Module 3(b): Inference for Blocked and Matched Pair Designs

Fall 2021

Matthew Blackwell

Gov 2003 (Harvard)

1/ Block randomized experiments

Block randomized experiments

- Basic idea: run completely randomized experiments within strata defined by covariates.
- Main motivation: **more efficient** than standard design (ie, lower SEs)
- George Box: “Block what you can and randomize what you cannot.”
- We will compare variance of blocked designs to complete randomization.
    - Some confusion in the literature: can blocking hurt?
    - Care needed: comparison depends on sample assumptions (Pashley & Miratrix, 2021, JEBS)

Simple two block example

- GOTV mailer experiment:
  - We have n households with registered voters.
  - Complete randomization: choose n_1 households to get mailers.
  - Outcome, Y_i: turnout in election.
- What if we have data from the voter file: **previous turnout**.
  - Create blocks: $V_i = 1$ if voted in last election, $V_i = 0$ otherwise.
  - n_{v} is the number of previous voters,
  - $n_{\text{nv}} = n - n_{\text{nv}}$ is the number of previous nonvoters.
- SATEs within blocks defined by V_i:

$$\tau_{\text{v, fs}} = \frac{1}{n_{\text{v}}} \sum_{i:V_i=1} \{Y_i(1) - Y_i(0)\} \qquad \tau_{\text{nv, fs}} = \frac{1}{n_{\text{nv}}} \sum_{i:V_i=0} \{Y_i(1) - Y_i(0)\}$$

- Iterated expectations gives us:

$$\tau_{\text{fs}} = \underbrace{\left(\frac{n_{\text{v}}}{n_{\text{v}} + n_{\text{nv}}}\right)}_{\text{fraction voters}} \tau_{\text{v, fs}} + \underbrace{\left(\frac{n_{\text{nv}}}{n_{\text{v}} + n_{\text{nv}}}\right)}_{\text{fraction nonvoters}} \tau_{\text{nv, fs}}$$

Block randomized design

- **Block/stratified randomized experiment**:
    - Completely randomized experiment in each block.
    - Choose $n_{1,\text{v}}$ voters to be treated, $n_{0,\text{v}} = n_{\text{v}} - n_{1,\text{v}}$ control.
    - Choose $n_{1,\text{nv}}$ nonvoters to be treated, $n_{0,\text{nv}} = n_{\text{nv}} - n_{1,\text{nv}}$ control.
- Probability of treatment in each group called the **propensity score**:
    - Prob. of treatment for voters: $\mathbb{P}(D_i = 1 \mid V_i = 1) = p_{\text{v}} = n_{1,\text{v}}/n_{\text{v}}$
    - Prob. of treatment for nonvoters: $\mathbb{P}(D_i = 1 \mid V_i = 0) = p_{\text{nv}} = n_{1,\text{nv}}/n_{\text{nv}}$
- Blocking ensures balance across blocks:
    - When $p_{\text{v}} = p_{\text{nv}}$, distribution of treatment is exactly the same in each block.
    - With complete randomization, treatment might be very imbalanced across V_i.
    - No possibility of "chance" imbalances skewing the estimates.

Estimators in blocked designs

- Within-strata difference in means:

$$\widehat{\tau}_{\text{v}} = \overline{Y}_{1,\text{v}} - \overline{Y}_{0,\text{v}} = \frac{1}{n_{1,\text{v}}} \sum_{i:V_i=1} D_i Y_i - \frac{1}{n_{0,\text{v}}} \sum_{i:V_i=1} (1 - D_i) Y_i$$

$$\widehat{\tau}_{\text{nv}} = \overline{Y}_{1,\text{nv}} - \overline{Y}_{0,\text{nv}} = \frac{1}{n_{1,\text{nv}}} \sum_{i:V_i=0} D_i Y_i - \frac{1}{n_{0,\text{nv}}} \sum_{i:V_i=0} (1 - D_i) Y_i$$

- Unbiased for the within-strata SATEs: $\mathbb{E}[\widehat{\tau}_{\text{v}} \mid \mathbf{O}] = \tau_{\text{v}}$
- $\rightsquigarrow$ unbiased estimator for the overall SATE:

$$\widehat{\tau}_{\text{b}} = \left(\frac{n_{\text{v}}}{n}\right) \widehat{\tau}_{\text{v}} + \left(\frac{n_{\text{nv}}}{n}\right) \widehat{\tau}_{\text{nv}}$$

  - Equivalent to the regular difference in means if $p_{\text{v}} = p_{\text{nv}} = 1/2$.
  - Otherwise, standard $\widehat{\tau}_{\text{diff}}$ under block design will be **biased**.

