Lecture 3 - MI for multiple variables and MI practicalities

Multiple imputation techniques for working with missing data

Jonathan Bartlett (thestatsgeek.com)

Copenhagen, March 2020

More on MAR

Imputation from a joint model

Imputation by chained equations

Variable selection, number of imputations, and model checking

A cautionary example

Overview

- MI provides valid estimates under the MAR assumption **and** provided the imputation model is reasonably correctly specified.
- If only one variable has missing values, this may be relatively simple.
- Commonly though, there may be missing values in many variables, and the variables may be a mixture of continuous and discrete, making the imputation process more difficult.
- In this session we'll look at the two main approaches to imputation in this setting, and some of the important practicalities which arise.

More on MAR

More on MAR

- With one variable partially observed variable Y, the definition of MAR is (hopefully) clear - the probability that Y is missing is independent of its value, conditional on other fully observed variables $\mathbf{X} = (X_1, X_2, .., X_q)$ which are being used in the MI analysis.
- Now let's consider the case where the partially observed variable $\mathbf{Y} = (Y_1, Y_2, .., Y_p)$ consists of multiple components.

MAR with monotone missingness

- A special type of missingness pattern is so called monotone missingness.
- This means that we can order the components of $\mathbf{Y}$ such that if Y_j is observed for an observation, all previous components are also observed.
- Monotone patterns most commonly occur in longitudinal studies subject to dropout:

id	1	2	3	4
1	x	x	x	x
2	x	.	.	.
3	x	x	.	.
4	x	x	x	.

MAR with monotone missingness

- In this case the ordering of variables is on the basis of time (it doesn't have to be ordered by time).
- When missingness is monotone, MAR can be shown (see for example (Tsiatis 2006) to mean that the probability of dropout at time t doesn't depend on the current or future values, conditional on the past (values before t).

MAR with non-monotone missingness

- Often in datasets (see the practical) the missingness pattern is not monotone.
- Even in longitudinal studies, in addition to dropout, one might have intermittent missingness.
- With non-monotone patterns the meaning of MAR becomes more complex - the probability of each pattern occurring only depends on the data observed under that pattern (Robins and Gill 1997).
- We will not dwell further on this here, but see the practical for more discussion of this issue.

Imputation from a joint model

Imputation with one partially observed variable

- In the previous session, we considered a situation where one variable Y has missing values, but other variables $\mathbf{X}$ are fully observed.

Y **need not be the outcome in our final model of interest!**

- We used a linear regression model for $Y|\mathbf{X}$ to impute the missing values of Y. This assumes $Y|\mathbf{X}$ is normal, with mean a linear function of $\mathbf{X}$.
- If Y were binary, we can similarly use a logistic regression model to impute the missing values of Y, given $\mathbf{X}$ (e.g. `sen` in the class size study)
- For categorical Y, we can use an ordered or multinomial logistic model for $Y|\mathbf{X}$.

Imputation with multiple partially observed variables

- More typically in epidemiology, we may have more than one variable with missing values.
- Extending our previous notation, we now let $\mathbf{Y}$ denote the vector of variables which have missing values, and let $\mathbf{X}$ denote the fully observed variables.
- To perform multiple imputation, we must specify a model for the joint or multivariate distribution, $f(\mathbf{Y}|\mathbf{X})$.
- This is tricky in general, particularly if $\mathbf{Y}$ contains a mixture of continuous and categorical variables.

Imputation with multiple partially observed variables

- One of the most important multivariate distributions is the multivariate normal (MVN).
- MI using the MVN was developed early on.

$$\mathbf{Y} \sim N(\boldsymbol{\mu}(\mathbf{X}), \boldsymbol{\Sigma})$$

- For the MVN, each component of Y is normal, and $Y_j | Y_{-j}$ is a linear regression model, where Y_{-j} denotes all of the components of $\mathbf{Y}$ except the jth.
- See (Schafer 1997) for more details.

