Section: Module 4

Linear Regression and Randomized Experiments

Kentaro Nakamura

GOV 2002

October 3rd, 2025

Logistics

- Midterm: October 15th, 2 hours
  - Review Section: October 8th
  - Extra Office Hour: October 14th 1:30-3:00pm

- What you should do before midterm
  - Solve review questions
  - Solve practice midterm
  - Check solutions of problem sets
  - Understand *details* of class materials

Today's Agenda

- Review of Linear Regression
  - Basic Properties
  - Best linear approximation
  - Ordinal Least Squares (OLS)
  - Frisch-Waugh-Lowell (FWL) theorem
  - Asymptotic Variance (Extra slides, optional)

- Regression and Causal Inference
  - Case 1: No Control
  - Case 2: With control (no interaction)
  - Case 3: With control (with interaction)
    - Recommended reading: Ch7 of Imbens and Rubin (2015)
  - Case 4: Stratified Design

- Cluster Randomized Experiment
  - Related to Problem Set 4

Review: Linear Regression (1)

- Linear regression with K predictors is written as

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_K X_{Ki} + \epsilon_i, \quad \underbrace{\mathbb{E}[\epsilon_i \mid X_{1i}, \cdots, X_{Ki}] = 0}_{\text{Strict Exogeneity}}$$

- Let $\mathbf{X}_i = [1, X_{1i}, \cdots, X_{Ki}]^\top$ and $\boldsymbol{\beta} = [\beta_0, \cdots, \beta_K]^\top$. Then,

$$Y_i = \mathbf{X}_i^\top \boldsymbol{\beta} + \epsilon, \quad \mathbb{E}[\epsilon_i \mid \mathbf{X}_i] = 0$$

- Equivalently,

$$\mathbb{E}[Y_i \mid \mathbf{X}_i] = \mathbf{X}_i^\top \boldsymbol{\beta}$$

- Regression is conditional expectation function of Y_i given $\mathbf{X}_i$

Review: Linear Regression (2)

- Note that $\mathbb{E}[\epsilon_i \mid \mathbf{X}_i] = 0$ implies (i) $\mathbb{E}[\epsilon_i] = 0$ and (ii) $\mathbb{E}[\epsilon_i \mathbf{X}_i] = \mathbf{0}$
  - $\mathbb{E}[\epsilon_i] = \mathbb{E}[\mathbb{E}[\epsilon_i \mid \mathbf{X}_i]] = \mathbb{E}[0] = 0$
  - $\mathbb{E}[\epsilon_i \mathbf{X}_i] = \mathbb{E}[\mathbb{E}[\epsilon_i \mathbf{X}_i \mid \mathbf{X}_i]] = \mathbb{E}[\mathbf{X}_i \mathbb{E}[\epsilon_i \mid \mathbf{X}_i]] = \mathbb{E}[\mathbf{X}_i \cdot 0] = \mathbf{0}$

- Thus, the error ϵ_i is uncorrelated with regressor $\mathbf{X}_i$
  - $\mathrm{Cov}(\mathbf{X}_i, \epsilon_i) = \mathbb{E}[\epsilon_i \mathbf{X}_i] - \mathbb{E}[\epsilon_i]\mathbb{E}[\mathbf{X}_i] = 0$

Review: Projection Coefficient

- Consider minimizing the mean squared prediction error

$$\begin{aligned} S(\beta) &= \mathbb{E}[(Y_i - \mathbf{X}_i^\top \beta)^2] \\ &= \mathbb{E}[Y_i^2] - 2\beta^\top \mathbb{E}[\mathbf{X}_i Y_i] + \beta^\top \mathbb{E}[\mathbf{X}_i \mathbf{X}_i^\top]\beta \end{aligned}$$

- Taking the derivative, we get

$$\frac{\partial S(\beta)}{\partial \beta} = -2\mathbb{E}[\mathbf{X}_i Y_i] + 2\mathbb{E}[\mathbf{X}_i \mathbf{X}_i^\top]\beta$$

- This gives the projection coefficient β

$$\beta = \left(\mathbb{E}[\mathbf{X}_i \mathbf{X}_i^\top]\right)^{-1} \mathbb{E}[\mathbf{X}_i Y_i]$$

- Check this by inserting $Y_i = \mathbf{X}_i^\top \beta + \epsilon$.

