Compatible Imputation of Missing Covariates in a Meta-Regression

Jacob M. Schauer

1 Introduction

Meta-regression analyses frequently contend with missing covariates (Pigott, 2001b; Tipton et al., 2019). 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 (Pigott, 2001a, 2019; Rubin, 1976). Multiple imputation (MI) is a somewhat flexible approach to analyses of incomplete data that has become popular in a variety of fields (Little & Rubin, 2002; Rubin, 1987; van Buuren, 2018). Though it has yet to find broad use in meta-analysis, there has been some initial research into its potential applications in systematic reviews (Ellington et al., 2015).

MI divides the process of analysis 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 (Fay, 1992; Meng, 1994). When an imputation model is incorrectly specified, it can lead to biased estimators, incorrect standard errors, and/or confidence intervals with poor coverage probabilities (Murray, 2018; Reiter, 2017). 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.

2 Example

[HOLD FOR SA INTERVENTION EXAMPLE]

3 Model and Notation

Suppose there are k effects $\theta_1, \ldots, \theta_k$ of interest in a meta-regression. Let T_i be the estimate of the ith effect. In this article, we assume effect estimates are independent. A common assumption in meta-analysis is that T_i is unbiased and normally distributed with known variance v_i. This is exactly true for some effect size indices, and is a very accurate approximation for others (Cooper et al., 2019). Finally, let $X_i = [X_{i1}, \ldots, X_{ip}]$ be a vector of covariates for the ith effect. Then, we can write the standard mixed-effects meta-regression model as:

$$T_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + e_i + u_i \tag{1}$$

Here, β_j are the regression coefficients, e_i is the estimation error for effect i, and u_i is the random effect. We assume that $e_i \sim N(0, v_i)$, $u_i \sim N(0, \tau^2)$, and that $e_i \perp u_i$. The model in (1) is equivalent to the standard mixed-effects meta-regression model, and it contains to sources of variance, v_i is the variance of the estimation error e_i and τ^2 is the variation of the θ_i around the regression line (Berkey et al., 1995; Hedges & Vevea, 1998).

The substantive model in a meta-regression concerns the distribution of effect estimate T given the covariates X and estimation error variance v. The parameters that index this model are the regression coefficients $\beta = [\beta_0, \ldots, \beta_p]$ and the variance component τ^2. Denote η as the parameter(s) of interest, so that $\eta = [\beta, \tau^2]$. Then we may write $p(T|X, v, \eta)$, which we will refer to as the *substantive model* but is also referred to in missing data literature as the *complete-data model* (Gelman et al., 2014). There is a substantial literature on how to estimate meta-regression models when no data are missing, including weighted least-squares estimators to estimate β, and a variety of methods for estimating τ^2 (Borenstein, 2009; Cooper et al., 2019; Viechtbauer, 2007).

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

$$p(T|X, v, \eta) = \frac{1}{\sqrt{2\pi(\tau^2 + v)}} e^{-\frac{(T - \beta_0 - \ldots - \beta_p X_p)^2}{2(\tau^2 + v)}} \tag{2}$$

Let $R = [R_1, \ldots, R_p]$ be a p-dimensional vector of response indicators for the covariates in X. This means that $R_j = 1$ if X_j is observed and $R_j = 0$ if X_j is missing. For a given effect, the set of observed covariates is $O = \{j : R_j = 1\}$ and missing covariates is $M = \{j : R_j = 0\}$. Note that the

substantive meta-regression model can be written:

$$p(T|X, v, \eta) = p(T|X_O, X_M, v, \eta) \tag{3}$$

In (3), we see that the model depends on both the observed covariates X_O and the unobserved covariates X_M.

How to handle estimating a meta-regression model with missing covariates will depend on why covariates are missing, often referred to as the *missingness mechanism* (Rubin, 1976). The missingness mechanism is typically defined in terms of the distribution of R. Let ξ index the distribution of R given the observed and unobserved data: $p(R|T, X_O, X_M, v, \xi)$. In this article, we assume that covariates are missing at random, so that $p(R|T, X_O, X_M, v, \xi) = p(R|T, X_O, v, \xi)$. This is a standard assumption of many missing data methods, including multiple imputation (Rubin, 1987). However, we note that multiple imputation can still provide accurate inferences if the imputation model takes into account the distribution of R (Gelman et al., 2014; Little & Rubin, 2002; van Buuren, 2018).

