

Module 4: Linear Regression and Randomized Experiments

Fall 2021

Matthew Blackwell

Gov 2003 (Harvard)

Where are we? Where are we going?

- So far: analysis of experiments from Fisher and Neyman approaches.
 - Neyman: Unbiased estimators, (conservative) variances.
 - Fisher: exact tests of the sharp null.
- Today: how does the workhorse estimator, OLS, fit into this story?
- Why might we use regression?
 - **Simplicity**: known tool that is already very common.
 - **Increased precision**: we may want to add covariates for more precise effect estimates.

1/ Regression with no covariates

Analyzing experiments with regression?

- Under complete randomization, can we use OLS to estimate ATEs?
 - Literally just $\text{lm}(y \sim d)$?
- Recall that the OLS estimator solves the least squares problem:

$$(\hat{\tau}_{\text{ols}}, \hat{\alpha}_{\text{ols}}) = \arg \min_{\tau, \alpha} \sum_{i=1}^n (Y_i - \alpha - \tau D_i)^2$$

- Remember coefficient on a binary r.v. is mechanically the diff. in means:

$$\hat{\tau}_{\text{ols}} = \bar{Y}_1 - \bar{Y}_0 = \hat{\tau}_{\text{diff}}$$

- \rightsquigarrow standard Neyman analysis for unbiasedness, sampling variance.
- Generalizes to discrete treatments with > 2 levels.

Justifying the linear model

- Mechanically the same, but can we justify the linear model itself?
 - Key assumptions: **linearity** and **mean independence of errors**.
- Some simple manipulations of the consistency assumption:

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

- “Linear” functional form fully justified by consistency alone with:
 - Intercept $\alpha = \mathbb{E}[Y_i(0)]$ average control outcome.
 - Slope $\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$ the PATE.
 - Error is deviation for control PO + treatment effect heterogeneity.

Mean independent errors

$$\varepsilon_i = (Y_i(0) - \mathbb{E}[Y_i(0)]) + D_i \cdot (\tau_i - \tau)$$

- What about mean independence $\mathbb{E}[\varepsilon_i | D_i] = 0$?
- Using randomization, we know D_i is independent of $Y_i(1), Y_i(0)$, so:

$$\begin{aligned}\mathbb{E}[\varepsilon_i | D_i] &= \mathbb{E} \{ [Y_i(0) - \mathbb{E}[Y_i(0)] + D_i \cdot (\tau_i - \tau) | D_i \} \\ &= \mathbb{E}[Y_i(0) | D_i] - \mathbb{E}[Y_i(0)] + D_i (\mathbb{E}[\tau_i | D_i] - \tau) \\ &= \mathbb{E}[Y_i(0)] - \mathbb{E}[Y_i(0)] + D_i \underbrace{(\mathbb{E}[\tau_i] - \tau)}_{\tau = \mathbb{E}[\tau_i]} \\ &= 0\end{aligned}$$

- Randomization + consistency \rightsquigarrow linear model.
 - Does not imply homoskedasticity or normal errors, though!

Homoskedasticity

- Software default assumption: **Homoskedasticity**

$$\mathbb{V}[\varepsilon_i | \mathbf{D}] = \sigma^2, \quad \forall i$$

- But in general, based on previous error definition:

$$\mathbb{V}[\varepsilon_i | \mathbf{D}] = \mathbb{V}[\varepsilon_i | D_i] = D_i\sigma_1^2 + (1 - D_i)\sigma_0^2$$

- \rightsquigarrow homoskedasticity true when $\sigma_1^2 = \mathbb{V}[Y_i(1)] = \mathbb{V}[Y_i(0)] = \sigma_0^2$
 - True under constant treatment effects!
- Under homoskedasticity, variance of the OLS estimator is:

$$\mathbb{V}[\widehat{\tau}_{\text{ols}} | \mathbf{D}] = \frac{\sigma^2}{\sum_{i=1}^n (D_i - \bar{D})^2}$$

Variance estimation

- “Standard” variance estimator under homoskedasticity:

$$\hat{V}_{const} = \frac{\frac{1}{n-2} \sum_{i=1}^n \hat{\epsilon}_i^2}{\sum_{i=1}^n (D_i - \bar{D})^2} = \frac{\frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{\alpha}_{ols} - \hat{\tau}_{ols} D_i)^2}{\sum_{i=1}^n (D_i - \bar{D})^2}$$

