Simple Linear Regression

Kosuke Imai

Harvard University

Spring 2021

Agenda

- Linear regression is commonly used in applied research
- We will explore how to use linear regression for causal effect estimation
- To build intuition, we focus on the application of simple linear regression

Linear Regression and Causality

- Regression $\rightsquigarrow$ conditional expectation function of Y given $\mathbf{X}$

$$\mathbb{E}(Y \mid \mathbf{X}) \;=\; f(\mathbf{X}) \;=\; \boldsymbol{\beta}^\top \mathbf{X}$$

- When can we interpret coefficients as causal effects?
- Causal model as structural equation model

$$Y_i(t) \;=\; \alpha + \beta t + \epsilon_i \quad \text{for} \;\; t = 0, 1, \;\; \text{where } \mathbb{E}(\epsilon_i) = 0$$

1. No interference between units
2. $\mathbb{E}(Y_i(0)) = \alpha$
3. $Y_i(1) - Y_i(0) = \beta$ for all $i \iff$ constant additive unit causal effect

- Heterogenous treatment effect model:

$$Y_i(t) \;=\; \alpha + \beta_i t + \epsilon_i \;=\; \alpha + \beta t + \underbrace{(\beta_i - \beta)t + \epsilon_i}_{=\epsilon_i(t)}$$

where $\mathbb{E}(\epsilon_i) = 0$ and $\beta = \mathbb{E}(\beta_i) = \mathbb{E}(Y_i(1) - Y_i(0))$

- $\mathbb{E}(\epsilon_i(t)) = 0$ for $t = 0, 1$
- $\alpha = \mathbb{E}(Y_i(0))$

Identification Assumption

- Strict exogeneity assumption:

$$\mathbb{E}(\epsilon_i \mid \mathbf{T}) \;=\; \mathbb{E}(\epsilon_i) \;=\; 0 \quad \text{where} \quad \mathbf{T} = (T_1, T_2, \ldots, T_n)^\top$$

- Under this assumption, least squares estimate $\hat{\beta}$ is unbiased for β

- Randomization of treatment: $\{Y_i(1), Y_i(0)\}_{i=1}^n \perp\!\!\!\perp \mathbf{T}$
  - $\mathbb{E}(Y_i(t)) = \mathbb{E}(Y_i(t) \mid T_i = t) = \mathbb{E}(Y_i \mid T_i = t) = \alpha + \beta t$
  - Recall that $\epsilon_i(t) = Y_i(t) - \alpha - \beta t$. Then, $\mathbb{E}(\epsilon_i(t) \mid \mathbf{T}) \;=\; \mathbb{E}(\epsilon_i(t)) = 0$

- Random sampling of units: $\{Y_i(1), Y_i(0)\} \overset{\text{i.i.d.}}{\sim} \mathcal{P}$
  - $\{\epsilon_i(1), \epsilon_i(0)\} \overset{\text{i.i.d.}}{\sim} \mathcal{P}^*$
  - independence across observations

Unbiasedness of Least Squares Estimator

- Consider a RCT with a binary treatment
- Least squares estimators:

$$
\begin{aligned}
\hat{\alpha} &= \frac{1}{n_0}\sum_{i=1}^{n}(1-T_i)Y_i, \\
\hat{\beta} &= \frac{1}{n_1}\sum_{i=1}^{n}T_iY_i - \frac{1}{n_0}\sum_{i=1}^{n}(1-T_i)Y_i.
\end{aligned}
$$

- Thus, $\hat{\alpha}$ and $\hat{\beta}$ are unbiased for $\mathbb{E}(Y(0))$ and $\mathbb{E}(Y_i(1) - Y_i(0))$

- A similar conclusion holds if T is a categorical treatment
- When T is continuous, the model assumes $Y_i(t)$ is linear in t

- What about standard errors?

Homoskedasticity

- Homoskedastic error:

$$\mathbb{V}(\boldsymbol{\epsilon} \mid \mathbf{T}) \;=\; \sigma^2 I_n$$

1. Equal variance of potential outcomes

$$\mathbb{V}(\epsilon_i(0)) = \mathbb{V}(\epsilon_i(1))$$

2. Random sampling of units implies $\epsilon_i \perp\!\!\!\perp \epsilon_j$ and $\mathbb{V}(\epsilon_i) = \mathbb{V}(\epsilon_j)$

- Under the homoskedasticity assumption,
  - model-based variance is,

$$\mathbb{V}(\hat{\beta} \mid \mathbf{T}) \;=\; \frac{\sigma^2}{\sum_{i=1}^n (T_i - \overline{T})^2}$$