4 Multiple Imputation

To simplify discussion of imputation, we focus on estimating a single parameter $\zeta \in \eta$ $\zeta \in \eta$, such as a regression coefficient or the variance component in the meta-regression model. Note, however, the principles presented in this section can be extended to estimating the vector of parameters η. MI works by filling in missing values—in this case, missing covariates—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 m "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 $\hat{\zeta}^{(i)}$ as the estimate of ζ from the ith imputed dataset, and let $\hat{V}[\hat{\zeta}^{(i)}]$ be its estimated variance. Then, the MI estimate of ζ is given by the average of the $\hat{\zeta}^{(i)}$:

$$\hat{\zeta}_{MI} = \sum_{i=1}^{m} \frac{\hat{\zeta}^{(i)}}{m} \tag{4}$$

The variance of this estimator is is a function of the variances of each estimate $\hat{V}[\hat{\zeta}^{(i)}]$, as well as

the variance between estimates. If we denote the following quantities

$$\bar{U}_m = \sum_{i=1}^{m} \frac{\hat{V}[\hat{\zeta}^{(i)}]}{m} \tag{5}$$

$$B_m = \sum_{i=1}^{m} \frac{(\hat{\zeta}^{(i)} - \hat{\zeta}_{MI})^2}{m-1} \tag{6}$$

$$\tag{7}$$

then the variance of $\hat{\zeta}_{MI}$ is given by:

$$\hat{V}[\hat{\zeta}_{MI}] = \bar{U}_m + \left(1 + \frac{1}{m}\right) B_m \tag{8}$$

Various researchers have discussed methods for constructing confidence intervals based on equations (4–8). Rubin (1987) proposed using the normal distribution for constructing confidence intervals of the form

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

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

$$\hat{\zeta}_{MI} \pm t_{\nu,1-\alpha/2}\sqrt{\hat{V}[\hat{\zeta}_{MI}]}, \quad \nu = (m-1)\left(1 + \frac{m}{m+1}\frac{\bar{U}_m}{B_m}\right) \tag{9}$$

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.

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.

4.1 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 T, variances v, and observed covariates X_O to predict values of unobserved covariates X_M.

In particular, we are interested in the predictive distribution

$$p(X_M|X_O, T, v) \equiv p(x|X_O, T, v)$$

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

$$p(X_M|X_O, T, v) = \int p(X_M|X_O, T, v, \gamma)p(\gamma|X_O, T, v)d\gamma \tag{10}$$

Here, $p(X_M|X_O, T, v, \gamma)$ is the imputation model indexed by γ, and $p(\gamma|X_O, T, v)$ is the posterior distribution of γ given the observed data.

Imputations can be generated (10) by the method of composition. To generate m imputations, for $t = 1, \ldots, m$ we:

1. Draw $\gamma^{(t)} \sim p(\gamma|X_O, T, v)$
2. Draw $X_M^{(t)} \sim p(X_M|X_O, T, v, \gamma^{(t)})$

The remainder of this article concerns how to specify models relevant to $p(\gamma|X_O, T, v)$ for meta-regression datasets and how to sample from the distributions in steps 1 and 2.

5   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 T to predict X in an imputation model and then turn around and use X to predict T in the meta-regression model, those two conditional models should come from the same joint distribution, and that distribution actually exists.

Mathematically, compatibility can be defined this way. Suppose $A = [A_1, \ldots, A_p]$ is a random vector, $A_{-j} = [A_1, \ldots, A_{j-1}, A_{j+1}, \ldots, A_{p+1}]$ is the vector that excludes A_j the jth element of A, and B is some random vector. In the context of meta-regression $T \equiv A_1$ and $X_j \equiv A_{j+1}$. Denote the set of $p + 1$ conditional distributions $C = \{p_j(A_j|A_{-j}, B, \lambda_j) : \lambda_j \in \Lambda_j\}$. The set of distributions in C are said to be compatible if there exists a joint distribution $p(A|B, \lambda)$ for $\lambda \in \Lambda$ and a series of surjective maps $\{f_j : \Lambda \mapsto \Lambda_j\}$ such that

$$p_j(A_j|A_{-j}, B, \lambda_j) = p(A_j|A_{-j}, B, \lambda)$$

for each j, $\lambda_j \in \Lambda_j$, and $\lambda \in f_j^{-1}(\lambda_j)$. 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 $p(A|B, \lambda)$ is correctly specified.