Why linear regression? Why projection?

- Linear regression is a **best linear approximation** of conditional expectation function $m(\mathbf{X}_i) = \mathbb{E}[Y_i \mid \mathbf{X}_i]$
- To see this, consider minimizing the mean squared loss between $m(\mathbf{X}_i)$ and $\mathbf{X}_i^\top \beta$

$$d(\beta) = \mathbb{E}[(m(\mathbf{X}_i) - \mathbf{X}_i^\top \beta)^2]$$

- Notice that this minimization problem is just replacing Y_i with $m(\mathbf{X}_i)$ from the previous one.
- Thus, the best linear approximation $\tilde{\beta}$ is given by

$$\begin{aligned}
\tilde{\beta} &= \left(\mathbb{E}[\mathbf{X}_i \mathbf{X}_i^\top]\right)^{-1} \mathbb{E}[\mathbf{X}_i m(\mathbf{X}_i)] \\
&= \left(\mathbb{E}[\mathbf{X}_i \mathbf{X}_i^\top]\right)^{-1} \mathbb{E}[\mathbf{X}_i \mathbb{E}[Y_i \mid \mathbf{X}_i]] \\
&= \left(\mathbb{E}[\mathbf{X}_i \mathbf{X}_i^\top]\right)^{-1} \mathbb{E}[\mathbb{E}[\mathbf{X}_i Y_i \mid \mathbf{X}_i]] \\
&= \left(\mathbb{E}[\mathbf{X}_i \mathbf{X}_i^\top]\right)^{-1} \mathbb{E}[\mathbf{X}_i Y_i] = \beta
\end{aligned}$$

- Projection coefficient is equal to the best linear approximation of $m(\mathbf{X}_i) = \mathbb{E}[Y_i \mid \mathbf{X}_i]$

Review: Ordinary Least Squares (OLS)

- Let's use matrix notation.

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \ \mathbf{X} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{K1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{Kn} \end{bmatrix}, \ \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{bmatrix}$$

- OLS estimator is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} = \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^\top \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i Y_i \right)$$

- This estimator is unbiased (Hint: show $\mathbb{E}[\hat{\boldsymbol{\beta}} \mid \mathbf{X}] = \boldsymbol{\beta}$)
  - Once you show $\mathbb{E}[\hat{\boldsymbol{\beta}} \mid \mathbf{X}] = \boldsymbol{\beta}$, then use the law of iterated expectation!
- Also, residual is orthogonal: $\mathbf{X}^\top \hat{\boldsymbol{\epsilon}} = \mathbf{X}^\top (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = 0$
  - Hint: Insert the definition of $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$

Frisch-Waugh-Lowell (FWL) theorem

- Consider multiple regressions with K regressors

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_K X_{Ki} + \epsilon_i, \quad \mathbb{E}[\epsilon_i \mid X_{1i}, \cdots, X_{Ki}] = 0$$

- Estimator for β_k needs matrix $\rightarrow$ often hard to make a proof

- **FWL theorem** gives another formula for k-th coefficient

$$\beta_k = \frac{\mathrm{Cov}(Y_i, \tilde{X}_{ki})}{\mathbb{V}[\tilde{X}_{ki}]} \quad \text{for } k \in \{1, \cdots, K\}$$

