Introduction to Causal Inference

Lecture 3

Regression Methods for Completely Randomized Experiments

邓婉璐

WANLUDENG@TSINGHUA.EDU.CN



清华大学统计与数据科学系

Completely Randomized Experiment

- A leading case of a classical randomized experiment
- The number of treated units N_t is fixed
- N_t units are randomly selected, from a population of N units, to receive the active treatment, with the remaining N_c assigned to the control group.
- Each unit has unit assignment probability $q = N_t / N$
- The assignment mechanism:

$$\Pr(\mathbf{W}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = \begin{cases} 1 \Big/ \binom{N}{N_t} & \text{if } \sum_{i=1}^{N} W_i = N_t, \\ 0 & \text{otherwise,} \end{cases}$$

$$\binom{N}{N_t} = \frac{N!}{N_t!(N - N_t)!}$$

- subset of the set of possible values for $\mathbf{W}$ with positive probability

$$\mathbb{W}^+ = \left\{ \mathbf{W} \in \mathbb{W} \,\middle|\, \sum_{i=1}^{N} W_i = N_t \right\}$$



LRC-CPPT Cholesterol Data Analysis

Table 7.1. ***Summary Statistics for PRC-CPPT Cholesterol Data***

	Variable	Control (N_c =172)		Treatment (N_t =165)			
		Average	Sample (S.D.)	Average	Sample (S.D.)	Min	Max
Pre-treatment	chol1	297.1	(23.1)	297.0	(20.4)	247.0	442.0
	chol2	289.2	(24.1)	287.4	(21.4)	224.0	435.0
	cholp	291.2	(23.2)	289.9	(20.4)	233.0	436.8
Post-treatment	cholf	282.7	(24.9)	256.5	(26.2)	167.0	427.0
	chold	−8.5	(10.8)	−33.4	(21.3)	−113.3	29.5
	comp	74.5	(21.0)	59.9	(24.4)	0	101.0

Y, W, X

Thinking: why do we need placebo?



Is a Covariate Useful?

Leading terms of the variance → Reduce them can reduce the variance

Theorem 6.2 *The sampling variance of* $\hat{\tau}^{\text{dif}} = \overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}}$ *is*

$$\mathbb{V}_W\left(\overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}}\right) = \frac{S_c^2}{N_{\text{c}}} + \frac{S_t^2}{N_{\text{t}}} - \frac{S_{tc}^2}{N},$$

variance of $Y_i(0)$ *in the finite population*

variance of the unit-level treatment effects in the finite population

$$S_c^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i(0) - \overline{Y}(0)\right)^2,$$

$$S_t^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i(1) - \overline{Y}(1)\right)^2,$$

$$S_{tc}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i(1) - Y_i(0) - (\overline{Y}(1) - \overline{Y}(0))\right)^2$$

$$= \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i(1) - Y_i(0) - \tau_{\text{fs}}\right)^2.$$

variance of $Y_i(1)$ *in the finite population*

Is a Covariate Useful?

- Scenario A: the covariate is totally independent to the potential outcomes

- Scenario B: the covariate is predictive to or highly correlated with the potential outcomes (and thus, Y^{obs})



Inference of CREs

- Fisher’s Exact P-Values
  - whether there exists a non-zero unit-level causal effect?
- Neyman’s Repeated Sampling Approach
  - how to estimate the average causal effect?
- Regression Methods
  - how covariates help estimate the average causal effect?

Outline

- Settings and Notations
- Linear regression with no covariates
- Linear regression with covariates
- Extensions
  - Linear regression with covariates and interactions
  - Testing
- Summary

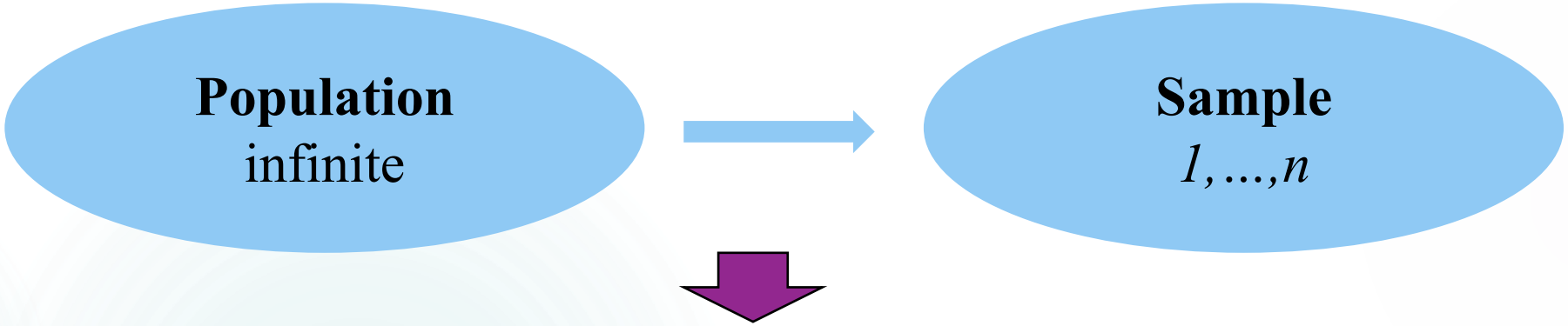
- Application

Settings and Notations



Problem Settings

- An infinite super-population of units.



- The potential outcomes in the sample are random.
- Properties of the estimators are assessed by resampling from that population, sometimes conditional on the predictor variables including the treatment indicator.

