Simple Linear Regression for Randomized Experiments

Yuki Shiraito

POLSCI 699 Statistical Methods in Political Research II
University of Michigan

Simple Linear Regression Model

- Setup (same as before):
  1. Units $i = 1, \dots, n$; random sample from superpopulation
  2. Potential outcomes $\left(Y_i(0), Y_i(1)\right)$
  3. Treatment $T_i \in \{0, 1\}$; completely random assignment

- Simple linear regression model:
$$Y_i = \alpha + \beta T_i + \varepsilon_i, \quad \mathbb{E}(\varepsilon_i) = 0$$
- Y_i: *observed* (not potential) outcome
- Parameters: intercept α, slope β
- ε_i: error term, disturbance, ~~residual~~
  - $\mathbb{E}(\varepsilon_i) = 0$ is not really an assumption because we have α
- Ordinary least squares (OLS) estimator:
$$(\hat{\alpha}_{\text{OLS}}, \hat{\beta}_{\text{OLS}}) = \underset{a,b}{\operatorname{argmin}} \sum_{i=1}^{n} (Y - a - bT_i)^2$$

Regression and Conditional Expectation

- Conditional expectation minimizes the mean squared error:

$$\mathbb{E}[Y_i \mid T_i] = \underset{f(T_i)}{\operatorname{argmin}}\, \mathbb{E}\left[\left(Y_i - f(T_i)\right)^2\right]$$

- Linear predictor that minimizes the mean squared error:

$$(\alpha, \beta) \equiv \underset{a,b}{\operatorname{argmin}}\, \mathbb{E}\left[\left(Y_i - a - bT_i\right)^2\right]$$

1. $T_i \in \{0, 1\} \implies \mathbb{E}[Y_i \mid T_i]$ is either $\mathbb{E}[Y_i \mid T_i = 0]$ or $\mathbb{E}[Y_i \mid T_i = 1]$
2. $\mathbb{E}[Y_i \mid T_i] = \mathbb{E}[Y_i \mid T_i = 0] + ([Y_i \mid T_i = 1] - \mathbb{E}[Y_i \mid T_i = 0])T_i$
3. Population regression parameter $\beta = [Y_i \mid T_i = 1] - \mathbb{E}[Y_i \mid T_i = 0]$

- Population regression parameter β is PATE:

1. $\beta = [Y_i \mid T_i = 1] - \mathbb{E}[Y_i \mid T_i = 0] = [Y_i(1) \mid T_i = 1] - \mathbb{E}[Y_i(0) \mid T_i = 0]$
2. $(Y_i(1), Y_i(0)) \perp\!\!\!\perp T_i \implies \mathbb{E}[Y_i(1) \mid T_i = 1] - \mathbb{E}[Y_i(0) \mid T_i = 0] = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)] = \text{PATE}$

- OLS estimator as sample analog

More Causal Interpretation

- Association: you can always regress Y_i on T_i and vice versa
- Causal model as structural equation model

- Linear model in terms of potential outcomes:

$$Y_i(t) = \alpha + \beta t + \varepsilon_i, \quad \mathbb{E}(\varepsilon_i) = 0$$

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

- A more general model with heterogeneous treatment effects:

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

  where $\mathbb{E}[\varepsilon_i] = 0$ and $\beta = \mathbb{E}[\beta_i] = \text{PATE}$

  - Relax the assumption of constant additive unit causal effect
  - $\mathbb{E}[\varepsilon_i(t)] = 0$ for $t = 0, 1$
  - $\alpha = \mathbb{E}(Y_i(0))$ as before

Assumptions for Linear Regression

1. Random assignment, $(Y_i(1), Y_i(0)) \perp\!\!\!\perp T_i$ for all i, implies:
$$\mathbb{E}[Y_i(t) \mid T_i] = \mathbb{E}[Y_i(t)] \iff \mathbb{E}[\varepsilon_i(t) \mid T_i] = \mathbb{E}[\varepsilon_i(t)] = 0$$
2. Random sampling of units, $(Y_i(1), Y_i(0)) \perp\!\!\!\perp (Y_j(1), Y_j(0))$ for any (i, j) s.t. $i \neq j$, implies:
$$(\varepsilon_i(1), \varepsilon_i(0)) \perp\!\!\!\perp (\varepsilon_j(1), \varepsilon_j(0)) \text{ for any } (i, j) \text{ s.t. } i \neq j$$