where $\tilde{X}_{ik}$ is the residual obtained by

$$X_{ki} = \gamma_0 + \sum_{j \neq k} \gamma_j X_{ji} + \tilde{X}_{ki}$$

Frisch-Waugh-Lowell (FWL) theorem: Proof

$$
\begin{aligned}
\frac{\mathrm{Cov}(Y_i, \tilde{X}_{ki})}{\mathbb{V}[\tilde{X}_{ki}]} &= \frac{\mathrm{Cov}(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_K X_{Ki} + \epsilon_i, \tilde{X}_{ki})}{\mathbb{V}[\tilde{X}_{ki}]} && (\because \text{def of } Y_i) \\
&= \frac{\mathrm{Cov}(\beta_k X_{ki}, \tilde{X}_{ki})}{\mathbb{V}[\tilde{X}_{ki}]} \\
&= \frac{\mathrm{Cov}(\beta_k \{\gamma_0 + \textstyle\sum_{j \neq k} \gamma_j X_{ji} + \tilde{X}_{ki}\}, \tilde{X}_{ki})}{\mathbb{V}[\tilde{X}_{ki}]} && (\because \text{def of } X_{ki}) \\
&= \frac{\mathrm{Cov}(\beta_k \tilde{X}_{ki}, \tilde{X}_{ki})}{\mathbb{V}[\tilde{X}_{ki}]} && (\because \mathrm{Cov}(\tilde{X}_{ki}, X_{ji}) = 0 \text{ for } j \neq k) \\
&= \beta_k
\end{aligned}
$$

where the second line is because

- $\mathrm{Cov}(\tilde{X}_{ki}, X_{ji}) = 0$ for any $j \neq k$
- $\mathrm{Cov}(X_{li}, \epsilon) = 0$ for all $l \in \{1, \cdots, K\} \rightarrow \mathrm{Cov}(\tilde{X}_{ki}, \epsilon) = 0$

Frisch-Waugh-Lowell (FWL) theorem: Remark

- Note that you can also residualize the outcome; i.e.,

$$\beta_k = \frac{\mathrm{Cov}(\tilde{Y}_i, \tilde{X}_{ki})}{\mathbb{V}[\tilde{X}_{ki}]} \quad \text{for } k \in \{1, \cdots, K\}$$

where $\tilde{Y}_i$ is the residual obtained by

$$Y_i = \alpha_0 + \sum_{j \neq k} \alpha_j X_{ji} + \tilde{Y}_i$$

- FWL Theorem is a tool for you to make the proof of multiple regressions without using matrix
  - This can be a helpful tool for exam / problem set

Extra slides: Tools for Asymptotic Variance

- **Law of Large Numbers (LLN)**: If $X_1, \cdots, X_n$ are i.i.d.,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mathbb{E}[X_i]$$

- **Central Limit Theorem (CLT)**: If $X_1, \cdots, X_n$ are i.i.d.,

$$\sqrt{n}(\bar{X} - \mathbb{E}[X_i]) \xrightarrow{d} \mathcal{N}(0, \mathbb{V}[X_i])$$

- **Slutsky's Lemma**[1]: If $X_n \xrightarrow{d} X$ for some random variable X and $Y_n \xrightarrow{p} c$ for some constant c,

$$X_n Y_n \xrightarrow{d} cX$$

[1] If you haven't seen it, don't worry. Not required for midterm.

Extra slides: Asymptotic Variance (1)

- **Goal**: To understand where sandwich comes from.
  - Don't worry even if you can't get all the details (it is optional)
- Recall that the estimator for $\hat{\beta}$ is given by

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} = \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^\top \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i Y_i \right)$$

- Because $\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}$,

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\boldsymbol{X}\beta + \boldsymbol{\epsilon}) = \beta + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\epsilon}$$

Thus,

$$\sqrt{n}(\hat{\beta} - \beta) = \sqrt{n}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\epsilon} = \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^\top \right)^{-1} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{X}_i \epsilon_i \right)$$

Extra slides: Asymptotic Variance (2)

- Notice that by law of large numbers, the first term converges in probability to

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^\top \xrightarrow{p} \mathbb{E}[\mathbf{X}^\top\mathbf{X}]$$

- On the other hand, because each $\mathbf{X}_i\epsilon_i$ are i.i.d., by central limit theorem, the second term converges in distribution to

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{X}_i\epsilon_i = \sqrt{n}\underbrace{\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\epsilon_i}_{\text{Form of Avg!}} \xrightarrow{d} \mathcal{N}(0, \mathbb{E}[(\mathbf{X}^\top\boldsymbol{\epsilon}(\mathbf{X}^\top\boldsymbol{\epsilon})^\top)])$$

- Note that mean of normal distribution here is 0 because $\mathbb{E}[\mathbf{X}_i\epsilon_i] = 0$

Extra slides: Asymptotic Variance (3)