Notations

$$\tau_{\text{fs}} = \frac{1}{N}\sum_{i=1}^{N}(Y_i(1) - Y_i(0)) \qquad \tau_{\text{sp}} = \mathbb{E}_{\text{sp}}\left[Y_i(1) - Y_i(0)\right]$$

$$\mu_c(x) = \mathbb{E}_{\text{sp}}\left[Y_i(0)|X_i = x\right], \quad \mu_t(x) = \mathbb{E}_{\text{sp}}\left[Y_i(1)|X_i = x\right],$$

$$\sigma_c^2(x) = \mathbb{V}_{\text{sp}}\left(Y_i(0)|X_i = x\right), \quad \sigma_t^2(x) = \mathbb{V}_{\text{sp}}\left(Y_i(0)|X_i = x\right),$$

$$\tau(x) = \mathbb{E}_{\text{sp}}(Y_i(1) - Y_i(0)|X_i = x] \quad \sigma_{ct}^2(x) = \mathbb{V}_{\text{sp}}\left(Y_i(1) - Y_i(0)|X_i = x\right)$$

$$\mu_c = \mathbb{E}_{\text{sp}}\left[\mu_c(X_i)\right], \qquad \mu_t = \mathbb{E}_{\text{sp}}\left[\mu_t(X_i)\right],$$

$$\sigma_c^2 = \mathbb{E}_{\text{sp}}\left[\sigma_c^2(X_i)\right] + \mathbb{V}_{\text{sp}}\left(\mu_c(X_i)\right) \quad \sigma_t^2 = \mathbb{E}_{\text{sp}}\left[\sigma_t^2(X_i)\right] + \mathbb{V}_{\text{sp}}\left(\mu_t(X_i)\right)$$

$$\mu_X = \mathbb{E}_{\text{sp}}\left[X_i\right] \qquad \Omega_X = \mathbb{V}_{\text{sp}}(X_i) = \mathbb{E}_{\text{sp}}\left[(X_i - \mu_X)^T(X_i - \mu_X)\right]$$



LINEAR REGRESSION WITH NO COVARIATES



LRC-CPPT Cholesterol Data Analysis

Table 7.1. ***Summary Statistics for PRC-CPPT Cholesterol Data***

	Variable	Control (N_c =172)		Treatment (N_t =165)			
		Average	Sample (S.D.)	Average	Sample (S.D.)	Min	Max
Pre-treatment	chol1	297.1	(23.1)	297.0	(20.4)	247.0	442.0
	chol2	289.2	(24.1)	287.4	(21.4)	224.0	435.0
	cholp	291.2	(23.2)	289.9	(20.4)	233.0	436.8
Post-treatment	cholf	282.7	(24.9)	256.5	(26.2)	167.0	427.0
	chold	−8.5	(10.8)	−33.4	(21.3)	−113.3	29.5
	comp	74.5	(21.0)	59.9	(24.4)	0	101.0

Y, W, X

Y	W	X
250	1	⋮
280	0	⋮
265	0	⋮
⋮	⋮	⋮

Estimator

Regression model: $Y_i^{\text{obs}} = \alpha + \tau \cdot W_i + \varepsilon_i,$

Least square estimate: $(\hat{\tau}^{\text{ols}}, \hat{\alpha}^{\text{ols}}) = \arg\min_{\tau,\alpha} \sum_{i=1}^{N} \left(Y_i^{\text{obs}} - \alpha - \tau \cdot W_i\right)^2$

$$\hat{\tau}^{\text{ols}} = \frac{\sum_{i=1}^{N} (W_i - \overline{W}) \cdot (Y_i^{\text{obs}} - \overline{Y}^{\text{obs}})}{\sum_{i=1}^{N} (W_i - \overline{W})^2}, \quad \text{and} \quad \hat{\alpha}^{\text{ols}} = \overline{Y}^{\text{obs}} - \hat{\tau}^{\text{ols}} \cdot \overline{W},$$

$$\sum_{i=1}^{N} W_i Y_i^{obs} - N\overline{W}\overline{Y}$$

$$\sum_{i=1}^{N} W_i - N\overline{W}^2 = \frac{N_t N_c}{N}$$

$$\overline{Y}^{\text{obs}} = \frac{1}{N}\sum_{i=1}^{N} Y_i^{\text{obs}}$$

$$\overline{W} = \frac{1}{N}\sum_{i=1}^{N} W_i = \frac{N_t}{N}$$

Interpretation (1): its Causal Meaning

Regression model: $Y_i^{\text{obs}} = \alpha + \tau \cdot W_i + \varepsilon_i,$

Least square estimate: $(\hat{\tau}^{\text{ols}}, \hat{\alpha}^{\text{ols}}) = \arg\min_{\tau,\alpha} \sum_{i=1}^{N} \left(Y_i^{\text{obs}} - \alpha - \tau \cdot W_i\right)^2$

$$\hat{\tau}^{\text{ols}} = \frac{\sum_{i=1}^{N} (W_i - \overline{W}) \cdot (Y_i^{\text{obs}} - \overline{Y}^{\text{obs}})}{\sum_{i=1}^{N} (W_i - \overline{W})^2}, \quad \text{and} \quad \hat{\alpha}^{\text{ols}} = \overline{Y}^{\text{obs}} - \hat{\tau}^{\text{ols}} \cdot \overline{W},$$

$$\hat{\tau}^{\text{ols}} = \overline{Y}_{\text{t}}^{\text{obs}} - \overline{Y}_{\text{c}}^{\text{obs}} = \hat{\tau}^{\text{dif}}$$

An unbiased estimator of τ_{sp} of in CREs

$$\overline{Y}_{\text{c}}^{\text{obs}} = \sum_{i:W_i=0} Y_i^{\text{obs}}/N_{\text{c}}$$
$$\overline{Y}_{\text{t}}^{\text{obs}} = \sum_{i:W_i=1} Y_i^{\text{obs}}/N_{\text{t}}$$

The key from association statement to causal statement

Justification for Linear Regression

$$\alpha = \mu_c = \mathbb{E}_{\text{sp}}[Y_i(0)] \qquad \tau_{\text{sp}} = \mu_t - \mu_c = \mathbb{E}_{\text{sp}}[Y_i(1) - Y_i(0)]$$

Regression model: $Y_i^{\text{obs}} = \alpha + \tau \cdot W_i + \varepsilon_i,$ $\varepsilon_i = \begin{cases} Y_i(0) - \mu_c & \text{if } W_i = 0 \\ Y_i(1) - \mu_t & \text{if } W_i = 1 \end{cases}$

