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(dfc("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 is missing | X, Y) = P(Y 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

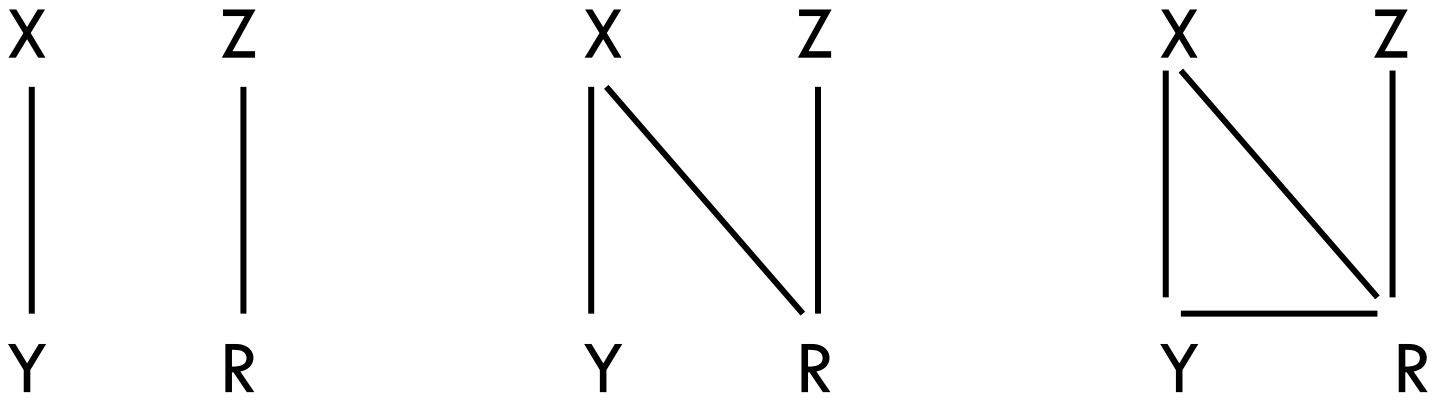
Missing at Random (MAR)

- 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= Yobs, Y = Ymis , R = response indicator

Z = '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

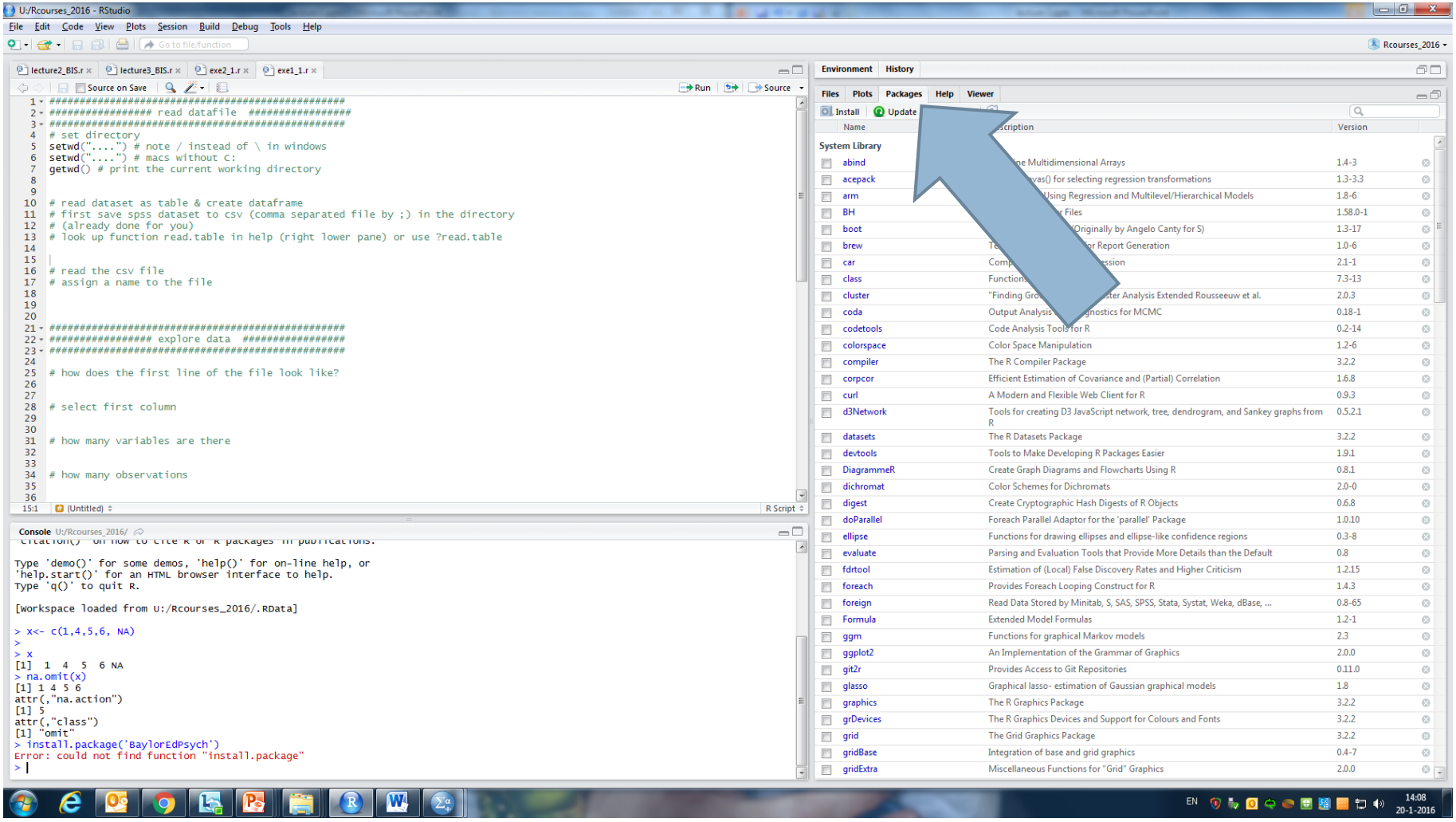
First install the package:

