Notes on Complete- and Available-Case Analyses

Model and Notation

Let T_i be the estimate of the effect parameter θ_i, and let v_i be the estimation error variance of T_i. Denote a vector of covariates that pertain to T_i as $X_i = [1, X_{i1}, \ldots, X_{i,p-1}]^T$. Note that the first element of X_i is a 1, which corresponds to an intercept term in a meta-regression model, and that model can be expressed as:

$$T_i | X_i, v_i, \eta = X_i^T \beta + r_i + e_i$$

Here, $\beta \in \mathbb{R}^p$ is the vector of regression coefficients, $r_i \perp e_i$ and r_i is the random effect of effect i such that $V[r_i] = \tau^2$. The term e_i is the estimation error for effect i and $V[e_i] = v_i$. This is the standard random effects meta-regression model, is also consistent with subgroup analysis models. The parameter $\eta = [\beta, \tau^2]$ refers to the parameters of model. Note that a under a fixed-effects model, it is assumed that $\tau^2 = 0$, in which case $\eta = \beta$.

A common assumption in random effects meta-regression is that the random effects r_i are independent and normally distributed with mean zero and variance τ^2: $r_i \sim N(0, \tau^2)$. In that case, the distribution $p(T|X, v, \eta)$ can be written as

$$p(T_i | X_i, v_i, \eta) = \frac{1}{\sqrt{2\pi(\tau^2 + v_i)}} e^{-\frac{(T_i - X_i^T \beta)^2}{2(\tau^2 + v_i)}}$$

Note that this assumes that all covariates are observed, and is referred to as the complete data likelihood function.

Let R_i be a vector of response indicators for effect i. Each element R_{ij} of R_i corresponds to a variable in a meta-analytic dataset. The R_{ij} take a value of either 0 or 1: $R_{ij} = 1$ indicates a given variable is observed and $R_{ij} = 0$, indicates a given variable is not observed. For the data $[T_i, v_i, X_i]$, $R_i \in \{0, 1\}^{p+1}$ is a vector of 0s and 1s of length $p + 1$. If v_i were missing for some effect, then $R_{i2} = 0$.

Our focus is on missing covariates, and thus, this article assumes that T_i and v_i are observed for every effect of interest in a meta-analysis. Thus, we amend the notation so that $R_i \in \{0, 1\}^{p-1}$ and $R_{ij} = 1$ if X_{ij} is observed and $R_{ij} = 0$ if X_{ij} is missing. Note that this omits the intercept term described above. For instance if $X_i = [1, X_{i1}]$ with $X_{i1} \in \mathbb{R}$, then R_i is a scalar such that $R_i = 1$ if X_{i1} is observed, and $R_i = 0$ if it is missing. Denote $O = \{(i, j) : R_{ij} = 1\}$ as the indices of covariates that are observed and $M = \{(i, j) : R_{ij} = 0\}$ be the set of indices for unobserved covariates. Then, the complete-data model can be written as $p(T_i | X_i, v_i, \eta) = p(T_i | X_{iO}, X_{iM}, v_i, \eta)$. Note that the complete-data model depends on X_{iM}, which are unobserved.

Conditional Incomplete Data Meta-Regression

When meta-analytic datasets are missing covariates, analyses involve incomplete data. In practice, meta-regressions of incomplete data have largely relied on one of two approaches: complete-case analyses and available case analyses. A complete-case analysis includes only effects for which all covariates of interest are observed, which means that $R_i = [1, \ldots, 1] = \mathbf{1}$ for all effects included in the analysis. In such cases, inferences are based on the conditional distribution of $T_i | X_i, v_i, R_i = \mathbf{1}$ for some missing data pattern $r \in \{0, 1\}^{p-1}$. As we argue in a later section, available-case meta-regressions typically amount to including a subset of relevant covariates that are completely observed. This is equivalent to including observations

for which R_i fall into some set of missingness patterns $\mathcal{R}_0$. In this formulation, available-case inferences are based on the distribution of $T_i|X_i, v_i, R_i \in \mathcal{R}$.

