Regression with Observational Data

Kosuke Imai

Harvard University

Spring 2021

Difficulties of Observational Studies

- Observational studies $\rightsquigarrow$ No randomized treatment assignment
- Confounding:

$$\{Y_i(1), Y_i(0)\} \not\perp\!\!\!\perp T_i$$

- Treatment assignment mechanism is often unknown
- Possible existence of observed and unobserved confounders
- Credible causal inference in observational studies

What is your identification assumption/strategy?

- In causal inference, identification precedes statistical inference:
  1. Identification: How much can you learn about the estimand if you had an infinite amount of data?
  2. Statistical Inference: How much can you learn about the estimand from a finite sample?

Identification of the Average Treatment Effect

- Identification assumptions:
  1. Overlap / Positivity (i.e., no extrapolation):

$$0 < \Pr(T_i = 1 \mid \mathbf{X}_i = \mathbf{x}) < 1 \text{ for any } \mathbf{x}$$

  2. Unconfoundedness (exogeneity, ignorability, no omitted variable, selection on observables, etc.)

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp T_i \mid \mathbf{X}_i = \mathbf{x} \text{ for any } \mathbf{x}$$

- Under these assumptions:

$$\begin{aligned}
\tau &= \mathbb{E}\{Y_i(1) - Y_i(0)\} \\
&= \mathbb{E}[\mathbb{E}\{Y_i(1) - Y_i(0) \mid \mathbf{X}_i\}] \\
&= \mathbb{E}\left[\mathbb{E}\{Y_i(1) \mid T_i = 1, \mathbf{X}_i\} - \mathbb{E}\{Y_i(0) \mid T_i = 0, \mathbf{X}_i\}\right] \\
&\quad \text{(overlap + unconfoundedness)} \\
&= \mathbb{E}\{\mu_1(\mathbf{X}_i) - \mu_0(\mathbf{X}_i)\}
\end{aligned}$$

where $\mu_t(\mathbf{x}) = \mathbb{E}(Y_i \mid T_i = t, \mathbf{X}_i = \mathbf{x})$ for $t = 0, 1$

Regression-based Causal Estimation

- Two general regression-based estimators:
  1. Plug-in estimator:

$$\hat{\tau}_{\text{reg}} = \frac{1}{n}\sum_{i=1}^{n}\{\hat{\mu}_1(\mathbf{X}_i) - \hat{\mu}_0(\mathbf{X}_i)\}$$

  2. Imputation estimator:

$$\hat{\tau}_{\text{reg-imp}} = \frac{1}{n}\sum_{i=1}^{n}[T_i\{Y_i - \hat{\mu}_0(\mathbf{X}_i)\} + (1 - T_i)\{\hat{\mu}_1(\mathbf{X}_i) - Y_i\}]$$

- Linear regressions (with/without interactions) $\rightsquigarrow$ use coefficients
- Nonlinear regressions: e.g., Logistic regression

$$\hat{\tau}_{\text{reg}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\exp(\hat{\alpha} + \hat{\beta}\cdot 1 + \mathbf{X}_i^\top\hat{\gamma})}{1 + \exp(\hat{\alpha} + \hat{\beta}\cdot 1 + \mathbf{X}_i^\top\hat{\gamma})} - \frac{\exp(\hat{\alpha} + \hat{\beta}\cdot 0 + \mathbf{X}_i^\top\hat{\gamma})}{1 + \exp(\hat{\alpha} + \hat{\beta}\cdot 0 + \mathbf{X}_i^\top\hat{\gamma})}\right\}$$

Asymptotic Variance Calculation

- Delta method for the conditional variance:

$$\mathbb{V}(\hat{\tau}_{\text{reg}} \mid \mathbf{X}) = \frac{1}{n^2}\mathbb{V}\left(\sum_{i=1}^{n} \hat{\mu}_1(\mathbf{X}_i) - \hat{\mu}_0(\mathbf{X}_i) \mid \mathbf{X}\right)$$

$$= \frac{1}{n^2}\left[\sum_{i=1}^{n}\{\mathbb{V}(\hat{\mu}_1(\mathbf{X}_i) \mid \mathbf{X}) + \mathbb{V}(\hat{\mu}_0(\mathbf{X}_i) \mid \mathbf{X}) - 2\text{Cov}(\hat{\mu}_1(\mathbf{X}_i), \hat{\mu}_0(\mathbf{X}_i) \mid \mathbf{X})\} + \sum_{i=1}^{n}\sum_{i' \neq i}\text{Cov}(\hat{\mu}_1(\mathbf{X}_i) - \hat{\mu}_0(\mathbf{X}_i), \hat{\mu}_1(\mathbf{X}_{i'}) - \hat{\mu}_0(\mathbf{X}_{i'}) \mid \mathbf{X})\right]$$

- Bootstrap for the unconditional variance:
  1. Independently sample n observations with replacement
  2. Fit a regression model and compute $\hat{\tau}_{\text{reg}}$

- Quasi-Bayesian Monte Carlo (Zelig; King et al. 2000. *Amer. J. Political Sci*):
  1. Sample (α, β, γ) from $\mathcal{N}((\hat{\alpha}, \hat{\beta}, \hat{\gamma}), \widehat{\mathbb{V}((\hat{\alpha}, \hat{\beta}, \hat{\gamma}))})$
  2. Compute $\hat{\tau}_{\text{reg}}$

Sensitivity Analysis for Linear Regression

- Linear regression model:

$$Y_i = \alpha + \beta T_i + \gamma^\top \mathbf{X}_i + \delta U_i + \epsilon_i$$

where U_i is an unobserved (scalar) confounder

- Recall the omitted variable bias formula:

$$\hat{\beta} \xrightarrow{p} \beta + \delta \times \underbrace{\frac{\mathrm{Cov}(T_i^{\perp \mathbf{X}}, U_i^{\perp \mathbf{X}})}{\mathbb{V}(T_i^{\perp \mathbf{X}})}}_{\text{regression of } U_i^{\perp \mathbf{X}} \text{ on } T_i^{\perp \mathbf{X}}}$$

