Lecture 4 - MI with derived variables, survival outcomes, dependent data and survey data

Multiple imputation for missing data

Jonathan Bartlett (thestatsgeek.com)

Oslo, November 2019

Derived variables

Survival outcomes

Imputation with dependent data

Survey data

Summary

Derived variables

Derived variables

- If our substantive model of interest includes derived variables, like non-linear effects and/or interactions, our imputation model should respect these.
- In the next practical, we will look at a substantive model for `sbp` in NHANES which includes the interaction between `waist_circum` and `ALQ150`.
- The imputation model should be 'compatible' or congenial with the substantive model (Meng 1994).
- e.g. suppose our model of interest is a linear regression of Y on X and X^2, if we impute missing values of X using a linear regression of X on Y, the imputed data will not have the correct quadratic relationship between Y and X.

Interactions

- Suppose the outcome/substantive model contains an interaction between two predictors, X_1 and X_2, one of which (X_1) is categorical (e.g. `ALQ150`).
- If X_1 is fully observed, a convenient approach to allow for the interaction is to impute separately in different levels of X_1.
- Stata's `mi` commands make this very easy: simply add `by(x1)` at the end of the command.
- In R, we could split the data into multiple data frames, run `mice` on each, and then recombining the imputed datasets.
- If X_1 itself has missing values (as in `ALQ150` variable), we cannot use this approach.
- It also does not work if both X_1 and X_2 are continuous, or we want to allow for multiple interactions.

Impute then transform

- The simplest approach to handling derived variables is to perform imputation as normal, then create the derived variables (e.g. interactions) in the imputed datasets.
- This is not a good idea.
- The imputation models will not be compatible with what the substantive model.
- Biased estimates will be obtained.

Passive imputation

- Passive imputation involves adding the derived variable(s) to the data frame and updating its value during the imputation process.
- e.g. we can add a variable `waist_circum*AQL150` and tell `mice` how to update its value.
- This interaction term can be used as a covariate in the imputation models.

Limitations of passive imputation

- Passive imputation has limitations - it is not always obvious how to specify imputation models which are compatible with the substantive model.
- e.g. when imputing `ALQ150`, we need to ensure its imputation model is compatible with the presence of an interaction between it and `waist_circum` in the model for `sbp`.
- Used naively it will usually lead to biased estimates.

'Just another variable' approach

- The 'transform then impute' or 'just another variable' (JAV) approach recently proposed by von Hippel (Hippel 2009) involves treating derived variables as if they were just any other variables and includes them in the imputation process.
- e.g. we include `waist_circum*AQL150` in the imputation process and impute it as if it were a regular continuous variable, and ignore the deterministic relationship between it and `waist_circum` and `ALQ150`.
- An unappealing feature of this is that we have imputed values of `waist_circum*AQL150` which are not equal to the product of the values of `waist_circum` and `ALQ150`.

Statistical properties of the 'just another variable' approach

- For linear models where data are MCAR, the JAV approach gives consistent point estimates, but Rubin's rules may not be valid.
- With data MAR, JAV gives biased estimates, since it consists of fitting a mis-specified parametric model by maximum likelihood.
- For logistic regression models JAV can be badly biased.
- For more on this, see (S. R. Seaman, Bartlett, and White 2012).

Substantive model compatible FCS (SMC-FCS)

- We developed a modified version of MICE/FCS, which imputes each covariate compatibly with a user-specified substantive model (SM) (Bartlett et al. 2015).
- Suppose we have an outcome of interest Y, partially observed covariates $X_1, X_2, .., X_p$, and fully observed covariates $\mathbf{Z}$.
- We specify a substantive model (SM) for $f(Y|X_1, .., X_p, \mathbf{Z}, \psi)$, with parameters ψ.
- e.g. linear regression of Y, with covariate vector some function of $X_1, .., X_p$ and $\mathbf{Z}$.
- e.g. covariates include $X_1 \times X_2$, or X_1^2, or $X_1/X_2^2 \ldots$
- The covariates $X_1, .., X_p$ have missing values.

Substantive model compatible FCS

- We must impute from a model for $f(X_j|X_{-j}, \mathbf{Z}, Y)$.
- This can be expressed as

$$\frac{f(Y|X_j, X_{-j}, \mathbf{Z})f(X_j|X_{-j}, \mathbf{Z})}{\int f(Y|X_j^*, X_{-j}, \mathbf{Z})f(X_j^*|X_{-j}, \mathbf{Z})dX_j^*}.$$

- The SM is a model for $f(Y|X_j, X_{-j}, \mathbf{Z})$.
- We can thus specify an IM for X_j which is compatible with the SM by additionally specifying a model for $f(X_j|X_{-j}, \mathbf{Z})$.