- Therefore,

$$\sqrt{n}(\hat{\beta} - \beta) = \sqrt{n}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \boldsymbol{\epsilon}$$

$$\rightarrow \mathcal{N}\left(0, \underbrace{\mathbb{E}[\mathbf{X}^\top \mathbf{X}]^{-1}\mathbb{E}[\mathbf{X}^\top \boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \mathbf{X}]\mathbb{E}[\mathbf{X}^\top \mathbf{X}]^{-1}}_{\text{Asymptotic Variance}}\right)$$

- **Homoskedasticity**: $\mathbb{V}[\epsilon_i \mid \mathbf{X}_i] = \sigma^2$
  - This gives $\mathbb{V}[\epsilon_i \mid \mathbf{X}_i] = \mathbb{E}[\epsilon_i^2 \mid \mathbf{X}_i] - \underbrace{\mathbb{E}[\epsilon_i \mid \mathbf{X}_i]^2}_{=0} = \mathbb{E}[\epsilon_i^2 \mid \mathbf{X}_i]$

$$\mathbb{E}[\mathbf{X}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \mathbf{X}^\top] = \mathbb{E}[\mathbb{E}[\mathbf{X}\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \mathbf{X}^\top \mid \mathbf{X}]] \quad (\because \text{Law of Iterated Expectation})$$

$$= \mathbb{E}[\mathbf{X}\underbrace{\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \mid \mathbf{X}]}_{=\sigma^2 I_n}\mathbf{X}^\top] = \sigma^2\mathbb{E}[\mathbf{X}^\top \mathbf{X}]$$

and thus the asymptotic variance is simplified to

$$\mathbb{E}[\mathbf{X}^\top \mathbf{X}]^{-1}\mathbb{E}[\mathbf{X}^\top \boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \mathbf{X}]\mathbb{E}[\mathbf{X}^\top \mathbf{X}]^{-1} = \sigma^2\mathbb{E}[\mathbf{X}^\top \mathbf{X}]^{-1}$$

- Homoskedasticity is usually not plausible

Extra slides: Eicker-Huber-White (EHW) robust variance estimator (HC0)

- **Eicker-Huber-White (EHW) robust variance estimator** (or HC0)

$$\underbrace{(\mathbf{X}^\top \mathbf{X})^{-1}}_{\text{Bread}} \underbrace{\mathbf{X}^\top \operatorname{diag}(\hat{\epsilon}_i^2)\, \mathbf{X}}_{\text{Meat}} \underbrace{(\mathbf{X}^\top \mathbf{X})^{-1}}_{\text{Bread}}$$

- Asymptotically consistent
  - i.e., as N increases, it approaches to the true value
- But this has some bias in finite sample
  - We will see it with simulation (two slides later)

Extra slides: HC2 variance estimator

- **HC2 heteroskedasticity-robust variance estimator**
  - Make finite sample bias smaller

$$\underbrace{(\mathbf{X}^\top\mathbf{X})^{-1}}_{\text{Bread}}\underbrace{\mathbf{X}^\top \operatorname{diag}\left(\frac{\hat{\epsilon}_i^2}{1-p_{ii}}\right)\mathbf{X}}_{\text{Meat}}\underbrace{(\mathbf{X}^\top\mathbf{X})^{-1}}_{\text{Bread}}$$

where $p_{ii} = \mathbf{X}_i(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_i$

- When $\mathbf{X}_i = [1, T_i]^\top$ (i.e., no control),

$$\begin{aligned} p_{ii} &= \begin{bmatrix} 1 \\ T_i \end{bmatrix} \left(\begin{bmatrix} \sum_{i=1}^n 1 & \sum_{i=1}^n T_i \\ \sum_{i=1}^n T_i & \sum_{i=1}^n T_i^2 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 \\ T_i \end{bmatrix} \\ &= \frac{1}{n_1 n_0}\Big(n_1(1-T_i) + n_0 T_i\Big) = \begin{cases} 1/n_1 & (\text{if } \; T_i = 1) \\ 1/n_0 & (\text{if } \; T_i = 0) \end{cases} \end{aligned}$$