- We can rewrite this as a function of the **pooled** variance $\hat{\sigma}_{Y|D}^2$:

$$\hat{V}_{const} = \hat{\sigma}_{Y|D}^2 \left(\frac{1}{n_0} + \frac{1}{n_1} \right)$$

$$\hat{\sigma}_{Y|D}^2 = \frac{1}{n-2} \left(\sum_{i:D_i=0} (Y_i - \bar{Y}_0)^2 + \sum_{i:D_i=1} (Y_i - \bar{Y}_1)^2 \right)$$

- **Inconsistent:** $\hat{V}_{const} - \mathbb{V}[\hat{\tau}] \xrightarrow{P} c \neq 0$ unless

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

Robust SEs

- Eicker-Huber-White (EHW) robust/sandwich variance estimator:

$$\begin{aligned}\hat{V}_{\text{EHW}} &= \left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left(\sum_{i=1}^n \hat{\varepsilon}_i^2 \mathbf{x}_i \mathbf{x}_i' \right) \left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \\ &= (\mathbb{X}'\mathbb{X})^{-1} \left(\sum_{i=1}^n \hat{\varepsilon}_i^2 \mathbf{x}_i \mathbf{x}_i' \right) (\mathbb{X}'\mathbb{X})^{-1} \quad \text{where } \mathbb{X} = [1 \ \mathbf{D}]\end{aligned}$$

- Recall the PATE-targeted variance of the difference-in-means:

$$\mathbb{V}(\hat{\tau}_{\text{diff}}) = \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} = \frac{\mathbb{V}[Y_i(0)]}{n_0} + \frac{\mathbb{V}[Y_i(1)]}{n_1}$$

- To see this, we can derive \hat{V}_{EHW} under our case:

$$\hat{V}_{\text{EHW}} = \frac{\tilde{\sigma}_1^2}{n_1} + \frac{\tilde{\sigma}_0^2}{n_0}, \quad \text{where } \tilde{\sigma}_d^2 = \frac{1}{n_d} \sum_{i:D_i=d} (Y_i - \bar{Y}_d)^2$$

- $\tilde{\sigma}_0^2, \tilde{\sigma}_1^2$ consistent for $\sigma_0^2, \sigma_1^2 \rightsquigarrow \hat{V}_{\text{EHW}}$ consistent for $\mathbb{V}(\hat{\tau}_{\text{diff}})$

Better robust SEs

- Many different “improved” versions of robust variances proposed.
 - Almost all are “finite-sample corrections” (no asymptotic effects)
- HC2 estimator normalizes residuals by the leverage, h_{ii} :

$$\hat{\mathbf{V}}_{\text{HC2}} = (\mathbf{X}'\mathbf{X})^{-1} \left(\sum_{i=1}^n \frac{\hat{\varepsilon}_i^2}{1 - h_{ii}} \mathbf{x}_i \mathbf{x}_i' \right) (\mathbf{X}'\mathbf{X})^{-1}$$

- Leverage: $h_{ii} = \mathbf{x}_i (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_i'$
- In this setting, $h_{ii} = n_1^{-1}$ if $D_i = 1$ and n_0^{-1} if $D_i = 0$
- Samii & Aronow (2012): HC2 is exactly the Neyman variance estimator:

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

- \rightsquigarrow simple OLS + HC2 = unbiased point and variance estimator.

2/ Linear regression with covariates

Adding covariates

- What if we add covariates to our regression model?

$$(\hat{\tau}_{\text{adj}}, \hat{\alpha}_{\text{adj}}, \hat{\beta}_{\text{adj}}) = \arg \min_{\tau, \alpha, \beta} \sum_{i=1}^n (Y_i - \alpha - \tau D_i - \tilde{\mathbf{X}}_i' \beta)^2$$

- $(\tilde{\mathbf{X}}_i - \bar{\mathbf{X}})$ are **centered** covariates for notational ease.
- Why might we do this? To increase **precision** of our estimates.
 - We hope $\mathbb{V}[\hat{\tau}_{\text{adj}}] < \mathbb{V}[\hat{\tau}_{\text{diff}}]$ so we have smaller CIs, more powerful tests, etc
 - Intuition: less residual variation in Y_i after accounting for \mathbf{X}_i
- Questions:
 - Is $\hat{\tau}$ still unbiased? Consistent?
 - Should we expect an increase in precision?
 - Controversial! Freedman (2008): “Randomization does not justify the regression model”

