Missing data in health surveys

Kyrre Breivik

Regional Centre for Child and Youth Mental Health and Child Welfare, NORCE Norwegian Research

Centre,Centre on Patient-Reported Outcomes, Department of Research and Development, Haukeland University

Prevalance of missing data

- Political survey data 50 % missing (King et al., 2001)
- Family research. Often as much as 50% missing(Acock, 2005).
- The studies had on average more than 30% missing in a prominent psychological journal (Mcknight et. al., 2007)
- There has genereally been a decline in response rate of surveys during the last decades (Luiten et. al., 2020)
- The response rate in the Trøndelag Health Study (The HUNT Study) has declined from 88% in HUNT1(1984-86) , 71% in HUNT2 (1995-97) to 54% in HUNT3(2006-08) ,

Is this worrying?

- YES

a) Biased parameters (means, variances, correlations etc)

b) less statistical power

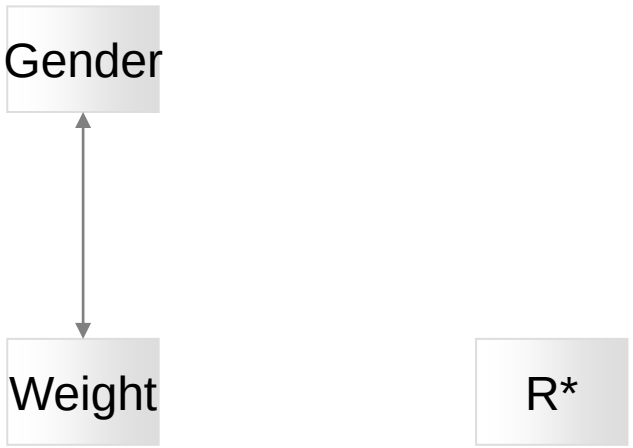
c) technical problems when running analyses

- BUT some (sane) researchers actually want missingness and plan for it.

- Bias depends on degree of control of missing data mechanisms and how you handle it.

Missing mechanisms (Rubin, 1976)

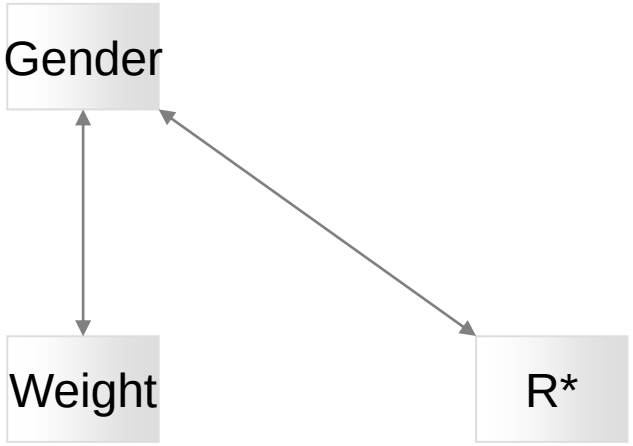
- Missing completely at random (MCAR)
  1. Cases that are missing are a random sample of all the cases in the study (Graham, 2009)
  2. Cause not correlated with variables of interest (Schafer & Graham, 2002)



* Missing on Weight

Missingness mechanisms (Rubin, 1976)

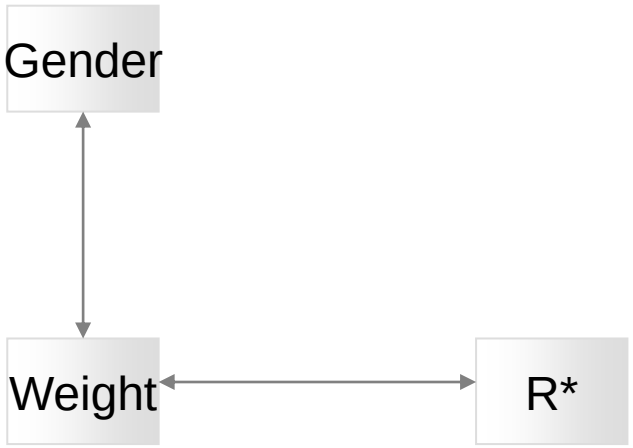
- Missing at random (MAR)
  - MAR if missingness is related to other measured variables in the analysis model, but not to the underlying values of the incomplete variable (Baraldi & Enders, 2010)
  - Example: Girls less prone to answer items measuring weight.