Drawing imputations

- Having specified a model for $f(X_j|X_{-j}, \mathbf{Z})$, the implied imputation model $f(X_j|X_{-j}, \mathbf{Z}, Y)$ will in general not belong to a standard distributional family.
- We appeal to the Monte-Carlo method of rejection sampling to generate draws.
- Rejection sampling involves drawing from an easy-to-sample (candidate) distribution until a particular criterion/bound is satisfied.
- Deriving this bound is relatively easy if we use our model for $f(X_j|X_{-j}, \mathbf{Z})$ as the candidate distribution.

smcfcs

- smcfcs implements the SMC-FCS approach in R. smcfcs in Stata.
- Linear, logistic and Cox proportional hazards outcome models are supported.
- It also supports competing risks outcomes, and nested case-control and case-cohort studies.
- Normal linear, logistic, Poisson, proportional odds and multinomial logistic imputation methods are provided.
- The SM can contain essentially any function of the variables, e.g. squares, cubes, interactions, logarithms of variables, etc etc.
- The approach can also be used when imputing components of a ratio variable, e.g. BMI.
- In the practical we will see how smcfcs can be used to accommodate an interaction, and be used to impute missing covariates in a Cox model analysis.

Survival outcomes

Incorporating the outcome in imputation

- As we noted earlier, the outcome variable in the final model of interest *must* be included in the imputation model.
- If we do not, imputed values will not have the correct associations with the outcome.
- How to incorporate the outcome in an imputation model depends on the type of variable being imputed and the type of outcome / outcome model.

Survival outcomes

- A common outcome type is time to some event of interest (often called survival outcomes).
- Sometimes we do not observe the event occurring for every subject in the available follow-up, leading to censoring.
- The outcome then consists of a variable T representing time to the event of interest and an event indicator D ($D = 1$ if event occured, $D = 0$ otherwise).
- If $D = 0$, T records the censoring time – the last time at which a subject was seen, and had still not had the event.
- If we have some missing values in the covariates X in our survival model, how should we impute them?

Incorporating survival outcomes in imputation models

- Early recommendations were to impute X by putting T (or $\log(T)$ and D as covariates).
- More recently, White and Royston (2009) investigated theoretically how the imputation model for X should be specified when a Cox proportinal hazards model is used:

$$h(t|\mathbf{X}) = h_0(t)\exp(\beta^T\mathbf{X}) \tag{1}$$

where $h_0(t)$ denotes an arbitrary baseline hazard function and β a vector of (log) hazard ratios.

Incorporating survival outcomes in imputation models

- White and Royston showed that when imputing a normally distributed variable X one should use a linear regression imputation model, with D and $H_0(T) = \int_0^T h_0(u)du$ (baseline cumulative hazard function) as covariates.
- For binary X, one should use a logistic regression imputation model, again with D and $H_0(T)$ as covariates.
- Their results are exact for binary X, but are approximate for normal X.
- The approximation for normal X should work well provided the covariate X does not have a large effect on hazard or if the incidence of the event of interest is low.

Incorporating survival outcomes in imputation models

- $H_0(t) = \int_0^t h_0(u)du$ is the baseline cumulative hazard function.
- White and Royston suggest a number of approaches to estimating $H_0(t)$:
  - Substantive knowledge - e.g. it may be reasonable to assume constant baseline hazard so that $H_0(t) \propto t$. In this case, we just include D and T as covariates in our imputation model(s).
    - When covariate effects are small, one could approximate $H_0(t)$ by the Nelson-Aalen (marginal) cumulative hazard estimator $H(t)$, which ignores covariates and thus can be estimated using all subjects.
    - Estimating $H_0(t)$ within the FCS algorithm by fitting the Cox proportional hazards model to the current imputed dataset.

Substantive model compatible FCS

- Our SMC-FCS approach can also 'solve' the problem.
- Each partially observed covariate is imputed compatibly with the specified Cox model.
- Our approach is particularly attractive if there are additionally interactions or non-linear covariate effects in the Cox model.
- If censoring mechanism is related to partially observed covariate(s), then censoring should be treated as a competing risk at the imputation stage.

Imputation with dependent data

Example - longitudinal data

- Suppose some quantity y was intended to be measured repeatedly on subjects over time.
- There are some missing values of y.
- How should we impute these missing values?

Imputing in 'long form'

id	gender	time	y
1	m	0	4.5
1	m	1	3.9
1	m	2	4.1
1	m	3	.
1	m	4	4.2

- If we impute the dataset is 'long' form, we treat each observation as independent.
- This is clearly inappropriate - observations from the same subject are usually correlated.
- The observed values of y on a subject contain information about missing y at $t = 3$.
- If we ignore the longitudinal structure, imputations will not only be inefficient, they will not have the correct correlation structure.

Imputing in 'wide form'

id	gender	y0	y1	y2	y3	y4
1	m	4.5	3.9	4.1	.	4.2

- If measurement times are common to all subjects, we may be able to impute with the data in 'wide' form.
- e.g. we could apply `mice` to gender, y0, y1, y2, y3, y4.
- This uses available longitudinal information to impute missing value at $t = 3$ for id $= 1$.
- Note that this strategy generally cannot be applied if observations take place at different times for different subjects.
- You may run into co-linearity issues when y is highly correlated within subjects over time.

