Lecture 2 - Shortcomings of ad-hoc methods, complete case analysis, & introduction to multiple imputation (MI)

Multiple imputation techniques for working with missing data

Jonathan Bartlett (thestatsgeek.com)

Copenhagen, March 2020

Ad-hoc methods

Complete case analysis

Multiple imputation

Ad-hoc methods

Ad-hoc methods

- Ad-hoc methods are simple and easy apparent solutions to handling missing data.
- Commonly used examples are: simple mean imputation, missing category method, last observation carried forward.
- All (except missing category) are examples of single imputation methods.
- Question: in general will we get valid answers using these ad-hoc methods?

Issues with ad-hoc methods

- Answer: No in general.
- They can introduce bias into estimates.
- They can lead to confidence intervals that are too narrow.
- The latter is true of all single imputation methods, unless special procedures are used to allow for uncertainty due to imputation (e.g. bootstrapping)

Simple mean imputation

- Replaces missing values with mean of observed.
- Variance of the variable is artificially reduced.
- Associations with other variables are distorted.
- **It's a bad idea!**

Regression mean imputation

- Replaces missing values with prediction based on observed variables.
- Better than mean imputation.
- Variance of the variable is still too small.
- Associations with other variables may still be distorted.
- **It's better, but still a bad idea!**

Missing category method

- For categorical variables with missing values, create a new missing category.
- In general regression coefficients after using this method are biased.
- To see why, think about the case where the variable is a confounder. . .
- An exception is with missing baseline in randomized trials, see (White and Thompson 2005)

Last observation carried forward (LOCF)

- In longitudinal studies, an approach that was historically popular is last observation carried forward (LOCF).
- Makes strong, implausible assumptions.
- In general neither conservative or liberal for treatment effects.
- Bias depends on unknown treatment effect!
- See (Molenberghs et al. 2004; Cook, Zeng, and Yi 2004; Carpenter et al. 2004)

Ad-hoc methods summary

- Ad-hoc methods are an attempt to ‘solve’ the problem of missing data.
- They avoid any serious thinking about the issues raised by missing data.
- They do not utilize statistical principles.
- Generally they result in misleading conclusions.

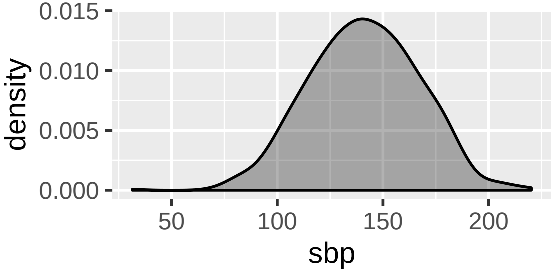
Complete case analysis

Complete case analysis

- Complete case analysis (CCA) ignores all units/observations with incomplete data in those variables involved in analysis.
- It is the default of most (all?!) statistical packages when presented with missing data.
- We will lose precision in estimates (compared to full data).
- We explore biases of CCA in different situations. . .

Marginal estimands

- Suppose we were interested in estimating the (marginal) mean systolic blood pressure (SBP) in a population.
- The plot below shows the complete sample (n=1000) of SBP values:



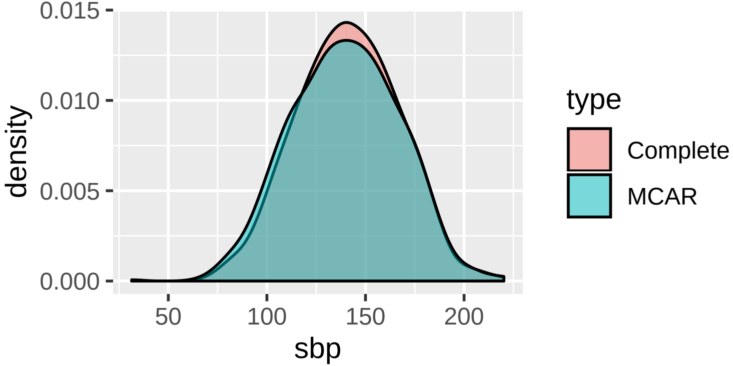
The mean is 141.1

MCAR - what will happen?

- Now we will make 50% of values missing.
- If we make them missing completely at random, will the complete case distribution and mean go up or down?

MCAR - what will happen?

