Cluster RCT Estimators Technical Document

Miratrix

2022-11-16

Introduction

This gives technical details for the identified estimators for estimating average effects for Blocked, Cluster-Randomized RCTs.

Once an estimand is identified, a researcher must decide how to estimate it, given the observed data, and obtain uncertainty estimates (standard errors) for those estimates. We call the ATE estimators "effect estimators" and the uncertainty estimators "standard error estimators." In this section, we systematically go through three general classes of effect estimators, discussing variants of implementing each. As in our work focused on multisite randomized trials, we divide our estimators into three general classes of linear modeling estimators, multilevel (random effects) estimators, and design-based estimators.

Because blocking is so prevalent for cluster randomized trials, we proceed with this full blocking incorporated in our discussion, even though it increases some of the technical detail and complexity. When there is no blocking, one can simply assume a single block, holding all the clusters, in most cases. We will note departures from this as we go along.

Our notation is to discuss individual students (i) in clusters (e.g., schools) j, nested in blocks (e.g., districts or sites) k.

Before we look at estimators, let's spend a moment thinking about estimands. In particular, for each individual we have $Y_{ijk}(0)$ and $Y_{ijk}(1)$, with an individual treatment effect of $\tau_{ijk} = Y_{ijk}(1) - Y_{ijk}(0)$. We can never see both potential outcomes, and thus will never know τ_{ijk} for any individual, regardless of how we assign treatment.

Similarly, for each cluster we have the average treatment effect in that cluster as

$$\beta_{jk} = \frac{1}{N_{jk}} \sum_{i=1}^{N_{jk}} Y_{ijk}(1) - Y_{ijk}(0).$$

The β_{jk} are the average treatment effect of the individuals in cluster jk. Similar to the τ_{ijk}, we have no ability to estimate a cluster's β_{jk} because we treat the entire cluster, or not. This is very different from multisite randomized trials, where we get to see some of the individuals in a site treated and some not treated, allowing for (a perhaps noisy) estimate of the site's average treatment impact.

Because of this, a cluster-randomized experiment can be thought of, in a sense, as an individually randomized experiment where the individuals are the clusters themselves. In fact, this view is prevalent in much of the design based causal inference literature. It is also the intuition why we might think of our very large (in terms of N) experiment as quite small (in terms of J, the total number of clusters).

Within a district k we have different possible estimands of interest. Similar to Miratrix, Weiss and Henderson (CITE), we can have the simple average of the β_{jk} (the cluster-average effect) of

$$\beta_k = \sum_j \beta_{jk} / J_k,$$

or the person-weighted average of the β_{jk} (the person-average effect) of

$$\beta_k = \frac{1}{N_k} \sum_j N_{jk} \beta_{jk}.$$

The latter is the average effect across all individuals in site k. The former is the average effect of the clusters, regardless of their size, in site k.

We can then weight these β_k across the districts in different ways: a simple average, giving the district-average of the district-level estimands, regardless of district size, a cluster-weighted average, weighing each district by the number of clusters in it, or a person-weighted average, weighing each district by the number of people in it.

We argue it is not typically sensible to, for example, calculate a person-weighted average of cluster-average impacts. The weighting at level 3 should be aligned with the weighting at level 2. Furthermore, we generally think it is a rare instance where a simple average across the districts is a desired quantity. That all being said, there are many weighting options one might select.

Acknowledgements

Parts of this document, in particular some notation and mathematical formulation, are taken from the technical supplement to the Hunter et al. paper on PUMP.

Linear Regression Estimators

One of the most common estimation strategies for cluster RCTs in the social sciences is linear regression. For a two-level context (no districts), a natural choice is to fit a linear regression of the outcome onto the treatment indicator (and some individual or cluster-level control variables) and then use cluster robust standard errors (CRVE).

Here the individual nature of the data will naturally target a person-average estimand, as larger clusters have larger weight. We cannot have fixed effects at the cluster level, as these would be completely confounded with treatment assignment, but we can have them at the district level.

