Simulation Options for OOS Estimation

Carly Lupton Brantner

April 2023

1 Introduction

- Idea: Study-specific CATE estimates are randomly distributed around some true CATE value.
- We want to disentangle the study effect / marginalize out study and get an interval for the CATE independent of study membership
- Start with random effects meta-analysis
  - A typical formula for IPD MA could be $E(Y) = \alpha_s + \delta_s A + \beta_s^T X$. Then a prediction interval can be calculated for δ under the assumption that $\delta_s \sim N(\delta, \sigma_\delta^2)$.

> After a random effects meta-analysis, a prediction interval can be calculated to give a range for the predicted parameter value in a new study. Assuming the random effects (that is, the individual study parameter values) are normally distributed with between-study standard deviation (τ), then the prediction interval is approximately[6]
>
> $$\hat{\mu} - t_{k-2}\sqrt{\hat{\tau}^2 + \mathrm{SE}(\hat{\mu})^2}\ , \quad \hat{\mu} + t_{k-2}\sqrt{\hat{\tau}^2 + \mathrm{SE}(\hat{\mu})^2}$$
>
> where $\hat{\mu}$ is the estimate of the average parameter value across studies; $\mathrm{SE}(\hat{\mu})$ is the standard error of $\hat{\mu}$; $\hat{\tau}$ is the estimate of between study standard deviation; t_{k-2} is the 100(1−α/2) percentile of the t distribution with k−2 degrees of freedom, where k is the number of studies in the meta-analysis and α is usually chosen as 0.05, to give a 5% significance level and thus 95% prediction interval. A t distribution, rather than a normal distribution, is used to help account for the uncertainty of $\hat{\tau}$. The correct number of degrees of freedom for this t distribution is complex, and we use a value of k−2 largely for pragmatic reasons.

  - What if we add in treatment effect heterogeneity, so the model looks more like $E(Y) = \alpha_s + \delta_s A + \beta_s^T X + \theta_s^T AZ$? Then the CATE is $\theta(Z) = \delta_s + \theta_s^T Z$.
  - Let's create a prediction interval for the CATE
  - Compare the prediction interval with some bootstrapping approaches and imputation options
- Can we then carry these ideas to nonparametric approaches?
  - Would likely have to just do bootstrapping or use the variance somehow; can't assume distribution of parameters

2 Setup

- Assume we fit a model with random effects for the treatment ($c_s \sim N(0, \sigma_c^2)$) and for the treatment-covariate interaction ($d_s \sim N(0, \sigma_d^2)$) by study that are allowed to be correlated.

$$y_{si} = (\beta_0 + a_s) + (\beta_1 + b_s)x_{si} + (\beta_2 + c_s)w_{si} + (\beta_3 + d_s)x_{si}w_{si} + \epsilon_{si}$$

- Then we can define the CATE as

$$\theta(x_{si}) = (\beta_2 + c_s) + (\beta_3 + d_s)x_{si}$$

- Let's assume x is one covariate for now and s is some fixed value. Let's also assume that the random effects are iid normal and that the random effects are independent from the fixed effects. Then:

$$\begin{aligned} Var(\hat{\theta}(x_{si})) &= Var((\hat{\beta}_2 + c_s) + (\hat{\beta}_3 + d_s)x_{si}) \\ &= Var(\hat{\beta}_2) + Var(\hat{\beta}_3)x_{si}^2 + 2x_{si}Cov(\hat{\beta}_2, \hat{\beta}_3) + \\ &\quad \sigma_c^2 + x_{si}^2\sigma_d^2 + 2x_{si}Cov(c_s, d_s) \end{aligned}$$

- We can also define in matrix notation:

$$Y = X\beta + Zu + \epsilon$$

where

$$X = \begin{bmatrix} 1 & x_{11} & w_{11} & x_{11}w_{11} \\ . & . & . & . \\ 1 & x_{kn} & w_{kn} & x_{kn}w_{kn} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

$$Z_s = \begin{bmatrix} 1 & x_{11} & w_{11} & x_{11}w_{11} \\ . & . & . & . \\ 1 & x_{sn} & w_{sn} & x_{sn}w_{sn} \end{bmatrix} \quad Z = diag\{Z_s\}$$

$$u_s = \begin{bmatrix} a_s \\ b_s \\ c_s \\ d_s \end{bmatrix} \quad u = \begin{bmatrix} u_1 \\ ... \\ u_k \end{bmatrix}$$

