Simple Linear Regression

Kosuke Imai

Harvard University

STAT186/GOV2002 CAUSAL INFERENCE

Fall 2018

Linear Regression and Causality

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

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

- Question: When can we interpret coefficients as causal effects?
- Causal model $\rightsquigarrow$ 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. $\alpha = \mathbb{E}(Y_i(0))$
3. $\beta = Y_i(1) - Y_i(0)$ for all $i \Longleftrightarrow$ Constant additive unit causal effect

- Heterogenous treatment effect model:

$$Y_i(t) \;\; = \;\; \alpha + \beta_i t + \epsilon_i^* \; = \; \alpha + \beta t + \epsilon_i(t)$$

where $\mathbb{E}(\epsilon_i^*) = 0$ and $\epsilon_i(t) = \epsilon_i^* + (\beta_i - \beta)t$ for $t = 0, 1$

  - ATE: $\beta = \mathbb{E}(\beta_i) = \mathbb{E}(Y_i(1) - Y_i(0))$
  - $\mathbb{E}(\epsilon_i(t)) = 0$ for $t = 0, 1$ and $\alpha = \mathbb{E}(Y_i(0))$

Identification Assumption

- Strict exogeneity assumption:

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

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

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

- Random sampling of units:
  - $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp \{Y_j(1), Y_j(0)\}$ for any $i \neq j$
  - $\{\epsilon_i(1), \epsilon_i(0)\} \perp\!\!\!\perp \{\epsilon_j(1), \epsilon_j(0)\}$ for any $i \neq j$

Unbiasedness of Least Squares Estimator

- Let $(\hat{\alpha}, \hat{\beta})$ be the least squares estimators
- When the treatment is binary,

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

- So, $\hat{\alpha}$ and $\hat{\beta}$ are unbiased for $\mathbb{E}(Y(0))$ and $\mathbb{E}(Y_i(1) - Y_i(0))$ from the randomization inference perspective
- The same conclusion holds if T is categorical
- When T is continuous, the model assumes $Y_i(t)$ is linear in T

Homoskedasticity

- Homoskedastic error:

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

  1. Random sampling of units implies $\epsilon_i \perp\!\!\!\perp \epsilon_j$
  2. Equal variance of potential outcomes

$$\mathbb{V}(\epsilon_i(t)) = \mathbb{V}(\epsilon_i(t) \mid \mathbf{T}) = \mathbb{V}(Y_i(t) \mid \mathbf{T}) = \mathbb{V}(Y_i(t)) = \sigma_t^2 \quad \text{for } t = 0, 1$$

- 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: $\sigma_1^2 \neq \sigma_0^2$

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

- 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 $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^\top$

- When the treatment is binary,

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

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

Improved Robust Variance Estimators

- HC2 heteroskedasticity-robust variance estimator:

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

  where $p_i = \mathbf{X}_i^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}_i$ is the leverage
- When the treatment is binary,

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

- Behrens-Fisher problem:
  - even under normality, $(\hat{\beta} - \beta) / \sqrt{\widehat{\mathbb{V}_{\mathsf{HC2}}(\hat{\beta} \mid \mathbf{T})}}$ does not follow the Student's t-distribution
  - Degrees of freedom adjustments by Welch (1951) and Bell and McCaffery (2002) to improve approximation

Reducing Transphobia (Brookman and Kalla. 2016. *Science*)

- Can perspective-taking – "imagining the world from another's perspective" – reduce transphobia?

- Treatment group ($n_1 = 912$): canvasser asking people to think about a time when they were judged unfairly and were guided to translate that experience to a transgender individual's experience
- Control group ($n_0 = 913$): canvasser asking people to recycle

- Outcome variable: feeling thermometer towards transgender people ($0 - 100$)
  1. baseline
  2. 3 day
  3. 3 week
  4. 6 week
  5. 3 month

Durable Effects

1. 3 day effect
   - $n_1 = 291$, $n_0 = 309$, $\hat{\sigma}_1^2 = 933.87$, $\hat{\sigma}_0^2 = 666.45$
   - Neyman: est $= 6.76$, s.e. $= 2.32$, 95% CI $= [2.23,\ 11.28]$
   - Regression: est $= 6.76$, s.e. $= 2.30$, 2.31 (HC), 2.32 (HC2)
2. 3 month effect
   - $n_1 = 179$, $n_0 = 206$, $\hat{\sigma}_1^2 = 764.77$, $\hat{\sigma}_0^2 = 714.76$
   - Neyman: est $= 5.30$, s.e. $= 2.78$, 95% CI $= [-0.15,\ 10.75]$
   - Regression: est $= 5.30$, s.e. $= 2.78$, 2.78 (HC), 2.78 (HC2)