We give some notation to connect to other regression estimators discussed below. Define K site dummies $S_{q,ijk}$, $q = 1, \dots, K$ with $S_{q,ijk} = 1$ if individual ijk is in site k; in other words, $S_{q,ijk} = 1\{q = k\}$.

The simplest OLS working model for estimating an impact would then arguably be

$$Y_{ijk} = \sum_{q=1}^{K} \alpha_q S_{q,ijk} + \beta T_{ij} + \epsilon_{ij},$$

where the α_q are the K fixed effects for the districts. This model has $K + 1$ coefficients in total. The covariance of the errors would then be a block-diagonal matrix, one block per cluster (not district). We exclude an overall intercept to have individual fixed effects for each district and no reference district.

One advantage of the fixed effects model compared to design-based is the single impact parameter reduces the degrees of freedom used by the model, which could improve asymptotic inference in smaller studies. Given how few clusters exist in many education experiments, this may indeed be worthwhile.

For simplicity, we can sometimes leave the dummy variables implicit, simply writing

$$Y_{ijk} = \alpha_k + \beta T_{ijk} + \epsilon_{ijk}.$$

OLS interacted by district

We can extend the above model to estimate impacts within each district as so:

$$\begin{aligned} Y_{ijk} &= \sum_{q=1}^{K} \alpha_q S_{q,ijk} + \sum_{q=1}^{K} \beta_k S_{q,ijk} T_{ij} + \epsilon_{ij} \\ &= \alpha_k + \beta_k T_{ijk} + \epsilon_{ijk}. \end{aligned}$$

Now we have treatment-by-district interaction terms. This model will estimate some sort of average within each district, giving K estimates, $\hat{\beta}_k$, $k = 1, \dots, K$, in total.

To get an overall estimate we would then average these. This opens the door to needing to make explicit choices about weighting. For example if we are wanting to weight each cluster equally, we would weight each district by the number of clusters it has, as follows:

$$\hat{\beta} = \sum_{k=1}^{K} \frac{J_k}{J} \hat{\beta}_k, \tag{1}$$

where J_k are the number of clusters in district k.

If we wanted to tend towards the individual-average effect we would weight differently. E.g., we might weight as

$$\hat{\beta} = \sum_{k=1}^{K} \frac{N_k}{N} \hat{\beta}_k,$$

where N_k are the number of individuals in district k.

Both of these weighting schemes are somewhat suspect: we are weighting the $\hat{\beta}_k$, but what are the $\hat{\beta}_k$ themselves estimating? If they are estimating the average treatment impact for the individuals in site k, then the cluster-weighted expression (Equation 1) is going to weight each district by the number of clusters, but each district estimate is the person-average treatment effect. The final implied estimand is not easily considered. If we truly want the average cluster impact, we would need to ensure the $\hat{\beta}_k$ are aligned with our final district weighting step.

For our linear regression model, we argue that the $\hat{\beta}_k$ are estimating

Aggregation Estimators

Another regression approach is to first aggregate the data within each cluster to obtain the average cluster outcome ($\bar{Y}_{jk}$), defined in terms of the individual potential outcomes as

$$\bar{Y}_{jk} = \frac{1}{N_{jk}} \sum_{i=1}^{N_{jk}} Y_{ij} = Z_{jk} \bar{Y}_{jk}(1) + (1 - Z_{jk}) \bar{Y}_{jk}(0)$$

Here Z_{jk} is an 0/1 treatment indicator for cluster j, and we observe either the average of the cluster's treatment or control potential outcomes. We then run a subsequent regression of the aggregated outcomes onto their treatment indicators. This approach collapses the individual-level data to get measures at the level of randomization. We can then weight these averages differently for different estimands. If cluster sizes differ, we will also potentially have to account for heteroskedasticity, where some estimated cluster averages (the average of N_{jk} individuals) will be more uncertain (due to smaller samples of individuals) than others. In this case, we can use heteroskedastic robust standard errors to account for this. See Schochet (2020) for more.

