SEM 1: Confirmatory Factor Analysis

Week 4 - Missing data

Sacha Epskamp

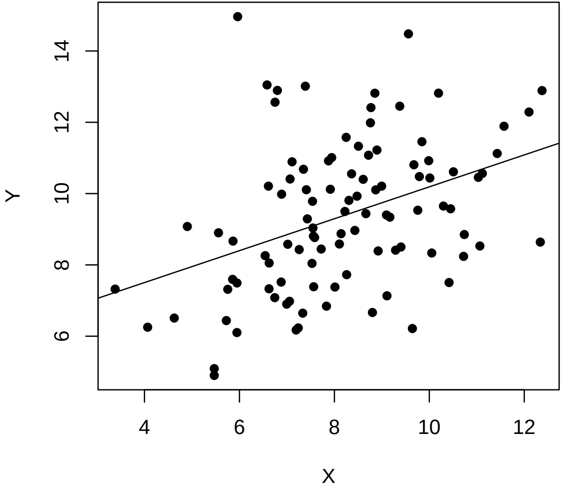
2020

Why are data missing? In a general X predicts Y case:

- **Missing completely at random** (MCAR)
  - Missingness is independent of Y or X
  - Everything is fine!
- **Missing at random** (MAR)
  - Missingness is independent of Y, but not of X
  - Example: Men less willing to respond to mental health questionnaire
  - Not a big problem
- **Missing not at random** (MNAR)
  - Missingness depends on Y
  - Example: People with severe mental health problems fill in less questionnaires
  - This is bad :(

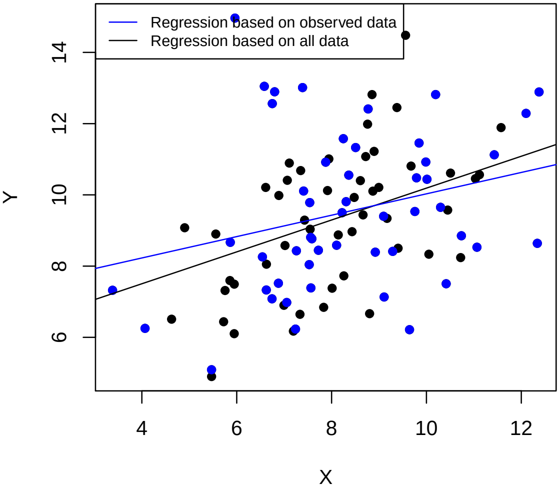
Unfortunatly, there is no way to know exactly how your data is missing.

X	Y
5	5
6	5
5	6
8	5
6	7
7	7
6	9
9	8
9	9
12	9
⋮	⋮



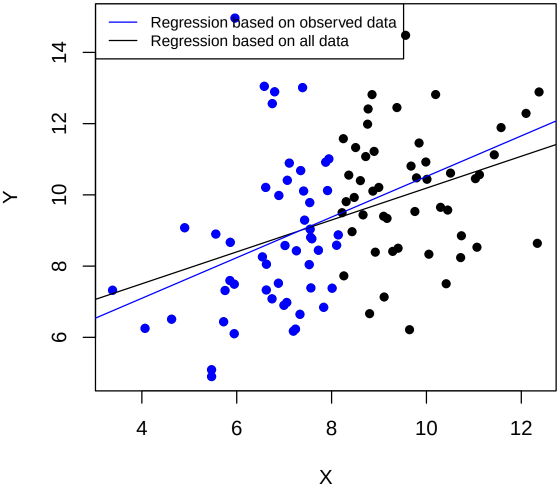
Missing completely at random (MCAR)

X	Y
5	5
6	5
5	6
8	5
6	7
7	7
6	9
9	8
9	9
12	9
⋮	⋮



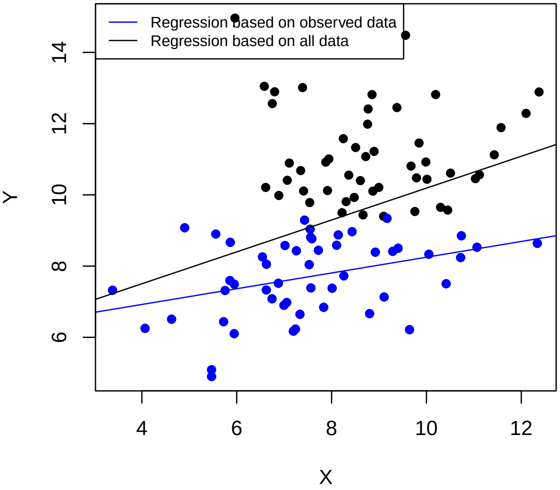
Missing at random (MAR)