An imputation model is correctly specified only if it is valid semi-compatible with the substantive model. Thus, the imputation model $p(X_M|X_O, T, v, \gamma)$ for $\gamma = [\gamma_1, \ldots, \gamma_m]$ would be valid-semi compatible with the meta-regression model $p(T|X, v, \eta)$ if $p(T|X, v, \eta)$ and $p(X_M|X_O, T, v, \gamma*)$ are compatible, where γ^* is the vector γ with some γ_j set to zero, and if the meta-regression model is $p(T|X, v, \eta)$ correctly specified.

Bartlett et al. (2015) show that we can ensure a compatible imputation model of missing covariates by setting

$$p(x|X_O, T, v, \gamma) \propto p(T|x, X_O, v, \eta)p(x|X_O, v, \psi) \tag{11}$$

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.

6   Compatible Imputations of Missing Covariates in Meta-Regression

Equation (11) for a compatible imputation model actually depends on two different functions. First, the imputation model depends on the substantive model $p(T|x, v, \eta)$. Second, it depends on a model of the missing covariates given the observed covariates and the estimation variances $p(x|X_O, v, \psi)$. Taken together, the imputation model parameter is $\gamma = [\eta, \psi]$, which contains the parameters in the meta-regression model. If we treat η and ψ as independent, then sampling γ^* from its posterior distribution is equaivalent to sampling η from the posterior distribution $p(\eta|X_O, T, v) = p(T|X_O, v, \eta)p(\eta)$ and sampling ψ from its posterior $p(x|X_O, v, \psi)p(\psi)$. Given a draw γ^*, we can then impute missing values by drawing $X_M^* \sim p(x|X_O, T, v, \gamma^*)$.

6.1   Missingness in One Covariate

If there is only missingness in a single covariate X_j, generating imputations is relatively straightforward. Samples $\eta^{(1)}, \ldots, \eta^{(m)}$ can be drawn from the posterior distribution of $p(\eta|X_{-j}, X_{jO}, T, v)$, which is equivalent to the posterior distribution when fitting the meta-regression model on complete-cases. Samples $\psi^{(1)}, \ldots, \psi^{(m)}$ can be drawn from the posterior distribution of $p(\psi|X_{-j}, X_{jO}, T, v)$. Assuming the model

$p(X_j|X_{-j}, v, \psi)$ is a linear regression for continuous X_j and log-linear for categorical X_j, this procedure has been well-studied.

Given $\gamma^{(t)} = [\eta^{(t)}, \psi^{(t)}]$, we then generate imputations by drawing $X_j^{(t)}$ for each missing X_j. Recall that the distribution of X_j is proportional to $f(X_j) = p(T|X_j, X_{-j}, v, \eta^{(t)})p(X_j|X_{-j}, v, \psi^{(t)})$. Bartlett et al. (2015) propose a rejection sampling approach to generating draws of $X_j^{(t)}$ from this distribution. However it may be possible to sample directly from $f(X_j)$ when the auxiliary model $p(X_j|X_{-j}, v, \psi^{(t)})$ follows specific parametric distributions. When X_j is continuous and the auxiliary model is a normal linear model, then $f(X_j)$ can be written as a normal model with mean XXX and variance XXX (NOTE: perhaps an appendix). Conversely if X_j is categorical and the auxiliary model is a log-linear model (i.e., logit or multinomial logit), it can be shown that $f(X_j)$ is also a log-linear model with probabilities XXX (NOTE: appendix?).

6.2 Missingness in Multiple Covariates: Conditional Models

If there are multiple missing covariates (and potentially different missing data patterns), specifying $p(X_M|X_O, T, v, \gamma)$ becomes more difficult. This is because it requires the specification of joint distributions across the X_j that are missing values. While it may be possible to specify a joint model across all missing covariates $p(X_M|X_O, v, \psi)$ doing so in a way that takes into account different variable types of the missing X_j and different missingness patterns can be cumbersome, and it is seldom obvious what type of joint distribution may be appropriate.