When we have a blocking variable, there are a few options. First, one can simply include district-level fixed effects as above. Alternatively, one can include district by treatment interactions, allowing different districts to have different average ATEs, and then these district ATEs can be aggregated to estimate the overall ATE. These can give different weighted averages of impacts across units. As part of this work, we will identify which specific estimands are being targeted by all these different choices.

Hierarchical linear models (aka random effects models)

Multilevel modeling explicitly captures the hierarchical nature of the data and allows for estimation of the degree of variation in outcomes at each level (and variation in impacts in the three-level case) as well as the overall ATE. Multilevel modeling is also probably the most common tool used to analyze data from CRT experiments in education contexts. As a hint as to typical practice, we discuss the models identified in the popular PowerUp! power calculator (Dong & Maynard, 2013). This tool explicitly notes different couplings of design and modeling choices, and provides specific uncertainty formulas for several contexts.

For the two-level case, PowerUp! suggests a random intercept model, with a single coefficient for average treatment effect. This assumes a constant treatment effect for all schools, and we can also include school covariates for the intercept.

The model for estimating impacts on outcome m is given by:

$$\begin{aligned} Y_{ijk} &= \theta_{0,jk} + \gamma_{mp}^{\top} C_{ijkp} + r_{ijk} \\ \theta_{0,jk} &= \mu_{0,k} + \psi_1 T_{jk} + \sum_{r=1}^{g_2} \delta_{mr} X_{jkr} + u_{0,jk} \end{aligned}$$

with reduced form:

$$Y_{ijk} = \psi_1 T_{jk} + \mu_{0,k} + \sum_{r=1}^{g_2} \delta_{mr} X_{jkr} + \sum_{p=1}^{g_1} \gamma_{mp} C_{ijkp} + u_{0,jk} + r_{ijk}$$

and distributions:

$$\begin{aligned} u_{0,jk} &\sim N(0, \tau_0^2) \\ r_{ijk} &\sim N(0, \sigma_m^2)\,. \end{aligned}$$

The standard error of the treatment effect estimate is given by:

$$Q_m = \sqrt{\frac{\mathrm{ICC}_2(1 - R_2^2)}{\bar{T}(1 - \bar{T})J} + \frac{(1 - \mathrm{ICC}_2)(1 - R_1^2)}{\bar{T}(1 - \bar{T})J\bar{n}}}.$$

The degrees of freedom are

$$\mathrm{df}_m = J - g_1 - 2.$$

In R, fitting this multilevel model would be have a `lmer()` formula along the lines of:

```
Yobs ~ 1 + T.x + X.jk + C.ijk + (1 | S.id)
```

The constant effects model means that we assume no treatment variation across our sites, i.e., that $\omega_2 = 0$. The resulting impact estimate when there is impact heterogeneity across clusters is likely to be a form of precision weighting.

The modeling options multiply once we allow for three levels, as we then can potentially model cross-site means and impact variation using either fixed or random effects at level 3. PowerUp! identifies two estimators for this context, which they call fixed and random effects. The random effect model allows the treatment coefficient to vary across sites (implying a superpopulation of sites). We discuss these two in the following

Fixed blocks, random effects for clusters

In this model we have fixed intercepts for districts, fixed treatment effects for districts, random intercepts for schools, constant effects for schools within a district, school covariates for intercept.

The model for estimating impacts on outcome m is given by:

$$\begin{aligned}
Y_{ijk} &= \theta_{0,jk} + \sum_{p=1}^{g_1} \gamma_{mp} C_{ijkmp} + r_{ijk} \\
\theta_{0,jk} &= \mu_{0,k} + \psi_{1,jk} T_{jk} + \sum_{r=1}^{g_2} \delta_{mr} X_{jkmr} + u_{0,jk} \\
\mu_{0,k} &= \Xi_0 + w_{0,k} \\
\psi_{1,k} &= \Xi_1 + w_{1,k}
\end{aligned}$$

with reduced form:

$$Y_{ijk} = \left(\Xi_1 + w_{1,k}\right) T_{jk} + \Xi_0 + \sum_{r=1}^{g_2} \delta_{mr} X_{jkmr} + \sum_{p=1}^{g_1} \gamma_{mp} C_{ijkmp} + w_{0,k} + u_{0,jk} + r_{ijk}$$

and distributions:

$$\begin{aligned}
u_{0,jk} &\sim N\left(0, \tau_0^2\right) \\
r_{ijk} &\sim N\left(0, \sigma_m^2\right).
\end{aligned}$$