Completely randomized experiment

$W_i \perp\!\!\!\perp (Y_i(0), Y_i(1))$

$\mathbb{E}_{\text{sp}}[\varepsilon_i | W_i = w] = 0$ for $w = 0, 1$

$\mathbf{V}_{sp}[\varepsilon_i | W_i] = \sigma_t^2 W_i + \sigma_c^2 (1 - W_i)$

If $\sigma_t^2 = \sigma_c^2 = \sigma^2$, $\mathbf{V}_{sp}[\varepsilon_i | W_i] = \sigma^2$

Standard regularity conditions for linear model

We can use all theoretical tools for inferring linear model coefficients (α, τ, σ^2)

Inference

$$\alpha = \mu_c = \mathbb{E}_{\text{sp}}[Y_i(0)] \qquad \tau_{\text{sp}} = \mu_t - \mu_c = \mathbb{E}_{\text{sp}}[Y_i(1) - Y_i(0)]$$

Regression model: $Y_i^{\text{obs}} = \alpha + \tau \cdot W_i + \varepsilon_i,$ $\varepsilon_i = \begin{cases} Y_i(0) - \mu_c & \text{if } W_i = 0 \\ Y_i(1) - \mu_t & \text{if } W_i = 1 \end{cases}$

$$Y^{obs} = X\beta + \varepsilon \quad \beta = (\alpha, \tau)^T \quad X = \begin{pmatrix} 1_{N_c} & 0_{N_c} \\ 1_{N_t} & 1_{N_t} \end{pmatrix} \quad \begin{matrix} \mathbb{E}[\varepsilon|X] = 0 \\ \mathbb{V}_{sp}[\varepsilon|X] = \Sigma = \text{diag}(\sigma_c^2 \cdot 1_{N_c}, \sigma_t^2 \cdot 1_{N_t}) \end{matrix}$$

Assume $\sigma_t^2 = \sigma_c^2 = \sigma^2$

$$\boldsymbol{Y^{obs}} = \boldsymbol{X\beta + \varepsilon} \quad \beta = (\alpha, \tau)^T \quad \boldsymbol{X} = \begin{pmatrix} 1_{N_c} & 0_{N_c} \\ 1_{N_t} & 1_{N_t} \end{pmatrix} \quad \begin{matrix} \mathbb{E}[\boldsymbol{\varepsilon}|\boldsymbol{X}] = \mathbb{E}[\boldsymbol{\varepsilon}] = \boldsymbol{0} \\ \mathbb{V}_{sp}[\boldsymbol{\varepsilon}|\boldsymbol{X}] = \Sigma = diag(\sigma^2, \cdots, \sigma^2) \end{matrix}$$

Inference: Assume $\sigma_t^2 = \sigma_c^2 = \sigma^2$

$$\boldsymbol{Y}^{obs} = \boldsymbol{X}\beta + \boldsymbol{\varepsilon} \quad \beta = (\alpha, \tau)^T \quad \boldsymbol{X} = \begin{pmatrix} 1_{N_c} & 0_{N_c} \\ 1_{N_t} & 1_{N_t} \end{pmatrix}$$

$$\mathbb{E}[\boldsymbol{\varepsilon}|\boldsymbol{X}] = \mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$$
$$\mathbb{V}_{sp}[\boldsymbol{\varepsilon}|\boldsymbol{X}] = \Sigma = diag(\sigma^2, \cdots, \sigma^2)$$

Asymptotically normal (think why)

$$\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Y^{obs} = \beta + H\boldsymbol{\varepsilon}$$

$$H = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = \begin{pmatrix} \frac{1_{N_c}}{N_c} & \frac{-1_{N_c}}{N_c} \\ \frac{0_{N_t}}{N_t} & \frac{1_{N_t}}{N_t} \end{pmatrix}^T$$

$$Var(\hat{\beta}) = H\Sigma H^T$$

$$Var(\hat{\tau}^{ols}) = \sigma^2\left(\frac{1}{N_c} + \frac{1}{N_t}\right) \quad \& \quad \hat{\sigma}^2 = \frac{1}{N-2}\sum_{i=1}^{N}\hat{\varepsilon}_i^2$$

$$\widehat{Var}(\hat{\tau}^{ols}) = \hat{\sigma}^2\left(\frac{1}{N_c} + \frac{1}{N_t}\right) \Leftrightarrow \hat{\mathbb{V}}^{\text{const}} = s^2 \cdot \left(\frac{1}{N_c} + \frac{1}{N_t}\right)$$

The inference based on regression model is identical to Neyman approach!

Inference: Allow $\sigma_t^2 \neq \sigma_c^2$

$$\boldsymbol{Y}^{obs} = \boldsymbol{X}\beta + \boldsymbol{\varepsilon} \quad \beta = (\alpha, \tau)^T \quad \boldsymbol{X} = \begin{pmatrix} 1_{N_c} & 0_{N_c} \\ 1_{N_t} & 1_{N_t} \end{pmatrix} \qquad \begin{aligned} \mathbb{E}[\varepsilon|\boldsymbol{X}] &= 0 \\ \mathbb{V}_{sp}[\varepsilon|\boldsymbol{X}] &= \Sigma = \mathrm{diag}(\sigma_c^2 \cdot 1_{N_c}, \sigma_t^2 \cdot 1_{N_t}) \end{aligned}$$

$$\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Y^{obs} = \beta + H\varepsilon$$

$$H = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = \begin{pmatrix} \frac{1_{N_c}}{N_c} & \frac{-1_{N_c}}{N_c} \\ \frac{0_{N_t}}{N_t} & \frac{1_{N_t}}{N_t} \end{pmatrix}^T$$

$$Var(\hat{\beta}) = H\Sigma H^T$$

$$Var(\hat{\tau}^{ols}) = \frac{\sigma_c^2}{N_c} + \frac{\sigma_t^2}{N_t}$$

&

$$\begin{cases} \hat{\sigma}_t{}^2 = \dfrac{1}{N_t - 1}\sum_{i=1}^{N} \hat{\varepsilon}_i^2 W_i \\ \hat{\sigma}_c{}^2 = \dfrac{1}{N_c - 1}\sum_{i=1}^{N} \hat{\varepsilon}_i^2 (1 - W_i) \end{cases}$$

$$\widehat{Var}(\hat{\tau}^{ols}) = \frac{\hat{\sigma}_c{}^2}{N_c} + \frac{\hat{\sigma}_t{}^2}{N_t} \quad \Leftrightarrow \quad \hat{\mathbb{V}}^{\text{neyman}} = \frac{s_c^2}{N_c} + \frac{s_t^2}{N_t}.$$

The inference based on regression model is identical to Neyman approach!