$\Longrightarrow$ Strict exogeneity: $\mathbb{E}[\varepsilon_i \mid \mathrm{T}] = \mathbb{E}[\varepsilon_i] = 0$ where $\mathrm{T} = (T_1, T_2, \ldots, T_n)$

1. Orthogonality: $\mathbb{E}[\varepsilon_i T_j] = 0$ for any (i, j) (not limited to $i \neq j$)
2. Zero correlation: $\mathrm{Cov}(\varepsilon_i, T_j) = 0$ for any (i, j) (not limited to $i \neq j$)

3. Variance of potential outcomes:
$$\mathbb{V}(\varepsilon_i(t)) = \mathbb{V}(\varepsilon_i(t) \mid T_i) = \mathbb{V}(Y_i(t) \mid T_i) = \mathbb{V}(Y_i(t)) = \sigma_t^2 \text{ for } t = 0, 1$$
   - $\sigma_0 = \sigma_1 = \sigma$ if constant additive unit causal effect
   - $\varepsilon_i(t) = (\beta_i - \beta)t + \varepsilon_i = \varepsilon_i$ if $\beta_i = \beta$ for all i

$\Longrightarrow$ Homoskedasticity: $\mathbb{V}(\varepsilon_i \mid \mathrm{T}) = \mathbb{V}(\varepsilon_i) = \sigma^2$

Least Squares Estimation

- Model parameters for population regression
$$(\alpha, \beta) \equiv \operatorname*{argmin}_{a,b} \mathbb{E}\left[(Y_i - a - bT_i)^2\right]$$
- Minimization of the sum of squared residuals (SSR):
$$(\hat{\alpha}_{\text{OLS}}, \hat{\beta}_{\text{OLS}}) = \operatorname*{argmin}_{a,b} \sum_{i=1}^{n} (Y - a - bT_i)^2$$
$$= \operatorname*{argmin}_{(a,b)} \sum_{i=1}^{n} \hat{\varepsilon}_i^2$$
    - Predicted (fitted) value: $\widehat{Y}_i = \hat{\alpha}_{\text{OLS}} + \hat{\beta}_{\text{OLS}} T_i$
    - Residual: $\hat{\varepsilon}_i = Y_i - \widehat{Y}_i = Y_i - \hat{\alpha}_{\text{OLS}} - \hat{\beta}_{\text{OLS}} T_i$
- OLS estimator (Adam will derive in the section):
$$\hat{\beta}_{\text{OLS}} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(T_i - \overline{T})}{\sum_{i=1}^{n}(T_i - \overline{T})^2}$$
$$\hat{\alpha}_{\text{OLS}} = \overline{Y} - \hat{\beta}_{\text{OLS}} \overline{T}$$

Unbiasedness of OLS Estimator

- When T_i is binary, $\hat{\beta}_{\text{OLS}} =$ Difference-in-Means estimator (Adam's section)
- So, $\hat{\beta}_{\text{OLS}}$ is unbiased for PATE from the design-based perspective

- Is $\hat{\beta}_{\text{OLS}}$ unbiased for β, population regression parameter?
  - Yes if T_i is binary, because β is PATE
  - More generally yes, under strict exogeneity and linearlity
- Model-based estimation error:

$$\hat{\beta}_{\text{OLS}} - \beta = \frac{\sum_{i=1}^{n}(T_i - \overline{T})\varepsilon_i}{\sum_{i=1}^{n}(T_i - \overline{T})^2}$$

- Thus, the exogeneity assumption implies,

$$\mathbb{E}\left[\hat{\beta}_{\text{OLS}}\right] - \beta = \mathbb{E}\left[\mathbb{E}\left[\hat{\beta}_{\text{OLS}} - \beta \mid \mathbf{T}\right]\right] = 0$$