The standard error of the treatment effect estimate is given by:

$$Q_m = \sqrt{\frac{\mathrm{ICC}_2(1-R_2^2)}{\bar{T}(1-\bar{T})JK} + \frac{(1-\mathrm{ICC}_2-\mathrm{ICC}_3)(1-R_1^2)}{\bar{T}(1-\bar{T})JK\bar{n}}}.$$

The degrees of freedom are

$$\mathrm{df}_m = K(J-2) - g_2.$$

This model assumes no variation of impacts within schools, and no variation at the district level, i.e., that $\omega_2 = 0$.

The R model is

```
Yobs ~ 0 + T.x * D.id - T.x + X.jk + C.ijk + (1 | S.id)
```

The overall treatment effect is then the average of the T.x interaction terms.

Full random effects model

In this model we have random intercepts and impact effects for districts, random intercepts for schools, constant effects for schools within a district, school and district covariates for intercept. Powerup also allows for district covariates for treatment effects.

The model for estimating impacts on outcome m is given by:

$$\begin{aligned}
Y_{ijk} &= \theta_{0,jk} + \sum_{p=1}^{g_1} \gamma_{mp} C_{ijkmp} + r_{ijk} \\
\theta_{0,jk} &= \mu_{0,k} + \sum_{r=1}^{g_2} \delta_{mr} X_{jkmr} + u_{0,jk} \\
\mu_{0,k} &= \Xi_0 + \psi_{1,k} T_k + \sum_{s=1}^{g_3} \xi_{ms} V_{kms} + w_{0,k} \\
\psi_{1,jk} &= \Xi_1 + w_{1,k}
\end{aligned}$$

with reduced form:

$$\begin{aligned} Y_{ijk} &= \left(\Xi_1 + w_{1,k}\right) T_{jk} + \Xi_0 \\ &+ \sum_{s=1}^{g_3} \xi_{ms} V_{kms} + \sum_{r=1}^{g_2} \delta_{mr} X_{jkmr} + \sum_{p=1}^{g_1} \gamma_{mp} C_{ijkmp} \\ &+ w_{0,k} + u_{0,jk} + r_{ijk} \end{aligned}$$

and distributions:

$$\begin{aligned} u_{0,jk} &\sim N\left(0, \tau_0^2\right) \\ \begin{pmatrix} w_{0,k} \\ w_{1,k} \end{pmatrix} &\sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \eta_0^2 & \kappa_m^w \eta_0 \eta_1 \\ \kappa_m^w \eta_1 \eta_0 & \eta_1^2 \end{pmatrix} \right) \\ r_{ijk} &\sim N\left(0, \sigma_m^2\right). \end{aligned}$$

Similar to other blocked models model, in PowerUp! they further assume that district covariates also influence the treatment impact

$$\mu_{1,jk} = \xi_1 + \sum_{s=1}^{g_3} \zeta_{mr} V_{kms} w_{1,k}$$

but we do not make this assumption.

The standard error of the treatment effect estimate is given by:

$$Q_m = \sqrt{\frac{\text{ICC}_3 \omega_3}{K} + \frac{\text{ICC}_2(1 - R_2^2)}{\bar{T}(1 - \bar{T})JK} + \frac{(1 - \text{ICC}_2 - \text{ICC}_3)(1 - R_1^2)}{\bar{T}(1 - \bar{T})JK\bar{n}}}.$$

The degrees of freedom are often taken as $\text{df}_m = K - 1$. We usually further assume $\omega_2 = 0$.