  - standard model-based variance estimator is,

$$\widehat{\mathbb{V}(\hat{\beta} \mid T)} \;=\; \frac{\hat{\sigma}^2}{\sum_{i=1}^n (T_i - \overline{T})^2} \quad \text{where} \quad \hat{\sigma}^2 \;=\; \frac{1}{n-2}\sum_{i=1}^n \hat{\epsilon}_i^2$$

Violation of the Homoskedasticity Assumption

- Homoskedasticity assumption may be unrealistic: $\sigma_1^2 \neq \sigma_0^2$
- Constant additive treatment effect $\rightsquigarrow$ homoskedasticity
- Bias in the estimated variance when the assumption does not hold

$$\begin{aligned} \text{Bias} &= \underbrace{\mathbb{E}\left(\frac{\hat{\sigma}^2}{\sum_{i=1}^n (T_i - \overline{T})^2}\right)}_{\text{under complete randomization}} - \underbrace{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0}\right)}_{\text{true variance}} \\ &= \frac{(n_1 - n_0)(n-1)}{n_1 n_0 (n-2)}(\sigma_1^2 - \sigma_0^2) \end{aligned}$$

- Bias is zero when
  1. homoskedasticity assumption holds: $\sigma_1^2 = \sigma_0^2$
  2. design is balanced: $n_1 = n_0$
- Bias can be negative or positive
- Bias is typically small but does not go away asymptotically

Heteroskedasticity-Robust Variance Estimator

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

$$\underbrace{\widehat{\mathbb{V}_{\text{EHW}}((\hat{\alpha}, \hat{\beta}) \mid \mathbf{X})}}_{\text{sandwich}} = \underbrace{(\mathbf{X}^\top \mathbf{X})^{-1}}_{\text{bread}} \underbrace{(\mathbf{X}^\top \operatorname{diag}(\hat{\epsilon}_i^2) \mathbf{X})}_{\text{meat}} \underbrace{(\mathbf{X}^\top \mathbf{X})^{-1}}_{\text{bread}}$$

$$= \left(\sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^\top \right)^{-1} \left(\sum_{i=1}^n \hat{\epsilon}_i^2 \mathbf{X}_i \mathbf{X}_i^\top \right) \left(\sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^\top \right)^{-1}$$

where $\mathbf{X}_i = (1, T_i)$ and $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^\top$

- The EHW robust variance estimator is asymptotically unbiased

$$\widehat{\mathbb{V}_{\text{EHW}}(\hat{\beta} \mid \mathbf{T})} = \frac{\tilde{\sigma}_1^2}{n_1} + \frac{\tilde{\sigma}_0^2}{n_0} \quad \text{where} \quad \tilde{\sigma}_t^2 = \frac{1}{n_t} \sum_{i=1}^n \mathbf{1}\{T_i = t\}(Y_i - \overline{Y}_t)^2$$

$$\text{Bias} = \mathbb{E}(\widehat{\mathbb{V}_{\text{EHW}}(\hat{\beta} \mid \mathbf{T})}) - \underbrace{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0} \right)}_{\text{true variance}} = -\left(\frac{\sigma_1^2}{n_1^2} + \frac{\sigma_0^2}{n_0^2} \right)$$

Improved Robust Variance Estimators

- HC2 heteroskedasticity-robust variance estimator:

$$\widehat{\mathbb{V}_{\text{HC2}}((\hat{\alpha}, \hat{\beta}) \mid \mathbf{X})} = (\mathbf{X}^\top \mathbf{X})^{-1} \left\{ \mathbf{X}^\top \operatorname{diag}\left(\frac{\hat{\epsilon}_i^2}{1 - p_{ii}} \right) \mathbf{X} \right\} (\mathbf{X}^\top \mathbf{X})^{-1}$$

  where p_{ii} is the leverage

$$p_{ii} = \mathbf{X}_i^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}_i = \|\mathbf{P}_\mathbf{X} v(i)\|^2 = \begin{cases} \frac{1}{n_1} & \text{if } T_i = 1 \\ \frac{1}{n_0} & \text{if } T_i = 0 \end{cases}$$

  - $\mathbf{P}_\mathbf{X} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$: projection matrix
  - $v(i)$: a vector whose ith element is 1 and other elements are zero

- HC2 variance estimator is identical to the Neyman's variance estimator and is unbiased (Samii and Aronow. 2012. *Stat Probab Lett*)

$$\widehat{\mathbb{V}_{\text{HC2}}(\hat{\beta} \mid \mathbf{T})} = \frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_0^2}{n_0}$$

Summary

- Simple linear regression $=$ Difference-in-means estimator

- Homoskedasticity implies the equal variance of potential outcomes

- Heteroskedasticity-robust variance estimator relaxes this assumption