Note that both complete- and available-case analyses condition on the value of R_i. In that sense, we can see both of these approaches as *conditional on missingness* (COM). Because both models above are conditional, they are not necessarily identical to the complete-data model, which is the model of interest. Yet, the analytic approaches of complete- and available-case analyses proceed as if the complete-data and COM models are equivalent. Doing so ignores the sources and impacts of missingness, and can lead to inaccurate results.

The complete-data model can be related to the COM models by through the distribution of missingness. This approach is referred to as a *selection model* in the missing data literature. Let ψ parametrize the distribution of R given the observed and unobserved data. We can write a selection model as:

$$p(T_i|X_i, v_i, R_i = r, \eta, \psi) = \frac{p(R_i = r|T_i, X_i, v_i, \psi)p(T_i|X_i, v_i, \eta)}{p(R_i = r|X_i, v_i, \psi, \eta)}$$

This describes the conditional model as a function of the complete-data model $p(T|X, v, \eta)$ and a selection model $p(R_i = r|T_i, X_i, v_i, \psi)$ that gives the probability that a given effect and its covariates are used in the analysis. The denominator on the right hand side is the probability of a missingness pattern r given the estimation error variance v_i and the observed and unobserved covariates in the vector X_i, and can be written as

$$p(R_i = r|X_i, v_i, \psi, \eta) = \int p(R_i = r|T_i, X_i, v_i, \psi)p(T_i|X_i, v_i, \eta)dT$$

A standard approach for modelling missingness in covariates is to assume R_i follows some log-linear distribution. Suppose we can write

$$\text{logit}P[R_i = r|T_i, X_i, v_i] = \sum_{j=0}^{n} \psi_j f_j(T_i, X_i, v_i)$$

where $f_0(T_i, X_i, v_i) = 1$, so that ψ_0 would be the intercept term for this logit model. While log-linear models are not the only applicable or appropriate model for missingness, we make this assumption at points throughout this article in order to demonstrate conditions under which conditional meta-regressions are inaccurate, and how inaccurate they can be.

Issues with Conditional Inference on Incomplete Data

There are various concerns about the accuracy of inferences that condition on a given missingness pattern. Estimates based on a given missingness pattern may be biased. As well, because conditioning on a missingness pattern often invovles excluding data points, it is likely that estimates are also more variable (i.e., have higher standard errors).

Note that the bias induced by complete- or available-case analyses will depend on the expectation $E[T_i|X_i, v_i, R_i = r]$. It is possible to derive an approximation for $E[T_i|X_i, v_i, R_i = r]$ when $R_i|T_i, X_i, v_i$ follows a log-linear distribution as described in the previous section. Let $\mathcal{J} = \{j : f_j \text{ depends on } T_i\}$ so that we can write $f_j(T, X, v)$ for $j \in \mathcal{J}$ and $f_j(T, X, v) = f_j(X, v)$ for $j \notin \mathcal{J}$. Further, denote

$$G(X_i, v_i) = P[R_i = r|T_i, X_i, v_i]|_{T_i = X_i^T \beta}$$

Then an approximation for $E[T_i|X_i, v_i, R_i = r]$ is as follows:

$$
\begin{aligned}
& \text{omitting subscripts, Taylor series for exponents in } P[R_i = r|T_i, X_i, v_i] \text{ at } T = X^T\beta \\
E[T_i|X_i, v_i, R_i = r] &= \int \frac{T \exp\left\{-\frac{(T-X\beta)^2}{2(\tau^2+v)} + \sum_j \psi_j f_j(T, X, v)\right\}}{g(X, v)\sqrt{2\pi(\tau^2+v)} \log\left(1 + e^{\sum_j \psi_j f_j(T,X,v)}\right)} dT \\
&= \int \frac{T \exp\left\{-\frac{(T-X\beta)^2}{2(\tau^2+v)} + \sum_j \psi_j f_j(X^T\beta, X, v) + \sum_j \psi_j f_j'(X^T\beta, X, v)(T - X^T\beta)\right.}{g(X, v)\sqrt{2\pi(\tau^2+v)}} \\
&\qquad \frac{\left. - \log\left(1 + e^{\sum_j \psi_j f_j(X^T\beta,X,v)}\right) - G(X, v)(\sum_j \psi_j f_j'(X^T\beta, X, v))(T - X\beta) + O(T^2)\right\}}{g(X, v)\sqrt{2\pi(\tau^2+v)}} dT \\
&\approx \int \frac{T \exp\left\{-\frac{1}{2(\tau^2+v)}\left(T^2 - 2TX\beta - 2T(\tau^2+v)\sum_{j\in\mathcal{J}} \psi_j f_j'(X^T\beta, X, v)\right.\right.}{g(X, v)\sqrt{2\pi(\tau^2+v)}} \\
&\qquad \frac{\left.\left. + 2T(\tau^2+v)G(X, v)(\sum_{j\in\mathcal{J}} \psi_j f_j'(X^T\beta, X, v)) + \ldots\right)\right\}}{g(X, v)\sqrt{2\pi(\tau^2+v)}} dT \\
&= X\beta + (1 - G(X, v))(\tau^2+v)\sum_{j\in\mathcal{J}} \psi_j f_j'(X^T\beta, X, v)
\end{aligned}
$$