- Similarly, $\hat{\alpha}_{\text{OLS}} - \alpha = \overline{\varepsilon} - \left(\hat{\beta}_{\text{OLS}} - \beta\right)\overline{T}$
- Thus, $\mathbb{E}\left[\hat{\alpha}_{\text{OLS}}\right] - \alpha = 0$

Model-based Sampling Variance of OLS Estimator

- The homoskedasticity assumption implies
$$\mathbb{V}\left(\hat{\beta}_{\text{OLS}} \mid \mathbf{T}\right) = \frac{\sigma^2}{\sum_{i=1}^{n}\left(T_i - \overline{T}\right)^2}$$
- Standard model-based (conditional) variance estimator for $\hat{\beta}$:
$$\widehat{\mathbb{V}\left(\hat{\beta}_{\text{OLS}} \mid \mathbf{T}\right)} = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}\left(T_i - \overline{T}\right)^2} \quad \text{where} \quad \hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2$$
- (Conditionally) Unbiased: $\mathbb{E}\left[\hat{\sigma}^2 \mid \mathbf{T}\right] = \sigma^2$ implies
$$\mathbb{E}\left[\widehat{\mathbb{V}\left(\hat{\beta}_{\text{OLS}} \mid \mathbf{T}\right)} \mid \mathbf{T}\right] = \mathbb{V}\left(\hat{\beta}_{\text{OLS}} \mid \mathbf{T}\right)$$
- (Unconditionally) Unbiased: $\mathbb{V}\left(\mathbb{E}\left[\hat{\beta}_{\text{OLS}} \mid \mathbf{T}\right]\right) = 0$ implies
$$\mathbb{V}\left(\hat{\beta}_{\text{OLS}}\right) = \mathbb{E}\left[\mathbb{V}\left(\hat{\beta}_{\text{OLS}} \mid \mathbf{T}\right)\right] = \mathbb{E}\left[\mathbb{E}\left[\widehat{\mathbb{V}\left(\hat{\beta}_{\text{OLS}} \mid \mathbf{T}\right)} \mid \mathbf{T}\right]\right]$$
$$= \mathbb{E}\left[\widehat{\mathbb{V}\left(\hat{\beta}_{\text{OLS}} \mid \mathbf{T}\right)}\right]$$

Model-Based Asymptotic Inference

- Consistency: $\hat{\beta}_{\text{OLS}} \xrightarrow{p} \frac{\text{Cov}(T_i, Y_i)}{\mathbb{V}(T_i)} = \beta$ (c.f. Q3 of PS599 PSet 6)
- Asymptotic distribution and inference:

$$\sqrt{n}(\hat{\beta}_{\text{OLS}} - \beta) = \underbrace{\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}(T_i - \mathbb{E}[T_i])\,\varepsilon_i + \left(\mathbb{E}[T_i] - \overline{T}\right)\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\right)}_{\xrightarrow{d}\mathcal{N}\left(0,\sigma^2\mathbb{V}(T_i)\right)}$$

$$\times \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n}(T_i - \overline{T})^2\right)^{-1}}_{\xrightarrow{p}\mathbb{V}(T_i)^{-1}}$$

$$\xrightarrow{d} \mathcal{N}\left(0,\ \frac{\sigma^2}{\mathbb{V}(T_i)}\right)$$

$$\frac{\hat{\beta}_{\text{OLS}} - \beta}{\mathrm{s.e.}} \xrightarrow{d} \mathcal{N}(0, 1) \text{ where } \mathrm{s.e.} = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}\left(T_i - \overline{T}\right)^2}}$$