* Missing on Weight

Missingness mechanisms (Rubin, 1976)

- Missing not at random (MNAR)
  - Missingness depends on unobserved data after one has controlled for the variables in the analysis.
  - Example: Overweight persons less prone to answer items measuring weight.



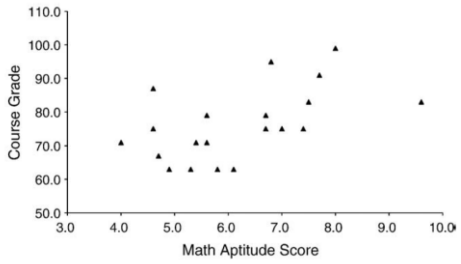
* Missing on Weight

Traditional approaches to handle missing data

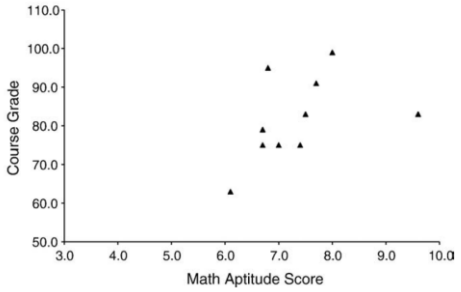
- Listwise deletion
  - Default in many programs
  - Better safe than sorry?
  - MCAR assumption
  - reduced statistical power
  - Muligens biased resultater (MCAR)

- Pairwise deletion
  - Usually more statistical power than listwise
  - Uncertain sample size
  - MCAR assumption
  - Possible convergence problems (matrix not positive definite)

Komplette data, Gjennomsn = 76.35, SD = 10,73, r=.55



Listwise deletion, Gjennomsn = 81.80, SD = 10,84, r=.44

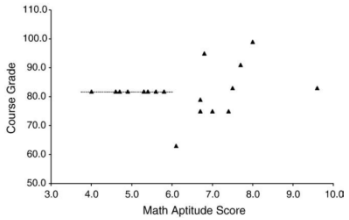


Baraldi & Enders, 2009

Traditional approaches to handle missing data

- Mean imputation
  - Several reasearchers regard this as the worst method
  - Possible bias even with MCAR!
  - Too low uncertainty (standardfeil)

Imputering gjennomsnitt, Gjennomsn = 81.80, SD = 7.46, r=.21



Baraldi & Enders, 2009

- Proration (person mean imputation)
  - Calculate scale scores by averaging the available items for each individual
  - Can work ok if missing item has similar values as the other available items in a scale + has the same correlation with the total as the other available items.
  - If this is not the case, the method can give strong bias also at MCAR (Mazza et. Al., 2015)
  - In most cases, it is better to use more modern methods

Mazza, G. L., Enders, C. K., & Ruehlman, L. S. (2015). Addressing Item-Level Missing Data: A Comparison of Proration and Full Information Maximum Likelihood Estimation. *Multivariate behavioral research*, *50*(5), 504-519.

Traditional approaches to handle missing data

- Other traditional approaches such as pairwise deletion, hot deck imputation, last observation carried forward, regression imputation either assumes MCAR or might even be biased under MCAR (e.g. Enders, 2010)

- American Psychological Association Task Force on Statistical Inference (1999). Traditional missing-data methods are ”among the worst methods available for practical applications”.

- Monitor on psychology (2002): ”Routine implimentation of these new methods for adressing missing data will be one of the major changes in research in the next decades.”

