# Section 4

## Linear Regression and Randomized Experiments

Sooahn Shin

GOV 2003

Sept 30, 2021

# Overview

- Logistics:
  - **Pset 4 released!** Due at 11:59 pm (ET) on Oct 6
  - Research project memo: Due at 11:59 pm (ET) on Oct 1
  - OH: Mondays **3-5pm**
- Today's topics:
  1. Linear regression and robust variance estimator
  2. Linear regression with covariates
  3. Block randomized trials
  4. Cluster randomized trials

# Recap: Linear Regression

- Using OLS to estimate ATEs
  - $\widehat{\tau}_{\text{ols}} = \arg\min_\tau \sum_{i=1}^n (Y_i - \alpha - \tau D_i)^2 = \widehat{\tau}_{\text{diff}} \rightsquigarrow$ unbiased
  - Linearity? $\rightsquigarrow$ justified by consistency assumption

$$\begin{aligned} Y_i &= D_i Y_i(1) + (1 - D_i) Y_i(0) \\ &= \mathbb{E}[Y_i(0)] + D_i\tau + \{Y_i(0) - \mathbb{E}[Y_i(0)]\} + D_i(\tau_i - \tau) \\ &= \alpha + D_i\tau + \epsilon_i \end{aligned}$$

  - Mean independent errors: $\mathbb{E}[\epsilon_i \mid D_i] = 0$? $\rightsquigarrow$ under randomization

# Linear regression and robust variance estimator

$\hat{\mathbb{V}}[\hat{\tau} - \tau]$

cf) $\hat{\beta} - \beta$

- Can we use "standard" variance estimator: $\mathbb{V}[\varepsilon_i \mid \mathbf{D}] = \sigma^2, \forall i$?
  - Inconsistent: $\widehat{\mathbb{V}}_{const} - \mathbb{V}[\widehat{\tau}] \xrightarrow{p} C \neq 0$ unless ...
  - Bias:

$$\mathbb{E}\left(\widehat{\mathbb{V}}_{const}\right) - \mathbb{V}[\widehat{\tau}]$$

$$= \mathbb{E}\left(\frac{\frac{1}{n-2}\sum_{i=1}^{n}\widehat{\varepsilon}_i^2}{\sum_{i=1}^{n}(D_i - \overline{D})^2}\right) - \left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0}\right)$$

$$= \frac{(n_1 - n_0)(n-1)}{n_1 n_0 (n-2)}(\sigma_1^2 - \sigma_0^2)$$

  - Unless
    - Homoskedasticity holds: $\sigma_1^2 = \sigma_0^2$
    - Design is balanced: $n_1 = n_0$

1. $\hat{\tau}_{ols} = \underset{\tau}{\arg\min}(\ \ )^2$ model

2. $V(\hat{\tau}_{ols}) =$ ~~~

cf) $V(\hat{\tau}_{diff})$

3. $\hat{V}(\hat{\tau}_{ols})$

e.g. "standard" est.

EHW

HC2 ...

Constant effect

$Y_i(1) - Y_i(0) = $ Const.

$V[Y_i(1)] = V[Y_i(0) + \text{const.}]$

$= V[Y_i(0)]$

# Linear regression and robust variance estimator

- Use robust variance estimator! [Pset4 Q1 (b)]
  - Eicker-Huber-White (EHW) estimator: consistent for $\mathbb{V}(\widehat{\tau}_{\text{diff}})$

$$\widehat{\mathbb{V}}_{\text{EHW}} = \frac{\widetilde{\sigma}_1^2}{n_1} + \frac{\widetilde{\sigma}_0^2}{n_0}, \quad \text{where} \quad \widetilde{\sigma}_d^2 = \frac{1}{n_d} \sum_{i:D_i=d} \left(Y_i - \overline{Y}_d\right)^2$$

  - HC2 estimator: exactly the Neyman variance estimator $\rightsquigarrow$ unbiased

$$\widehat{\mathbb{V}}_{\text{HC2}} = \frac{\widehat{\sigma}_0^2}{n_0} + \frac{\widehat{\sigma}_1^2}{n_1}$$

```r
# In R:
your_fitted_model <- lm(your_formula, data = your_data)
sandwich::vcovHC(your_fitted_model, type = 'HC2')
# Or
estimatr::lm_robust(your_formula, your_data, se_type = 'HC2')
```