- **Takeaway**: HC2 becomes Neyman Variance with no control

Extra slides: Many Robust Standard Errors

- Which robust SEs should we use in practice?
  - Do not use homoskedastic one unless you check it
  - When sample size is large, all robust SEs are consistent
  - For small sample, use degrees of freedom adjustment from Bell and McCaffrey (2002)
- Simulation results (from ECON2110 Lecture 12 Slide)
  - Homoskedasticity $\rightarrow$ Does not have a proper coverage (95%)
  - HC1 / HC2 $\rightarrow$ proper coverage for large sample, but some finite-sample bias

	$N = 1000$	$N = 100$	$N = 50$	$N = 20$
Homoskedasticity	**91.72%**	**90.77%**	**90.07%**	**88.75%**
HC1	**95.42%**	**93.68%**	**92.17%**	**88.62%**
HC2	**95.45%**	**93.97%**	**92.81%**	**90.38%**
HC2 w/ dof adjust	**95.51%**	**94.8%**	**94.58%**	**96.38%**

Regression and Causal Inference: No Control

- Consider the simple linear regression

$$Y_i = \alpha + \beta T_i + \epsilon_i \quad \text{equivalently} \quad \mathbb{E}[Y_i \mid T_i = t] = \alpha + \beta t$$

- Under complete randomization and consistency,

$$\mathbb{E}[Y_i \mid T_i = t] = \mathbb{E}[Y_i(t) \mid T_i = t] = \mathbb{E}[Y_i(t)]$$

- The regression above is written as

$$\mathbb{E}[Y_i(t)] = \alpha + \beta t$$

- Thus

$$\alpha = \mathbb{E}[Y_i(0)], \quad \beta = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] \; (= \text{Average Treatment Effect})$$

- **Takeaway**: Regression estimates ATE (same as diff-in-means)
  - Linearity does not matter
  - HC2 variance is numerical identical to Neyman's variance

No Interaction model: Consistency (1)

- After randomization, there might be imbalance in covariates
  - In this case, covariate adjustment helps improving efficiency
- Consider the multiple regression with demeaned pre-treatment covariates

$$Y_i = \alpha + \beta T_i + \boldsymbol{\gamma}^\top \tilde{\mathbf{x}}_i + \epsilon_i$$

- OLS estimator minimizes the mean squared error

$$\sum_{i=1}^{n} \{Y_i - \alpha - \beta T_i - \boldsymbol{\gamma}^\top \tilde{\mathbf{x}}_i\}^2$$

$$= \sum_{i=1}^{n} \{Y_i - \alpha - \beta T_i\}^2 + \sum_{i=1}^{n} \{\boldsymbol{\gamma}^\top \tilde{\mathbf{x}}_i\}^2 - 2 \sum_{i=1}^{n} \{(Y_i - \alpha - \beta T_i) \boldsymbol{\gamma}^\top \tilde{\mathbf{x}}_i\}$$

No Interaction model: Consistency (2)

- Notice that the last term converges to

$$\mathbb{E}[(Y_i - \alpha - \beta T_i)\gamma^\top \tilde{\mathbf{X}}_i] = \mathbb{E}[Y_i \gamma^\top \tilde{\mathbf{X}}_i] - \alpha\gamma^\top \overbrace{\mathbb{E}[\tilde{\mathbf{X}}_i]}^{=0} - \beta \overbrace{\mathbb{E}[T_i \gamma^\top \tilde{\mathbf{X}}_i]}^{=\mathbb{E}[T_i]\gamma^\top \mathbb{E}[\tilde{\mathbf{X}}_i]=0}$$
$$= \mathbb{E}[Y_i \gamma^\top \tilde{\mathbf{X}}_i]$$

where $\mathbb{E}[T_i \boldsymbol{X}_i] = \mathbb{E}[T_i]\mathbb{E}[\tilde{\boldsymbol{X}}_i]$ by randomization of treatment and $\mathbb{E}[\tilde{\boldsymbol{X}}_i] = 0$ thanks to demeaning.