Violation of Homoskedasticity

- The design-based perspective: use Neyman's exact variance
  - Not relying on constant additive unit causal effect
- Constant additive unit causal effect $\Longrightarrow$ homoskedasticity
- Heterogeneous effects $\Longrightarrow$ violation of homoskedasticity $\Longrightarrow$ bias of model-based variance estimator
- Finite sample bias:

$$\text{Bias} = \underbrace{\mathbb{E}\left(\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(T_i - \overline{T})^2}\right)}_{\text{expectation of variance estimator}} - \underbrace{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0}\right)}_{\text{true variance}}$$

$$= \frac{(n_1 - n_0)(n-1)}{n_1 n_0 (n-2)}(\sigma_1^2 - \sigma_0^2)$$

1. zero if homoskedasticity holds: $\sigma_1^2 - \sigma_0^2 = 0$
2. zero if design is balanced: $n_1 - n_0 = 0$
3. not asymptotically zero
4. can be negative or positive

Eicker-Huber-White (EHW) Variance Estimator

- Heteroskedasticity-consistent (HC) variance estimators
  - also known as "robust" or "sandwich" estimators
  - implemented in `sandwich` package in **R**
- EHW (or simply "Huber-White") robust variance estimator:

$$\widehat{\underset{(\text{EHW})}{\mathbb{V}}\left(\hat{\tilde{\beta}}_{\text{OLS}} \mid \mathsf{T}\right)} \equiv \left(\sum_{i=1}^{n} \widetilde{\mathsf{T}}_i \widetilde{\mathsf{T}}_i^\top\right)^{-1} \left(\sum_{i=1}^{n} \hat{\varepsilon}_i^2 \widetilde{\mathsf{T}}_i \widetilde{\mathsf{T}}_i^\top\right) \left(\sum_{i=1}^{n} \widetilde{\mathsf{T}}_i \widetilde{\mathsf{T}}_i^\top\right)^{-1}$$

  where

$$\tilde{\beta} \equiv \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \text{ and } \widetilde{\mathsf{T}}_i \equiv \begin{pmatrix} 1 \\ T_i \end{pmatrix}$$

- Design-based evaluation:

$$\text{Bias} = \mathbb{E}\left[\widehat{\underset{(\text{EHW})}{\mathbb{V}}\left(\hat{\beta}_{\text{OLS}} \mid \mathsf{T}\right)}\right] - \left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0}\right) = -\left(\frac{\sigma_1^2}{n_1^2} + \frac{\sigma_0^2}{n_0^2}\right)$$

- Negative bias, but vanishes asymptotically

HC2 Variance Estimator

- HC2 robust variance estimator: $\underset{(\text{HC2})}{\widehat{\mathbb{V}}\left(\hat{\hat{\beta}}_{\text{OLS}} \mid \mathbf{T}\right)}$

$$\equiv \left(\sum_{i=1}^{n} \widetilde{\mathrm{T}}_i \widetilde{\mathrm{T}}_i^\top\right)^{-1} \left(\sum_{i=1}^{n} \frac{\hat{\varepsilon}_i^2}{1-p_{ii}} \widetilde{\mathrm{T}}_i \widetilde{\mathrm{T}}_i^\top\right) \left(\sum_{i=1}^{n} \widetilde{\mathrm{T}}_i \widetilde{\mathrm{T}}_i^\top\right)^{-1}$$

  where

$$p_{ii} = \widetilde{\mathrm{T}}_i^\top \left(\widetilde{\mathbf{T}}^\top \widetilde{\mathbf{T}}\right) \widetilde{\mathrm{T}}_i = \begin{cases} \frac{1}{n_1} & \text{if } T_i = 1 \\ \frac{1}{n_0} & \text{if } T_i = 0 \end{cases}$$

- p_{ii} is the (i, i) element of the *projection matrix* (discussed later)
- Samii and Aronow (2012):