# Linear regression with covariates

- What if we add covariates to increase **precision** of our estimates?
  - Intuition: less residual variation in $Y_i$ after accounting for $\mathbf{X}_i$
  - Use **centered** covariates: $\widetilde{\mathbf{X}}_i = \mathbf{X}_i - \overline{\mathbf{X}}$

$$(\widehat{\tau}_{\text{adj}}, \widehat{\alpha}_{\text{adj}}, \widehat{\beta}_{\text{adj}}) = \underset{\tau,\alpha,\beta}{\arg\min} \sum_{i=1}^{n} \left(Y_i - \alpha - \tau D_i - \widetilde{\mathbf{X}}_i' \beta\right)^2$$

  - $\widehat{\tau}_{\text{adj}}$ now **biased** but **consistent** for $\tau$.

# Linear regression with covariates

- Variance of adjustment estimator if $p = \frac{1}{2} \rightsquigarrow \frac{\sigma_{0x}^2}{np(1-p)} > 0 \rightsquigarrow$ always help
  - Usually will help precision, but can hurt (Freedman 2008):

$$\mathbb{V}[\widehat{\tau}_{\text{diff}}] - \mathbb{V}[\widehat{\tau}_{\text{adj}}] = \frac{\sigma_{0x}\left\{\sigma_{0x} + 2(1-2p)\sigma_{1x}\right\}}{np(1-p)}$$

  - If fully interacted, will never hurt precision (Lin 2013) [Pset4 Q1 (c)]

$$Y_i = \alpha + \tau D_i + \widetilde{\mathbf{X}}_i'\beta + D_i\widetilde{\mathbf{X}}_i'\gamma + \varepsilon_i$$

  - Estimation: EHW robust variance estimators are consistent or asymptotically conservative for $\mathbb{V}[\widehat{\tau}_{\text{adj}}]$

# Linear regression with covariates

```r
# Step 1: Compute centered covariates
your_data$Xtilde <- NULL
# Step 2: Write down your formula
your_formula <- NULL
# Step 3: Fit the model using lm() or estimatr::lm_robust()
your_fitted_model <- lm(your_formula, data = your_data)
# Step 4: Compute robust standard errors (skip if you used lm_robust)
your_vcov <- sandwich::vcovHC(your_fitted_model, type = 'HC2')
# Step 5: Check the point and se estimate of your coefficients
#         (look for tau hat!)
est  <- cbind("coef" = your_fitted_model$coef,
              "se" = sqrt(diag(your_vcov)))
```

# Block randomized trials

- Setup: block randomized experiment with block indicators $W_{ij}$.
  - Block "fixed effects" $W_{ij} = 1$ if $i$ is in block $j$, 0 otherwise.
  - Blocks $j \in \{1, \dots, J\}$ with sizes $w_j = n_j/n$ and propensity scores $p_j = n_{1,j}/n_j$
- Recall STAR project: within each school $(j)$ (block), classes $(i)$ were randomized.

$p_j = \frac{n_{1,j}}{n_j} = $ # treated classes in school j / # total classes in school j

- Naive approach: just include the block FEs in OLS [Pset4 Q2 (a)]

$$(\widehat{\tau}_{\text{b,fe}}, \widehat{\alpha}_1, \dots, \widehat{\alpha}_J) = \underset{(\tau, \alpha_1, \dots, \alpha_J)}{\arg\min} \sum_{i=1}^{n} \left( Y_i - \tau D_i - \sum_{j=1}^{J} \alpha_j W_{ij} \right)^2$$

e.g. School 1 ← Class $D_1 = 1$, $D_2 = 1$, 0, 0 → $p_1 = \frac{1}{2}$

2 ← 1, 1, 0, 0 → $p_2 = \frac{1}{2}$

  - $\widehat{\tau}_{\text{b,fe}}$ **not consistent** for the PATE unless ...

$$\widehat{\tau}_{\text{b,fe}} \xrightarrow{p} \frac{\sum_{j=1}^{J} \omega_j \tau_j}{\sum_{j=1}^{J} \omega_j} \quad \text{where} \quad \omega_j = w_j p_j (1 - p_j)$$

$p_1 = p_2 = \frac{1}{2}$

    - Propensity scores are equal across blocks: $p_j = p$ for all $j$.
    - ATEs are equal across strata $\tau_j = \tau$ for all $j$.