- Then the CATE is

$$\theta = \tilde{X}\tilde{\beta} + \tilde{Z}\tilde{u} = \begin{bmatrix} \tilde{X}_1 \\ ... \\ \tilde{X}_k \end{bmatrix} \tilde{\beta} + \begin{bmatrix} \tilde{Z}_1 & ... & 0 \\ 0 & ... & \tilde{Z}_k \end{bmatrix} \begin{bmatrix} \tilde{u}_1 \\ ... \\ \tilde{u}_k \end{bmatrix}$$

  where

$$\tilde{X}_s = \tilde{Z}_s = \begin{bmatrix} 1 & x_{s1} \\ . & . \\ 1 & x_{sn} \end{bmatrix} \quad \tilde{\beta} = \begin{bmatrix} \beta_2 \\ \beta_3 \end{bmatrix} \quad \tilde{u}_s = \begin{bmatrix} c_s \\ d_s \end{bmatrix}$$

- Then we can calculate the variance of the CATE *estimate* as

$$\begin{aligned} Var(\hat{\theta}(\tilde{X})) &= Var(\tilde{X}\hat{\beta} + \tilde{Z}u) \\ &= \tilde{X}Var(\hat{\beta})\tilde{X}^T + \tilde{Z}Var(u)\tilde{Z}^T \end{aligned}$$

  Note that $Var(\hat{\beta})$ includes the square of the standard errors of the coefficient estimates. We can take the diagonals of this matrix to get the relevant variances for each $\hat{\theta}_{si}$. Then the prediction interval can be written as

$$\hat{\theta} \pm t_{K-2}\sqrt{Var(\hat{\theta})}$$

- If we have multiple moderators $x_1, ..., x_p$, then we can instead write

$$\tilde{X}_s = \tilde{Z}_s = \begin{bmatrix} 1 & x_{1s1} & ... & x_{ps1} \\ . & . & & \\ 1 & x_{1sn} & ... & x_{psn} \end{bmatrix} \quad \tilde{\beta} = \begin{bmatrix} \beta_2 \\ \beta_3 \\ ... \\ \beta_{p+2} \end{bmatrix} \quad \tilde{u}_s = \begin{bmatrix} c_s \\ d_s \\ ... \\ p_s \end{bmatrix}$$

- We also do not have to require for $\tilde{X}_s = \tilde{Z}_s$ and instead can have some moderators with only a fixed interaction effect. Then we would just write the two matrices out separately.

- Note: Our approach for estimating the CATE will be different in the training versus target sample.
  - **Matrix formulation:** $\tilde{X}$ will be the same in both samples, but $\tilde{Z}$ will be different. In the training sample, $\tilde{Z}$ will be a block diagonal matrix with $\tilde{Z}_s$ in each block. In the target sample, $\tilde{Z}$ will equal $\tilde{X}$.
  - **Mean CATE:** In the training sample, we will use the study-specific random effects such that: $\theta = \tilde{X}\tilde{\beta} + \tilde{Z}\tilde{u}$. In the target sample, we do not know S, so we will have to assume the random effects are 0 and instead use: $\theta = \tilde{X}\tilde{\beta}$.
  - **Variance of CATE:** In the training sample, we will use our block diagonal $\tilde{Z}$ and the variance matrix of $\hat{u}$ since we know each study-specific estimate of the random effects. In the target sample, we will use our different $\tilde{Z}$ where we just have all of the data in a long matrix, and we will use the variance matrix of u since we do not know the study that each individual belongs to.

3 Methods to Compare

- Fit a **correct** meta-analysis, then estimate a prediction interval by manually calculating the variance of the estimated CATE
  - Use vcov() to get the fixed effect covariance matrix and VarCorr() to get the random effect covariance matrix; then calculate variance as in above section
- Fit an **incorrect** meta-analysis, then estimate a prediction interval by manually calculating the variance of the estimated CATE
- Fit a causal forest (with pooling with trial indicator), then estimate a confidence interval using each of three bootstrap approaches
  - Randomly assign study for each target individual 1,000 times, predict CATE, derive 95% CI
  - Randomly assign study for each target individual 1,000 times based on a study membership model, predict CATE, derive 95% CI
  - Use default causal forest method by assigning all individuals one side of a split, estimate variance and get 95% CI

  ****Currently unused:**

- Estimate a regular confidence interval for the CATE using glht()

  - Just uses the covariance matrix of the fixed effects to get a 95% CI of the mean CATE

- Estimate a prediction interval through a bootstrap procedure
  - For each covariate value in the dataset, we will iterate 1000 times.
  - For each iteration, we will randomly draw the fixed coefficients from their mean and covariance matrix, and then the random effects as well
  - We will then calculate the linear combination of these with the covariate to get a CATE estimate
  - We can then get an interval based on the distribution of predictions across many iterations (mean +/- 1.96*sd)

