Compatible Imputation of Missing Covariates in a Meta-Regression

Jacob M. Schauer

Introduction

Why do we need MI, and why do we need adaptations for MR

Example & Model

Give an example of a dataset one might use.

Point out that many common statistical corrections and software assume data are MAR/MCAR, and say what those mean.

Status Quo: Complete Cases & OVAAT

Note that that the most common way to handle missing data is doing complete cases/one variable at a time (OVAAT).

Complete case analyses are only unbiased if the data are MCAR.

OVAAT is only appropriate when covariates are completely independent and data are MCAR.

Multiple Imputation

Suppose we are insterested in estimating some quantity η, such as the regression coefficients or the variance component in the meta-regression model. If we are not missing any data, then there are methods for estimating η, including weighted least squares (WLS). Denote $\hat{\eta}$ as the estimate of η and $\hat{V}[\hat{\eta}]$ as the estimated variance of $\hat{\eta}$. However, in practice, we may be missing some data, and in this article, we assume that one or more covariates of an effect estimate may be missing.

MI works by filling in those missing values with a variety values that we might have observed had not data been missing. The model used to fill in those missing values is called the *impuation model.* This effectively creates m complete datasets where any missing values have been imputed. Analyses are then conducted on each of the imputed datasets, and their results are pooled (see Rubin, 1987; Schaefer, 1997).

Denote $\hat{\eta}^{(i)}$ as the estimate of η from the ith imputed dataset, and let $\hat{V}[\hat{\eta}^{(i)}]$ be its estimated variance. For instance, $\hat{\eta}^{(i)}$ and $\hat{V}[\hat{\eta}^{(i)}]$ may arise from the same WLS approach as $\hat{\eta}$ and $\hat{V}[\hat{\eta}]$. Then, the MI estimate of η is given by

$$\hat{\eta}_{MI} = \sum_{i=1}^{m} \frac{\hat{\eta}^{(i)}}{m} \tag{1}$$

The variance of this estimator is given by

$$\bar{U}_m = \sum_{i=1}^{m} \frac{\hat{V}[\hat{\eta}^{(i)}]}{m} \tag{2}$$

$$B_m = \sum_{i=1}^{m} \frac{(\hat{\eta}^{(i)} - \hat{\eta}_{MI})^2}{m-1} \tag{3}$$

$$\hat{V}[\hat{\eta}_{MI}] = \bar{U}_m + \left(1 + \frac{1}{m}\right) B_m \tag{4}$$

Various researchers have discussed methods for constructing confidence intervals based on equations (1–4). Rubin (1987) proposed using the normal distribution for constructing confidence intervals of the form

$$\hat{\eta}_{MI} \pm z_{1-\alpha/2}\sqrt{\hat{V}[\hat{\eta}_{MI}]}$$

where z_x is the xth percentile of the standard normal distribution. Alternatively, one may also use a reference t-distribution to construct confidence intervals of the form

$$\hat{\eta}_{MI} \pm t_{\nu,1-\alpha/2}\sqrt{\hat{V}[\hat{\eta}_{MI}]}, \quad \nu = (m-1)\left(1 + \frac{m}{m+1}\frac{\bar{U}_m}{B_m}\right) \tag{5}$$

where $t_{a,b}$ is the bth percentile of the t-distribution with a degrees of freedom, and the expression for the degrees of freedom is given above.

Justifications for MI

The original derivation of MI assumed a Bayesian approach to analysis (Rubin, 1987) and is explained nicely by Murray (2018). The idea behind the Bayesian approach to MI is that we are after the posterior distribution

$$p(\eta|Y, X_{obs}, v) = \int p(\eta|Y, X_{obs}, X_{mis}, v)p(X_{mis}|Y, X_{obs}, v)dX_{mis} \tag{6}$$

More specifically, we may wish to the posterior mean and variance:

$$E[\eta|Y, X_{obs}, v] = E[E[\eta|Y, X_{obs}, X_{mis}, v]|Y, X_{obs}, v] \tag{7}$$
$$V[\eta|Y, X_{obs}, v] = E[V[\eta|Y, X_{obs}, X_{mis}, v]|Y, X_{obs}, v] + V[E[\eta|Y, X_{obs}, X_{mis}, v]|Y, X_{obs}, v] \tag{8}$$
$$\tag{9}$$

Generating the missing values X_{mis} via Monte Carlo estimates means that the statistics involve in computing quantities in MI are Monte Carlo estimates of the posterior mean and variance.