Imputation with the multivariate normal model

- What if we have missing values in non-normal, or even categorical variables?
- Early advice was that skewed variables could be transformed to (approximate) normality before imputation and then back transformed afterwards for analysis.
- However, it is important to be aware that if you do this, the functional relationships assumed at the imputation stage are changed.
- This has led to one paper recommending against the approach, and instead suggesting that if the MVN model is being used to just impute ignoring the skewness (Hippel 2013).
- Of course this approach isn't perfect either – imputing from a mis-specified model will in general result in biased inferences, but perhaps less biased than if we had used the transformation approach.

Imputation with the multivariate normal model

- For binary variables, some work has been done investigating imputation assuming normality, comparing various rounding strategies. See (Bernaards, Belin, and Schafer 2007) and (Lee and Carlin 2010).
- If the rounding is done carefully, the results can be (somewhat surprisingly) quite good.
- However, if $\mathbf{Y}$ contains a mixture of continuous and categorical variables, using the MVN model is tricky.

Joint models for mixtures of continuous and categorical data

- Log-linear models were proposed for imputing categorical data.
- With a mixture of continuous and categorical data, the general location model was proposed.
- This is available in the R package mix.
- These approaches have not however been widely adopted by researchers. This may be because for log-linear models one must usually carefully choose which interaction parameters to set to zero, and this is quite tricky.
- For more details on the model theory, again see (Schafer 1997).

Joint models for mixtures of continuous and categorical data

- Recently there has been further development of more flexible joint models for imputation.
- jomo uses a joint model with a latent multivariate normal structure.
- jointAI uses a joint model factorised as a product of univariate conditional models.
- I expect both to be increasingly used in the coming years in applied research.

Imputation by chained equations

Imputation by chained equations / full conditional specification

- Imputation by chained equations (MICE) or full conditional specification (FCS) is an alternative to joint model imputation.
- It was proposed independently by (van Buuren, Boshuizen, and Knook 1999) and (Raghunathan et al. 2001).
- Rather than directly specify a joint/multivariate model, we specify a series of conditional models.

Imputation by chained equations / full conditional specification

- e.g. suppose Y_1, Y_2 and Y_3 have missing values, and we have fully observed variables $\mathbf{X}$.
- Rather than specify a joint imputation model for $f(Y_1, Y_2, Y_3|\mathbf{X})$ directly, we specify models for:

$$f(Y_1|Y_2, Y_3, \mathbf{X})$$
$$f(Y_2|Y_1, Y_3, \mathbf{X})$$
$$f(Y_3|Y_1, Y_2, \mathbf{X})$$

MICE/FCS algorithm

For imputation $m = 1, .., M$:

- Initially impute missing values in Y_1, Y_2 and Y_3 by randomly sampling from the observed values.
- For iteration $t = 1, .., T$:
  - Impute missing values in Y_1 **once** using model for $f(Y_1|Y_2, Y_3, \mathbf{X})$ (using obs. Y_1 values and observed and imputed values of Y_2 and Y_3).
  - Impute missing values in Y_2 **once** using model for $f(Y_2|Y_1, Y_3, \mathbf{X})$ (using obs. Y_2 values and observed and imputed values of Y_1 and Y_3).
  - Impute missing values in Y_3 **once** using model for $f(Y_3|Y_1, Y_2, \mathbf{X})$ (using obs. Y_3 values and observed and imputed values of Y_1 and Y_2).
- Current imputed values of missing values used to form mth imputed dataset.

Strengths of MICE/FCS imputation

- The major advantage of MICE/FCS imputation (over 'joint model' imputation) is the ability to specify different model types for each variable.
- It has become an extremely popular approach for performing MI (Buuren 2007).
- e.g. logistic for binary variables, Poisson for count variables (in Stata), multinomial logistic for unordered categorical variables.
- It can do things more easily than joint model imputation, such as imputing certain variables only in subgroups.