Inference: Discussion for Case $\sigma_t^2 \neq \sigma_c^2$

$$\boldsymbol{Y}^{obs} = \boldsymbol{X}\beta + \boldsymbol{\varepsilon} \quad \beta = (\alpha, \tau)^T \quad \boldsymbol{X} = \begin{pmatrix} 1_{N_c} & 0_{N_c} \\ 1_{N_t} & 1_{N_t} \end{pmatrix} \qquad \begin{aligned} \mathbb{E}[\varepsilon|\boldsymbol{X}] &= 0 \\ \mathbb{V}_{sp}[\varepsilon|\boldsymbol{X}] &= \Sigma = \mathrm{diag}(\sigma_c^2 \cdot 1_{N_c}, \sigma_t^2 \cdot 1_{N_t}) \end{aligned}$$

$\sigma_t^2 \neq \sigma_c^2$:

$$\widehat{Var}(\hat{\tau}^{ols}) = \frac{\hat{\sigma}_c{}^2}{N_c} + \frac{\hat{\sigma}_t{}^2}{N_t}$$

&

$$\begin{cases} \hat{\sigma}_t{}^2 = \dfrac{1}{N_t - 1}\sum_{i=1}^{N} \hat{\varepsilon}_i^2 W_i \\ \hat{\sigma}_c{}^2 = \dfrac{1}{N_c - 1}\sum_{i=1}^{N} \hat{\varepsilon}_i^2 (1 - W_i) \end{cases}$$

R: `car::hccm`

$\neq$

$\sigma_t^2 = \sigma_c^2$:

$$\widehat{Var}(\hat{\tau}^{ols}) = \hat{\sigma}^2\left(\frac{1}{N_c} + \frac{1}{N_t}\right)$$

&

$$\hat{\sigma}^2 = \frac{1}{N - 2}\sum_{i=1}^{N} \hat{\varepsilon}_i^2$$

R: `lm`

Aside:

$$\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Y^{obs} = \beta + H\varepsilon \qquad H = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$$

$$Var(\hat{\beta}) = H\Sigma H^T = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\Sigma\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \qquad \widehat{Var}(\hat{\beta}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathrm{diag}(\hat{\varepsilon}_1^2, \ldots, \hat{\varepsilon}_n^2)\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

The Eicker–Huber–White (EHW) robust covariance estimator
(also known as the heteroskedasticity-consistent covariance matrix estimator (HCCME)):
a moment estimator for the asymptotic variance of $\hat{\beta}$

Interpretation (2): under Imputation

- Key challenge: Missing data
- Another interpretation, in lights of imputation-based methods, the topic of the next lecture.

$$\mathbb{E}_{\mathrm{SP}}[Y_i^{obs} | W_i = w]$$

unconfoundedness

$$\mathbb{E}_{\mathrm{SP}}[Y_i(0)]$$

$$\mathbb{E}_{\mathrm{SP}}[Y_i(1)]$$

$$\hat{\tau}_i = Y_i(1) - \hat{Y}_i(0) \quad \text{if } W_i = 1$$

$$\hat{\tau}_i = \hat{Y}_i(1) - Y_i(0) \quad \text{if } W_i = 0$$

$$\frac{1}{N}\sum_{i=1}^{N}\hat{\tau}_i = \frac{1}{N}\sum_{i=1}^{N}\left\{W_i \cdot \left(Y_i(1) - \hat{Y}_i(0)\right) + (1 - W_i) \cdot \left(\hat{Y}_i(1) - Y_i(0)\right)\right\}$$

Interpretation (3): under Limitation

Regression model: $Y_i^{\text{obs}} = \alpha + \tau \cdot W_i + \varepsilon_i,$

Least square estimate: $(\hat{\tau}^{\text{ols}}, \hat{\alpha}^{\text{ols}}) = \arg\min_{\tau,\alpha} \sum_{i=1}^{N} \left(Y_i^{\text{obs}} - \alpha - \tau \cdot W_i\right)^2$

Least square estimate when sample size goes to infinity:

$$(\alpha^*, \tau^*) = \arg\min_{\alpha,\tau} \mathbb{E}_{\text{sp}} \left[\frac{1}{N} \sum_{i=1}^{N} \left(Y_i^{\text{obs}} - \alpha - \tau \cdot W_i\right)^2\right]$$

$$= \arg\min_{\alpha,\tau} \mathbb{E}_{\text{sp}} \left[\left(Y_i^{\text{obs}} - \alpha - \tau \cdot W_i\right)^2\right].$$

$$\tau^* = \mathbb{E}_{\text{sp}}[Y_i^{\text{obs}} | W_i = 1] - \mathbb{E}_{\text{sp}}[Y_i^{\text{obs}} | W_i = 0]$$

Completely randomized experiment

$$\tau^* = \tau_{\text{sp}}$$



Regression Methods in CREs

- Advantages of regression methods
  - ✓ Well known & easy to understand
  - ✓ Reach to identical results to Neyman approach
  - ✓ Easy to extend to scenarios with covariates

- Critical thinking
  - ✓ Regression method only provide association statements
  - ✓ We can run regression to any data & get some results
  - ✓ In general the results do **not** have causal meaning
  - ✓ CREs warrantee the causal meaning of the results