Example - clustered data

- Another form of dependent data is clustered of multi-level data.
- The clustering should be accounted for in the imputation process.

Including fixed cluster effects

- One approach is to include cluster id as a fixed effect covariate in imputation models.
- Standard MI software can be used.
- If each cluster has a large number of (observed) units, this could work well.
- But if the substantive model is a random effects model, it has been shown to lead to invalid inferences (Andridge 2011).

Imputation by cluster

- Yet another approach is to impute separately in each cluster, thus allowing parameters to vary by cluster.
- This should work well when there are a small number of large clusters.
- An advantage of this approach is that it allows all the imputation model parameters to differ between clusters.
- Conversely, a disadvantage is that information is not borrowed between clusters.
- If there are many clusters, and/or clusters are small, imputation by cluster may perform poorly.

Random-effects imputation

- If your substantive analyses would treat the dependency in the data through random effects, you should probably impute mising data using random effects models.
- The principles of MI remain the same - all that changes is that we have random effects in our imputation model(s).

Random-effects imputation software

- `mice` can impute variables using FCS with certain random effects models.
- `jomo` can impute using joint random effects models based on latent multivariate normal structure.
- `jointAI` can impute using joint random effects models based on factorising the joint distribution as a product of conditionals.
- For further details, see Buuren (2011) and Audigier et al. (2018).

Likelihood based approaches

- If missingness is confined to the outcome variable, likelihood based approaches are statistically efficient and valid under MAR.
- In such cases, there is little point in trying to attempt imputation.
- This is the case for both dependent data situations and simpler independent data situations.

Survey data

MI with survey data

- Rubin's original aim was for MI to be used in the context of large survey datasets.
- However, it is not clear how 'proper' imputations can be generated when data were collected using a complex survey design.
- S. R. Seaman et al. (2012) derived some important results concerning this.

Recommendations for MI with survey data

Seaman et al's results imply that when performing MI with weighted survey data we should:

1. include the sample design weights as a covariate (not as a weight!) in the imputation model(s),
2. when analysing the imputed datasets, the completed data analyses should be weighted using the survey design weights.

Variance estimation

- One issue is that Rubin's variance estimator can be biased upwards (conservative inference) if the imputer makes an assumption which the analyst doesn't.
- In simple settings, Seaman et al showed that the upward bias in variance estimates could be avoided by including interactions between weights and fully observed variables.
- Even when Rubin's variance estimator was biased (upwards), Seaman et al found that the bias was small.
- In practice therefore, we might worry less about this issue.

Summary

Summary

- Care must be taken that variables are imputed compatibly/congenially with subsequent analyses (substantive models).
- In particular, imputing derievd variables involved in interactions or non-linear effects requires care.
- e.g. with a Cox model, we must account for the outcome appropriately if imputing covariates.
- e.g. with dependent data, our imputation model should ideally account for the dependency.

References I

Andridge, Rebecca R. 2011. "Quantifying the Impact of Fixed Effects Modeling of Clusters in Multiple Imputation for Cluster Randomized Trials." *Biometrical Journal* 53 (1). Wiley Online Library: 57–74.

Audigier, Vincent, Ian R White, Shahab Jolani, Thomas PA Debray, Matteo Quartagno, James Carpenter, Stef Van Buuren, Matthieu Resche-Rigon, and others. 2018. "Multiple Imputation for Multilevel Data with Continuous and Binary Variables." *Statistical Science* 33 (2). Institute of Mathematical Statistics: 160–83.

Bartlett, J W, S R Seaman, I R White, and J R Carpenter. 2015. "Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model." *Statistical Methods in Medical Research* 24 (4): 462–87. https://doi.org/10.1177/0962280214521348.

References II

Buuren, S van. 2011. "Multiple Imputation of Multilevel Data." In *The Handbook of Advanced Multilevel Analysis*, 173–96. Routledge.

Hippel, P T von. 2009. "How to Impute Interactions, Squares, and Other Transformed Variables." *Sociological Methodology* 39: 265–91.

Meng, X L. 1994. "Multiple-Imputation Inferences with Uncongenial Sources of Input (with Discussion)." *Statistical Science* 10: 538–73.

Seaman, S. R., J. W. Bartlett, and I. R. White. 2012. "Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods." *BMC Medical Research Methodology* 12: 46.

Seaman, S. R., I. R. White, A. J. Copas, and L. Li. 2012. "Combining Multiple Imputation and Inverse-Probability Weighting." *Biometrics* 68: 129–37.

References III

White, I. R., and P. Royston. 2009. "Imputing missing covariate values for the Cox model." *Statistics in Medicine* 28: 1982–98.