Sampling variance of blocking estimator

- Each block is a completely randomized experiment so we have:

$$\mathbb{V}(\widehat{\tau}_{\mathrm{v}} \mid \mathbf{O}) = \frac{S^2_{1,\mathrm{v}}}{n_{1,\mathrm{v}}} + \frac{S^2_{0,\mathrm{v}}}{n_{0,\mathrm{v}}} - \frac{S^2_{\tau_i,\mathrm{v}}}{n_{\mathrm{v}}}$$

  - $S^2_{d,\mathrm{v}}$ are the within-block sample variances of the potential outcomes

- Finite sample variance of the blocked estimator:

$$\mathbb{V}(\widehat{\tau}_b \mid \mathbf{O}) = \left(\frac{n_{\mathrm{v}}}{n}\right)^2 \mathbb{V}(\widehat{\tau}_v \mid \mathbf{O}) + \left(\frac{n_{\mathrm{nv}}}{n}\right)^2 \mathbb{V}(\widehat{\tau}_{\mathrm{nv}} \mid \mathbf{O})$$

- Use the conservative variance estimators from each strata:

$$\hat{\mathbb{V}}_b = \left(\frac{n_{\mathrm{v}}}{n}\right)^2 \left(\frac{\widehat{\sigma}^2_{1,\mathrm{v}}}{n_{1,\mathrm{v}}} + \frac{\widehat{\sigma}^2_{0,\mathrm{v}}}{n_{0,\mathrm{v}}}\right) + \left(\frac{n_{\mathrm{nv}}}{n}\right)^2 \left(\frac{\widehat{\sigma}^2_{1,\mathrm{nv}}}{n_{1,\mathrm{nv}}} + \frac{\widehat{\sigma}^2_{0,\mathrm{nv}}}{n_{0,\mathrm{nv}}}\right)$$

  - $\widehat{\sigma}^2_{d,\mathrm{v}}$ are the within-strata **observed outcome variances**

General blocking notation

- Blocks, $j \in \{1, \dots, J\}$.
  - Block indicator $B_i = j$ if i is in block j.
  - Sizes: $n_j > 2$ and proportions $w_j = n_j / n$.
  - Number treated in each block: $n_{1,j}$ and $n_{0,j} = n_j - n_{1,j}$
- Within-block estimators:

$$\widehat{\tau}_j = \frac{1}{n_{1,j}} \sum_{i:B_i=j} D_i Y_i - \frac{1}{n_{0,j}} \sum_{i:B_i=j} (1 - D_i) Y_i, \qquad \hat{\mathbb{V}}(\widehat{\tau}_j) = \frac{\widehat{\sigma}^2_{1,j}}{n_{1,j}} + \frac{\widehat{\sigma}^2_{0,j}}{n_{0,j}}$$

- Aggregate blocking estimators:

$$\widehat{\tau}_b = \sum_{j=1}^{J} w_j \widehat{\tau}_j, \qquad \hat{\mathbb{V}}(\widehat{\tau}_b) = \sum_{j=1}^{J} w_j^2 \hat{\mathbb{V}}(\widehat{\tau}_j)$$

Efficiency of blocking

- Efficiency of block versus CR depends on the sampling scheme.
    - Usually blocking will be more efficient/lower variance, but not always.
- Finite sample difference in sampling variances:

$$\mathbb{V}(\widehat{\tau}_{CR} \mid \mathbf{O}) - \mathbb{V}(\widehat{\tau}_b \mid \mathbf{O}) = \frac{1}{n-1}\left[B - W\right]$$

- Measures of between- and within-block variation:

$$B = \sum_{j=1}^{J} \left(\frac{n_j}{n}\right) \left\{\overline{Y}_j(1) + \overline{Y}_j(0) - (\overline{Y}(1) + \overline{Y}(0))\right\}^2$$

$$W = \sum_{j=1}^{J} \frac{n_j}{n} \frac{n - n_j}{n} \mathbb{V}(\widehat{\tau}_k \mid \mathbf{O})$$

- Difference can be positive or negative (blocking can hurt or help)
    - **Blocking is better when outcomes vary a lot across blocks, not much within blocks** (blocks are predictive of outcome, so usually the case)
    - Blocking always more efficient for PATE under stratified sampling

How to block

- Discrete covariates $\rightsquigarrow$ blocks by unique combinations.
- Alternative: create blocks by creating homogeneous groups in $\mathbf{X}$.
  - Choose distance metric such as Mahalanobis distance:

$$M(\mathbf{X}_i, \mathbf{X}_k) = \sqrt{(\mathbf{X}_i - \mathbf{X}_k)\widehat{\mathbb{V}}(\mathbf{X})^{-1}(\mathbf{X}_i - \mathbf{X}_k)}$$

  - Difficult/impossible to find optimal blocks in general, but "greedy" algorithms exist.
  - Possible to get optimal blocks with **pair matching** ($J = n/2$).

Matched pair design

- Keep blocking for efficiency until each block is size 2.
- **Matched pair design**:
    - Create $J = n/2$ pairs of similar units with outcomes (Y_{1j}, Y_{2j})
    - Random assignment:
        - $W_j = 1$ if first unit is treated
        - $W_j = -1$ if second unit is treated
- Unbiased difference in means estimator:

$$\widehat{\tau}_p = \frac{1}{J} \sum_{j=1}^{J} W_j (Y_{1j} - Y_{2j})$$

- Within-pair variance estimator not feasible (why?)
- Across-pair variance estimator (conservative for SATE):

$$\widehat{\mathbb{V}}(\widehat{\tau}_p) = \frac{1}{J(J-1)} \sum_{j=1}^{J} \{W_j(Y_{1j} - Y_{2j} - \widehat{\tau}_p)\}^2$$