Linear Regression with Covariates



LRC-CPPT Cholesterol Data Analysis

Table 7.1. ***Summary Statistics for PRC-CPPT Cholesterol Data***

	Variable	Control (N_c =172)		Treatment (N_t =165)			
		Average	Sample (S.D.)	Average	Sample (S.D.)	Min	Max
Pre-treatment	chol1	297.1	(23.1)	297.0	(20.4)	247.0	442.0
	chol2	289.2	(24.1)	287.4	(21.4)	224.0	435.0
	cholp	291.2	(23.2)	289.9	(20.4)	233.0	436.8
Post-treatment	cholf	282.7	(24.9)	256.5	(26.2)	167.0	427.0
	chold	−8.5	(10.8)	−33.4	(21.3)	−113.3	29.5
	comp	74.5	(21.0)	59.9	(24.4)	0	101.0

Y, W, X

Y	W	X
250	1	⋮
280	0	⋮
265	0	⋮
⋮	⋮	⋮

Model and Estimators

Some covariates, could be a vector

Regression model: $Y_i^{\text{obs}} = \alpha + \tau \cdot W_i + X_i\beta + \varepsilon_i$

Least square estimate: $(\hat{\tau}^{\text{ols}}, \hat{\alpha}^{\text{ols}}, \hat{\beta}^{\text{ols}}) = \arg\min_{\tau,\alpha,\beta} \sum_{i=1}^{N} \left(Y_i^{\text{obs}} - \alpha - \tau \cdot W_i - X_i\beta\right)^2$

Interpretation (1): under Limitation

Regression model: $Y_i^{\text{obs}} = \alpha + \tau \cdot W_i + X_i\beta + \varepsilon_i$

Least square estimate: $(\hat{\tau}^{\text{ols}}, \hat{\alpha}^{\text{ols}}, \hat{\beta}^{\text{ols}}) = \arg\min_{\tau,\alpha,\beta} \sum_{i=1}^{N} \left(Y_i^{\text{obs}} - \alpha - \tau \cdot W_i - X_i\beta\right)^2$

Least square estimate when sample size goes to infinity: $(\alpha^*, \tau^*, \beta^*) = \arg\min_{\alpha,\beta,\tau} \mathbb{E}\left[\left(Y_i^{\text{obs}} - \alpha - \tau \cdot W_i - X_i\beta\right)^2\right]$



$(\hat{\tau}^{ols}, \hat{\alpha}^{ols}, \hat{\beta}^{ols}) \to (\tau^*, \alpha^*, \beta^*)$

Consistency of τ^*

- Objective function: $Q(\alpha, \tau, \beta) = \mathbb{E}[(Y_i^{\text{obs}} - \alpha - \tau \cdot W_i - X_i\beta)^2]$

$$= \mathbb{E}\left[\left(Y_i^{\text{obs}} - \tilde{\alpha} - \tau \cdot W_i - (X_i - \mu_X)\beta\right)^2\right]$$

$$Q(\tilde{\alpha}, \tau, \beta) = \mathbb{E}_{\text{sp}}\left[\left(Y_i^{\text{obs}} - \tilde{\alpha} - \tau \cdot W_i - (X_i - \mu_X)\beta\right)^2\right]$$

$$= \mathbb{E}_{\text{sp}}\left[\left(Y_i^{\text{obs}} - \tilde{\alpha} - \tau \cdot W_i\right)^2\right] + \mathbb{E}_{\text{sp}}\left[((X_i - \mu_X)\beta)^2\right]$$

$$- 2 \cdot \mathbb{E}_{\text{sp}}\left[\left(Y_i^{\text{obs}} - \tilde{\alpha} - \tau \cdot W_i\right) \cdot (X_i - \mu_X)\beta\right]$$

$$= \mathbb{E}_{\text{sp}}\left[\left(Y_i^{\text{obs}} - \tilde{\alpha} - \tau \cdot W_i\right)^2\right] + \mathbb{E}_{\text{sp}}\left[((X_i - \mu_X)\beta)^2\right]$$

$$- 2 \cdot \mathbb{E}_{\text{sp}}\left[Y_i^{\text{obs}} \cdot (X_i - \mu_X)\beta\right],$$

$$\mathbb{E}_{\text{sp}}[(X_i - \mu_X)\beta] = 0,$$

$$\mathbb{E}_{\text{sp}}[\tau \cdot W_i \cdot (X_i - \mu_X)\beta] = 0,$$

Consistency and Asymptotic Normality

Theorem 7.1 Suppose we conduct a completely randomized experiment in a sample drawn at random from an infinite population. Then

$$\sigma_i^2 = Var(\varepsilon_i|W_i, X_i) = Var(Y_i(W_i)|X_i)$$

(1) $\tau^* = \tau_{\text{sp}}$,

(2) $$\sqrt{N} \cdot \left(\hat{\tau}^{\text{ols}} - \tau_{\text{sp}}\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\mathbb{E}\left[(W_i - p)^2 \cdot \left(Y_i^{\text{obs}} - \alpha^* - \tau_{\text{sp}} \cdot W_i - X_i\beta^*\right)^2\right]}{p^2 \cdot (1-p)^2}\right)$$

Smaller than $\hat{\mathbb{V}}^{\text{neyman}} = \frac{s_c^2}{N_c} + \frac{s_t^2}{N_t}$, if covariates X_i are predictive to Y_i^{obs}

Note:

- The Theorem is still correct even if the regression model is wrong.

