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 $X_i = [1, X_{i1}, \ldots, X_{i,p-1}]^T$. Then we can write the meta-regression model 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:

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

Note that the complete data model appears in the numerator of the right hand size of the equation. 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

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

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.

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).

- 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.

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 $R \perp (T, X, v)$. Most effect size indices, variances v are functions of the sample sizes within studies n. For some effect sizes, such as the z-transformed correlation coefficient, v depends 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.

- 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 how the distribution of R depends on T and X.

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

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

Therefore,

$$E[T|X,v,R=1] = \int T \frac{\exp\left\{-\frac{(T-\beta_0-\beta_1 X)^2}{2(v+\tau^2)} + \psi_0 + \psi_1 T\right\}}{\sqrt{2\pi(v+\tau^2)}g(X,v,\psi,\eta)(1+\exp\{\psi_0+\psi_1 T\}} dT$$

$$= \int T \frac{\exp\left\{-\frac{\left[T-\left(\beta_0+\beta_1 X-\psi_1(v+\tau^2)\left(1-\frac{e^{\psi_0}}{1+e^{\psi_0}}\right)\right)\right]^2}{2(v+\tau^2)}\right\}}{\sqrt{2\pi(v+\tau^2)}} dT$$

$$= \beta_0 + \beta_1 X - \psi_1(v+\tau^2)\left(1-\frac{e^{\psi_0}}{1+e^{\psi_0}}\right)$$

Suppose the first m effects have $X_i = 0$ with $R_i = 1$, then the ML and least squares estimator for β_0 is given by

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

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

$$\hat{\beta}_0 = \beta_0 - \frac{\psi_1\left(1-\frac{e^{\psi_0}}{1+e^{\psi_0}}\right)}{\sum_{i=1}^m m/(v_i+\tau^2)}$$

Note that the second term on the right hand side constitutes the bias of $\hat{\beta}_0$. The bias is negative when $\gamma_1 > 0$ and hence when larger effect estimates are more likely to be missing covariates, and it is positive when $\gamma_0 < 0$, which occurs when larger effect estimates are less likely to be missing covariates.

The ML estimator of β_1 is given by