Note that this uses a first order Taylor expansion of the exponentiated terms of the log-linear model. However, if it is assumed that the f_j are linear in T, then only a Taylor expansion of the denominator of the log-linear model is required.

Complete-Case Analyses

A common approach in meta-regression with missing covariates is to use a complete-case analysis. There are conditions under which the complete case analysis will lead to unbiased estimates. First, if the covariates are MCAR, so that

$$P[R|T, X, v, \psi] = \psi$$

then

$$
\begin{aligned}
p(T|X, v, R = r, \eta) &= \frac{p(R = r|T, X, v, \psi)p(T|X, v, \eta)}{p(R = r|X, v, \psi)} \\
&= \frac{\psi p(T|x, v, \eta)}{\psi} \\
&= p(T|x, v, \eta)
\end{aligned}
$$

Thus, likelihood-based estimation should be consistent, assuming it can be done when X is MCAR. This is consistent with broader results on analyses of MCAR data.

However, it would appear that a complete-case analysis is also valid under slightly less restrictive assumptions. Suppose that $R \perp (X, T)|v$, then

$$
\begin{aligned}
p(T|X, v, R = r) &= \frac{p(R = r|T, X, v, \psi)p(T|X, v, \eta)}{p(R = r|X, v, \psi)} \\
&= \frac{p(R = r|v, \psi)p(T|x, v, \eta)}{p(R = r|v, \psi)} \\
&= p(T|x, v, \eta)
\end{aligned}
$$

The assumption that $R \perp (X, T)|v$ implies that if missingness only depends on the estimation error variances, then a complete case analysis may be appropriate. This is a weaker assumption than MCAR, which requires $R \perp (T, X, v)$. For most effect size indices, variances v are functions of the sample sizes within studies n. Some effect sizes, such as the z-transformed correlation coefficient, have variances v that depend entirely on the sample size of a study, while for other effect sizes this is approximately true, such as the standardized mean difference. For such effect sizes, this assumption implies that missingness depends only on the sample size of the study. This may be true, for instance, if smaller studies are less likely to report more fine-grained demographic information regarding their sample out of concern for the privacy of the subjects who participated in the study (and that no other factors affect missingness).

- Note about reduction in precision.

However, when R is not independent of X or T (given v), then analyses can be biased. Precisely how biased will depend on the distribution of R and its relationship to effect estimates T and their covariates X. A general result based on the approximation in the previous section follows from the fact that a complete-case analysis involves $R = \mathbf{1}$. Further, denote $\mathbf{X}_R$ as the matrix of covariates such $R_i = \mathbf{1}$. If we denote $\mathbf{W}_R = diag(v_i + \tau^2) : R_i = \mathbf{1}$, the complete-case analysis will estimate coefficients as:

$$\begin{aligned}
\hat{\beta} &= (\mathbf{X}_R^T W_R \mathbf{X}_R)^{-1} \mathbf{X}_R^T \mathbf{W}_R E[T|X, v, R = r] \\
&= (\mathbf{X}_R^T \mathbf{W}_R \mathbf{X}_R)^{-1} \mathbf{X}_R^T \mathbf{W}_R \left[X_i^T \beta + \left((1 - G(X_i, v_i))(\tau^2 + v_i) \sum_{j \in \mathcal{J}} \psi_j f_j'(X_i^T \beta, X_i, v_i) \right) \right] \\
&= \beta + (\mathbf{X}_R^T \mathbf{W}_R \mathbf{X}_R)^{-1} \mathbf{X}_R^T \mathbf{W}_R \left[(1 - G(X_i, v_i))(\tau^2 + v_i) \sum_{j \in \mathcal{J}} \psi_j f_j'(X_i^T \beta, X_i, v_i) \right]
\end{aligned}$$