- So, the parameter of interest β only depends on the first term
  - Specification of covariates $\gamma^\top \tilde{\mathbf{X}}_i$ does not matter

- **Takeaway**: OLS estimator is consistent for PATE
  even if model is incorrect

No Interaction model: Efficiency

- Theorem 7.1 of Imbens and Rubin (2015):

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}\left(0, \frac{\mathbb{E}[(T_i - p)^2(Y_i - \alpha - \beta T_i - \boldsymbol{\gamma}^\top \tilde{\mathbf{X}}_i)^2]}{p^2(1-p)^2}\right)$$

where $p = \frac{n_1}{n} = \mathbb{P}(T_i = 1)$

- If covariates predict outcome well, $Y_i - \alpha - \beta T_i - \boldsymbol{\gamma}^\top \tilde{\mathbf{X}}_i$ becomes smaller

- **Takeaway**: Controlling $\tilde{\mathbf{X}}$ improves efficiency <u>if model is correct</u>
  - Efficiency gains can be lost under misspecifications (Freedman, 2008)
  - Though in most cases efficiency improves...

Fully-interacted model (1)

- Consider the model with interaction

$$Y_i = \alpha + \beta T_i + \boldsymbol{\gamma}^\top \tilde{\mathbf{X}}_i + \boldsymbol{\delta}^\top T_i \tilde{\mathbf{X}}_i + \epsilon_i$$

- Notice that

$$\mathbb{E}[Y_i(1) \mid \tilde{\mathbf{X}}_i] = \mathbb{E}[Y_i \mid T_i = 1, \tilde{\mathbf{X}}_i] = \alpha + \beta + (\boldsymbol{\gamma} + \boldsymbol{\delta})^\top \tilde{\mathbf{X}}_i$$
$$\mathbb{E}[Y_i(0) \mid \tilde{\mathbf{X}}_i] = \mathbb{E}[Y_i \mid T_i = 0, \tilde{\mathbf{X}}_i] = \alpha + \boldsymbol{\gamma}^\top \tilde{\mathbf{X}}_i$$

where $\mathbb{E}[Y_i(t) \mid \tilde{\mathbf{X}}_i] = \mathbb{E}[Y_i \mid T_i = t, \tilde{\mathbf{X}}_i]$ by complete randomization (i.e., $\{Y_i(t), \mathbf{X}_i\} \perp T_i$)

Fully-interacted model (2): Imputation Estimator

- Consider treated unit (i.e., $T_i = 1$). Then, $Y_i = Y_i(1)$, and $Y_i(0)$ is missing.
  - But we can impute $\widehat{Y_i(0)}$ using the fully interacted model
- In this case, the estimated treatment effect for that unit (τ_i) is written as

$$\hat{\tau}_i = Y_i(1) - \widehat{Y_i(0)} = Y_i - \hat{\alpha} - \hat{\gamma}^\top \tilde{\mathbf{X}}_i$$

- On the other hand, for the control unit,

$$\hat{\tau}_i = \widehat{Y_i(1)} - Y_i(0) = \hat{\alpha} + \hat{\beta} + \hat{\gamma}^\top \tilde{\mathbf{X}}_i + \hat{\boldsymbol{\delta}}^\top \tilde{\mathbf{X}}_i - Y_i$$

Fully-interacted model (3): Imputation Estimator

- Thus, the ATE estimator is written as

$$\frac{1}{N}\sum_{i=1}^{n}\hat{\tau}_i = \overbrace{\frac{1}{N}\sum_{i=1}^{n}\left\{ T_i\left(Y_i(1) - \widehat{Y_i(0)} \right) + (1-T_i)\left(\widehat{Y_i(1)} - Y_i(0) \right) \right\}}^{\text{Imputation Estimator}}$$

$$= \frac{1}{N}\sum_{i=1}^{n}\left\{ T_i(Y_i - \hat{\alpha} - \hat{\boldsymbol{\gamma}}^\top \tilde{\mathbf{X}}_i) + (1-T_i)(\hat{\alpha} + \hat{\beta} + \hat{\boldsymbol{\gamma}}^\top \tilde{\mathbf{X}}_i + \hat{\boldsymbol{\delta}}^\top \tilde{\mathbf{X}}_i - Y_i) \right\}$$