- Various reviews throughout the lates decade: Deletion methods continues to be the most common way handling missing data in various research fields such as counseling psychology, pediatric psychology, epidemiology , and prevention science (Riox & Little, 2021)

Methods based on MAR

- Full Information Maximum Likelihood (FIML)
- Multiple imputation
- Weighting (Vansteelandt et. al., 2010)

State of the art (Graham & Schafer, 2002)

- These methods are biased when missing is MNAR
- MNAR methods (e.g. selection models, pattern mixture models), is an important area of research. The models which exists
  - are often complicated to set up/run
  - rely on strong untestable assumptions which can lead to strong bias if not met

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. Psychological methods, 7(2), 147

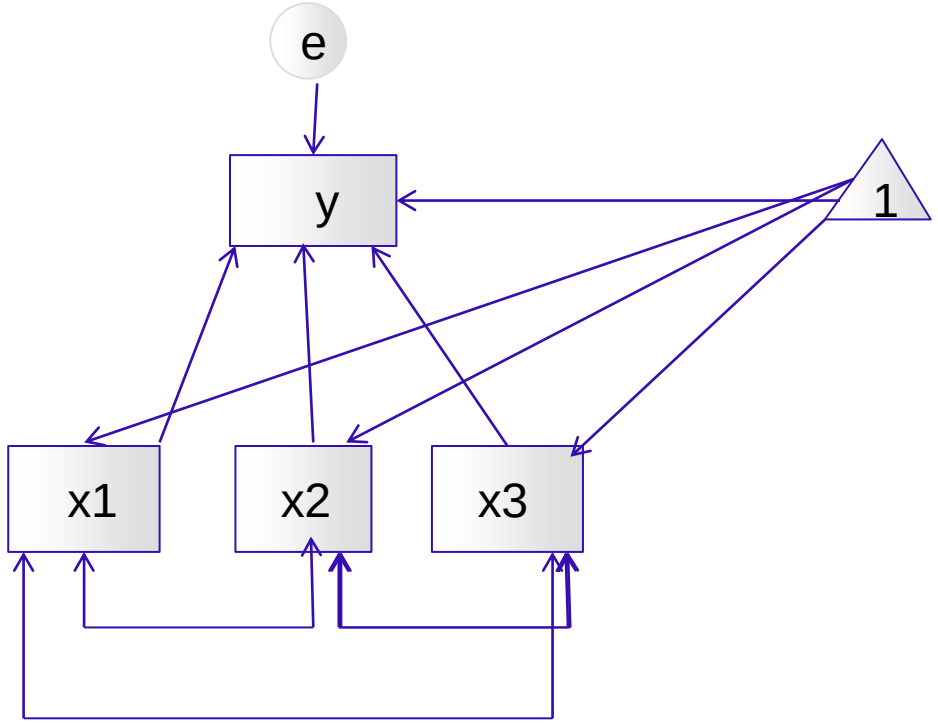
Vansteelandt S, Carpenter J and Kenward MG. Analysis of incomplete data using inverse probability weighting and doubly robu
estimators. Methodology 2010; 6: 37–48.

FIML

- Does not impute data
- Uses all available information from the individuals to find the most probable parameter values in the population
- Auditions different parameter values (means, correlations) to find out which parameter values have the greatest probability (maximum likelihood) of being given e.g. a multivariate normal distribution

$$logL = \sum_{i=1}^{N} log\left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-.5\left(\frac{y_i-\mu}{\sigma}\right)^2}\right].$$

FIML



- A structural equation model is based on a covariance matrix (variance and covariance). When using FIML, it also includes a mean vector
- Individuals with varying degrees of missing on the variables can contribute varying degrees of information about the overall model.
  - Individuals only answering x1 will contribute to find the most probable average value of x.
  - Individuals with complete data will be able to contribute information about covariances, variances and averages on the entire model

FIML

Ex: Formula for individual who has missing all variables except one.

$$(\mathbf{Y}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}) = (4.0 - \mu_{MA})^T \frac{1}{\sigma^2_{MA}} (4.0 - \mu_{MA}),$$