Theoretical deficiency of MICE/FCS

- A theoretical issue with MICE/FCS is that there is no guarantee that the algorithm draws imputations from a well defined joint/multivariate model.
- Recent work by two groups have identified certain conditions when it does (Hughes et al. 2014; Liu et al. 2013).
- The key condition is that the conditional models are **compatible**.
- This means that there exist multivariate distributions whose conditionals are those specified in MICE/FCS.
- Checking compatibility is not easy. In practice, we should be aware of the issue, and to situations where incompatibly could seriously mislead (see next session).

Joint modelling versus MICE/FCS

- A number of papers have compared the joint modelling approach with chained equations with real examples.
- (Buuren 2007) applied both methods to some growth data, and concluded that chained equations was preferable to joint modelling.
- (Lee and Carlin 2010) found that both methods worked well in a realistic epidemiological setting.
- In settings with continuous and categorical variables with missing values, at least in terms of availability of flexible software, MICE/FCS seems preferable (in my opinion).

MICE/FCS with monotone missingness

- If the pattern is monotone, there is no need to 'cycle' or iterate in MICE/FCS.
- This is because one can first impute $Y2|Y1$, then $Y3|Y2, Y1$, ...
- Stata's MICE/FCS command `mi impute chained` checks for this, and if it finds a monotone pattern, imputes sequentially.
- In R, `mice` has an option (visitSequence) to impute in order of increasing missingness (i.e. the same thing).
- The advantage of this is that we do not need to worry about convergence of MICE/FCS.
- We also don't have to worry about incompatibility between the conditional models, and the theoretical weakness of MICE/FCS.

Variable selection, number of imputations, and model checking

Which variables should be included in the imputation model?

- Usually, all variables which will be used in our model of interest / analysis model should be included in the imputation model.
- In terms of creating the imputations, there is no conceptual distinction between variables which are covariates or the outcome in your final model of interest.
- If we are imputing missing covariates, the outcome variable **must** be included, to ensure that the imputed covariate values have the correct association with the outcome.

Auxiliary variables

- Often we may have variables Z which are not involved in our model of interest.
- Recall that MI is only valid under MAR, which we cannot verify based on the observed data.
- If a variable Z is predictive of missingness in another variable we are imputing, Z should be included in the imputation model, to increase the likelihood that the MAR assumption is satisfied.
- Even if Z is not predictive of missingness, if it is predictive of the partially observed variables $\mathbf{Y}$, we should include it in the imputation model. Doing so will reduce the uncertainty in imputing missing values, thus increasing statistical efficiency.

Auxiliary variables

- The option to include auxiliary variables in the imputation model at the imputation stage but omit it from the analysis stage is a big advantage of MI
- e.g. we may have a variable on the causal pathway which we do not want to condition on in the analysis
- But we can include it in the imputation stage if we think it will improve MAR or is correlated with variables we are imputing
- Some datasets now have many hundreds of variables. In these cases one may have to be more judicious in selecting auxiliary variables

How many imputations?

- Earlier papers/books suggested M could be a small as 3-5
- Validity of inferences is not affected by choice of M
- Efficiency is improved (somewhat) by increasing M
- Choose M so that Monte-Carlo error is sufficiently small %(see Practical for how to assess this in Stata)

Model checking

- For valid inferences we need the imputation models to be correctly specified.
- Checking this is not easy.
- One approach, when using MICE/FCS, is to check the fit of each conditional model based on its complete case fit.
- Being based on the complete cases, these fits may themselves be biased.
- However, we will probably be able to detect and rectify any grossly mis-specified models.
- Examining plots of imputed values is also sensible, and can be used to diagnose serious issues.

