Causal inference: for statistics, social, and biomedical sciences

Chapter 7: Regression methods for completely randomized experiments

Kunwoong Kim

2022.7.15.

Contents

Introduction

In this section, we maintain an assumption of a completely randomized experiment.

- Consider models for the observed outcomes.
- The average treatment effect is a parameter of the statistical model (linear model in this section).
- The estimates with finite samples are consistent, that is, converge to the true average causal effect.

The LRC-CPPT cholesterol data

Later, we will use LPC-CPPT (Lipid Research Clinics Coronary Primary Prevention Trial) data from a randomized experiment.

- For 337 individuals, $N_t = 165$ are treated randomly (received cholestyramine) and $N_c = 172$ are controlled.
- Variables
  - Pre-treatment
    - `chol1`: initial cholesterol level
    - `chol2`: cholesterol level after a communication about benefits of a low-cholesterol diet
    - $\texttt{cholp} = 0.25 \cdot \texttt{chol1} + 0.75 \cdot \texttt{chol2}$
  - Post-treatment
    - `cholf`: cholesterol level averaged over 2 month for 7.3 years
    - $\texttt{chold} = \texttt{cholf} - \texttt{cholp}$
    - `comp`: taken dose of either treatment or placebo.
- We will see the differences induced by the treatment (or placebo).

The super-population average treatment effects

This section introduces some notations.

- Assume we have N random samples from the super-population.
- Also denote N_t the number of treated individuals, and N_c the number of controlled individuals with $N = N_t + N_c$.
- Abbreviate "fs" as finite sample and "sp" super-population.
- Then, for average effects of the treatment,

$$\tau_{fs} = \frac{1}{N} \sum_{i=1}^{N} (Y_i(1) - Y_i(0)) \text{ for finite sample,} \quad (1)$$

and

$$\tau_{sp} = \mathbb{E}_{sp} (Y_i(1) - Y_i(0)) \text{ for super-population.}$$

The super-population average treatment effects

For means and variances of $Y|X, Y$, and X, and for the average causal effects, we denote as the followings.

- $Y|X$
  - $\mu_c(x) = \mathbb{E}_{sp}(Y_i(0)|X_i = x)$
  - $\mu_t(x) = \mathbb{E}_{sp}(Y_i(1)|X_i = x)$
  - $\sigma_c^2(x) = \mathbb{V}_{sp}(Y_i(0)|X_i = x)$
  - $\sigma_t^2(x) = \mathbb{V}_{sp}(Y_i(1)|X_i = x)$
- Y
  - $\mu_c = \mathbb{E}_{sp}(Y_i(0)) = \mathbb{E}_{sp}\left(\mu_c(X_i)\right)$
  - $\mu_t = \mathbb{E}_{sp}(Y_i(1)) = \mathbb{E}_{sp}\left(\mu_t(X_i)\right)$
  - $\sigma_c^2 = \mathbb{V}_{sp}(Y_i(0)) = \mathbb{E}_{sp}\left(\sigma_c^2(X_i)\right) + \mathbb{V}_{sp}\left(\mu_c(X_i)\right)$
  - $\sigma_t^2 = \mathbb{V}_{sp}(Y_i(1)) = \mathbb{E}_{sp}\left(\sigma_t^2(X_i)\right) + \mathbb{V}_{sp}\left(\mu_t(X_i)\right)$
- X
  - $\mu_X = \mathbb{E}_{sp}(X_i)$
  - $\Omega_X = \mathbb{V}_{sp}(X_i) = \mathbb{E}_{sp}((X_i - \mu_X)^\top (X_i - \mu_X))$
- Average causal effect
  - $\tau(x) = \mathbb{E}_{sp}\left(Y_i(1) - Y_i(0)|X_i = x\right)$
  - $\sigma_{ct}^2(x) = \mathbb{V}_{sp}\left(Y_i(1) - Y_i(0)|X_i = x\right)$

Linear regression with no covariates

- Let $W_i \in \{0, 1\}$ the indicator for the receipt of treatment, and Y_i^{obs} the observed outcome of the ith individual.
- For the model, we consider a linear regression function as

$$Y_i^{obs} = \alpha + \tau \cdot W_i + \epsilon_i$$

  where ϵ_i is the unobserved error independent to W_i.
- The least squares estimate of τ is interpreted as an estimate of the causal effect of the treatment:

$$\hat{\tau}^{ols} = \frac{1}{N_t} \sum_{i:W_i=1} Y_i^{obs} - \frac{1}{N_c} \sum_{i:W_i=0} Y_i^{obs}.$$

- Moreover, $\hat{\tau}^{ols}$ is unbiased for τ_{fs} as well as τ_{sp}.