The complete case mean is 139.9



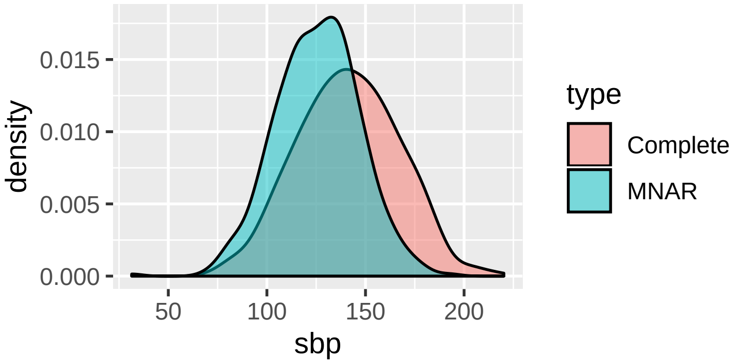
No bias

MNAR - what will happen?

- Now we will make higher values of SBP more likely to be missing.
- Will the complete case distribution and mean go up or down?

MNAR - what will happen?

The complete case mean is 125.6



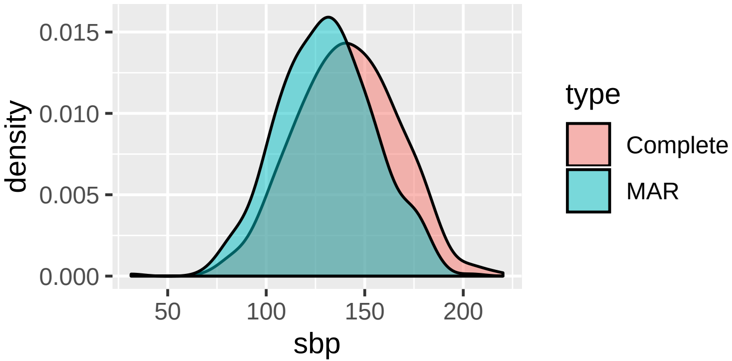
Biased downwards

MAR - what will happen?

- Now we will make values of SBP more likely to be missing if the person's age (assumed fully observed) is high.
- Will the complete case distribution and mean go up or down?

MAR - what will happen?

The complete case mean is 130.6



Biased downwards

Marginal estimands - conclusions

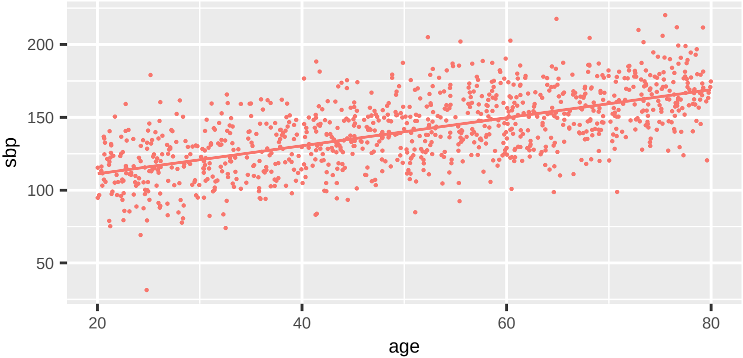
- For marginal estimands like means, we get bias under both MAR and MNAR mechanisms.
- Whether we get bias just depends on whether missingness is MCAR or not.
- Since in practice MCAR often doesn't hold, CCA for marginal estimands will usually be biased.

Complete case analysis - regression analyses

- Often we are interested in fitting a regression model for an outcome Y on covariates $X_1, .., X_p$.
- A CCA drops any observations which have one or more values missing in the variables used in the regression.
- With many variables in the regression and sporadic missingness, the complete cases can be a small subset, leading to big loss in information.
- What about bias?

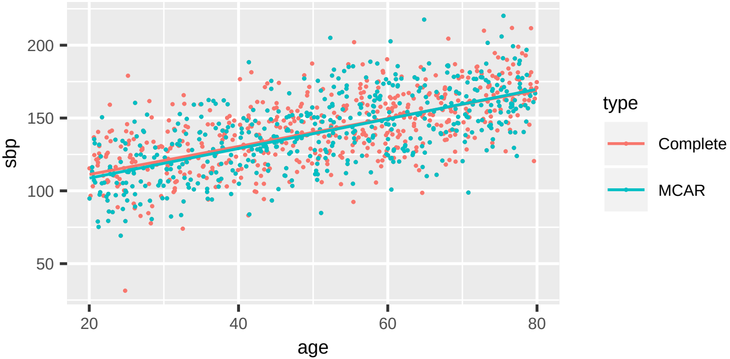
Linear regression complete case analysis

This is the full (complete) SBP against age data.



