Estimation of Causal Effect in the Absence of Treatment Observability

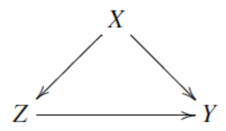
Joon Sup Park *supervised by* Dr. Fan Li

Department of Statistical Science, Duke University

Joonsup.park@duke.edu

Motivations & Backgrounds

- Causal Inference: Estimating the counterfactual effect of treatment Z_i on outcome Y_i: $Y_i(1) - Y_i(0)$
- Average Treatment Effect (ATE): $\delta = \mathrm{E}[Y_i(1) - Y_i(0)]$
- Need to control for confounders X_i that may affect both treatment and outcome:



- But treatment observability is not guaranteed in many practices
- Non-compliance: the actual treatment a patient takes is different from the treatment assigned to her

Research Goals

- Propose a method that recuperates the treatment Z from the information provided by the outcome Y and confounders X
- Propose a heuristic that summarizes when the method is applicable
- Conduct simulations to check the performance, robustness, and sensitivity of the method

(Fairly Standard) Assumptions

- Two ways to control for confounders:
  1. Outcome model-based (Parametric)
  2. Propensity score-based (Non-parametric)
- Assume the following outcome model as data generating process:
  $$Y_i = f(X_i, Z_i) + \epsilon_i = X_i^T\beta + Z_i\delta + \epsilon_i, \quad \epsilon_i \sim_{iid} N(0, \sigma_y^2)$$
- A1 (Ignorability): $\{Y_i(0), Y_i(1)\} \perp Z_i \mid X_i, \quad \forall i$
  → No unmeasured confounders!
- A2 (Overlap): $0 < \Pr(Z_i \mid X_i) < 1, \quad \forall i$
- A3 (Binary Treatment): $Z_i \in \{0,1\}, \quad \forall i$
- A4 (Positive Treatment Effect): $\delta > 0$
- A5 (Knowledge of the functional form $f(X_i, Z_i)$)

Method

- In matrix form: $Y = X\beta + Z\delta + \epsilon, \quad \epsilon \sim N(0, \sigma_y^2 I_n)$
- Step 1: Regress Y on X alone to obtain our first estimate of β

$$\hat{\beta}^{(1)} = (X^TX)^{-1}X^Ty = \beta + (X^TX)^{-1}X^T(Z\delta + \epsilon)$$

which is biased. Get the residual

$$\hat{\epsilon}^{(1)} = Y - X\hat{\beta}^{(1)} = (I_n - X(X^TX)^{-1}X^T)(Z\delta + \epsilon)$$

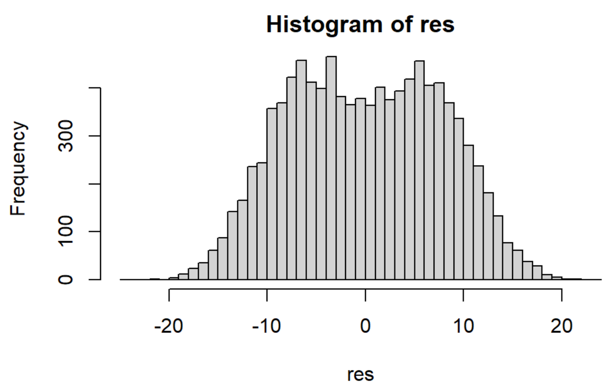
We use the bias in the residual to reconstruct our estimate of Z

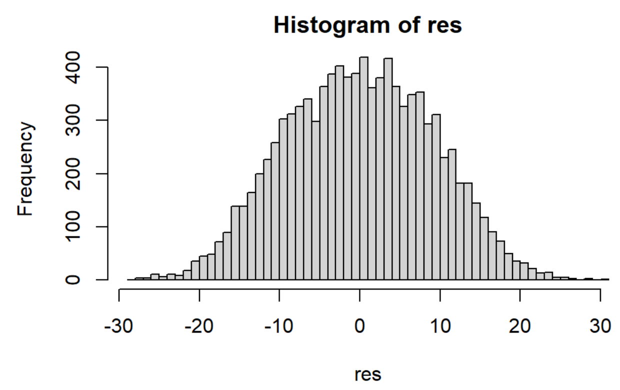
Method (Cont'd)

- Step 2: Since $Z_i = 0$ or $Z_i = 1$,

$$\hat{\epsilon}^{(1)} = (I_n - X(X^TX)^{-1}X^T)(Z\delta + \epsilon)$$

will be distributed around 2 centers 0 and $(I_n - X(X^TX)^{-1}X^T)Z\delta$





→ Clustering $\hat{\epsilon}^{(1)}$ into 2 groups may be feasible!