A cautionary example

The QRISK study

- The QRISK study aimed to derive a new cardiovascular disease (CVD) risk score for the UK, based on routinely collected data from general practice (Hippisley-Cox et al. 2007a)
- The score was derived using data from 1.28 million patients registered at UK GP practices between 1995 and 2007, who were free from CVD at registration
- The outcome of interest was time to first recorded diagnosis of CVD
- Cox proportional hazards models were used to model time to CVD, as a function of risk factors measured at registration

Missing data in QRISK

- Inevitably there was substantial missingness in 'baseline' risk factor data
- In particular, 70% of subjects had HDL cholesterol missing
- The investigators used MI to deal with missing baseline data, using the `ice` (the forerunner to `mi impute chained`) command in Stata

Cholesterol and CVD

- In the final model, the adjusted hazard ratio for the ratio of total to HDL cholesterol was 1.001 (95% 0.999 to 1.002)
- This suggested that, after adjusting for other baseline risk factors, cholesterol had no effect on CVD risk
- Given that cholesterol has been shown to have an independent effect on CVD risk in many previous studies, this result was unexpected

Cholesterol and CVD

- A complete case analysis did show evidence for an effect of cholesterol
- It turned out that when imputing the missing values, although the time to CVD or censoring was included in the imputation model, the censoring indicator (1=CVD, 0=censored) had inadvertently not been used
- The imputed cholesterol values thus did not have the correct association with time to CVD, resulting in there being no evidence of an independent effect
- Re-running with a more appropriate imputation model, an independent effect of cholesterol was found (Hippisley-Cox et al. 2007b)

Summary

- The meaning of MAR is clear with monotone patterns, but is more complex with non-monotone patterns.
- Joint modelling and MICE/FCS are the two broad imputation approaches with multivariate data.
- We have discussed practical issues, including number of imputations, variable choice, and model checking
- Important to remember: obtaining reasonable results depends on
  - the MAR assumption holding (at least approximately)
  - the imputation models used being correctly (at least approximately) specified

References I

Bernaards, C A, T R Belin, and J L Schafer. 2007. “Robustness of a Multivariate Normal Approximation for Imputation of Incomplete Binary Data.” *Statistics in Medicine* 26: 1368–82.

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

Hippel, P T von. 2013. “Should a Normal Imputation Model Be Modified to Impute Skewed Variables?” *Sociological Methods & Research* 42 (1): 105–38.

Hippisley-Cox, J, C Coupland, Y Vinogradova, J Robson, M May, and P Brindle. 2007a. “Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study.” *British Medical Journal* 335: 136.

———. 2007b. “QRISK authors response [electronic response].” *British Medical Journal* 335: 136.

References II

Hughes, R A, I R White, S R Seaman, J R Carpenter, K Tilling, and J A C Sterne. 2014. "Joint Modelling Rationale for Chained Equations." *BMC Medical Research Methodology* 14 (1). BioMed Central Ltd: 28.

Lee, K J, and J B Carlin. 2010. "Multiple Imputation for Missing Data: Fully Conditional Specification Versus Multivariate Normal Imputation." *American Journal of Epidemiology* 171: 624–32.

Liu, J., A. Gelman, J. Hill, Y. S. Su, and J. Kropko. 2013. "On the Stationary Distribution of Iterative Imputations." *Biometrika* 101: 155–73.

Raghunathan, Trivellore E, James M Lepkowski, John Van Hoewyk, and Peter Solenberger. 2001. "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models." *Survey Methodology* 27: 85–95.

References III

Robins, J M, and R D Gill. 1997. "Non-Response Models for the Analysis of Non-Monotone Ignorable Missing Data." *Statistics in Medicine* 16: 39–56.

Schafer, J L. 1997. *Analysis of incomplete multivariate data*. London: Chapman; Hall.

Tsiatis, A A. 2006. *Semiparametric Theory and Missing Data*. Springer, New York.

van Buuren, S., H C Boshuizen, and D L Knook. 1999. "Multiple Imputation of Missing Blood Presure Covariates in Survival Analysis." *Statistics in Medicine* 18: 681–94.