- You can show $\frac{1}{N}\sum_{i=1}^{n}\hat{\tau}_i = \hat{\beta}$ (below I show it)
  - Only takeaway is "regression gives you imputation estimator"
- Recall that residuals are orthogonal with regressor, which suggests

$$\underbrace{\begin{bmatrix} 1 & T_1 & \tilde{\mathbf{X}}_1 & T_1\tilde{\mathbf{X}}_1 \\ \vdots & & & \vdots \\ 1 & T_n & \tilde{\mathbf{X}}_n & T_n\tilde{\mathbf{X}}_n \end{bmatrix}^\top}_{=\text{Transpose of Regressor}} \underbrace{\begin{bmatrix} Y_1 - \hat{\alpha} - \hat{\beta}T_1 - \hat{\boldsymbol{\gamma}}^\top\tilde{\mathbf{X}}_1 - \hat{\boldsymbol{\delta}}^\top T_1\tilde{\mathbf{X}}_1 \\ \vdots \\ Y_n - \hat{\alpha} - \hat{\beta}T_n - \hat{\boldsymbol{\gamma}}^\top\tilde{\mathbf{X}}_n - \hat{\boldsymbol{\delta}}^\top T_n\tilde{\mathbf{X}}_n \end{bmatrix}}_{=\text{Residuals}} = 0$$

Fully-interacted model (4): Imputation Estimator

- The previous matrix calculation gives you

$$\begin{cases} \sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta} T_i - \hat{\gamma}^\top \tilde{\mathbf{X}}_i - \hat{\boldsymbol{\delta}}^\top T_i \tilde{\mathbf{X}}_i) = 0 \\ \sum_{i=1}^{n} T_i(Y_i - \hat{\alpha} - \hat{\beta} T_i - \hat{\gamma}^\top \tilde{\mathbf{X}}_i - \hat{\boldsymbol{\delta}}^\top \tilde{\mathbf{X}}_i) = 0 \end{cases}$$

- The previous two formulas give you

$$\sum_{i=1}^{n}(1 - T_i)(Y_i - \hat{\alpha} - \hat{\beta} T_i - \hat{\gamma}^\top \tilde{\mathbf{X}}_i - \hat{\boldsymbol{\delta}}^\top T_i \tilde{\mathbf{X}}_i) = 0$$

- Therefore, by noticing $T_i(1 - T_i) = 0$,

$$\sum_{i=1}^{n} T_i(Y_i - \hat{\alpha} - \hat{\beta} - \hat{\gamma}^\top \tilde{\mathbf{X}}_i - \hat{\boldsymbol{\delta}}^\top \tilde{\mathbf{X}}_i) = 0$$

$$\sum_{i=1}^{n}(1 - T_i)(Y_i - \hat{\alpha} - \hat{\gamma}^\top \tilde{\mathbf{X}}_i) = 0$$

Fully-interacted model (5): Imputation Estimator

- Using these equalities, the expression above simplifies to

$$\frac{1}{N}\sum_{i=1}^{N}\hat{\tau}_i = \frac{1}{N}\sum_{i=1}^{N}(\hat{\beta} + \hat{\boldsymbol{\delta}}^\top \tilde{\mathbf{X}}_i) = \hat{\beta}$$

where tha last equality is because $\tilde{\mathbf{X}}_i$ is demeaned.

- **Takeaway**: Fully-interacted model gives you imputation estimator
  - You can similarly show that it can be interpreted as projection estimator

$$\hat{\beta} = \frac{1}{N}\sum_{i=1}^{N}\left\{\widehat{Y_i(1)} - \widehat{Y_i(0)}\right\}$$

- These are about the interpretation of $\hat{\beta}$ (coefficient in regression)

Fully-interacted model (6): Misspecification

- **Takeaway 1**: $\hat{\beta}$ from fully-interacted model is consistent for PATE even if model is incorrect
  - You can prove it using the same logic as in no-interaction case

- **Takeaway 2**: $\hat{\beta}$ from fully-interacted model is at least as efficient as the difference-in-means estimator even if model is incorrect
  - In `estimatr` package, there is `lm_lin` function for the implementation

- But these are under complete randomization!

