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

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

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. The elements R_{ij} of R_i take values of either 1, indicating a given variable is observed, or 0, indicating that 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. For instance if $X_i \in \mathbb{R}$, then R_i is a scalar such that $R_i = 1$ if X_i is observed, and $R_i = 0$ if it is missing. Denote $O = \{(i, j) : R_{ij} = 1\}$ be the set of covariates that are observed and $M = \{(i, j) : R_{ij} = 0\}$ be the set of unobserved covariates. Then, the complete-data model can be written as $p(T|X, v, \eta) = p(T|X_O, X_M, v, \eta)$.

Conditional Meta-Regression

Only including certain data points in an analysis conditional on the missingness pattern is a common approach in meta-regression. For instance, complete-case analyses involve 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 $T|X, v, R = r$ for some missing data pattern $r \in \{0, 1\}^{p-1}$.

There are a few ways to relate the the complete-data model and the model that conditions on a missingness pattern in a few ways. Let ψ parametrize the distribution of R given the observed and unobserved data. We can write a **selection model**:

$$p(T|X, v, R = r, \eta, \psi) = \frac{p(R = r|T, X, v, \psi) p(T|X, v, \eta)}{p(R = r|X, v, \psi)}$$

This describes the conditional model as a function of the complete-data model $p(T|X, v, \eta)$ and a selection model $p(R = r|T, X, v, \psi)$ 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 and the observed and unobserved covariates in the vector X, and can be written as

$$p(R = r|X, v, \psi) = \int p(R = r|T, X, v, \psi)p(T|X, v, \eta)dT$$

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

$$\text{logit}P[R = r|T, X, v] = \sum_{j=0}^{n} \psi_j f_j(T, X, v)$$

Typically, it is assumed that $f_0(T, X, v) = 1$. While this is not necessary to model missingness, we make this assumption at points throughout this article in order to demonstrate conditions under which conditional meta-regressions are inaccurate.

Alternatively, the complete data model $p(T|X, v, \eta)$ can be expressed as a mixture over the missingness patterns:

$$p(T|X, v, \eta) = \sum_{r \in \{0,1\}^{p-1}} p(T|X, v, R = r, \eta)p(R = r|X, v, \eta, \psi)$$

Thus, for a specific $R = \tilde{r}$, we can write

$$p(T|X, v, R = \tilde{r}, \eta) = \frac{p(T|X, v, \eta)}{p(R = \tilde{r}|X, v, \eta, \psi)} - \sum_{r \neq \tilde{r}} p(T|X, v, R = r, \eta)\frac{p(R = r|X, v, \eta, \psi)}{p(R = \tilde{r}|X, v, \eta, \psi)}$$

- PMM and selection models can be shown to be equivalent.
- Under certain conditions, we can write the expectation of PMM models as (approximately) linear combinations of covariates analogous to the linear model of interest.

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).

It is possible to derive a somewhat general result for the bias of conditional meta-regressions when $R|T, X, v$ follows a log-linear distribution as described in the previous section. We must further assume that $f_j(T, X, v)$ in that model either do not depend on T or are linear in T. Thus, for some $f_j(T, X, v)$ we have $f_j = Tf_j(X, v)$ and for other f_j that do not depend on T, we can write $f_j = f_j(X, v)$. Let $\mathcal{T} = \{j : f_j(T, X, v) = Tf_j(X, v)\}$ and $G = P[R = r|T, X, v]|_{T=X\beta}$.

Then an approximation for $E[T|X, v, R = r]$ is as follows:

$$\begin{aligned}
E[T|X, v, R = r] &= \int \frac{T \exp\left\{-\frac{(T-X\beta)^2}{2(\tau^2+v)} + \sum_j \psi_j f_j(T, X, v)\right\}}{g(X, v)\sqrt{2\pi(\tau^2 + v)} \log\left(A + e^{\sum_j \psi_j f_j(T,X,v)}\right)} dT \\
&= \int \frac{T \exp\left\{-\frac{(T-X\beta)^2}{2(\tau^2+v)} + \sum_j \psi_j f_j(T, X, v) - \log\left(A + e^{\sum_j \psi_j f_j(X\beta,X,v)}\right)\right.}{g(X, v)\sqrt{2\pi(\tau^2 + v)}} \\
&\qquad \left. \frac{-G(\sum_{j\in\mathcal{T}} \psi_j f_j(X, v))(T - X\beta) + O(T^2)\right\}}{g(X, v)\sqrt{2\pi(\tau^2 + v)}} dT \\
&\approx \int \frac{T \exp\left\{-\frac{1}{2(\tau^2+v)}\left(T^2 - 2TX\beta - 2T(\tau^2 + v)\sum_{j\in\mathcal{T}} \psi_j f_j(X, v)\right.\right.}{g(X, v)\sqrt{2\pi(\tau^2 + v)}} \\
&\qquad \left.\left. \frac{+2T(\tau^2 + v)G(\sum_{j\in\mathcal{T}} \psi_j f_j(X, v)) + \ldots\right)\right\}}{g(X, v)\sqrt{2\pi(\tau^2 + v)}} dT \\
&= X\beta + (1 - G)(\tau^2 + v)\sum_{j\in\mathcal{T}} \psi_j f_j(X, v)
\end{aligned}$$

Note that this uses a first order Taylor expansion of the denominator in the log-linear model.