MCAR complete case regression

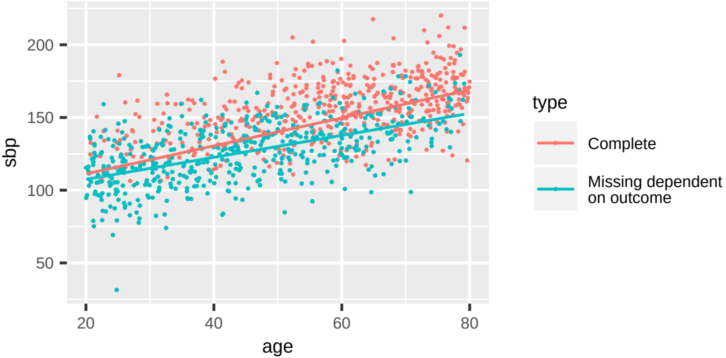
This is the CCA of the MCAR dataset.



CCA is unbiased, as we should expect.

Missingness dependent on outcome

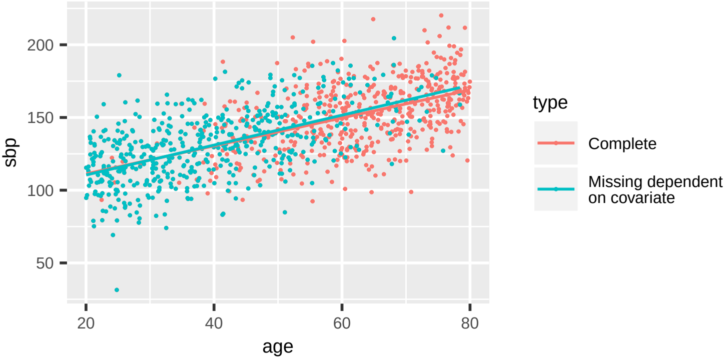
When missingness depends on outcome (SBP):



CCA is now biased.

Missingness dependent on covariate

When missingness depends on the covariate (age):



CCA is unbiased.

Complete case analysis - regression analyses

- CCA regression analyses are unbiased if probability of being a complete case is independent of outcome variable, conditional on covariates.
- This condition doesn't related to which variable(s) (outcome or covariates) have missing values.
- It is not true that CCA is valid under MAR and invalid under MNAR.
- It can be valid under both types - the key is whether missingness is conditionally (on covariates) independent of outcome.

Justification

Why does this result hold in general?

- Let R denote whether a subject is a complete case ($R = 1$ for complete cases, $R = 0$ for incomplete cases)
- Our assumption for missingness is that $f(R|Y, \mathbf{X}) = f(R|\mathbf{X})$
- A CCA involves fitting the conditional model for $f(Y|\mathbf{X})$ in the subset of subjects with $R = 1$:

$$\begin{aligned} f(Y|\mathbf{X}, R = 1) = \frac{f(Y, \mathbf{X}, R = 1)}{f(\mathbf{X}, R = 1)} &= \frac{f(R = 1|\mathbf{X}, Y)f(\mathbf{X}, Y)}{f(R = 1|\mathbf{X})f(\mathbf{X})} \\ &= \frac{f(R = 1|\mathbf{X})f(\mathbf{X}, Y)}{f(R = 1|\mathbf{X})f(\mathbf{X})} \\ &= f(Y|\mathbf{X}) \end{aligned}$$

- Thus the conditional distribution $Y|X$ in the complete cases is the same as in the complete data.

Complete case validity - Example

- (Bartlett et al. 2014) reported results of an illustrative analysis based on cross-sectional data from the US NHANES 2003-2004 study.
- They fitted a regression model for systolic blood pressure (SBP) with no. of alcoholic drinks, BMI, and age as covariates.
- No. of alcoholic drinks was missing for 34.1% of individuals.
- Missingness in this variable may well be related to level of alcohol consumption (i.e. MNAR), age, (and maybe) BMI, but given these is probably unrelated to SBP.
- If this assumption is true, the CCA is valid, even though the covariate is (assumed to be) MNAR.