A common way to simplify the problem of specifying a joint model is to instead use a series of conditional models. Let $X_{-j} = [X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p]$ be the vector X that omits X_j. Then, we can specify a series of conditional models $p(X_j|X_{-j}, v, \psi_j)$ and iteratively sample from each. If we let X_{jO} be the observed values of X_j in the data, the at iteration (t) we draw $\gamma_j^{(t)} \sim p(\gamma_j|X_{jO}, X_{-j}^{(t)}, T, v)$ for $j = 1, \ldots, p$, where

$$p(\gamma_j|X_{jO}, X_{-j}^{(t)}, T, v) = p(X_{jO}|X_{-j}^{(t)}, T, v, \gamma_j)p(\gamma_j) \tag{12}$$

is the posterior of γ_j given observed values of X_j and all observed or imputed values of X_{-j}. So long as the $p(X_j|X_{-j}, v, \gamma_j)$ are valid semi-compatible, this will provide consistent MI estimates.

In the algorithm above, we have not accommodated the substantive model. Doing so would involve factoring the left-hand side of (12) into

$$p(\gamma_j|X_j, X_{-j}^{(t)}, T, v) = p(\eta, \psi_j)p(T|X_j, X_{-j}^{(t)}, v, \eta)p(X_j|X_{-j}^{(t)}, v, \psi_j) \tag{13}$$

Again, assuming that $\psi_j \perp \eta$, we can simplify this function so that

$$p(\gamma_j | X_j, X_{-j}^{(t)}, T, v) = [p(\eta)p(T|X_j, X_{-j}^{(t)}, v, \eta)]p(\psi_j)p(X_j|X_{-j}^{(t)}, v, \psi_j) \tag{14}$$

Note that (14) is the product of two posterior distributions: one for η and one for ψ_j.

Based on (14), we can generate imputations of missing covariates via an algorithm analogous to a Gibbs sampler. The procedure doing so at iteration t is as follows. For $j = 1, \ldots, p$:

1. Draw $\psi_j^{(t)} \sim p(X_j^{(t-1)}|X_{-j}^{(t)}, v, \psi_j)p(\psi_j) = p(\psi_j|X_j^{(t-1)}, X_{-j}^{(t)}, v)$

2. Draw $\eta^{(t)} \sim p(T|X_j(t-1), X_{-j}^{(t)}, v, \eta)p(\eta) = p(\eta|X_j^{(t-1)}, X_{-j}^{(t)}, T, v)$

3. Draw $X_j^{(t)} \sim p(X_j|X_{-j}^{(t)}, T, v, \psi_j^{(t)}, \eta_j^{(t)}) \propto p(T|X_j^{(t)}, X_{-j}^{(t)}, v, \eta^{(t)})p(X_j^{(t)}|X_{-j}^{(t)}, v, \psi_j^{(t)})$

Note that the algorithm above loops through each predictor X_j. It uses imputations for X_j from the prior iteration $X_j^{(t-1)}$ and imputations of the other covariates in the current iteration $X_{-j}^{(t)}$ to compute posterior distributions in steps 1 and 2. Step 3 imputes new values for missing X_j, denoted $X_j^{(t)}$.

The model used to generate draws of $X_j^{(t)}$ in step 3 is again written as the product of the meta-regression likelihood and the distribution function $p(X_j^{(t)}|X_{-j}^{(t)}, v, \psi_j^{(t)})$, where $p(X_j^{(t)}|X_{-j}^{(t)}, v, \psi_j^{(t)})$ is specified by the imputer. The expression for this model in step 3 follows from the results of Bartlett et al. (2015), and serves to induce compatibility in the conditional imputation models. Bartlett et al. suggest using a rejection sampler for step 3. However, as was the case with a single covariate, when the auxiliary model is either a normal linear model or a log-linear model, we can sample directly from the relevant normal or log-linear model described in the previous section.

This algorithm is consistent with a the fully conditional specification approach to multiple imputation, often referred to as multiple imputation with chained equations (MICE). The key adaptation is that it induces compatibility with the substantive model at each step. Because the process is analogous to a Gibbs sampler, it is recommended that the algorithm cycle through all covariates numerous times as a burn-in period. Typically, 10 iterations are required for burn in.