- For simplicity, we ignored attrition, which can be problematic
- The treatment group has more attrition and a larger variance
- Relatively balanced sample $\rightsquigarrow$ homoskedasticity assumption matters little

Cluster Randomized Experiments

- Setup:
  - Clusters of units: $j = 1, 2, \ldots, J$
  - Number of treated (control) clusters: J_1 (J_0)
  - Units within cluster j: $i = 1, 2, \ldots, m_j$
  - Total sample size: $n = \sum_{j=1}^{J} m_j$
  - Treatment assignment at cluster level: $T_j \in \{0, 1\}$
  - Outcome: $Y_{ij} = Y_{ij}(T_j)$
- Assumptions:
  1. Random assignment: $\{Y_{ij}(1), Y_{ij}(0)\} \perp\!\!\!\perp T_j$ and $\Pr(T_j = 1) = J_1/J$
     $\rightsquigarrow$ often easy to assign to each cluster
  2. No interference across clusters $\rightsquigarrow$ useful when interference within a cluster is likely
- Causal quantities of interest:

$$\text{SATE} \equiv \frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{m_j}(Y_{ij}(1) - Y_{ij}(0))$$

$$\text{PATE} \equiv \mathbb{E}(Y_{ij}(1) - Y_{ij}(0))$$

Randomization Inference

- Assume equal cluster size for simplicity, i.e., $m_j = m$ for all j
- Unbiased estimator:

$$\hat{\tau}_{\text{cluster}} \quad = \quad \frac{1}{J_1} \sum_{j=1}^{J} T_j \left(\frac{1}{m} \sum_{i=1}^{m} Y_{ij} \right) - \frac{1}{J_0} \sum_{j=1}^{J} (1 - T_j) \left(\frac{1}{m} \sum_{i=1}^{m} Y_{ij} \right)$$

- Easy to show $\mathbb{E}(\hat{\tau}_{\text{cluster}} \mid \mathcal{O}) = \text{SATE}$ and thus $\mathbb{E}(\hat{\tau}_{\text{cluster}}) = \text{PATE}$
- Population variance (random sampling of clusters):

$$\mathbb{V}(\hat{\tau}_{\text{cluster}} \mid \mathcal{O}) \quad \leq \quad \mathbb{V}(\hat{\tau}_{\text{cluster}}) \;=\; \frac{\mathbb{V}(\overline{Y_j(1)})}{J_1} + \frac{\mathbb{V}(\overline{Y_j(0)})}{J_0}$$

where for $t = 0, 1$

$$\overline{Y_j(t)} \;=\; \frac{1}{m} \sum_{i=1}^{m} Y_{ij}(t)$$

Intracluster Correlation Coefficient (ICC)

- Comparison with the standard variance:

$$\mathbb{V}(\hat{\tau}) = \frac{\sigma_1^2}{J_1 m} + \frac{\sigma_0^2}{J_0 m}$$

$$\begin{aligned}
\frac{\mathbb{V}(\overline{Y_j(t)})}{J_t} &= \frac{1}{J_t m^2}\mathbb{E}\left[\left(\sum_{i=1}^{m} Y_{ij}(t) - m\cdot\mu_t\right)^2\right] \\
&= \frac{1}{J_t m^2}\mathbb{E}\left[\sum_{i=1}^{m}(Y_{ij}(t)-\mu_t)^2 + \sum_{i'=1}^{m}\sum_{i'\neq i}(Y_{ij}(t)-\mu_t)(Y_{i'j}(t)-\mu_t)\right] \\
&= \frac{\sigma_t^2}{J_t m}\{1+(m-1)\rho_t\} \overset{\text{typically}}{\geq} \frac{\sigma_t^2}{J_t m}
\end{aligned}$$

- Suppose $m = 101$ and $\rho_t = 0.5 \rightsquigarrow$ more than 50 times greater
- Random effect model: $\eta_j + \epsilon_{ij}$ where $\rho = \mathbb{V}(\eta_j)/\{\mathbb{V}(\eta_j) + \mathbb{V}(\epsilon_{ij})\} \geq 0$

Cluster Robust Variance Estimator

- Cluster robust variance estimator (Liang and Zeger. 1986. *Biometrika*):

$$\widehat{\mathbb{V}_{\text{cluster}}((\hat{\alpha}, \hat{\beta}) \mid \mathbf{X})} = \left(\sum_{j=1}^{J} \mathbf{X}_j^\top \mathbf{X}_j\right)^{-1} \left(\sum_{j=1}^{J} \mathbf{X}_j^\top \hat{\epsilon}_j \hat{\epsilon}_j^\top \mathbf{X}_j\right) \left(\sum_{j=1}^{J} \mathbf{X}_j^\top \mathbf{X}_j\right)^{-1}$$

  where $\mathbf{X}_j = (\mathbf{1}_{m_j}^\top, \mathbf{T}_j^\top)^\top$ and $\hat{\epsilon}_j = (\hat{\epsilon}_{1j}, \ldots, \hat{\epsilon}_{m_j j})^\top$
- Consistent as the number of clusters goes to infinity