Linear regression with no covariates

Estimates of $\hat{\tau}_{ols}$ variances under

- Homoskedasticity ($\sigma^2_{Y|W} = \sigma^2_c = \sigma^2_t$):

$$\hat{\mathbb{V}}^{homosk} = \frac{s^2}{N_c} + \frac{s^2}{N_t} \tag{2}$$

- Heteroskedasticity ($\sigma^2_c \neq \sigma^2_t$):

$$\hat{\mathbb{V}}^{hetero} = \frac{s^2_c}{N_c} + \frac{s^2_t}{N_t} \tag{3}$$

Linear regression with additional covariates

The key insight is that, by randomizing treatment assignment, the super-population correlation between the treatment indicator W_i and the covariate X_i is 0.

- For the model, we consider a linear regression function with additional covariates as

$$Y_i^{obs} = \alpha + \tau \cdot W_i + X_i \beta + \epsilon_i$$

  where X_i is a row vector of covariates and ϵ_i is the unobserved error.

Linear regression with additional covariates

Consistency of least squares estimators

- $\tau^{ols} \to \tau_{sp}$ in probability.
- $$\sqrt{N}\left(\hat{\tau}^{ols} - \tau_{sp}\right) \to \mathcal{N}\left(0, \Sigma\right)$$

  for some Σ.

Linear regression with covariates and interactions

- As the last one, we consider a linear regression function with additional covariates as

$$Y_i^{obs} = \alpha + \tau \cdot W_i + X_i\beta + W_i \cdot (X_i - \bar{X})\gamma + \epsilon_i$$

where ϵ_i is the unobserved error.

Linear regression with covariates and interactions

We compute the unit-level causal effect of ith individual as the following two cases.

- Treated, i.e., $W_i = 1$
  1. $\hat{Y}_i(0) = \hat{\alpha}^{ols} + X_i\hat{\beta}^{ols}$: predicted
  2. $Y_i(1)$: observed
  3. $\hat{\tau}_i = Y_i(1) - \hat{Y}_i(0) = Y_i^{obs} - (\hat{\alpha}^{ols} + X_i\hat{\beta}^{ols})$
- Controlled, i.e., $W_i = 0$
  1. $Y_i(0)$: observed
  2. $\hat{Y}_i(1) = \hat{\alpha}^{ols} + \hat{\tau}^{ols} + X_i\hat{\beta}^{ols} + (X_i - \bar{X})\hat{\gamma}^{ols} - Y_i^{obs}$: predicted
  3. $\hat{\tau}_i = \hat{Y}_i(1) - Y_i(0) = \hat{\alpha}^{ols} + \hat{\tau}^{ols} + X_i\hat{\beta}^{ols} + (X_i - \bar{X})\hat{\gamma}^{ols} - Y_i^{obs}$

Linear regression with covariates and interactions

Consistency of least squares estimators

- $\tau^{ols} \rightarrow \tau_{sp}$ in probability.
- $$\sqrt{N}\left(\hat{\tau}^{ols} - \tau_{sp}\right) \rightarrow \mathcal{N}\left(0, \Sigma\right)$$

  for some Σ.

Linear regression with covariates and interactions

We can estimate the overall average treatment effect τ_{fc} by averaging the estimates of the unit-level causal effects $\hat{\tau}_i$.

$$\begin{aligned}\hat{\tau}^{ols} &= \frac{1}{N}\sum_{i=1}^{N}\hat{\tau}_i \\ &= \frac{1}{N}\sum_{i=1}^{N}\left(W_i\left(Y_i(1)-\hat{Y}_i(0)\right)+(1-W_i)\left(\hat{Y}_i(1)-Y_i(0)\right)\right)\end{aligned} \tag{4}$$

- Thus we can conclude that the least squares estimator $\hat{\tau}^{ols}$ can be interpreted as averaging estimated unit-level causal effects.

