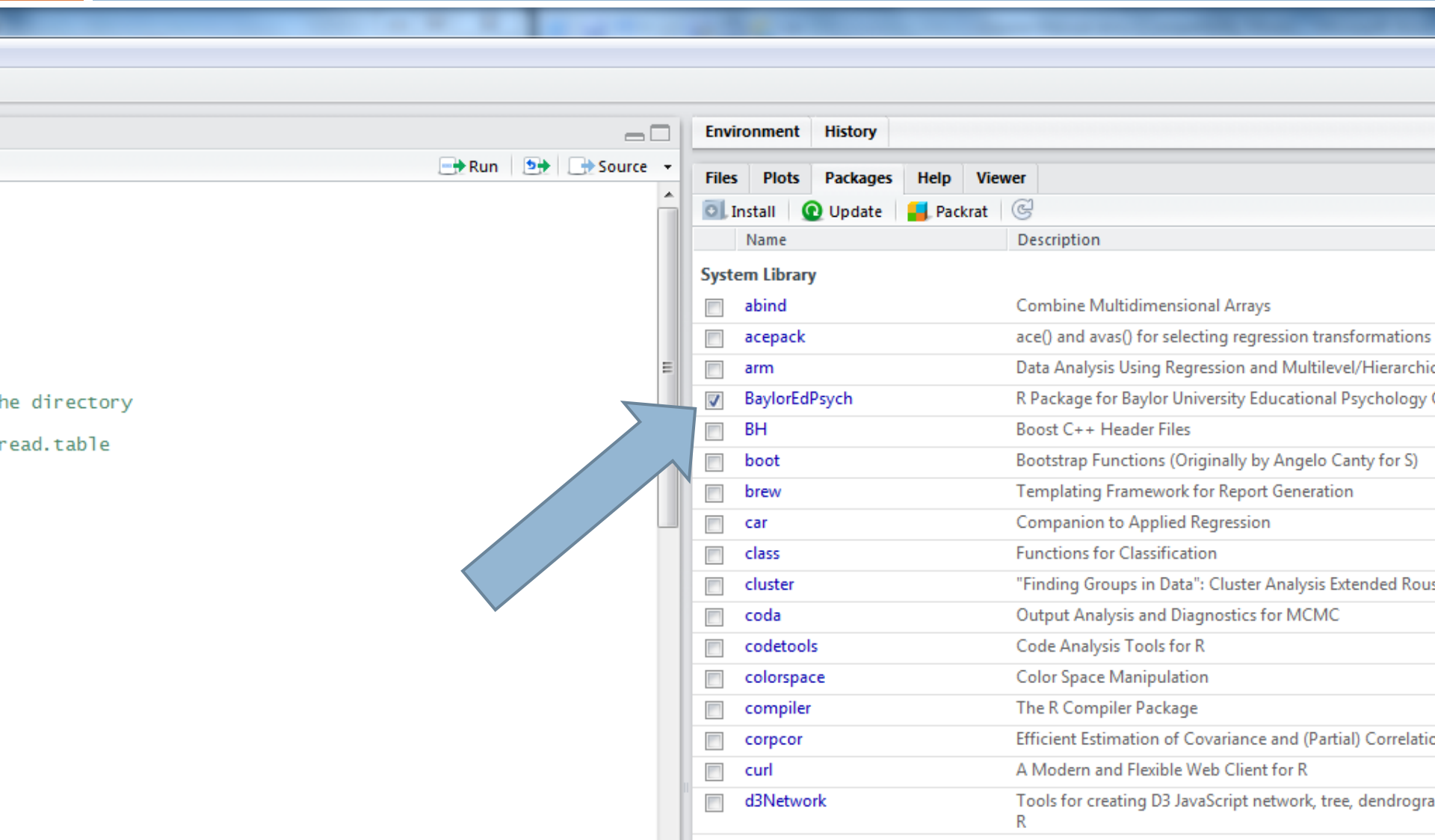
# Install packages into R (interface)

the directory  
?read.table

The screenshot shows the RStudio interface with the 'Install Packages' dialog box open. The dialog box has a dropdown menu for 'Install from:' set to 'Repository (CRAN, CRANextra)'. Below this, there is a text input field for 'Packages (separate names with space or comma):' containing the text 'Bay'. A list of suggested packages is shown below the input field, including BayClone2, BayesBD, BayesBridge, BayesComm, bayescount, BayesCR, BayesDA, bayesDccGarch, bayesDem, BayesFactor, bayesGARCH, bayesGDS, BayesGESM, BayesianAnimalTracker, Bayesianbetareg, BayesLCA, and bayesLife. The 'Install' button is visible at the bottom of the dialog box. In the background, the 'Packages' pane is visible, showing a list of installed and available packages. Two large blue arrows point from the right side of the image towards the 'Repository (CRAN, CRANextra)' dropdown and the 'Packages' input field in the dialog box.