μ_{MA} = mean, σ^2_{MA} = variance

- Formula for individual as 2 observed variables

$$(\mathbf{Y}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}) = \left(\begin{bmatrix} 9.6 \\ 83.0 \end{bmatrix} - \begin{bmatrix} \mu_{MA} \\ \mu_{CG} \end{bmatrix}^T \right) \begin{bmatrix} \sigma^2_{MA} & \sigma_{MA,CG} \\ \sigma_{CG,MA} & \sigma^2_{CG} \end{bmatrix}^{-1} \left(\begin{bmatrix} 9.6 \\ 83.0 \end{bmatrix} - \begin{bmatrix} \mu_{MA} \\ \mu_{CG} \end{bmatrix} \right).$$

Multippel imputasjon (MI)

1. Imputation of x number of datasets
   - Imputation phase (I): Mean values and the covariance matrix are used to predict the most probable value for missing (regression) + random variation (reminiscent of imputation by stochastic regression). The complete data set is forwarded to the P-phase
   - Posterior phase (P): Construction of a new covariance matrix and mean values using Baysian estimator. **Adds random variation to send back to I**
   - Repeats this process x times so that you get x datasets
2. Run the same analysis (eg multiple regression) on x number of datasets
3. Merging of results

FIML vs Multiple imputation

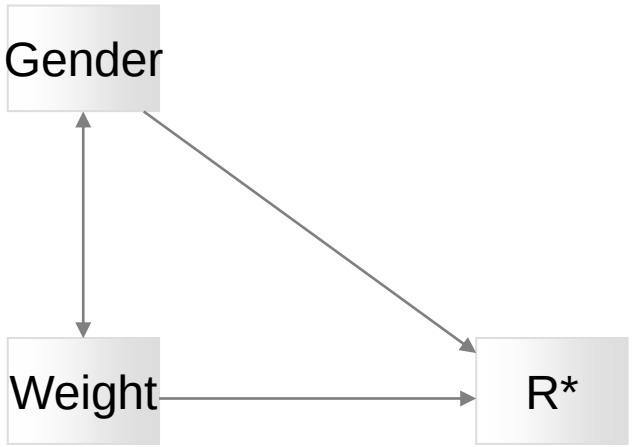
- Both work well and usually gives very similar results
- FIML is easier to use, but MI can be used in statistical programs that do not have FIML (e.g. SPSS)
- MI can be used on categorical outcomes
- MI probably works better than FIML if you have a small samples. Worked fine with n = 50 with 50% missing (Graham, 2009).

Auxiliary variabler

- MAR → Missing data depends on the variables <u>included</u> in the analysis
- Important to include variables that explain or correlate with missing or correlate with variables with missing.

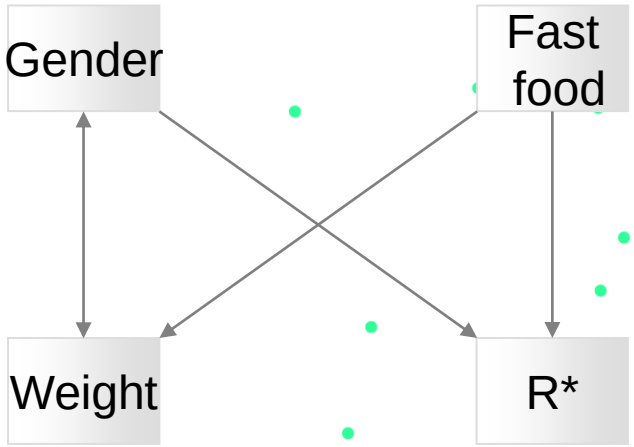
NMAR

a



MAR

b



* Missing on Weight

Auxiliary variabler

- Good auxiliary variables: Relatively strong correlation with variables with missing data (r> 0.40) (Collins et al, 2001; Enders, 2010)

- If they are not part of the analysis, they can be
  - Multiple imputation: part of the dataset used for imputation
  - FIML: included as satured auxillary variables

Satured correlates model (Graham, 2003)