Regression Adjustment under Stratified Design

- **Stratified Design**: Randomize treatment *within* strata
  - We have so far assumed complete randomization
  - Complete randomization $\rightarrow$ covariates and treatment are independent
  - Under stratified design, strata and treatment can be correlated
- Linear model with strata fixed effects

$$Y_i = \alpha_{\text{strata}_i} + \beta T_i + \epsilon_i$$

- The estimator $\hat{\beta}$ converges to the weighted average of strata specific ATE:

$$\hat{\beta} \xrightarrow{p} \frac{\sum_{j=1}^{J} w_j \overbrace{k_j(1-k_j)}^{\text{Strata } j\text{'s weight}} \cdot \overbrace{\mathbb{E}[Y_i(1) - Y_i(0) \mid \text{strata}_i = j]}^{\text{Strata } j\text{'s ATE}}}{\sum_{j=1}^{J} w_j k_j (1-k_j)}$$

where $w_j = n_j/n$ and $k_j = n_{j1}/n_j$

- $\hat{\beta}$ converges to PATE either (i) when k_j is identical across strata or (ii) when strata-specific ATE is identical across strata

Cluster Randomized Trials

- We assume *no interference* (implied by consistency)
  - Interference: potential outcome is function of other people's treatment status
- **Cluster randomized experiment**: assign treatment at the cluster level
  - We allow spillover *within* each cluster
  - We assume no spillover *across* clusters
- Notation
  - $j \in \{1, \cdots, J\}$: cluster indicator
  - $i \in \{1, \cdots, m_j\}$: individual indicator
  - T_j: treatment indicator for cluster j
  - Y_{ij}: outcome for individual i at cluster j
- Because everyone in each cluster is in the same treatment status,

$$\underbrace{Y_{ij}(T_{1j}, \cdots, T_{m_j j})}_{\text{Allowing interference within cluster } j} = Y_{ij}(T_j)$$

Efficiency Loss by Clustering

- Inference under clustering: regarding each cluster as unit

ATE Estimator: $\hat{\tau}_{\text{cluster}} = \underbrace{\frac{1}{J_1}\sum_{j=1}^{J} T_j \bar{Y}_j}_{\text{Avg. of Treated Cluster}} - \underbrace{\frac{1}{J_0}\sum_{j=1}^{J}(1-T_j)\bar{Y}_j}_{\text{Avg. of Control Cluster}}$

Variance Estimator: $\widehat{\mathbb{V}[\hat{\tau}_{\text{cluster}}]} = \frac{\text{var}(\bar{Y}_1(1))}{J_1} + \frac{\text{var}(\bar{Y}_0(t))}{J_0}$

- From the slide 5 of module 4.1, if we assume $m_j = m$ for all j,

$$\underbrace{\frac{\text{var}(\bar{Y}_j(t))}{J_t}}_{\text{Var w/ Clustering for } T_i=t} = \underbrace{\frac{\text{var}(Y_{ij}(t))}{J_t m}}_{\text{Variance w/o Clustering for } T_i=t} \underbrace{\Big(1+(m-1)\rho_t\Big)}_{\geq 1 \text{ if ICC is positive}}$$

where $\rho_t = \text{Corr}(Y_{ij}(t), Y_{i'j}(t))$ is ICC

- ICC is typically positive $\rightarrow$ Clustering typically loses efficiency

Extra Slides: Cluster Robust Standard Errors (CR)

- As with the previous case,

$$\sqrt{n}(\hat{\beta}-\beta) = \sqrt{n}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\epsilon} = \left(\frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{m_j}\mathbf{X}_{ij}\mathbf{X}_{ij}^\top\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{j=1}^{J}\sum_{i=1}^{m_j}\mathbf{X}_{ij}\epsilon_{ij}\right)$$

- Independence holds across clusters (not within clusters!)
  - Slight modification is needed for the previous proof (regard each cluster as unit)
- **Cluster Robust Variance Estimator**

$$\underbrace{\left(\sum_{j=1}^{J}\mathbf{X}_j^\top\mathbf{X}_j\right)^{-1}}_{\text{Bread}}\underbrace{\left(\sum_{j=1}^{J}\mathbf{X}_j^\top\hat{\boldsymbol{\epsilon}}_j\hat{\boldsymbol{\epsilon}}_j{}^\top\mathbf{X}_j\right)}_{\text{Meat}}\underbrace{\left(\sum_{j=1}^{J}\mathbf{X}_j^\top\mathbf{X}_j\right)^{-1}}_{\text{Bread}}$$

- **CR2** (Bias-adjustment): Same idea as HC2
  - Same degrees of freedom adjustment is available for small sample