- `install.packages("BaylorEdPsych")`

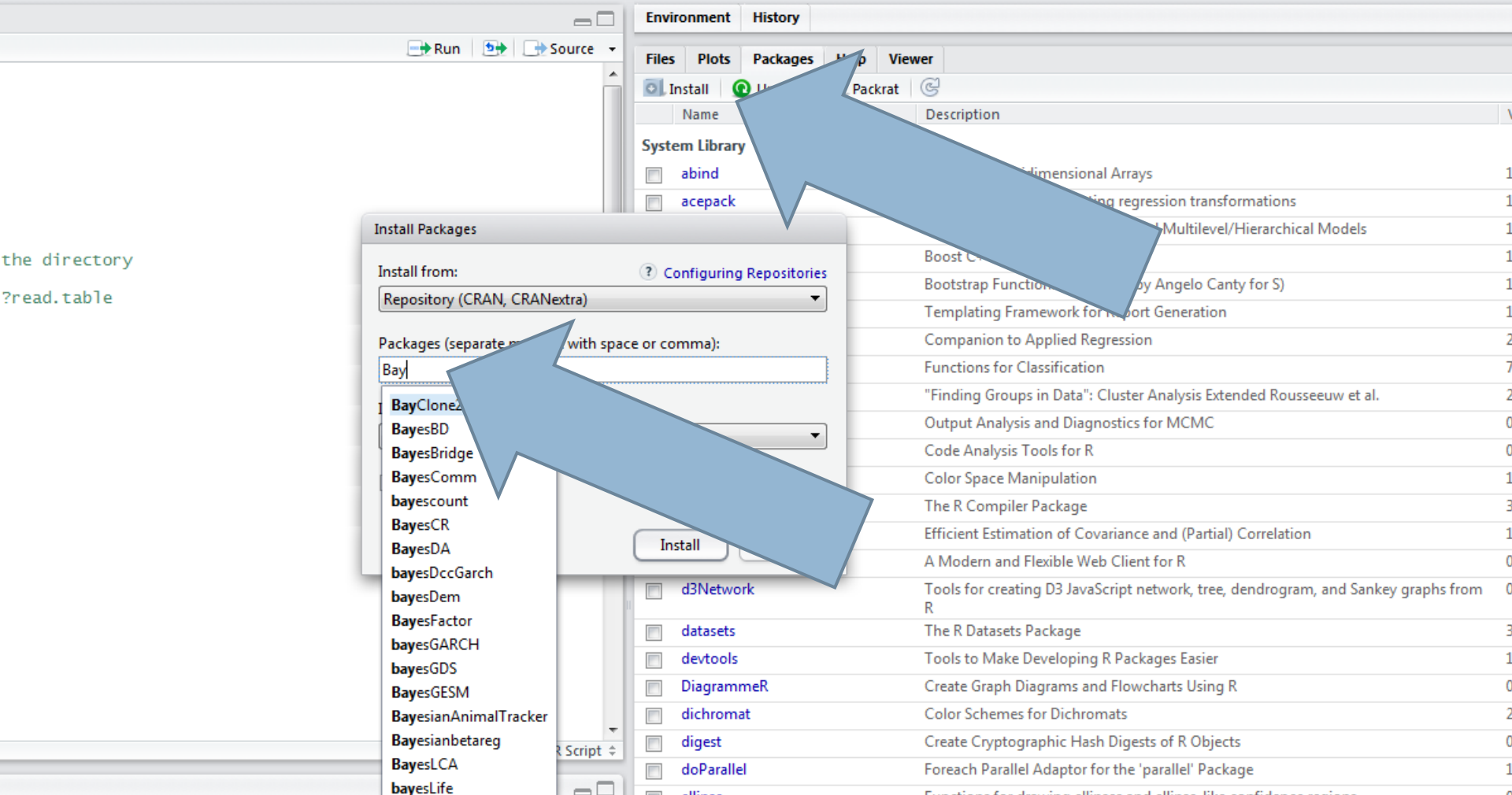
Then load the package into your environment so you can use the commands:

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

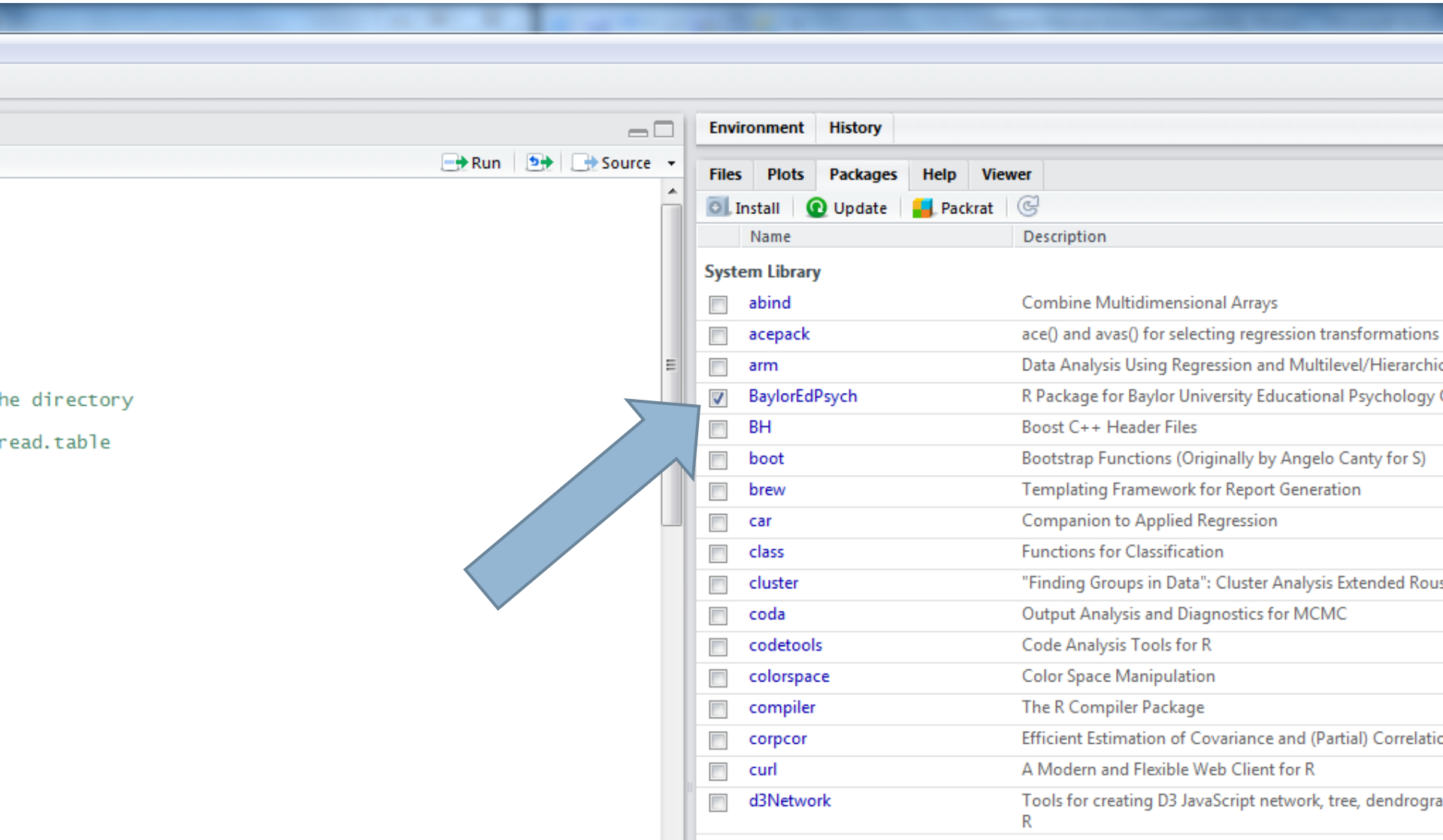
Install packages into R (interface)



Install packages into R (interface)



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 uncertaintly 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

MI 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 MI

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)

http://statisticalhorizons.com/listwise-deletion-its-not-evil

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