Logistic regression CCA

- If the outcome model is logistic regression, CCA can give valid estimates (of covariate effects) under even weaker missingness assumptions (Bartlett, Harel, and Carpenter 2015).
- This is due to the symmetry property of odds ratios (the same reason we can use odds ratios in case-control studies).
- For covariate effects (but not the intercept), we get consistent estimates if missingness is:
  - dependent on Y, or
  - dependent on $\mathbf{X} = (X_1, .., X_p)$
- Furthermore, missingness could be dependent on Y and $X_2, .., X_p$, and estimates of coefficient of X_1 are still consistent.

CCA - recommendations

- It is generally always a good idea to perform CCA for your analysis.
- The estimates you get can be compared with those from other analyses which make other assumptions.
- Important to remember that CCA might be valid in your situation, depending on the analysis you are performing and missingness assumptions.

Why are we wasting our time on MAR and MNAR?

- Validity of CCA doesn't fit neatly into the MCAR/MAR/MNAR framework.
- Why then did we spend time defining and thinking about MAR and MNAR?
- Answer: because an important collection of methods can give valid inferences under MAR mechanisms.
- One such method is multiple imputation. . .

Multiple imputation

Multiple imputation

- Multiple imputation (MI) is a flexible and increasingly popular approach to handling missing data.
- It relies (at least in its usual form) on assuming data are MAR.
- We will introduce it in a simple setting with two variables.
- Later we will look at extensions to more realistic situations.

Intuition for MI

- Suppose our data set has variables X and Y, with some Y values MAR given X.
- Our aim is to impute missing values in Y, taking X into account.
- In parametric imputation, we specify a regression model for $f(Y|X)$.
- We want to impute the missing Y values from this model.
- Y **need not necessarily be the outcome in our final analysis**.

Intuition for MI

- MAR here means missingness in Y is independent of Y, given X.
- This means that if we fit the model for $f(Y|X)$ using complete cases, estimates are valid.
- Using the fitted model, we can then impute Y for the incomplete cases.
- With the imputed data set, we can calculate our statistic of interest (e.g. sample mean, variance, regression of X on Y).

Why multiple imputation?

In multiple imputation we create a number M imputed datasets, estimate our parameter(s) of interest from each imputed dataset, and then calculate the average across imputations

There are two main reasons why we create *multiple* imputed datasets:

1. We reduce Monte-Carlo error which is introduced through using a simulation based method
2. Estimating variances and finding confidence intervals is relatively easy if we create multiple imputations, but is rather difficult with only a single imputation

Multiple imputation for one continuous variable

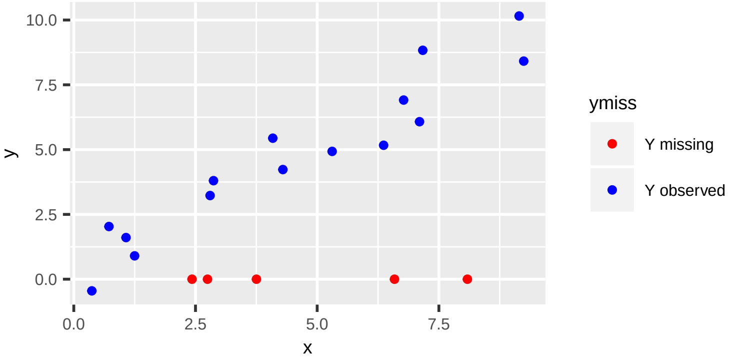
- Next describe the details/steps for linear regression imputation of one variable.
- Later on, we will see that application of MI in practice requires careful considerations of a number of aspects (e.g. missingness assumptions, model specification).
- For now, we will put these to one side.

Multiple imputation for one continuous variable

- X is fully observed.
- Y contains missing values, and we assume Y is MAR given X.
- We want to create multiple imputations of the missing values in Y, using X.
- We will create M imputations - we will come back to the choice of M later.