- Use pimeta package in R to try different types of PIs
  - I believe this package only works for two-stage meta-analysis and is not great for dealing with linear combinations

4 Steps of Simulation

1. Simulate training data using gen-mdd() function using one of four options
   (a) Same covariate distribution
   (b) Madrs varies
   (c) Madrs and age vary
   (d) Age nearly distinguishable across studies
2. Simulate test data using one of three options
   (a) Random sample of 100 individuals from the training data but with missing study
   (b) Different sample of 100 individuals who are younger and have less severe depression according to their madrs score
3. Simulate m and τ function to get Y for training and target data
   (a) We are allowing for between-study heterogeneity for the intercept, treatment, and treatment-moderator interaction
   (b) Options (see table):
      i. Age is only moderator, has fixed and random interactions with treatment

      $$E_1(Y_s) = (\beta_0 + a_s) + \beta_1 * Age + \beta_2 * MADRS + \beta_3 * Sex + (\beta_4 + b_s) * Trt + (\beta_5 + c_s) * Trt * Age$$

      where $a_s \sim N(0, \sigma_a^2)$, $b_s \sim N(0, \sigma_b^2)$, $c_s \sim N(0, \sigma_c^2)$
      ii. Age and MARDS are both moderators, have fixed and random interactions with treatment

      $$E_2(Y_s) = E_1(Y_s) + (\beta_6 + d_s) * Trt * MADRS$$

      iii. Age^2 and Sex are both moderators, but Sex only has a fixed interaction while age has fixed and random

      $$E_4(Y_s) = E_1(Y_s) + (\beta_6) * Trt * Sex$$

      but replace Age with Age^2.
      iv. Non-linear relationship:

      $$E_5(Y_s) = (\beta_0 + a_s) + \beta_1 * Age + (\zeta_0 + b_s) * Trt * \frac{2}{1 + \exp[(\zeta_1 + c_s) * Age]}$$

   (c) Correct meta-analysis parametrization: Including random effects for all relevant moderators
   (d) Incorrect meta-analysis parametrization: Removing study-level heterogeneity (random effects) for the interaction term(s)
4. Fit IPD mixed effects meta-analysis models - correct and incorrect - and pull relevant coefficient estimates and covariance matrices
5. Predict 95% interval for CATE of all training and target individuals
6. Calculate measures of accuracy in training and target data
   (a) MSE
   (b) Confidence interval coverage
   (c) Confidence interval length
   (d) Percent statistically significant

Fixed moderators	**Random moderators**	**L/NL**	σ_τ	σ_{inter}
Age	Age	Linear	0.05	0.05
Age	Age	Linear	0.5	0.05
Age	Age	Linear	1	0.5
Age	Age	Linear	1	1
Age, MADRS	Age, MADRS	Linear	0.05	0.05, 0.05
Age, MADRS	Age, MADRS	Linear	0.5	0.05, 0.05
Age, MADRS	Age, MADRS	Linear	0.5	0.5, 0.05
Age^2, Sex	Age^2	Linear	0.5	0.05
Age^2, Sex	Age^2	Linear	0.5	0.5
Age	Age	Non-linear	0.05	0.05
Age	Age	Non-linear	0.5	0.05
Age	Age	Non-linear	0.5	0.5

Table 1: Options for moderators and study-level variation

5 Simulation Setup

Training data setup (based on MDD data):

- 10 studies with 500 people in each
- Probability of treatment is 1/2
- 5 covariates (sex, smoking, weight, age, baseline MADRS) per person defined by multivariate normal distribution within each study; continuous variables are standardized to have mean 0 and variance 1
- Test data is one of:
  - Random sample of 100 individuals from the training data but with missing study
  - Different sample of 100 individuals who are younger and have less severe depression according to their madrs score
- Randomly sample study-level variation terms for the intercept (ϵ_m), treatment (ϵ_τ), and treatment-covariate interaction terms (i.e., ϵ_{age})
- m and τ depending on scenario:
  - Age only moderator

$$m = (-17.40 + \epsilon_m) - 0.13 * age - 2.05 * madrs - 0.11 * sex$$

$$\tau = (2.505 + \epsilon_\tau) + (0.82 + \epsilon_{age}) * age$$

  - Age and MADRS both moderators
  - Age and Sex both moderators
  - Age and Sex both moderators, Sex only has fixed effect
- Define $Y = m + W * \tau + \epsilon$ where $\epsilon \sim N(0, 0.05^2)$