Transformations of the outcome variable

- One can be interested in the average effect of the treatment on a *transformation* of the outcome.
- For example, assume

$$\ln(Y_i^{obs}) = \alpha + \tau W_i + X_i\beta + \epsilon_i. \tag{5}$$

Then, least squares estimates of τ are consistent for the average effect $\mathbb{E}\left(\ln(Y_i(1)) - \ln(Y_i(0))\right)$.

The limits on increases in precision due to covariates

Including covariates in the linear regression model would increase the precision of the estimator for the average treatment effect.

- $N \cdot \mathbb{V}_{nocov} = \frac{\sigma_c^2}{1-p} + \frac{\sigma_t^2}{p}$: with no covariates
- $N \cdot \mathbb{V}_{bound} = \frac{\mathbb{E}_{sp}(\sigma_c^2(X_i))}{1-p} + \frac{\mathbb{E}_{sp}(\sigma_t^2(X_i))}{p}$: with additional covariates
- The difference between the two variances are:

$$\mathbb{V}_{nocov} - \mathbb{V}_{bound} = \frac{\mathbb{V}_{sp}(\mu_c(X_i))}{1-p} + \frac{\mathbb{V}_{sp}(\mu_t(X_i))}{p}. \qquad (6)$$

- Additional covariates X_i increase the precision and decrease the variance.

Testing for the presence of treatment effects

In addition, not only estimating average treatment effect, but we can also test for the presence of treatment effects.

$$\begin{aligned} H_0 &: \mathbb{E}_{sp}\left(Y_i(1) - Y_i(0) | X_i = x\right) = 0, \forall x, \\ &\text{vs.} \\ H_a &: \mathbb{E}_{sp}\left(Y_i(1) - Y_i(0) | X_i = x\right) \neq 0, \text{ for some } x. \end{aligned} \tag{7}$$

Estimates for LRC-CPPT cholesterol data

- Here, we return to the LRC-CPPT cholesterol data and look at estimates for two average effects: (1) the effect on `cholf`, and (2) the effect on `comp`.
- For detailed descriptions of variables, please revisit 7.2.

Estimates for LRC-CPPT cholesterol data

- Including more covariates in model improves the precision.
- Estimates of τ are negative for all cases, that is, treatment reduces cholesterol levels.

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		Compliance	
	Est	$\widehat{(\text{s.e.})}$	Est	$\widehat{(\text{s.e.})}$
No covariates	−26.22	(3.93)	−14.64	(3.51)
`cholp`	−25.01	(2.60)	−14.68	(3.51)
`chol1, chol2`	−25.02	(2.59)	−14.95	(3.50)
`chol1, chol2`, interacted with W	−25.04	(2.56)	−14.94	(3.49)

Estimates for LRC-CPPT cholesterol data

- Transformed cholesterol levels (logarithm).

Table 7.3. ***Regression Estimates for Average Treatment Effects on Post-Cholesterol Levels for the PRC-CPPT Cholesterol Data from Table 7.1***

Covariates	Model for Levels		Model for Logs	
	Est	$\widehat{(\text{s. e.})}$	Est	$\widehat{(\text{s. e.})}$
Assignment	−25.04	(2.56)	−0.098	(0.010)
Intercept	−3.28	(12.05)	−0.133	(0.233)
`chol1`	0.98	(0.04)	−0.133	(0.233)
`chol2-chol1`	0.61	(0.08)	0.602	(0.073)
`chol1` × Assignment	−0.22	(0.09)	−0.154	(0.107)
(`chol2-chol1`) × Assignment	0.07	(0.14)	0.184	(0.159)
R-squared	0.63		0.57	

Conclusion

- Linear regression models with complete randomization for three cases:

Section	Covariates	Interactions
7.4	X	X
7.5	O	X
7.6	O	O

- The randomization is a necessary condition for the consistency of the least squares estimator.
- A bridge from exact results based on randomization inference to the model-based methods: we will see in the next chapter.