The observed data

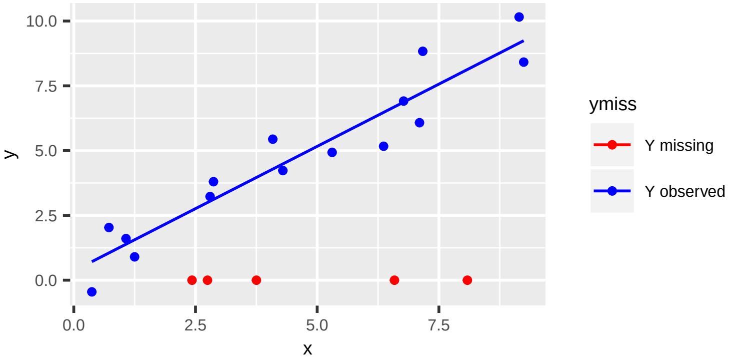
The plot shows the complete cases (where Y and X observed) and five subjects with X observed but Y missing.



Step 1 - fit the imputation model

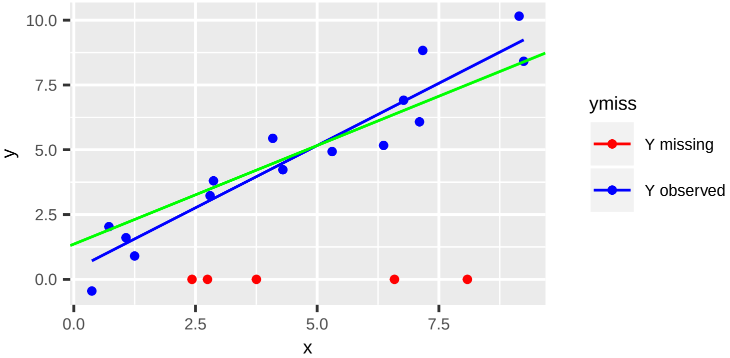
We first fit the imputation model.

This is a model for the partially observed variable (Y) on the fully observed variable (X), using a CCA.



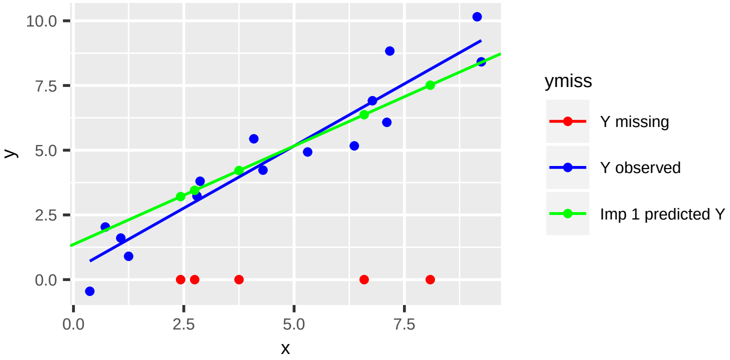
Step 2 - draw new imp. model parameter values

Next we perturb the fitted line to account for uncertainty in its estimation (we take draws from the Bayesian posterior of the regression model parameters).



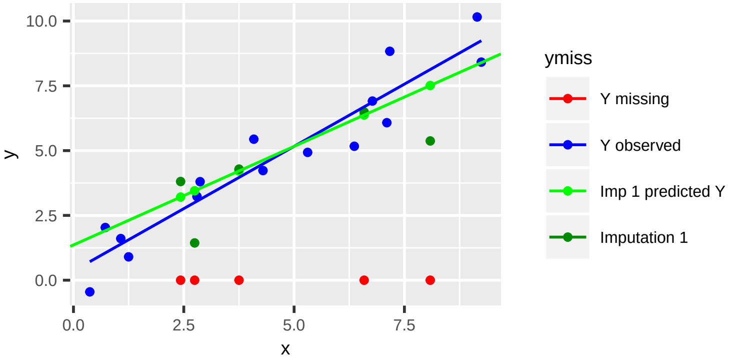
Step 3 - calculate predicted values

We then calculate predicted value of Y for those with Y missing.



Step 4 - create imputed values

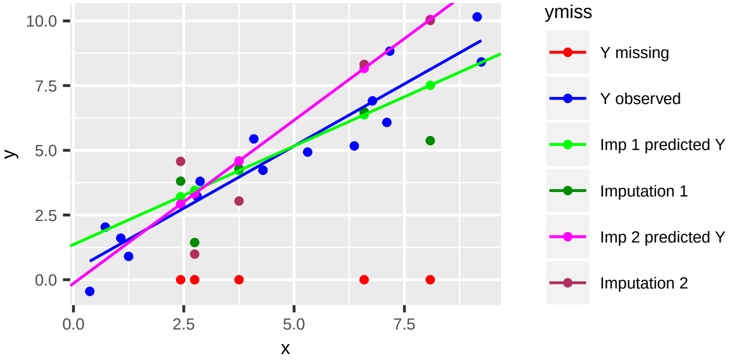
Imputed values are random draws centred at predicted Y values, with error variance as drawn in earlier Bayesian posterior draw step.