Bartlett et al. argue that if the individual conditional models are valid semi-compatible and are compatible with the substantive model, then the entire process should provide valid MI inference for meta-regression. By using imputation models that are compatible with the substantive model at step 3 for each X_j, this ensures compatibility with the substantive model. The conditional models will be semi-compatible with each other if they follow the same linear exponential family, such as linear regression models for $p(X_j^{(t)}|X_{-j}^{(t)}, v, \psi_j^{(t)})$ when X_j is continuous, and log-linear models when X_j is categorical. An implementation of this algorithm is included as a supplement to this article.

7   Simulation 1: Single Covariate

- Single binary covariate
- Single continuous covariate

8   Simulation 2: Multiple Covariates

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

9   Discussion

- Augment models to incorporate auxiliary variables.

10   References

Bartlett, J. W., Seaman, S. R., White, I. R., Carpenter, J. R., & for the Alzheimer's Disease Neuroimaging Initiative*. (2015). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research*, *24*(4), 462–487. https://doi.org/10.1177/0962280214521348

Berkey, C. S., Hoaglin, D. C., Mosteller, F., & Colditz, G. A. (1995). A random-effects regression model for meta-analysis. *Statistics in Medicine*, *14*(4), 395–411. https://doi.org/10.1002/sim.4780140406

Borenstein, M. (2009). *Introduction to meta-analysis.* Chichester, U.K.: John Wiley & Sons. Retrieved from http://public.ebookcentral.proquest.com/choice/publicfullrecord.aspx?p=427912

Cooper, H. M., Hedges, L. V., & Valentine, J. C. (Eds.). (2019). *Handbook of research synthesis and meta-analysis* (3rd edition). New York: Russell Sage Foundation.

Ellington, E. H., Bastille-Rousseau, G., Austin, C., Landolt, K. N., Pond, B. A., Rees, E. E., . . . Murray, D. L. (2015). Using multiple imputation to estimate missing data in meta-regression. *Methods in Ecology and Evolution*, *6*(2), 153–163. https://doi.org/10.1111/2041-210X.12322

Fay, R. E. (1992). When are inferences from multiple imputation valid? *Proceedings of the Survey Research Methods Section, American Statistical Association*, 227–232.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (Third edition). Boca Raton: CRC Press.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*(4), 486–504. https://doi.org/10.1037/1082-989X.3.4.486

Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data.* Hoboken, NJ, USA: John Wiley & Sons, Inc. Retrieved from http://doi.wiley.com/10.1002/9781119013563

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, *9*(4), 538–558. https://doi.org/10.1214/ss/1177010269

Murray, J. S. (2018). Multiple imputation: A review of practical and theoretical findings. *arXiv:1801.04058 [Stat].* Retrieved from http://arxiv.org/abs/1801.04058

Pigott, T. D. (2001a). A review of methods for missing data. *Educational Research and Evaluation*, *7*(4), 353–383. https://doi.org/10.1076/edre.7.4.353.8937

Pigott, T. D. (2001b). Missing predictors in models of effect size. *Evaluation & the Health Professions*, *24*(3), 277–307. https://doi.org/10.1177/01632780122034920

Pigott, T. D. (2019). Handling missing data. In Harris Cooper, Larry V. Hedges, & Jeffrey C. Valentine (Eds.), *The Handbook for Research Synthesis and Meta-analysis* (3rd ed.). New York: Russell Sage.

Reiter, J. (2017). Discussion: Dissecting Multiple Imputation from a Multi-phase Inference Perspective: What Happens When God's Imputer's and Analyst's Models are Uncongenial? *Statistica Sinica*. https://doi.org/10.5705/ss.202016.0287

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Tipton, E., Pustejovsky, J. E., & Ahmadi, H. (2019). Current practices in meta-regression in psychology, education, and medicine. *Research Synthesis Methods*, *10*(2), 180–194. https://doi.org/10.1002/jrsm.1339

van Buuren, S. (2018). *Flexible Imputation of Missing Data, Second Edition* (2nd ed.). Second edition. | Boca Raton, Florida : CRC Press, [2019] |: Chapman and Hall/CRC. Retrieved from https://www.taylorfrancis.com/books/9780429492259

Viechtbauer, W. (2007). Accounting for heterogeneity via random-effects models and moderator analyses in meta-analysis. *Zeitschrift Für Psychologie / Journal of Psychology*, *215*(2), 104–121. https://doi.org/10.1027/0044-3409.215.2.104