Label the unit i's in the group with the larger center with $\hat{Z}_i^{(1)} = 1$
Label the unit i's in the group with the smaller center with $\hat{Z}_i^{(1)} = 0$

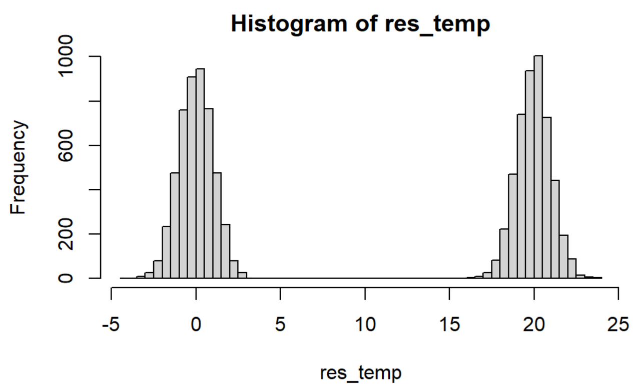
- Step 3: Regress Y on X and $\hat{Z}^{(1)}$ to obtain our second estimate of β, $\hat{\beta}^{(2)}$, and our first estimate of δ, $\hat{\delta}^{(1)}$
- Step 4: If the histogram of $\hat{\epsilon}^{(1)}$ suggests that clustering is promising, then we know that our estimate $\hat{Z}^{(1)}$ would provide good information about Z

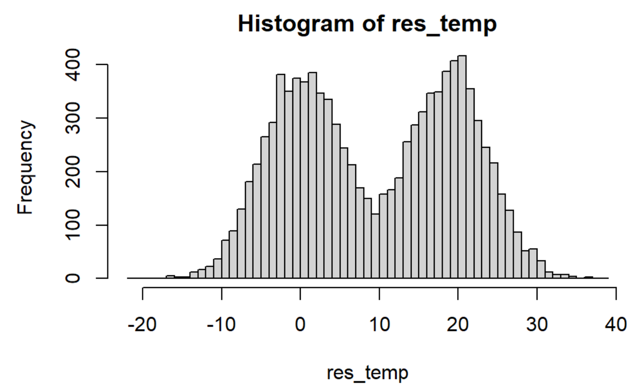
  → Our estimate $\hat{\beta}^{(2)}$ obtained from regressing Y on X and $\hat{Z}^{(1)}$ will be a better estimate of β than $\hat{\beta}^{(1)}$ obtained from regressing Y on X alone

  → The residual $\hat{\epsilon}^{(2)} = Y - X\hat{\beta}^{(2)}$ will provide more precise information about Z than $\hat{\epsilon}^{(1)} = Y - X\hat{\beta}^{(1)}$

  → Clustering $\hat{\epsilon}^{(2)}$ into 2 groups will result in finer labeling and a better estimate of Z, $\hat{Z}^{(2)}$

  → Regressing Y on X and $\hat{Z}^{(2)}$ will result in better estimates of β, $\hat{\beta}^{(3)}$, and of δ, $\hat{\delta}^{(2)}$
- Step 5: Iterate Step 4 until $\|\hat{\beta}^{(S+1)} - \hat{\beta}^{(S)}\| < t$ for some small threshold value t, and obtain our final estimates of Z and ATE, $\hat{Z}^{(S)}$ and $\hat{\delta}^{(S)}$





Simulation Results

- The data generating process for simulation:

$$X_i = (X_{i,1}, X_{i,2}, \cdots, X_{i,16}) \sim_{iid} Multivariate\ Normal(0_{16}, \sigma_x^2 I_{16})$$

$$Z_i \sim_{iid} Bernoulli(\pi_i), \quad where\ \pi_i = \frac{\exp\{X_i\theta\}}{1 + \exp\{X_i\theta\}}$$

where $\theta = (-1, 0.5, -0.25, -0.1, \ \cdots\ , -1, 0.5, -0.25, -0.1)$ and

$$Y_i = 210 + X_i^T\beta + Z_i\delta + \epsilon_i, \quad where\ \epsilon_i \sim_{iid} N(0, \sigma_y^2)$$

where $\beta = (27.4, 13.7, -10, 20, \cdots, 27.4, 13.7, -10, 20)$ and $\delta = 20$

for $i \in \{1, 2, \cdots, 10000\}$

Simulation Results (Cont'd)