# Block randomized trials: Correct analysis

1. Just use original Neyman analysis aggregating within-strata analyses. [Pset3 Q5]
2. Weight OLS by inverse of the propensity score. [Pset4 Q2 (b)]
3. Fully interact block FEs with treatment. [Pset4 Q2 (c)]

- Check Imbens and Rubin (2015) Ch.9.6.1, second model
- See this simulation study using `DeclareDesign`: https://declaredesign.org/blog/biased-fixed-effects.html

# Block randomized trials: Correct analysis

2. Weight OLS by inverse of the propensity score.

$$(\widehat{\tau}_{\text{b,w}}, \widehat{\alpha}_1, \ldots, \widehat{\alpha}_J) = \underset{(\tau, \alpha_1, \ldots, \alpha_J)}{\arg\min} \sum_{i=1}^{n} s_i \left( Y_i - \tau D_i - \sum_{j=1}^{J} \alpha_j W_{ij} \right)^2$$

where $s_i = \sum_{j=1}^{J} \left\{ \left( \frac{1}{p_j} \right) D_i + \left( \frac{1}{1-p_j} \right) (1 - D_i) \right\} W_{ij}$ and $p_j = n_{1,j}/n_j$.

```r
# In R
your_formula <- as.formula("outcome ~ treat + x_tilde1 + x_tilde2")
your_data <- data.frame(outcome, treat,
                        x_tilde1, x_tilde2,
                        weights, block)
your_fitted_model <- estimatr::lm_robust(your_formula, data = your_data,
                                         weights = weights, # s
                                         se_type = "HC2",
                                         fixed_effects = block)
```

| ID of i | ID of j (school) |
|---|---|
| 1 | 1 |
| 2 | 1 |
| 3 | 1 |
| 4 | 2 |
| 5 | 2 |
| 6 | 3 |
| ⋮ | ⋮ |

$\times$ Not $W_{ij}$!

# Cluster randomized trials

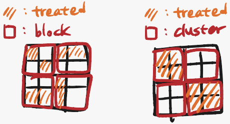
e.g. STAR

School < class < student

block    cluster

- Treatment allocated at a higher level than the data.
  - Suppose schools are randomized and all the classes in same school receives same treatment
  - Now school is not a block, but cluster!

e.g.




- Setup:
  - Clusters: $k \in \{1, \ldots, K\}$
  - Randomly choose $K_1$ treatment clusters, $K_0$ control.
  - Each cluster has units $i \in \{1, \ldots, m_k\}$ with $\sum_{k=1}^{K} m_k = n$
  - Treatment assignment at cluster level: $D_{ik} = D_k$
  - Potential outcomes $Y_{ik}(d)$
- Cost of clustering
  - More similarity $\rightsquigarrow$ each unit provides redundant information $\rightsquigarrow$ less efficiency under clustering

# Cluster randomized trials

- Use **cluster-robust variance estimator**

```r
# In R
your_formula <- as.formula("outcome ~ treat + x_tilde1 + x_tilde2")
your_data <- data.frame(outcome, treat,
                        x_tilde1, x_tilde2,
                        cluster)
your_fitted_model <- estimatr::lm_robust(your_formula, data = your_data,
                                         clusters = cluster,
                                         se_type = "CR2")
??estimatr::lm_robust # Check more options for se_type
# Or
your_model <- lm(your_formula, data = your_data)
your_vcov <- clubSandwich::vcovCR(your_model, cluster = your_data$cluster,
                                  type = "CR2")
```

- You may have block and cluster design at the same time! [Pset4 Q3]