Given this result, we can express the expected weighted least squares estimators to be:

$$\begin{aligned}
\hat{\beta} &= (X^T W X)^{-1} X^T W E[T|X, v, R = r] \\
&= (X^T W X)^{-1} X^T W \left(X\beta + \left[(1 - G_i)(\tau^2 + v_i)\sum_{j\in\mathcal{T}} \psi_j f_j(X_i, v_i)\right]\right) \\
&= \beta + (X^T W X)^{-1} X^T W \left[(1 - G_i)(\tau^2 + v_i)\sum_{j\in\mathcal{T}} \psi_j f_j(X_i, v_i)\right]
\end{aligned}$$

Thus, the bias of the regression coefficients is given by

$$(X^T W X)^{-1} X^T W \left[(1 - G_i)(\tau^2 + v_i)\sum_{j\in\mathcal{T}} \psi_j f_j(X_i, v_i)\right]$$

- Refine for matrix notation.
- 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.

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. While a more general result is provided in the previous section, some of that expression will be difficult to intuit. Thus, here we present some examples of models where a complete-case analysis that ignores the missingness mechanism can induce bias.

Example: Suppose there is only one covariate $X \in \mathbb{R}$, and that $p(R = 1|T, X, v) \propto \frac{\exp\{\psi_0+\psi_1 X+\psi_2 T+\psi_3 TX\}}{1+exp\{\psi_0+\psi_1 X+\psi_2 T+\psi_3 TX\}}$

Then

$$
\begin{aligned}
p(R=1|X,v) &= \int p(T|X,v,R=1)f(T)dT \\
&= \int \frac{\exp\left\{-\frac{(T-\beta_0-\beta_1 X)^2}{2(v+\tau^2)} + \psi_0 + \psi_1 X + \psi_2 T + \psi_3 XT\right\}}{\sqrt{2\pi(v+\tau^2)}(1+\exp\{\psi_0+\psi_1 X+\psi_2 T+\psi_3 XT\})}dT \\
&= \int \frac{\exp\left\{-\frac{(T-\beta_0-\beta_1 X)^2}{2(v+\tau^2)} + \psi_0 + \psi_1 X + \psi_2 T + \psi_3 XT - \log(1+\exp\{\psi_0+\psi_1 X+\psi_2 T+\psi_3 XT\})\right\}}{\sqrt{2\pi(v+\tau^2)}}dT \\
G(X) = G &= \frac{e^{\psi_0+\psi_1 X+\psi_2 B+\psi_3 XB}}{1+e^{\psi_0+\psi_1 X+\psi_2 B+\psi_3 XB}} \\
B &= \beta_0+\beta_1 X \\
\therefore &= \int \frac{\exp\left\{-\frac{(T-\beta_0-\beta_1 X)^2}{2(v+\tau^2)} + \psi_0 + \psi_1 X + \psi_2 T + \psi_3 XT\right\}}{\sqrt{2\pi(v+\tau^2)}}dT \\
&\quad + \int \frac{\exp\left\{-\log(1+e^{\psi_0+\psi_1 X+\psi_2 B+\psi_3 XB}) - G\psi_1 X - G\psi_2 B - G\psi_3 XB + O(T^2)\right\}}{\sqrt{2\pi(v+\tau^2)}}dT \\
&\approx \int \frac{\exp\left\{-\frac{T^2-2T\left[B+(v+\tau^2)\psi_2+(v+\tau^2)\psi_3 X-(v+\tau^2)G\psi_2-(v+\tau^2)G\psi_3 X\right]}{2(v+\tau^2)}\right\}}{\sqrt{2\pi(v+\tau^2)}}dT \\
&= g(X,v,\psi,\eta)
\end{aligned}
$$

Therefore,

$$E[T|X,v,R=1] = \beta_0+\beta_1 X+(1-G)(v+\tau^2)(\gamma_2+\gamma_3 X)$$

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 $i = 1, \ldots m_0$ and $X_i = 1$ (and $R_i = 1$) for $m_0+1, \ldots, M$. 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.