Thus, the bias of the regression coefficients is given by

$$(\mathbf{X}_R^T \mathbf{W}_R \mathbf{X}_R)^{-1} \mathbf{X}_R^T \mathbf{W}_R \left[(1 - G(X_i, v_i))(\tau^2 + v_i) \sum_{j \in \mathcal{J}} \psi_j f_j'(X_i^T \beta, X_i, v_i) \right]$$

- Refine for matrix notation. Let $\mathbf{I}$ be the identity matrix, $\mathbf{G} = diag(G(X_i, v_i))$ and $Y = [\sum_{j \in \mathcal{J}} \psi_j f_j'(X_i^T \beta, X_i, v_i), i = 1, \ldots]^T$. Then we can write:

$$(\mathbf{X}_R^T \mathbf{W}_R \mathbf{X}_R)^{-1} \mathbf{X}_R^T \mathbf{W}_R \mathbf{W}_R^{-1} (\mathbf{I} - \mathbf{G}) Y = (\mathbf{X}_R^T \mathbf{W}_R \mathbf{X}_R)^{-1} \mathbf{X}_R^T (\mathbf{I} - \mathbf{G}) Y$$

- Set up as proposition and proof.
- If the likelihood for conditional inference can be written free of R then conditional inference may be appropriate.
- If we're worried about bias, then the PMM or selection model can help us understand bias. PMM may be most useful for this. Show formulas.
- If we're worried about uncertainty, selection models can help unpack that.

Example: Suppose there is only one covariate $X_{i1} \equiv X_i \in \mathbb{R}$, and that $p(R = 1|T, X, v) \propto \frac{\exp\{\psi_0 + \psi_1 X + \psi_2 T + \psi_3 TX\}}{1 + exp\{\psi_0 + \psi_1 X + \psi_2 T + \psi_3 TX\}}$ Note that $G(X, v) \equiv G(X)$. Further, $\mathcal{J} = \{2, 3\}$ and $f_j(T, X, v) = T f_j(X, v)$, and hence $f_j'(T, X, v) = f_j(X, v)$ for $j \in \mathcal{J}$. Thus,

$$E[T_i|X_i, v_i, R_i = 1] = \beta_0 + \beta_1 X_i + (1 - G(X_i))(v_i + \tau^2)(\psi_2 + \psi_3 X_i)$$

Suppose X is a binary variable, so that $X \in \{0,1\}^{p-1}$, and that X is observed for $M \leq k$ effects. Denote m_0 as the nubmer of observed $X_i = 0$ and m_1 be the observed $X_i = 1$ so that $M = m_0 + m_1$. Further, assume that $X_i = 0$ (but $R_i = 1$) for $i = 1, \dots m_0$ and $X_i = 1$ (and $R_i = 1$) for $m_0 + 1, \dots, M$. Then then the ML and least squares estimator for β_0 is given by

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{m_0} T_i/(v_i + \tau^2)}{\sum_{i=1}^{m_0} 1/(v_i + \tau^2)}$$

This would imply that the complete-case estimator of β_0 under the missing data model specified would have expectation:

$$\begin{aligned} E[\hat{\beta}_0] &= \beta_0 + \frac{(1 - G(0))\, m_0}{\sum_{i=1}^{m_0} 1/(v_i + \tau^2)} \psi_2 \\ w_{j\cdot} &= \left[\sum_{i:X_i=j, R_i=1} 1/(v_i + \tau^2) \right] / m_j \\ E[\hat{\beta}_0] &= \beta_0 + \frac{1 - G(0)}{w_{0\cdot}} \psi_2 \end{aligned}$$

Note that the second term on the right hand side constitutes the bias of $\hat{\beta}_0$. The bias depends on a variety of quantitites. First, it depends on the selection model, solely through ψ_2 When $\psi_2 > 0$ and hence when larger effect estimates are more likely to be missing covariates, the bias is negative. The bias is positive when $\psi_0 < 0$, which occurs when larger effect estimates are less likely to be missing covariates.