- Bias adjustment a.k.a. CR2 (Bell and McCaffrey. 2002. *Surv. Methodol.*);

$$\text{meat} = \sum_{j=1}^{J} \mathbf{X}_j^\top (\mathbf{I}_{m_j} - \mathbf{P}_{\mathbf{X}_j})^{-1/2} \hat{\epsilon}_j \hat{\epsilon}_j^\top (\mathbf{I}_{m_j} - \mathbf{P}_{\mathbf{X}_j})^{-1/2} \mathbf{X}_j$$

  where $\mathbf{P}_{\mathbf{X}_j} = \mathbf{X}_j (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}_j$
- Degree of freedom adjustments are also available
- **Implication:** cluster standard errors by the unit of treatment assignment

Effects of Clientalistic Campaign (Wantchekon. 2003. *World Politics*)

- Does clientalism affect voting behavior?
- Experiment involving 4 main candidates during the 2001 presidential election in Benin
- 8 non-competitive districts (out of 84) are selected
- Random assignment within each district:
  1. one village: clientalistic campaign
  2. one village: public policy campaign
  3. the other villages: control group
- Many experimental villages were at least 25 miles apart: “The risk of contagion was thereby minimized”
- Post-election survey of 2344 registered voters to measure vote choice

Stratified Cluster Randomized Design

District	*Candidate*	*Exp. Villages*	*Reg. Voters*	*Sample Size*	*Sample Mean*	*Population Mean*
Kandi	Kerekou	clientelist	1133	61	1.00 (0)	0.81
		public policy	1109	60	0.49 (.50)	0.60
		control	3896	61	0.96 (.18)	0.75
Nikki	Kerekou	clientelist	462	60	0.95 (.21)	0.90
		public policy	1090	60	0.93 (.24)	0.85
		control	2979	60	0.95 (.20)	0.82
Bembereke	Lafia	clientelist	999	60	0.92 (.26)	0.94
		public policy	931	60	0.89 (.30)	0.93
		control	5204	61	0.91 (.28)	0.74
Perere	Lafia	clientelist	657	59	0.76 (.42)	0.81
		public policy	442	60	0.13 (.33)	0.25
		control	4477	61	0.52 (.40)	0.58
Abomey	Soglo	clientelist	1172	60	0.98 (.13)	0.91
		public policy	1199	60	0.98 (.13)	0.90
		control	5204	61	0.74 (.15)	0.86
Ouidah	Soglo	clientelist	321	60	0.93 (.25)	0.86
		public policy	701	61	0.92 (.26)	0.72
		control	2414	60	0.73 (0.44)	0.64
Aplahoue	Amoussou	clientelist	492	59	0.98 (.13)	0.87
		public policy	511	60	0.91 (.28)	0.77
		control	4037	61	0.98 (.20)	0.72
Dogbo	Amoussou	clientelist	1397	60	0.64 (.48)	0.65
		public policy	736	61	0.50 (.50)	0.47
		control	1161	59	0.45 (0.44)	0.84

Empirical Results

Neyman		public policy		
est.	−0.070	−0.074	−0.070	**−0.074**
s.e.	0.033	0.024	0.182	**0.093**
cluster assignment	No	No	Yes	**Yes**
stratification	No	Yes	No	**Yes**
		clientalism		
est.	0.127	0.109	0.127	**0.109**
s.e.	0.030	0.022	0.167	**0.043**
cluster assignment	No	No	Yes	**Yes**
stratification	No	Yes	No	**Yes**

Regression	public policy		clientalism	
	HC2	CR2	HC2	CR2
est.	−0.070		0.127	
s.e.	0.033	0.096	0.030	0.170

Summary

- Simple linear regression $\rightsquigarrow$ Difference-in-means estimator
- Homoskedasticity implies the equal variance of potential outcomes
- Heteroskedasticity-robust variance estimator relaxes this assumption

- Cluster randomized experiments:
  1. permits interference within cluster
  2. less powerful when intracluster correlation is high
  3. use cluster-robust standard variance estimator

- Suggested readings:
  1. ANGRIST AND PISCHKE, Chapter 2
  2. IMBENS AND RUBIN, Chapter 7