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

  - HC2 estimator is identical Neyman's estimator

Cluster Randomized Experiments

- Units: $i = 1, 2, \ldots, n_j$
- Clusters of units: $j = 1, 2, \ldots, m$
- Treatment at cluster level: $T_j \in \{0, 1\}$
- Outcome: $Y_{ij} = Y_{ij}\left(T_j\right)$
- Random assignment: $\left(Y_{ij}(1), Y_{ij}(0)\right) \perp\!\!\!\perp T_j$
- No interference between units of different clusters
- Possible interference between units of the same cluster
- Random sampling of clusters and units
- Estimands at unit level:

$$\begin{aligned}
\text{SATE} &\equiv \frac{1}{\sum_{j=1}^{m} n_j} \sum_{j=1}^{m} \sum_{i=1}^{n_j} \left(Y_{ij}(1) - Y_{ij}(0)\right) \\
\text{PATE} &\equiv \mathbb{E}\left[Y_{ij}(1) - Y_{ij}(0)\right]
\end{aligned}$$

Design-Based Inference

- For simplicity, assume the following:
  1. equal cluster size, i.e., $n_j = n$ for all j
  2. we observe all units for a selected cluster (no sampling of units)
- The difference-in-means estimator:

$$\hat{\tau} \equiv \frac{1}{m_1}\sum_{j=1}^{m} T_j \overline{Y}_j - \frac{1}{m_0}\sum_{j=1}^{m}(1-T_j)\overline{Y}_j$$

  where $\overline{Y}_j \equiv \sum_{i=1}^{n} Y_{ij}/n$

- Easy to show $\mathbb{E}(\hat{\tau} \mid \mathcal{O}) = \mathrm{SATE}$ and thus $\mathbb{E}(\hat{\tau}) = \mathrm{PATE}$
- Exact population variance:

$$\mathbb{V}(\hat{\tau}) = \frac{\mathbb{V}\left(\overline{Y_j(1)}\right)}{m_1} + \frac{\mathbb{V}\left(\overline{Y_j(0)}\right)}{m_0}$$

Intracluster Correlation Coefficient

- Comparison with the standard variance:

$$\mathbb{V}(\hat{\tau}) = \frac{\sigma_1^2}{m_1 n} + \frac{\sigma_0^2}{m_0 n}$$

- Correlation of potential outcomes across units within a cluster

$$\begin{aligned}
\mathbb{V}\left(\overline{Y_j(t)}\right) &= \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^{n} Y_{ij}(t)\right) \\
&= \frac{1}{n^2}\left\{\sum_{i=1}^{n}\mathbb{V}(Y_{ij}(t)) + \sum_{i\neq i'}\sum_{i'=1}^{n}\text{Cov}(Y_{ij}(t), Y_{i'j}(t))\right\} \\
&= \frac{\sigma_t^2}{n}\{1 + (n-1)\rho_t\} \overset{\text{typically}}{\geq} \frac{\sigma_t^2}{n}
\end{aligned}$$

Cluster Standard Error

- Cluster robust variance estimator: $\underset{\text{(Cluster)}}{\widehat{\mathbb{V}\left(\widehat{\beta}_{\text{OLS}} \mid \mathbf{T}\right)}}$

$$= \left(\sum_{j=1}^{m} \mathbf{T}_j^\top \mathbf{T}_j\right)^{-1} \left(\sum_{j=1}^{m} \mathbf{T}_j^\top \hat{\varepsilon}_j \hat{\varepsilon}_j^\top \mathbf{T}_j\right) \left(\sum_{j=1}^{m} \mathbf{T}_j^\top \mathbf{T}_j\right)^{-1}$$

where

$$\hat{\varepsilon}_j = \begin{pmatrix} \hat{\varepsilon}_{1j} \\ \vdots \\ \hat{\varepsilon}_{nj} \end{pmatrix} \text{ and } \mathbf{T}_j = \begin{pmatrix} 1 & T_{1j} \\ \vdots & \vdots \\ 1 & T_{nj} \end{pmatrix}$$

- Design-based evaluation:

$$\text{Bias} = -\left(\frac{\mathbb{V}\left(\overline{Y_j(1)}\right)}{m_1^2} + \frac{\mathbb{V}\left(\overline{Y_j(0)}\right)}{m_0^2}\right)$$

- Bias vanishes asymptotically as $m \to \infty$ with n fixed
- **Implication:** cluster by the unit of treatment assignment