Name	Description
abind	Bind multiple arrays into a multidimensional array
acepack	Accelerated regression transformations
	Multilevel/Hierarchical Models
	Boost Classification
	Bootstrap Functions (by Angelo Canty for S)
	Templating Framework for Report Generation
	Companion to Applied Regression
	Functions for Classification
	"Finding Groups in Data": Cluster Analysis Extended Rousseeuw et al.
	Output Analysis and Diagnostics for MCMC
	Code Analysis Tools for R
	Color Space Manipulation
	The R Compiler Package
	Efficient Estimation of Covariance and (Partial) Correlation
	A Modern and Flexible Web Client for R
	Tools for creating D3 JavaScript network, tree, dendrogram, and Sankey graphs from R
	The R Datasets Package
	Tools to Make Developing R Packages Easier
	Create Graph Diagrams and Flowcharts Using R
	Color Schemes for Dichromats
	Create Cryptographic Hash Digests of R Objects
	Foreach Parallel Adaptor for the 'parallel' Package
	Functions for deriving ellipses and ellipsoids like confidence regions

# Install packages into R (interface)



# T-test: MCAR vs MAR

- Create dummy variable of missingness on variable of interest
- Run t-tests (and x2 tests) between this variable and other variables in the dataset



# Regression

## Dummy variable adjustment

1. Create a dummy variable for missingness (0 = not missing, 1 = missing) for each predictor
2. Add these variables to the regression

PRODUCES BIASED ESTIMATES OF REGRESSION COEFFICIENTS

# Exercise 2\_2.r

- Analyze missing data
- Install package to conduct the Little MCAR's test
- (see slides on install packages – bb)
  
- Create own function to conduct t-test for missing data

IMPUTATION



# Imputation

- Single Methods
  - ▣ Mean substitution
  - ▣ Regression imputation
- Model based methods
  - ▣ Maximum likelihood
  - ▣ Multiple imputation
- Consequences

# Mean

Replace missing value with sample mean

Run analyses as if all complete cases

- Advantages:

- ▣ Can use complete case analysis methods

- Disadvantages:

- ▣ Reduces variability (variances underestimated)
  - ▣ Weakens covariance and correlation estimates in the data

# Regression imputation

- Replace missing values with predicted score
- Advantage:
  - ▣ use information from data
- Disadvantage:
  - ▣ Overestimates model fit (overfitting)
  - ▣ Downward SE estimates
  - ▣ Ignores uncertainty in imputed values

Predicts most likely value but does not show uncertainty about value

# Multiple imputation

- Data is imputed  $m$  times. Imputed values are estimated by regression to which an error term is added. The errors are drawn from distribution (and can be different every time).
- Analyze each  $m$  dataset
- Integrate  $m$  analysis results.

# Maximum likelihood

Likelihood function: Expresses the probability of the data as a function of the data and the unknown parameter values

- Advantages

- Uses full information (both complete and incomplete cases) to calculate log-likelihood
- Unbiased parameter estimates when MCAR/MAR

- Disadvantages

- SE biased downward
- Assumes multivariate normal distributions

Paul D. Allison (2012). Handling missing data by maximum likelihood.

<http://www.statisticalhorizons.com/wp-content/uploads/MissingDataByML.pdf>



# ML and ML

If assumptions are met, the estimators are

1. Consistent (approx. Unbiased)
2. Asumptotically efficient (minimal sampling variability)
3. Asymptotically normal (justifies p-values and confidence intervals)

But

ML is more efficient than ML

ML gives always the same result

# Assumptions & listwise deletion

- Listwise deletion will produce unbiased estimates when missing data is MAR
  - ▣ Most likely similar to multiple imputation
- Listwise deletion will produce unbiased estimates even if data is not MAR
  - ▣ This biased estimates could happen when using multiple imputation or ML (i.e. when data are missing on predictor variable in regression analysis, e.g. to predict income based on number of children)

When impute the data, make sure you have the correctly specified model! (if not, use listwise deletion)

# Exercise 2\_3.r

- Impute values to these missing values
  - ▣ Multiple imputation (amelia)
  - ▣ Maximum likelihood (mvnmle)
  
- Compare the two in a simple regression
  - ▣ Use 'lm' function
  - ▣ Look up function ?lm

# ASSIGNMENT



# Assignment 1

Each group needs to select two countries

- You will analyze the missing values of income
- Conduct multiple imputation to impute these values
- Estimate an regression model including income to investigate the effect of imputation (compare with listwise deletion)
- Compare the two countries (include moderator)

For detailed information, see [GitHub](#)