Model Comparisons

Regression model without X:

$$\hat{\sigma}^2_{Y|W} = \frac{1}{N-2}\sum_{i=1}^{N}\hat{\varepsilon}_i^2 = \frac{1}{N-2}\sum_{i=1}^{N}\left(Y_i^{\text{obs}} - \hat{Y}_i^{\text{obs}}\right)^2$$

$$= \frac{1}{N-2}\left(\sum_{i:W_i=0}\left(Y_i^{\text{obs}} - \overline{Y}_c^{\text{obs}}\right)^2 + \sum_{i:W_i=1}\left(Y_i^{\text{obs}} - \overline{Y}_t^{\text{obs}}\right)^2\right),$$

Variance:

Assume the variance of two potential outcomes does not vary by treatment.

$$\hat{\mathbb{V}}^{\text{homosk}} = \frac{\hat{\sigma}^2_{Y|W}}{\sum_{i=1}^{N}\left(W_i - \overline{W}\right)^2} = s^2 \cdot \left(\frac{1}{N_c} + \frac{1}{N_t}\right).$$

When sample size goes to infinity:

$$N \cdot \hat{\mathbb{V}}^{\text{homosk}} \xrightarrow{p} \frac{\sigma^2_{Y|W}}{p \cdot (1-p)}.$$

$$p = \lim_{N\to\infty}\frac{N_t}{N}$$

Regression model with X:

$$\mathbb{E}_{\text{sp}}\left[Y_i^{\text{obs}} \middle| X_i = x, W_i = t\right] = \alpha_c + \tau_{\text{sp}} \cdot t + \beta' x,$$

Variance:

Assume the variance of two potential outcomes does not vary by treatment or covariates:

$$\mathbb{V}_{\text{sp}}(Y_i(w)|X_i = x) = \sigma^2_{Y|W,X},$$

for $w = 0, 1$, and all x.

When sample size goes to infinity:

$$N \cdot \mathbb{V}_{\text{sp}}^{\text{homosk}} = \frac{\sigma^2_{Y|W,X}}{p \cdot (1-p)}.$$

Using a special case to see the gain in precision from including the covariates.

Key of the proof (not required)

Reparametrization: $(\alpha, \tau, \beta) \longrightarrow (\tilde{\alpha}, \tau, \beta)$ $\quad \tilde{\alpha} = \alpha - p \cdot \tau - \mathbb{E}_{\text{sp}}[X_i]\beta.$

$$\sum_{i=1}^{N} \left(Y_i^{\text{obs}} - \left(\tilde{\alpha} - p \cdot \tau - \mathbb{E}_{\text{sp}}[X_i]\beta\right) - \tau \cdot W_i - X_i\beta \right)^2$$

$$= \sum_{i=1}^{N} \left(Y_i^{\text{obs}} - \tilde{\alpha} - \tau \cdot (W_i - p) - \left(X_i - \mathbb{E}_{\text{sp}}[X_i]\right)\beta \right)^2$$

Estimation:

$$\sum_{i=1}^{N} \psi(Y_i^{\text{obs}}, W_i, X_i, \hat{\tilde{\alpha}}^{\text{ols}}, \hat{\tau}^{\text{ols}}, \hat{\beta}^{\text{ols}}) = 0,$$

$$\psi(y, w, x, \alpha, \tau, \beta) = \begin{pmatrix} y - \alpha - \tau \cdot (w - p) - \left(x - \mathbb{E}_{\text{sp}}[X_i]\right)\beta \\ (w - p) \cdot \left(y - \alpha - \tau \cdot (w - p) - \left(x - \mathbb{E}_{\text{sp}}[X_i]\right)\beta\right) \\ \left(x - \mathbb{E}_{\text{sp}}[X_i]\right) \cdot \left(y - \alpha - \tau \cdot (w - p) - \left(x - \mathbb{E}_{\text{sp}}[X_i]\right)\beta\right) \end{pmatrix}$$



Key of the proof (not required)

Asymptotically normal:

$$\sqrt{N}\cdot\begin{pmatrix}\hat{\tilde{\alpha}}^{\text{ols}}-\alpha^*\\ \hat{\tau}^{\text{ols}}-\tau_{\text{sp}}\\ \hat{\beta}^{\text{ols}}-\beta^*\end{pmatrix}\xrightarrow{d}\mathcal{N}\left(\begin{pmatrix}0\\0\\0\end{pmatrix},\Gamma^{-1}\Delta(\Gamma^T)^{-1}\right)$$

$$\Delta=\mathbb{E}_{\text{sp}}\left[\psi(Y_i^{\text{obs}},W_i,X_i,\tilde{\alpha}^*,\tau_{\text{sp}},\beta^*)\cdot\psi(Y_i^{\text{obs}},W_i,X_i,\tilde{\alpha}^*,\tau_{\text{sp}},\beta^*)^T\right]$$

$$=\mathbb{E}_{\text{sp}}\left[\left(Y_i^{\text{obs}}-\alpha^*-\tau_{\text{sp}}-X_i\beta^*\right)^2\cdot\begin{pmatrix}1\\W_i-p\\(X_i-\mathbb{E}_{\text{sp}}[X_i])^T\end{pmatrix}\begin{pmatrix}1\\W_i-p\\(X_i-\mathbb{E}_{\text{sp}}[X_i])^T\end{pmatrix}^T\right]$$

$$\Gamma=\mathbb{E}_{\text{sp}}\left[\frac{\partial}{\partial(\alpha,\tau,\beta)}\psi(Y_i^{\text{obs}},W_i,X_i,\alpha,\tau,\beta)\right]\Big|_{(\tilde{\alpha}^*,\tau_{\text{sp}},\beta^*)}$$

$$=\mathbb{E}_{\text{sp}}\left[\begin{pmatrix}-1 & -(W_i-p) & -(X_i-\mathbb{E}_{\text{sp}}[X_i])\\ -(W_i-p) & -(W_i-p)^2 & -(W_i-p)\cdot(X_i-\mathbb{E}_{\text{sp}}[X_i])\\ -(X_i-\mathbb{E}_{\text{sp}}[X_i])^T & -(W_i-p)\cdot(X_i-\mathbb{E}_{\text{sp}}[X_i])^T & -(X_i-\mathbb{E}_{\text{sp}}[X_i])^T\cdot(X_i-\mathbb{E}_{\text{sp}}[X_i])\end{pmatrix}\right]$$

$$=\mathbb{E}_{\text{sp}}\left[\begin{pmatrix}-1 & 0 & 0\\ 0 & -p(1-p) & 0\\ 0 & 0 & -\mathbb{E}_{\text{sp}}\left[(X_i-\mathbb{E}_{\text{sp}}[X_i])^T\cdot(X_i-\mathbb{E}_{\text{sp}}[X_i])\right]\end{pmatrix}\right]$$

