Compatible Imputation of Missing Covariates in a Meta-Regression

Jacob M. Schauer

# Introduction

Meta-regression analyses must frequently contend with missing covariates. This can occur if researchers do not report aspects of their study (e.g., sample composition), or if their reporting does not align with how variables are extracted or coded. How best to handle missing covariates will largely depend on why they are missing. Multiple imputation (MI) is a somewhat flexible approach to analyses of incomplete data that has become popular in a variety of fields. Though it has yet to find broad use in meta-analysis, there has been some initial research into its potential applications in systematic reviews.

MI divides the process of analyses with missing data into two steps. First, missing values are filled in with random draws from an *imputation model*. Multiple values are imputed for each missing field, which creates a series of “complete” datasets. Second, these datasets are analyzed individually by fitting a *substantive model* (i.e., the meta-regression model), and those results are then pooled across datasets to compute point estimates and standard errors.

The accuracy of inferences generated through MI will depend on a number of factors, including the imputation model. When an imputation model is incorrectly specified, it can lead to biased estimators, incorrect standard errors, and/or confidence intervals with poor coverage probabilities. In the context of meta-regression with missing covariates, this can occur if the imputation model is incompatible with the meta-regression model. However, which imputation models for covariates are compatible with meta-regression models has not been explicitly studied in the literature. Nor does there appear to be any implementation of compatible imputations of missing covariates in a meta-regression in most commonly available software.

This article extends existing findings on compatible imputations of covariates to meta-regression models.

# Example

# Model and Notation

In this article, we focus on effects in a meta-analysis that are independent. Suppose there are effects of interest in a meta-regression. Let be the estimate of the th effect. A common assumption in meta-analysis is that is unbiased and normally distributed with known variance . This is exactly true for some effect size indices, and is a very accurate approximation for others. Finally, let be a vector of covariates for the th effect. Then, we can write the standard mixed-effects meta-regression model as:

Here, are the regression coefficients, is the estimation error for effect , and is the random effect. Note that we assume that , , and that . This model contains to sources of variance, is the variance of the estimation error and is the variation of the around the regression line.

The substantive model in a meta-regression concerns the distribution of effect estimates given the covariates and estimation error variance . The parameters of interest that index this model are the regression coefficients and (potentially) the variance component . Denote as the parameter(s) of interest, so that . Then we may write , which we will refer to as the *substantive model* but is also referred to in missing data literature as the *complete-data model*.

Under the assumptions above, the likelihood for a single effect can be written as:

Let be a -dimensional vector of response indicators. This means that if is observed and if is missing. Denote the set of observed covariates and missing covariates . Note that the substantive meta-regression model can be written:

In this article, we assume that covariates are missing at random, so that . This is a standard assumption of many missing data methods, including multiple imputation. However, we note that multiple imputation can still provide accurate inferences if the imputation model takes into account the distribution of .

[NOTE: DESCRIBE COMPLETE DATA ESTIMATION]

## Multiple Imputation

Suppose we are insterested in estimating some quantity , such as a regression coefficient or the variance component in the meta-regression model. If the data are missing covariates, MI works by filling in those missing values with draws from an imputation model. The imputation model, in some sense, reflects what we might have observed had data *not* been missing. This process effectively creates complete datasets such that 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 as the estimate of from the th imputed dataset, and let be its estimated variance. Then, the MI estimate of is given by the average of the :

The variance of this estimator is is a function of the variances of each estimate , as well as the variance between estimates. If we denote the following quantities

then the variance of is given by:

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

where is the th percentile of the standard normal distribution. Alternatively, one may also use a reference -distribution to construct confidence intervals of the form $$ where is the th percentile of the -distribution with degrees of freedom, and the expression for the degrees of freedom is given above.

[NOTE: This is for a single parameter, point out relevant methods for vectors of parameters]

Various researchers have studied conditions under which MI provides accurate inferences. While there are several assumptions, including about the missing data mechanism, two key conditions concern the imputation and substantive models. If both the imputation and substantive models are correctly specified, then MI provides valid and accurate inferences. In this article, we assume that the substantive model is correctly specified.