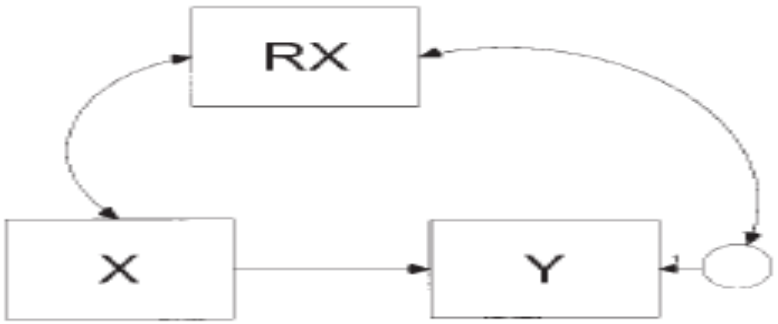


FIGURE 5 Model 3: "saturated correlates" model (manifest variable version).

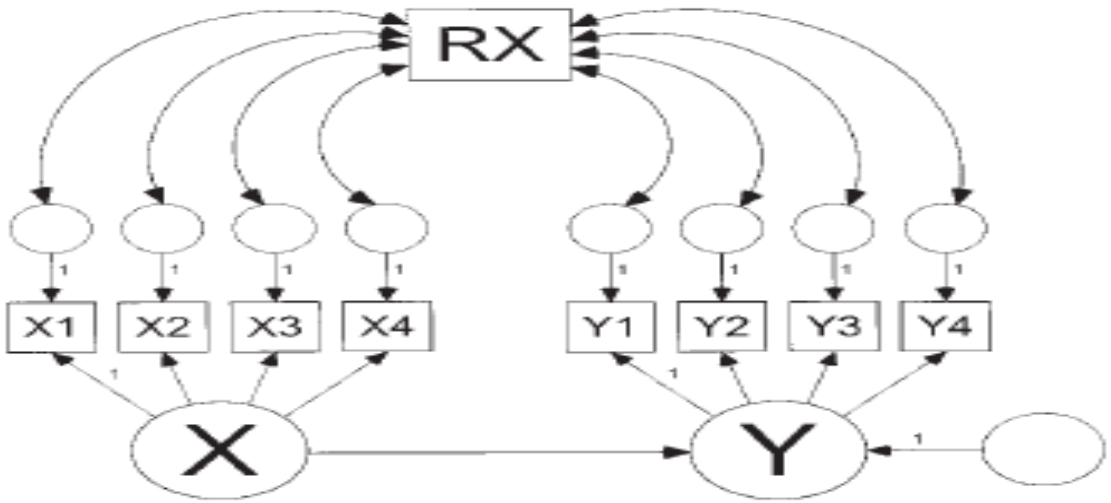


FIGURE 6 Model 3: "saturated correlates" model (latent variable version).

What about survey-level missingness?

- Multiple imputation and FIML relies on item-level or scale level missingness
- If survey-level missingness is in:
  - a longitudinal study – MI and FIML can be used as previous scores on the relevant variables are often good predictors of missing
  - A cross-sectional study – nonresponders are a kind of listwise deletion and bias depends whether the data is MCAR

Suervey-level missing and bias

- Low responserate does not necessarly equal biased results
- Degree of bias depends on the interaction between a) how much respondents and nonrepondents differ on the relevant parameters (e.g. means, prevalence, correlations) and response rate
- Researchers should aim at 1) getting a high responserate and 2)reporting all relevant information which can illuminate possible nonresponder bias.