Review: Consistency and Asymptotic Normality

Theorem 7.1 Suppose we conduct a completely randomized experiment in a sample drawn at random from an infinite population. Then

$$\sigma_i^2 = Var(\varepsilon_i|W_i, X_i) = Var(Y_i(W_i)|X_i)$$

(1) $\tau^* = \tau_{\text{sp}}$,

(2) $$\sqrt{N} \cdot \left(\hat{\tau}^{\text{ols}} - \tau_{\text{sp}}\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\mathbb{E}\left[(W_i - p)^2 \cdot \left(Y_i^{\text{obs}} - \alpha^* - \tau_{\text{sp}} \cdot W_i - X_i\beta^*\right)^2\right]}{p^2 \cdot (1-p)^2}\right)$$

Inference

- Estimator for the sampling variance for the average treatment effect
- Heteroskedasticity:

$$\widehat{\mathbb{V}}_{\text{SP}}^{\text{hetero}} = \frac{1}{N(N-1-\dim(X_i))} \cdot \frac{\sum_{i=1}^{N}(W_i-\bar{W})^2 \cdot \left(Y_i^{obs} - \hat{\alpha}^{ols} - \hat{\tau}^{ols}W_i - X_i\hat{\beta}^{ols}\right)^2}{(\bar{W}\cdot(1-\bar{W}))^2}$$

- Homoskedasticity

$$\widehat{\mathbb{V}}_{\text{SP}}^{\text{homo}} = \frac{1}{N(N-1-\dim(X_i))} \cdot \frac{\sum_{i=1}^{N}\left(Y_i^{obs} - \hat{\alpha}^{ols} - \hat{\tau}^{ols}W_i - X_i\hat{\beta}^{ols}\right)^2}{\bar{W}\cdot(1-\bar{W})}.$$



Interpretation (2): under Imputation

- This LSE can also be interpreted with imputation method.

$$\mathbb{E}_{\text{sp}}\left[Y_i^{\text{obs}} \middle| X_i = x, W_i = w\right] =$$

unconfoundedness

$$\mathbb{E}_{\text{sp}}\left[Y_i(0) \middle| X_i = x\right] =$$

$$\mathbb{E}_{\text{sp}}\left[Y_i(1) \middle| X_i = x\right] =$$

$$\hat{\tau}_i = Y_i(1) - \hat{Y}_i(0) \quad \text{if } W_i = 1$$

$$\hat{\tau}_i = \hat{Y}_i(1) - Y_i(0) \quad \text{if } W_i = 0$$

$$\frac{1}{N}\sum_{i=1}^{N}\hat{\tau}_i = \frac{1}{N}\sum_{i=1}^{N}\left\{W_i \cdot \left(Y_i(1) - \hat{Y}_i(0)\right) + (1 - W_i) \cdot \left(\hat{Y}_i(1) - Y_i(0)\right)\right\}$$



LRC-CPPT Cholesterol Data Analysis

Table 7.2. ***Regression Estimates for Average Treatment Effects for the PRC-CPPT Cholesterol Data from Table 7.1***

Covariates	Effect of Assignment to Treatment on Post-Cholesterol Level	
	Est	$\widehat{(\text{s.e.})}$
No covariates	−26.22	(3.93)
`cholp`	−25.01	(2.60)

The more covariates, the better? E.g. compliance?



投票 最多可选1项

设置

36

若满足无混杂假设，是否协变量越多越好?

A 是

B 否

提交

LRC-CPPT Cholesterol Data Analysis

Table 7.2. ***Regression Estimates for Average Treatment Effects for the PRC-CPPT Cholesterol Data from Table 7.1***

Covariates	Effect of Assignment to Treatment on Post-Cholesterol Level	
	Est	$\widehat{(\text{s.e.})}$
No covariates	−26.22	(3.93)
`cholp`	−25.01	(2.60)
`chol1, chol2`	−25.02	(2.59)

The more covariates, the better?

Extensions (Model with Interaction, Testing)

NOT REQUIRED



LRC-CPPT Cholesterol Data Analysis

Table 7.1. ***Summary Statistics for PRC-CPPT Cholesterol Data***

	Variable	Control (N_c =172)		Treatment (N_t =165)			
		Average	Sample (S.D.)	Average	Sample (S.D.)	Min	Max
Pre-treatment	chol1	297.1	(23.1)	297.0	(20.4)	247.0	442.0
	chol2	289.2	(24.1)	287.4	(21.4)	224.0	435.0
	cholp	291.2	(23.2)	289.9	(20.4)	233.0	436.8
Post-treatment	cholf	282.7	(24.9)	256.5	(26.2)	167.0	427.0
	chold	−8.5	(10.8)	−33.4	(21.3)	−113.3	29.5
	comp	74.5	(21.0)	59.9	(24.4)	0	101.0

Y, W, X

Y	W	X
250	1	⋮
280	0	⋮
265	0	⋮
⋮	⋮	⋮

Linear Regression with Covariates & Interaction

To further improve the precision of the estimator, we may include interaction:

$$Y_i^{\text{obs}} = \alpha + \tau \cdot W_i + X_i\beta + W_i \cdot (X_i - \overline{X})\gamma + \varepsilon_i.$$

Least square estimate: $$(\hat{\tau}^{\text{ols}}, \hat{\alpha}^{\text{ols}}, \hat{\beta}^{\text{ols}}, \hat{\gamma}^{\text{ols}}) = \arg\min_{\tau,\alpha,\beta,\gamma} \sum_{i=1}^{N} \left(Y_i^{\text{obs}} - \alpha - \tau \cdot W_i - X_i\beta - W_i \cdot (X_i - \overline{X})\gamma\right)^2$$