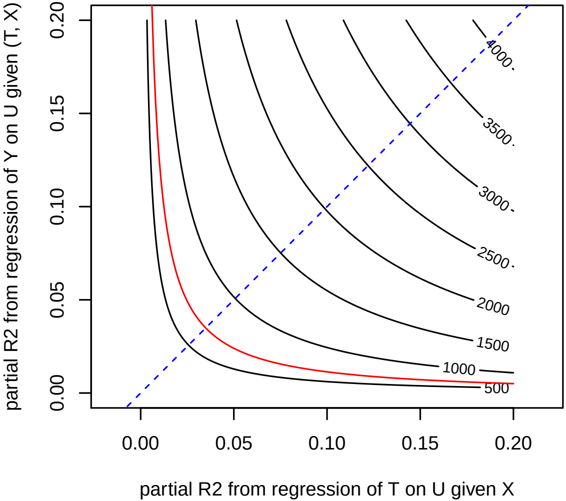
- Partial R^2 formulation (Cineli and Hazlett. 2020. *J. R. Stat. Soc. B*):

$$|\widehat{\text{bias}}| = \sqrt{R^2_{Y \sim U | T, \mathbf{X}} \times \frac{R^2_{T \sim U | \mathbf{X}}}{1 - R^2_{T \sim U | \mathbf{X}}} \times \frac{\widehat{\mathbb{V}(Y_i^{\perp \mathbf{X}, T})}}{\widehat{\mathbb{V}(T_i^{\perp \mathbf{X}})}}}$$

where e.g., $\underbrace{R^2_{Y \sim U | T, \mathbf{X}}}_{\text{partial } R^2} = \underbrace{(R^2_{Y \sim U + T + \mathbf{X}} - R^2_{Y \sim T + \mathbf{X}})}_{\text{additional variance explained by } U} \, / \, \underbrace{(1 - R^2_{Y \sim T + \mathbf{X}})}_{\text{unexplained by } T, \mathbf{X}}$

Sensitivity Analysis Results

- Linear regression estimate: \$1548 (s.e. = \$750)



Selection Bias as Misspecification (Heckman. 1978. *Econometrica*)

- The outcome model: $Y_i \;=\; \alpha + \beta T_i + \boldsymbol{\gamma}^\top \mathbf{X}_i + \epsilon_i$
- The selection model: $T_i = \mathbf{1}\{T_i^* > 0\}$ with $T_i^* \;=\; \lambda + \mathbf{X}_i^\top \boldsymbol{\delta} + \eta_i$ which equals the probit model if $\eta_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$
- Selection bias: $\mathbb{E}(\epsilon_i \mid T_i, \mathbf{X}_i) \neq 0$ if $\epsilon_i \not\perp\!\!\!\perp \eta_i \mid \mathbf{X}_i$

$$
\begin{aligned}
\mathbb{E}(Y_i \mid \mathbf{X}_i, T_i = 1) &= \alpha + \beta + \boldsymbol{\gamma}^\top \mathbf{X}_i + \mathbb{E}(\epsilon_i \mid \mathbf{X}_i, T_i = 1) \\
&= \alpha + \beta + \boldsymbol{\gamma}^\top \mathbf{X}_i + \mathbb{E}(\epsilon_i \mid \mathbf{X}_i, \eta_i > -\lambda - \boldsymbol{\delta}^\top \mathbf{X}_i) \\
\mathbb{E}(Y_i \mid \mathbf{X}_i, T_i = 0) &= \alpha + \boldsymbol{\gamma}^\top \mathbf{X}_i + \mathbb{E}(\epsilon_i \mid \mathbf{X}_i, \eta_i < -\lambda - \boldsymbol{\delta}^\top \mathbf{X}_i)
\end{aligned}
$$

- Selection bias as a specification error
  - Bivariate normal assumption:

$$
\begin{pmatrix} \epsilon_i \\ \eta_i \end{pmatrix} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix}\right]
$$

  - Inverse Mill's ratio

$$
\mathbb{E}(\epsilon_i \mid \mathbf{X}_i, T_i) \;=\; W_i \rho\sigma \frac{\phi(\lambda + \boldsymbol{\delta}^\top \mathbf{X}_i)}{\Phi(W_i(\lambda + \boldsymbol{\delta}^\top \mathbf{X}_i))} \quad \text{where } W_i = 2T_i - 1
$$

- Two-step estimation; Identification by parametric assumption

Control Function Method

- Control function: a variable that, when adjusted for, renders an otherwise endogenous treatment variable exogenous
- Instrumental variables needed for nonparametric identification
- An alternative formulation of the two-stage least squares
  1. Regress T_i on Z_i and $\mathbf{X}_i$ and get residuals $\hat{\eta}_i$
  2. Regress Y_i on T_i, $\mathbf{X}_i$, and residuals $\hat{\eta}_i$

  $\rightsquigarrow$ $\hat{\eta}_i$ is a control function
- Nonparametric identification (Imbens and Newey. 2009. *Econometrica*)
  - Triangular system:

$$\begin{aligned} Y_i &= f(T_i, \epsilon_i) \\ T_i &= g(Z_i, \eta_i) \end{aligned}$$

    where $Z_i \perp\!\!\!\perp \{\epsilon_i, \eta_i\}$
  - $C_i = \Pr(T_i \leq t \mid Z_i)$ is a control function: $\epsilon_i \perp\!\!\!\perp T_i \mid C_i$