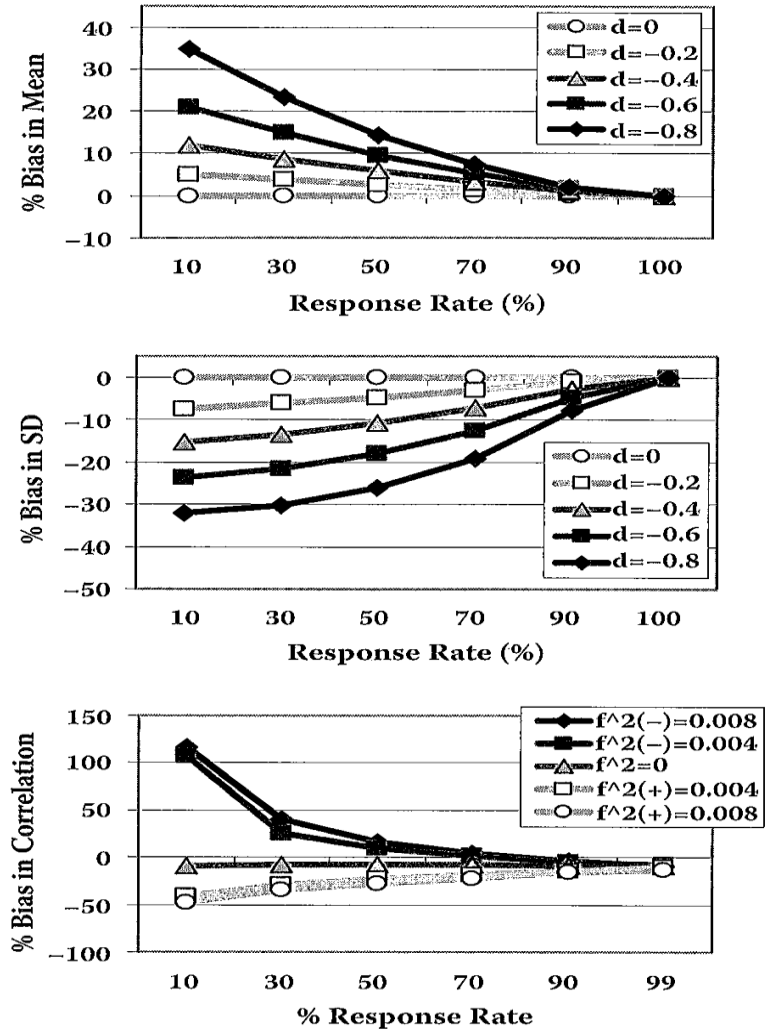


Figure 1.1 (a) Response rate bias in the mean. (b) Response rate bias in the standard deviation. (c) Response rate biases in the correlation. *Note.* Mean bias evaluated at $\bar{X}_{resp} = 4$; correlation bias at $r_{resp} = .3$; $d_{miss_x} =$ and $d_{miss_y} = -.4$.

d=standardized difference between responders and nonresponders (Cohen's d), d>0 is MNAR

f2= describes the strength of the moderating effect (ratio of systematic variance accounted for by the moderator relative to unexplained variance in the criterion) (Aguinis et. al., 2005)

Newman, D. A. (2009). Missing data techniques and low response rates: The role of systematic nonresponse parameters. In C. E. Lance & R. J. Vandenberg (Eds.), Statistical and methodological myths and urban legends: Doctrine, verity, and fable in the organizational and social sciences (pp. 7-36). New York, NY: Routledge

Conclusion

- Missing data is a frequent in health surveys due to survey, scale and item level nonresponding

- Should aim at keeping missing data as low as possible as it associated with bias. A high degree of survey-level nonresponding is especially problematic as we often lack information which can be used for imputing data

- Missing data is better handled by multiple imputation and FIML (+ weighting)than traditional methods such as listwise deletion, pairwise deletion and mean impution

Recommended litterature

- Enders, C. K. (2010). Applied missing data analysis. New York, NY: Guilford Press.

- Graham, J. W. (2009). Missing data analysis: Making it work in the real world. Annual review of psychology, 60, 549-576.

- Newman, D. A. (2009). Missing data techniques and low response rates: The role of systematic nonresponse parameters. In C. E. Lance & R. J. Vandenberg (Eds.), Statistical and methodological myths and urban legends: Doctrine, verity, and fable in the organizational and social sciences (pp. 7-36). New York, NY: Routledge

- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. Psychological Methods, 7, 147–177

- Vansteelandt S, Carpenter J and Kenward MG. Analysis of incomplete data using inverse probability weighting and doubly robust estimators. Methodology 2010; 6: 37–48.