LSE when sample size goes to infinity: $$(\alpha^*, \tau^*, \beta^*, \gamma^*) = \arg\min_{\alpha,\beta,\tau,\gamma} \mathbb{E}_{\text{sp}}\left[\left(Y_i^{\text{obs}} - \alpha - \tau \cdot W_i - X_i\beta - W_i \cdot (X_i - \mu_X)\gamma\right)^2\right]$$

Theorem 7.2 *Suppose we conduct a completely randomized experiment in a random sample from a super-population. Then* (*i*) $\tau^* = \tau_{\text{sp}}$, (*ii*),

$$\sqrt{N} \cdot \left(\hat{\tau}^{\text{ols}} - \tau_{\text{sp}}\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\mathbb{E}_{\text{sp}}\left[(W_i - p)^2 \cdot \left(Y_i^{\text{obs}} - \alpha^* - \tau_{\text{sp}} \cdot W_i - X_i\beta^* - W_i \cdot (X_i - \mu_X)\gamma^*\right)^2\right]}{p^2 \cdot (1-p)^2}\right)$$

Note: this LSE can also be interpreted with imputation method.

LRC-CPPT Cholesterol Data Analysis

Table 7.2. ***Regression Estimates for Average Treatment Effects for the PRC-CPPT Cholesterol Data from Table 7.1***

Covariates	Effect of Assignment to Treatment on Post-Cholesterol Level	
	Est	$\widehat{(\text{s. e.})}$
No covariates	−26.22	(3.93)
`cholp`	−25.01	(2.60)
`chol1, chol2`	−25.02	(2.59)
`chol1, chol2`, interacted with W	−25.04	(2.56)

Testing based on Model with Interaction

Testing for no existence of interaction and constant causal effect in each subgroup:

$$H_0': \mathbb{E}_{\text{sp}}[Y_i(1) - Y_i(0)|X_i = x] = \tau_{\text{sp}}, \quad \text{for all } x,$$

$$H_a': \exists\, x_0, x_1, \text{ such that } \mathbb{E}_{\text{sp}}[Y_i(1) - Y_i(0)|X_i = x_0] \neq \mathbb{E}_{\text{sp}}[Y_i(1) - Y_i(0)|X_i = x_1].$$

The normalized covariance matrix of the vector $(\hat{\tau}^{\text{ols}}, \hat{\gamma}^{\text{ols}})$ is $\mathbb{V}_{\tau,\gamma} = \begin{pmatrix} \mathbb{V}_\tau & \mathbb{C}_{\tau,\gamma} \\ \mathbb{C}_{\tau,\gamma}^T & \mathbb{V}_\gamma \end{pmatrix}$

Theorem 7.3 *Suppose we conduct a completely randomized experiment in a random sample from a large population. If* $Y_i(1) - Y_i(0) = \tau$ *for some value* τ *and all units, then* (i): $\gamma^* = 0$, *and* (ii)

$$Q_{\text{const}} = (\hat{\gamma}^{\text{ols}})^T \hat{\mathbb{V}}_\gamma^{-1} \hat{\gamma}^{\text{ols}} \xrightarrow{d} \mathcal{X}(\dim(X_i)).$$

Wald test for multiple parameters

Testing based on Model with Interaction

Testing for no existence of interaction and zero causal effect in each subgroup:

$$H_0: \mathbb{E}_{\text{sp}}[Y_i(1) - Y_i(0)|X_i = x] = 0, \quad \forall\, x,$$

$$H_a: \mathbb{E}_{\text{sp}}[Y_i(1) - Y_i(0)|X_i = x] \neq 0, \quad \text{for some } x.$$

The normalized covariance matrix of the vector $(\hat{\tau}^{\text{ols}}, \hat{\gamma}^{\text{ols}})$ is $\mathbb{V}_{\tau,\gamma} = \begin{pmatrix} \mathbb{V}_\tau & \mathbb{C}_{\tau,\gamma} \\ \mathbb{C}_{\tau,\gamma}^T & \mathbb{V}_\gamma \end{pmatrix}$

Theorem 7.3 *Suppose we conduct a completely randomized experiment in a random sample from a large population. If* $Y_i(1) - Y_i(0) = 0$ *for all units, then*
(i): $\gamma^* = 0$, *and (ii)*

$$Q_{\text{zero}} = \begin{pmatrix} \hat{\tau}^{\text{ols}} \\ \hat{\gamma}^{\text{ols}} \end{pmatrix}^T \hat{\mathbb{V}}_{\tau,\gamma}^{-1} \begin{pmatrix} \hat{\tau}^{\text{ols}} \\ \hat{\gamma}^{\text{ols}} \end{pmatrix} \xrightarrow{d} \mathcal{X}(\dim(X_i) + 1).$$



LRC-CPPT Cholesterol Data Analysis

Table 7.4. ***P-Values for Tests for Constant and Zero Treatment Effects, Using*** `chol1` ***and*** `chol2-chol1` ***as Covariates for the PRC-CPPT Cholesterol Data from Table 7.1***

		Post-Cholesterol Level
Zero treatment effect	$\chi^2(3)$ approximation	<0.001
	Fisher exact p-value	<0.001
Constant treatment effect	$\chi^2(2)$ approximation	0.029

- Conclusion:
  - ✓ The evidence for the presence of treatment effects is very strong
  - ✓ the evidence for heterogeneity in the treatment effect is weak.



Summary



Summary

- Motivation: include covariates for increasing precision of estimators.
- Settings: Super-population, CRE
- Method: linear regression estimand -- parameter
  - ✓ With no covariate
    - Causal interpretation
    - Connections with Neyman's approach
  - ✓ With covariates
    - Causal interpretation
    - Asymptotic theory
  - ✓ With covariates and interactions

Model

- ✓ Point estimation
- ✓ Confidence interval
- ✓ Hypothesis testing

Inference

Any limitations?



Reference

- Reference: [T1] Chapter 7