The ML estimator of β_1 is given by

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=m_0+1}^{M} T_i/(v_i + \tau^2)}{\sum_{i=m_0+1}^{M} 1/(v_i + \tau^2)} - \frac{\sum_{i=1}^{m_0} T_i/(v_i + \tau^2)}{\sum_{i=1}^{m_0} 1/(v_i + \tau^2)} \\ &= \frac{\sum_{i=m_0+1}^{M} T_i/(v_i + \tau^2)}{\sum_{i=m_0+1}^{M} 1/(v_i + \tau^2)} - \hat{\beta}_0 \end{aligned}$$

The expectation of this estimator is approximately

$$\begin{aligned} E[\hat{\beta}_1] &= \beta_0 + \beta_1 + \frac{(1 - G(1))m_1}{\sum_{i=m_0+1}^{M} 1/(v_i + \tau^2)} (\psi_2 + \psi_3) - \beta_0 - \frac{(1 - G(0))\, m_0}{\sum_{i=1}^{m_0} 1/(v_i + \tau^2)} \psi_2 \\ &= \beta_1 + \frac{(1 - G(1))m_1}{\sum_{i=m_0+1}^{M} 1/(v_i + \tau^2)} (\psi_2 + \psi_3) - \frac{(1 - G(0))\, m_0}{\sum_{i=1}^{m_0} 1/(v_i + \tau^2)} \psi_2 \\ &= \beta_1 + \left[\frac{1 - G(1)}{w_{1\cdot}} - \frac{1 - G(0)}{w_{0\cdot}} \right] \psi_2 + \frac{1 - G(1)}{w_{1\cdot}} \psi_3 \end{aligned}$$

- Plots?
- Continuous covariates?

Available-Cases and Shifting Units of Analysis

When there are multiple covariates of interest, each of which has some missing data, it may be that there are only a few effects for which all covariates of interest are observed. When that happens, a complete case analysis may be infeasible. A common solution to this in meta-analysis is to use an available-case analysis.

In meta-anlaysis, an *available-case analysis* typically takes the form of fitting several meta-regression models, each including a subset of the covariates of interest. Sometimes this even takes the form of regressing effect estimates on one covariate at a time. This "sifting units of analysis" (SUA) inherently conditions on a set of

missingness patterns: $R \in \{\tilde{r}_1, \ldots, \tilde{r}_n\}$. To see this, note that SUA amounts to a complete-case analysis on a subset of covariates. Thus $R_j = 1$ for those covariates X_j that are observed and included in the analyses, but $R_k \in \{0, 1\}$ for covariates X_k that are excluded; that is, excluded covariates may be observed or unobserved.

Let $O_S = \{i, j : R_{ij} = 1 \text{ and } X_j \text{ included in model}\}$, and let M_S be all of the indices i, j not in O_S. The set M_S would indicate observations i for which an X_{ij} would have been included in an analysis had it been observed, or covariates j that are excluded. Then, the SUA model can be written as:

$$p(T|X_{O_S}, v, R \in \mathcal{R}) = \frac{P[R \in \mathcal{R}|T, X_{O_S}, v]p(T|X_{O_S}, v)}{P[R \in \mathcal{R}|X_{O_S}, v]}$$

Note that $p(T|X_{O_S}, v)$ is not necessarily the same as $p(T|X, v)$, which includes all covariates of interest. Thus, the complete-SUA likelihood function is only equal to the complete-data likelihood if $T \perp X_E|X_I, v$. That is, the two models are equivalent only if the coefficients $\beta_j = 0$ for all excluded covariates (given the excluded covariates).

Example: Suppose $X = [1, X_1, X_2]$ and that there is missingness in both X_1 and X_2. Then $R \in \{0, 1\}^2$ so that $R = [1, 1]$ indicates both covariates are observed, and $R = [1, 0]$ indicates only the first covariate is observed.

If missingness is such that $R = [1, 1]$ for very few effect estimates, then a shifting units analysis might involve regressing T on the observed values of X_1 and then on the observed values of X_2.