From a frequentist perspective, suppose that a inferences using the complete data are confidence valid, which means that confidence intervals have the proper coverage probabilities (i.e., a 95% CI has a 95% coverage probability). Then, MI will also be valid if inferences based on it are also confidence valid. Assuming that estimates of η are (asymptotically) normal, then this would require

$$E[\hat{\eta}] = \eta \tag{10}$$
$$E[\hat{U}] \geq Var[\hat{\eta}] \tag{11}$$

Whether using a Bayesian or frequentist justification, the concept of validity will depend on how the missing data are imputed. We can see this directly with the Bayesian approach. If the imputation model $p(X_{mis}|Y, X_{obs}, v)$ is not consistent in some manner with the analytic model $p(Y|X_{obs},)$

Compatible Imputations

As argued in the previous section, we want the imputation model $p(X_{obs}|Y, v, \gamma)$ to be appropriate. There are various mathematical definitions of what *appropriate* might mean in multiple imputation, though one relevant concept is that of *compatibility.* The general idea is that we are imuting values of $X|Y, v$ and then turning around and using an analytical model $Y|X, v$.
Suppose that X, T have some joint distribution $g(X, Y|v, \eta, \gamma)$, then the imputation and analytic model are compatible if they both proceed from this joint distribution. The formal definition and conditions of compatibility are set out in the Appendix.

Bartlett et al. (2015) show that a natural approach to ensuring that the imputation model is comptabile with the analytic model is to set

$$p(X|T, v, \eta, \gamma) \propto p(Y|X, Z, \eta)p(X|v, \gamma)$$

Example: One Missing Categorical Covariate

Suppose that the meta-regression will involve only a single covariate X. Then the resulting model is

$$T_i = \beta_0 + \beta_1 X_i + u_i + e_i$$

Denote X_{obs} as the X_i we do observe, and X_{mis} as those we do not observe. According to the integral in (6), the inference we are interested in requires Monte Carlo simulations of $p(X_{mis}|T, X_{obs}, v)$. This means is that we require draws of X_{mis} given what we know about T, X_{obs}, and v. Note that we can decompose this into

$$p(X_{mis}|T, X_{obs}, v) = \int p(X_{mis}|\gamma, T, v)p(\gamma|T, X_{obs}, v)d\gamma$$

where γ is the parameter that describes the distribution of $X|T, v$:

$$p(\gamma|T, X_{obs}, v) \propto p(X_{obs}|T, v, \gamma)p(\gamma)$$

Note that the result by Bartlett et al. implies that $\gamma = \eta$.

When X is categorical with $j = 1, \ldots, c$ categories, then we may model it as multinomial. Using the result in () above, this means that

$$P[X = k|T, v, \gamma] = \frac{p(T|X = k, v, \eta)}{\sum_{j=1}^{c} p(T|X = j, v, \eta)} \tag{12}$$

Thus, we can generate imputations of X_{mis} that are compatible with $p(T|X, v, \eta)$ by the following algorithm:

1. Draw $\eta^{(i)}$ from the posterior distribution $p(\eta|T, X_{obs}, v)$
2. Draw X_{mis} from a multinomial distribution with probabilities given in (12).

Note that in this algorithm, η may corresond to β_0, β_1 if it is assumed $\tau^2 = 0$, or $\eta = (\beta_0, \beta_1, \tau^2$ if it is assumed $\tau^2 > 0$.

Note: when $\tau^2 = 0$ in the model, draws of $p(\beta|T, v, X)$ is pretty easy, since the posterior is normal under a flat prior. However, draws from the posterior are difficult. Is there a way to factor this? Is there a way to simplify this for the software?

Example 2: Multiple Missing Categorical Covariates

When there are multiple covariates, the model may be written as

$$T_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + u_i + e_i$$

Let $X_j = (X_{1j}, \dots X_{kj})$ denote the jth variable and $X_{-j} = \{X_1, \dots, X_{j-1}, X_{j+1}, \dots, X_p\}$ denote all of the covariates *except* the jth variable. Let $X_{j,mis}$ denote the missing observations for the jth variable.

Imputation is somewhat trickier here, since the entire the imputation of the joint distribution $p(X_1, \dots, X_p | T, v, \gamma)$ needs to be comptible with the analytic model $p(T|X, v, \eta)$. Not only are multivariate variables more difficult to work with than individual variables, but we would also need to ensure that that multivariate distribution has specific properties. However, as has been done in multiple imputation elsewhere, it may be simpler to break this problem down into a series of conditional distributions.

Example 3: Multiple Mixed-Type Covariates

Generating Imputations

Simulation 1: Amputing Existing Data

Simulation 2: Fully Simulated Data

Discussion