X	Y
5	5
6	5
5	6
8	5
6	7
7	7
6	9
9	8
9	9
12	9
⋮	⋮



Missing not at random (MNAR)

X	Y
5	5
6	5
5	6
8	5
6	7
7	7
6	9
9	8
9	9
12	9
⋮	⋮



"Older" methods of handling missing data:

"Older" methods of handling missing data:

- Compute $\boldsymbol{S}$ using listwise deletion
  - Assumes MCAR
  - Delete all rows with a missing value
  - Downside: deletes observed data

"Older" methods of handling missing data:

- Compute $\boldsymbol{S}$ using listwise deletion
  - Assumes MCAR
  - Delete all rows with a missing value
  - Downside: deletes observed data
- Compute $\boldsymbol{S}$ using pairwise estimation
  - Assumes MAR
  - Estimate each element of $\boldsymbol{S}$ using all available data
  - Downside: Each covariance is based on different n and variance–covariance matrix might not be positive definite

"Older" methods of handling missing data:

- Compute $\boldsymbol{S}$ using listwise deletion
  - Assumes MCAR
  - Delete all rows with a missing value
  - Downside: deletes observed data
- Compute $\boldsymbol{S}$ using pairwise estimation
  - Assumes MAR
  - Estimate each element of $\boldsymbol{S}$ using all available data
  - Downside: Each covariance is based on different n and variance–covariance matrix might not be positive definite
- (Multiple) inputations
  - Assumes MAR
  - Inpute missingness using mean scores or regression models
  - Downside: complicated, can increase bias if MNAR

Full-information maximum likelihood (FIML):

- Compute likelihood for each person or each data subset with the same missingness pattern
- Assumes MAR
- Uses the full data set and all observations
- Downside: full data needed (analysis can not be done using covariance matrix)
- Implemented in most software (e.g., lavaan, Mplus, psychonetrics)
  - Use `missing = "FIML"` in *lavaan* or `estimator = "FIML"` in *psychonetrics*

FIML estimator in *psychonetrics* for every subset of data i with same missingness pattern:

$$F_{\mathrm{FIML}} = \frac{1}{n} \sum_i n_i \left(\mathrm{trace}\left(\boldsymbol{S}_i \boldsymbol{\Sigma}_i^{-1} \right) + \left(\bar{\boldsymbol{y}}_i - \boldsymbol{\mu}_i \right)^\top \boldsymbol{\Sigma}_i^{-1} \left(\bar{\boldsymbol{y}}_i - \boldsymbol{\mu}_i \right) - \ln \left| \boldsymbol{\Sigma}_i^{-1} \right| \right)$$

- n_i: sample size of subset i
- $\boldsymbol{S}_i$: sample covariances (ML) of subset i (note, $\boldsymbol{S}_i = \boldsymbol{O}$ if $n_i = 1$)
- $\bar{\boldsymbol{y}}_i$: sample means of of subset i (note, same as the observed score if $n_i = 1$)
- $\boldsymbol{\Sigma}_i$: Subset of $\boldsymbol{\Sigma}$ containing only elements of observed data in subset i
- $\boldsymbol{\mu}_i$: Subset of $\boldsymbol{\mu}$ containing only elements of observed data in subset i

Downside: saturated model needs to be computed as well.

Assumptions of maximum likelihood estimation

1. Independence: Observations are a simple random sample from some population
   - Consequence: underestimated standard errors, inflated Type-I error rates
   - Solution 1: use SE correction for dependence structure
   - Solution 2: multilevel SEM
2. Multivariate Normality: Variables are univariate normally distributed at levels of all other variables, residuals are normal and homoscedastic, latent variables are normal, bivariate relations are linear
   - Consequence: standard errors are incorrect (probably too low), χ^2 test value is not accurate (probably too high)
   - Solution 1: use "robust" standard errors (`estimator = 'MLM'`, with complete data; `estimator = 'MLR'` with incomplete data
   - Solution 2: use bootstrapped standard errors & test statistic