- $\sigma_y^2 = 1$ and $\sigma_x^2 = 1$

  $\widehat{Pr}(\{\hat{Z}^{(1)} = Z\}) = 0.8797, \quad \hat{\delta}^{(1)} = 13.50$

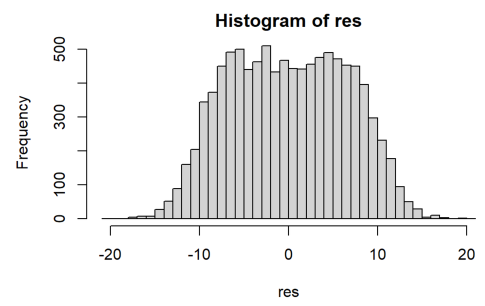
  → $\widehat{Pr}(\{\hat{Z}^{(S)} = Z\}) = 1, \quad \hat{\delta}^{(S)} = 20.002$
- $\sigma_y^2 = 5$ and $\sigma_x^2 = 1$

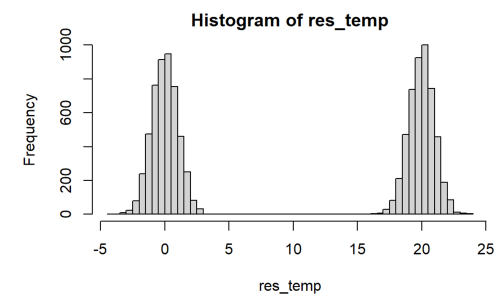
  $\widehat{Pr}(\{\hat{Z}^{(1)} = Z\}) = 0.7997, \quad \hat{\delta}^{(1)} = 15.07$

  → $\widehat{Pr}(\{\hat{Z}^{(S)} = Z\}) = 0.9675, \quad \hat{\delta}^{(S)} = 18.98$
- $\sigma_y^2 = 1$ and $\sigma_x^2 = 5$

  $\widehat{Pr}(\{\hat{Z}^{(1)} = Z\}) = 0.8081, \quad \hat{\delta}^{(1)} = 11.31$

  → $\widehat{Pr}(\{\hat{Z}^{(S)} = Z\}) = 1, \quad \hat{\delta}^{(S)} = 19.98$

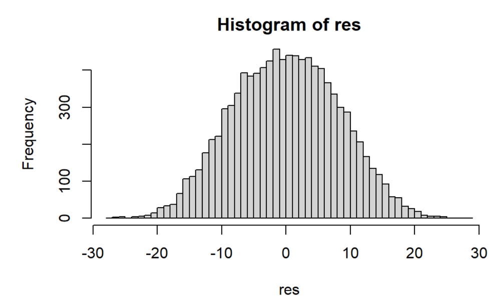


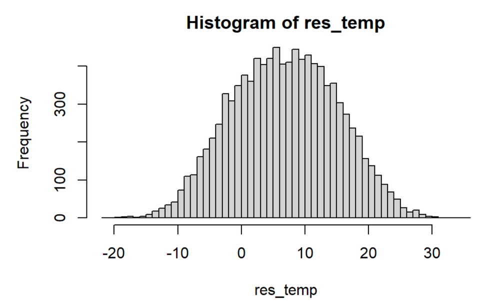


- $\sigma_y^2 = 5$ and $\sigma_x^2 = 5$

  $\widehat{Pr}(\{\hat{Z}^{(1)} = Z\}) = 0.7353, \quad \hat{\delta}^{(1)} = 13.44$

  → $\widehat{Pr}(\{\hat{Z}^{(S)} = Z\}) = 0.7641, \quad \hat{\delta}^{(S)} = 13.55$





Conclusion

- Successful clustering of $\hat{\epsilon}^{(1)}$ depends on how large δ is relative to σ_y^2 and σ_x^2, and their relative sizes are summarized well in the histogram of $\hat{\epsilon}^{(1)}$

→ This serves as a heuristic to see if the method would be applicable

- The method was robust to increasing the dimensions
- Uncertainty quantification of the estimate from the method is still under consideration
- Application of the method to non-compliance settings is straightforward: we would have an additional variable "assigned treatment" W distinguished from "actual treatment" Z, but W will affect Y only through Z and does not enter in the outcome model

→ No change in the method required to recuperate actual treatment Z

- Extension of the method to non-parametric cases is on the way: To relax the assumption of the knowledge in outcome model, we substitute machine learning algorithms for linear regression