OLS is biased, but consistent

- Agnostic approach: don't assume correctness of the linear model.
 - $\mathbf{X}'_i\beta$ could be badly misspecified.
 - No assumptions about homoskedasticity.
- Under minimal assumptions, OLS is consistent for the linear projection

$$(\hat{\tau}_{\text{adj}}, \hat{\alpha}_{\text{adj}}, \hat{\beta}_{\text{adj}}) \xrightarrow{p} (\tau_0, \alpha_0, \beta_0) = \arg \min_{(\tau, \alpha, \beta)} \mathbb{E} \left[\left(Y_i - \alpha - \tau D_i - \tilde{\mathbf{X}}'_i \beta \right)^2 \right]$$

- $\hat{\tau}_{\text{adj}}$ now **biased** for τ though bias should be small.
- But $\hat{\tau}_{\text{adj}}$ is **consistent** for τ .
 - Intuition: omitted variable bias. Since $D_i \perp\!\!\!\perp \mathbf{X}_i$, including $\tilde{\mathbf{X}}_i$ won't (asymptotically) affect coefficient on D_i .
- Freedman (2008) shows the same thing for finite-sample inference.

Variance of adjustment estimator

- Complete randomization + single, mean-zero covariate X_i
 - Generalizes easily to more covariates.
 - Let $\sigma_{0x} = \text{cov}(Y_i(0), X_i)$ and $\sigma_{1x} = \text{cov}(Y_i(1), X_i)$.
 - Probability of treatment $p = n_1/n$
- Freedman (2008) derived gains from adjusting for X_i using OLS:

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

- Will adjustment decrease the sampling variance?
 - If design is balanced, $p = 1/2$, then adjustment always helps.
 - Design imbalance and correlation with “smaller” potential outcome could lead to adjustment hurting.
- Estimation: EHW robust variance estimators are consistent or asymptotically conservative for $\mathbb{V}[\widehat{\tau}_{\text{adj}}]$

Regression with full interactions

- OLS estimator from fully interacted model, $\hat{\tau}_{\text{inter}}$:

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

- Equivalent to running separate Y_i on $\tilde{\mathbf{X}}_i$ in each D_i group
- As with non-interacted model, $\hat{\tau}_{\text{inter}}$ is consistent for τ and asymptotically normal.
- Lin (2013): **fully interacted model will never hurt precision asymptotically.**
 - Freedman critique was right, but Lin shows an easy way to resolve.
- EHW robust variance estimator is consistent or asymptotically conservative.

Summarizing regression

- Regression with no covariates: standard Neyman analysis.
- Regression with (uninteracted) covariates:
 - Consistent for SATE/PATE.
 - Usually will help precision, but can hurt.
- Regression with interacted covariates:
 - Consistent for SATE/PATE
 - Asymptotically will never hurt precision.
- Always use robust/HC2 variance estimators unless you have good reasons.

Regression for stratified experiments

- 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 $p_j = n_{1,j}/n_j$
- Can we just include the block FEs in OLS?

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

- Converges to a weighted average of block-specific effects, τ_j :

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

- $\hat{\tau}_{b,fe}$ not consistent for the PATE unless:
 - Propensity scores are equal across blocks: $p_j = p$ for all j .
 - ATEs are equal across strata $\tau_j = \tau$ for all j .

Correct analysis of block randomized trials

1. Just use original Neyman analysis aggregating within-strata analyses.
2. Weight OLS by inverse of the propensity score: $1/p_j$.
3. Fully interact block FEs with treatment.
 - Latter two allow for additional covariates to be added.

3/ Cluster randomized experiments

Clustering treatments

- Treatment often allocated at a higher level than the data.
 - Counties are treated, but we have individual-level data.
 - Classrooms are treated, but we have student data.
- Has considerable benefits:
 - Often cheaper/easier to implement than individual assignment.
 - Allows for interference within clusters without bias.
- But lots of confusion about how to analyze.
 - More valuable to add more individuals or clusters?
 - What to do with individual-level covariates?