Step 5 - repeat steps to create more imputations

We then repeat these steps to create as many imputations as desired:

- draw new parameter values from Bayesian posterior
- draw new predicted values
- draw new imputations around predicted values



Algorithm

- Estimate $\sigma^2, \beta_0, \beta_1$ using the n_0 complete case analysis, giving $\hat{\sigma}^2, \hat{\beta}_0, \hat{\beta}_1$.
- For $m = 1, .., M$
- Draw from posterior distribution of parameters:
  1. Draw a $\sigma^{2(m)}$ from $\hat{\sigma}^2(n_0 - 2)/\chi^2_{n_0-2}$.
  2. Draw (β_0^m, β_1^m) from

$$N\left\{\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix}, \sigma^{2(m)}(W^T W)^{-1}\right\}$$

- If Y is missing for subject i, impute Y_i by

$$Y_i^m = \beta_0^m + \beta_1^m X_i + \epsilon_i^m$$

where $\epsilon_i^m \sim N(0, \sigma^{2(m)})$

Things to note

- Imputations are constructed by adding normal errors to the predicted value of Y based on the value of X.
- The variance of these errors depends on the estimated error variance in the model fitted to the complete cases.
- For a given value of X, the predicted values are different for each imputation, because a different line is used for each imputation.
- The new imputation model parameter values are draws from their posterior distribution, under standard non-informative priors.

Imputation using other types of model

- Imputation can also be performed using other types of regression model.
- These can be chosen so that they are suitable for the variable being imputed.
- e.g. logistic regression for binary variables.
- The principles outlined remain the same.
- The only changes are that we take a draw from a different distribution depending on the type of regression model.

Analysis of imputed datasets

- As described above, we have imputed M complete data sets.
- We analyse each of them in the usual way (i.e. using the model intended for the complete data) giving us M estimates of the original quantity of interest, say θ. Denote these estimates $\hat{\theta}_1, \ldots, \hat{\theta}_M$.
- The analysis of each imputed data set will also give an estimate of the variance of the estimate $\hat{\theta}_m$, say $\hat{\sigma}^2_m$. Again, this is the usual variance estimate from the model.
- We combine these quantities to get our overall estimate and its variance using certain rules, developed by Rubin.

Combining the estimates - Rubin's rules

Let the multiple imputation estimate of θ be $\hat{\theta}_{MI}$. Then

$$\hat{\theta}_{MI} = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m.$$

Further define the within imputation and between imputation components of variance by

$$\hat{\sigma}_w^2 = \frac{1}{M} \sum_{m=1}^{M} \hat{\sigma}_m^2, \quad \text{and} \quad \hat{\sigma}_b^2 = \frac{1}{M-1} \sum_{m=1}^{M} (\hat{\theta}_m - \hat{\theta}_{MI})^2,$$

Then

$$\hat{\sigma}_{MI}^2 = \left(1 + \frac{1}{M}\right) \hat{\sigma}_b^2 + \hat{\sigma}_w^2,$$

so the estimated standard error of $\hat{\theta}_{MI}$ is $\hat{\sigma}_{MI}$.

Inference for θ

To test the null hypothesis $\theta = \theta_0$, compare

$$\frac{\hat{\theta}_{MI} - \theta_0}{\hat{\sigma}_{MI}} \quad \text{to} \quad t_\nu,$$

where

$$\nu = (M-1)\left[1 + \frac{\hat{\sigma}_w^2}{(1+1/M)\hat{\sigma}_b^2}\right]^2 .$$

Thus, if $t_{\nu,0.975}$ is the 97.5% point of the t distribution with ν degrees of freedom, the 95% confidence interval is

$$(\hat{\theta}_{MI} - \hat{\sigma}_{MI} t_{\nu,0.975}, \quad \hat{\theta}_{MI} + \hat{\sigma}_{MI} t_{\nu,0.975})$$

Software

- As we shall see, the software automates the previous steps.
- Although these steps are fairly automated, our input is critical.
- There are various modelling choices to be made, and poor choices can lead to invalid inferences.