The R model formula for the above is

```
Yobs ~ 1 + T.x + V.k + X.jk + C.ijk + (1 | S.id) + (1 + T.x | D.id)
```

Discussion

The fixed effect model puts no distribution on the u, but still has the random effect at level 2 for the intercept of each cluster within the sites (implying a finite sample of sites). There are other modeling options within this framing, such as an extension of the FIRC model (Bloom et al., 2017) with only a distribution on u_1k.

Design-based estimators

Design-based estimators, although not as common as linear regression or multi-level modeling, are usually directly and explicitly tied to an estimand of interest (Schochet, 2015). Design-based estimators focus on the random assignment mechanisms and the sampling mechanism as the sources of uncertainty, rather than specifying linear models with random residual components. They are generally considered to rely on weaker assumptions than other approaches; e.g., homoskedasticity assumptions, or assumptions of a random normal residual, are not needed.

Design-based estimators work with the cluster averages (sometimes adjusted with covariates). We first focus on a single site k. The simplest cluster-average estimator for site k would be

$$\hat{\beta}_{k-cluster} = \frac{1}{J_{k1}} \sum_{j=1}^{J_k} Z_{jk} \hat{Y}_{jk} - \frac{1}{J_{k0}} \sum_{j=1}^{J_k} (1 - Z_{jk}) \hat{Y}_{jk},$$

where $J_{k1} and J_{k0} are the number of clusters assigned to treatment or control status, respectively. The $\hat{Y}_{jk}$ are the observed mean outcomes of the clusters. This is a version of the aggregation approach, above. This estimator targets the average of the cluster average effects, in site k.

For district k, the person-average estimator would be

$$\hat{\beta}_{k-person} = \frac{1}{N_{k1}} \sum_{j=1}^{J_k} Z_{jk} N_{jk} \hat{Y}_{jk} - \frac{1}{N_{k0}} \sum_{j=1}^{J_k} (1 - Z_{jk}) N_{jk} \hat{Y}_{jk}$$

where $N_{k1} = \sum Z_{jk} N_{jk}$ are the number of individuals assigned to treatment, and N_{k0} is defined similarly. A site-size-weighted average of these gives the effect for the average person: For the site-average of cluster averages, we weight differently again:

$$\beta = \sum_{k=1}^{K} \frac{1}{K} \beta_{k-person}$$

There are even more weightings options possible, such as the district-average of the cluster-average impact.

Uncertainty estimation depends on the population model used. For example, we could focus entirely on the randomization that was conducted, holding individuals, clusters, and sites as fixed (i.e., finite). This fully finite-sample view is the most well-known for design-based inference. Here we use design-based uncertainty estimators within each site (with weights to align with the ATE estimator), and then aggregate based on site size, again using weights depending on our target estimand. For this general context, see Middleton and Aronow (2015) or Schochet (2015).

Alternatively, we could model the clusters in each site as sampled from infinite populations at each site. Students could then be finite, or not, within this framing. If we view the clusters as sampled from a fixed set of superpopulations, one for each site, with individuals fixed, we can use Schochet et al. (2021) for uncertainty estimation. Other models are possible; see Pashley and Miratrix (2021a, 2021b) for further discussion in the context of individual randomization with blocking. We will identify the different estimators for uncertainty estimation in the literature and tie them to their specific estimands and sampling frameworks.

Covariate adjustment

Cluster randomized experiments are notoriously difficult to power. To improve precision researchers frequently turn to level 1 and level 2 covariates to explain some of the outcome variation and thus reduce standard errors of an average effect estimator. While we focused on the no-covariate case above for clarity, the covariate-adjusted models parallel that discussion.

Most of the various estimation strategies, in particular the linear regression models, explicitly allow for covariate adjustment. The design-based estimators have covariate-based sister estimators that can be shown to generalize the special no-covariate case. E.g., Shochet et al (2020) and (2021) show how design-based approaches, rooted in randomization, can be expressed as a weighted linear regression, which naturally then allows for covariate adjustment. We use this work that shows equivalence to use weighted regression for our design based estimators.