## Generating Imputations

There is a vast literature on methods for generating imputations. In general, observed values are typically used to build models to predict unobserved variables, and then those models are used to randomly generate imputations. In the context of meta-regression, we might use effect estimates , variances , and observed covariates to predict values of unobserved covariates .

In particular, we are interested in the predictive distribution

There are various ways to specify this distribution, and for the sake of simplicity, we present a general parametric approach. Let denote the parameters that index the distribution of . Then the imputation model is given by:

Here, is the imputation model indexed by , and is the posterior distribution of given the observed data.

Imputations can be generated () by the method of composition:

* Repeat for :
  1. Draw
  2. Draw

# Compatibility

As noted in the previous sections, MI provides valid inferences if the imputation model is correctly specified. One way the imputation model may be incorrectly specified is if it is not *compatible* with the substantive model. Loosely, compatibility means that if we use to predict in an imputation model and then turn around and use to predict in the meta-regression model, those two conditional models should come from the same joint distribution that exists.

Mathematically, compatibility can be defined this way. Suppose is a random vector, is the vector that excludes , and is some random vector. In the context of meta-regression and . Denote the set of conditional distributions . The set of distributions in are said to be compatible if there exists a joint distribution for and a series of surjective maps such that

for each , , and . A set of distributions is said to be *semi-compatible* if they can be made to be compatible by setting some of the parameter values to zero. Finally, a set of distributions is *valid semi-compatible* if they are semi-compatible and their joint distribution is correctly specified.

An imputation model is correctly specified only if it is valid semi-compatible with the substantive model. If the imputation model for would be valid-semi compatible with the meta-regression model if and are compatible, where is the vector with some set to zero.

Bartlett et al. show that we can ensure a compatible imputation model of missing covariates by setting

That is, if we specify an imputation model whose likelihood is proportional to the meta-regression likelihood, that imputation model will be compatible with the meta-regression model. The remainder of this article applies this finding to meta-regression.

# Compatible Imputations of Missing Covariates in Meta-Regression

# Simulation 1: Single Covariate

* Single binary covariate
* Single continuous covariate

# Simulation 2: Multiple Covariates

* 3 continuous covariates
* 3 binary covariates
* 2 binary and 1 continuous covariate

# Discussion

# 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

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

Generating the missing values 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, the justification for MI assumes that 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 are (asymptotically) normal, then this would require

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 is not consistent in some manner with the analytic model

# Generating Imputations

In order for imputations to be proper, they should be generated from a predictive model for . $$ where is the parameter that describes the distribution of :

# Compatible Imputations

As argued in the previous section, we want the imputation model 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 behind compatibility is that we are imuting values of and then turning around and using an analytical model , and that these models should be in synch. Suppose that have some joint distribution , 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

# Compatible Imputation with a Single Categorical Covariate

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

Denote as the we do observe, and as those we do not observe. According to the integral in (), the inference we are interested in requires Monte Carlo simulations of . This means is that we require draws of given what we know about and . Note that we can decompose this into

where is the parameter that describes the distribution of :

Note that the result by Bartlett et al. implies that .

When is categorical with categories, then we may model it as multinomial. Using the result in () above, this means that

Thus, we can generate imputations of that are compatible with by the following algorithm:

Note that in this algorithm, may corresond to if it is assumed , or if it is assumed .

**Note: when in the model, draws of 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?**

# Compatible Imputation with a Single Continuous Covariate

# Multiple Missing Covariates: Fully Conditional Specifications

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

Let denote the th variable and denote all of the covariates *except* the th variable. Let denote the missing observations for the th variable.

Imputation is somewhat trickier here, since the entire the imputation of the joint distribution needs to be comptible with the analytic model . 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.

# Simulation 1: Amputing Existing Data

# Simulation 2: Fully Simulated Data

# Discussion