The attractions of MI

- MI is attractive, because once we have imputed the missing data, we can analyse the completed data sets as we would have done if no data were missing.
- It is particularly useful in messy complex datasets, with missing values in multiple variables, where alternative approaches are less readily applied.
- Compared to CCA, MI can often give estimates with improved precision.

When is MI is the same as complete case analysis?

- If missingness is only in the outcome, and the analysis model is the same as the imputation model (i.e. no auxiliary variables), MI gives you (essentially) the same estimates as complete case analysis.
- So in this special case, there is no point in doing MI.

Likelihood based analyses

- Also note that some methods (e.g. linear mixed models) analyse all observed data using maximum likelihood.
- They are valid under MAR, and are efficient.
- e.g. in longitudinal trials with missingness in outcomes, there may be no need to do MI.
- But MI can incorporate auxiliary variables, which is often very useful.

Some papers on MI

(Schafer 1999)

(Buuren 2007)

(Kenward and Carpenter 2007)

(Sterne et al. 2009)

There are of course many many more…

Some books on missing data and MI

Statistical Analysis with Missing Data (Little and Rubin 2019) – excellent book on analysis with missing data. 3rd edition recently released.

Flexible Imputation of Missing Data (Van Buuren 2018) – A particular focus on mice package in R. 2nd edition recently release. Free online version here

Multiple Imputation and its Application (Carpenter and Kenward 2013) – includes coverage of imputation with survival data, multi-level data, non-linearities and interactions, sensitivity analyses.

Summary

- Ad-hoc methods attempt to deal with the computational difficulty introduced by missing data.
- But they generally do not give valid inferences under plausible assumptions.
- MI gives valid inferences if data are MAR and the imp. model is correctly specified.
- So far though we have only considered the case of a single partially observed continuous variable.
- In the next session we will explore its extension to more realistic settings.

References I

Bartlett, J W, J R Carpenter, K Tilling, and S Vansteelandt. 2014. “Improving upon the efficiency of complete case analysis when covariates are MNAR.” *Biostatistics* 15: 719–30.

Bartlett, J W, O Harel, and J R Carpenter. 2015. “Asymptotically Unbiased Estimation of Exposure Odds Ratios in Complete Records Logistic Regression.” *American Journal of Epidemiology* 182 (8): 730–36.

Buuren, S van. 2007. “Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specification.” *Statistical Methods in Medical Research* 16: 219–42.

Carpenter, J, M Kenward, S Evans, and I White. 2004. “Letter to the editor: Last observation carry forward and last observation analysis by J. Shao and B. Zhong, Statistics in Medicine, 2003, **22**, 2429–2441.” *Statistics in Medicine* 23: 3241–4.

References II

Carpenter, J R, and M G Kenward. 2013. *Multiple Imputation and its Application*. John Wiley & Sons, Ltd, Chichester, U.K.

Cook, R J, L Zeng, and G Y Yi. 2004. "Marginal analysis of incomplete longitudinal binary data; a cautionary note on LOCF imputation." *Biometrics*, 820–28.

Kenward, M G, and J R Carpenter. 2007. "Multiple Imputation: Current Perspectives." *Statistical Methods in Medical Research* 16: 199–218.

Little, Roderick JA, and Donald B Rubin. 2019. *Statistical Analysis with Missing Data*. Vol. 793. John Wiley & Sons.

Molenberghs, G, H Thijs, I Jansen, C Beunkens, M G Kenward, C Mallinkrodt, and R J Carroll. 2004. "Analyzing Incomplete Longitudinal Clinical Trial Data." *Biostatistics* 5: 445–64.

Schafer, J L. 1999. "Multiple imputation: a primer." *Statistical Methods in Medical Research* 8: 3–15.

References III

Sterne, J A C, I R White, J B Carlin, M Spratt, P Royston, M G Kenward, A M Wood, and J R Carpenter. 2009. “Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls.” *British Medical Journal* 339: 157–60.

Van Buuren, Stef. 2018. *Flexible Imputation of Missing Data*. Chapman; Hall/CRC.

White, I R, and S G Thompson. 2005. “Adjusting for Partially Missing Baseline Measurements in Randomized Trials.” *Statistics in Medicine* 24: 993–1007.