Cluster randomized trials

- Setup:
 - Clusters: $k \in \{1, \dots, K\}$
 - Randomly choose K_1 treatment clusters, K_0 control.
 - Each cluster has units $i \in \{1, \dots, m_k\}$ with $\sum_{k=1}^K m_k = n$
 - Treatment assignment at cluster level: $D_{ik} = D_k$
 - Potential outcomes $Y_{ik}(d)$
- Random assignment at the cluster level: $\{Y_{ik}(1), Y_{ik}(0)\} \perp\!\!\!\perp D_j$.
- Quantity of interest still at individual level:

$$\text{SATE} = \frac{1}{n} \sum_{k=1}^K \sum_{i=1}^{m_k} \{Y_{ik}(1) - Y_{ik}(0)\}$$

Analysis of clustered experiments

- Simple setting: all clusters have the same size $m_k = m$ for all k .
- Simple difference in means is unbiased:

$$\begin{aligned}\widehat{\tau}_{\text{cl}} &= \frac{1}{mK_1} \sum_{k=1}^K \sum_{i=1}^{m_k} D_k Y_{ik} - \frac{1}{mK_0} \sum_{k=1}^K \sum_{i=1}^{m_k} (1 - D_k) Y_{ik} \\ &= \frac{1}{K_1} \sum_{k=1}^K D_k \bar{Y}_k - \frac{1}{K_0} \sum_{k=1}^K (1 - D_k) \bar{Y}_k\end{aligned}$$

- \bar{Y}_k is the cluster average: $\frac{1}{m} \sum_{i=1}^m Y_{ik}$
 - Unbiasedness follows from Neyman-style analysis at cluster level.
 - Estimator is biased, but consistent (in K) if cluster size varies.
- Neyman-style conservative variance:

$$\mathbb{V}[\widehat{\tau}_{\text{cl}} \mid \mathbf{0}] \leq \frac{\mathbb{V}[\bar{Y}_k(1)]}{J_1} + \frac{\mathbb{V}[\bar{Y}_k(0)]}{J_0} \quad \text{where for } d = 0, 1 \quad \bar{Y}_k(d) = \frac{1}{m} \sum_{i=1}^m Y_{ik}(d)$$

Cost of clustering

- Standard variance under **individual assignment**:

$$\mathbb{V}[\widehat{\tau}_{\text{diff}}] = \frac{\mathbb{V}[Y_{ik}(1)]}{mK_1} + \frac{\mathbb{V}[Y_{ik}(0)]}{mK_0}$$

- How different is variance under clustering compare to no clustering?

$$\frac{\mathbb{V}[\bar{Y}_k(1)]}{K_1} = \frac{\mathbb{V}[Y_{ik}(1)]}{mK_1} (1 + (m-1)\rho_1)$$

$$\rho_1 = \frac{1}{m(m-1)} \sum_{i=1}^m \sum_{j \neq i} \mathbb{E}\{(Y_{ik}(1) - \bar{Y}(1))(Y_{jk}(1) - \bar{Y}(1))\}$$

- ρ_1 is the intraclass correlation coefficient (ICC)
 - Measures how similar units are within clusters.
 - Usually cluster is less efficient because $\rho_1 > 0$
 - More similarity \rightsquigarrow each unit provides redundant information \rightsquigarrow less efficiency under clustering

Cluster robust standard errors

- What if we want to use OLS at individual level?
 - Adding individual and unit controls.
- **Cluster-robust variance estimator** for OLS:

$$\hat{V}_{cl}[\hat{\alpha}, \hat{\tau}] = (\mathbb{X}'\mathbb{X})^{-1} \left(\sum_{k=1}^K \mathbb{X}'_k \hat{\boldsymbol{\varepsilon}}_k \hat{\boldsymbol{\varepsilon}}_k' \mathbb{X}_k \right) (\mathbb{X}'\mathbb{X})^{-1}$$

- Here $\mathbb{X}' = [1 \ \mathbf{D}]'$, $\mathbb{X}'_k = [1 \ \mathbf{D}_k]'$, and $\hat{\boldsymbol{\varepsilon}}_k = (\hat{\varepsilon}_{1k}, \dots, \hat{\varepsilon}_{mk})$
- Consistent as the number of clusters grows.
- **Cluster at the treatment assignment level** (no higher or lower)
- Vanilla CRVE is biased, Bell & McCaffrey proposed CR2 adjustment similar to HC2 (usually preferable)