Segment 3: Causal Inference with Regression (in randomized studies)

Section 04: Regression for Causal Inference

Regression for Causal Inference: What's Different?

- Most often motivated as a *prediction* problem
  - Model how values of Y differ across values of X
  - Tailored to comparisons *between* units

$$Y_i = \beta_0 + \beta_1 Z_i + \beta_2 x_i + \cdots + \varepsilon_i$$

- But **causal inference** is all about comparing what would happen if different treatments were applied to the *same units*

$$\tau_{SATE} = E[Y_i^t - Y_i^c]$$

  - What is the average difference between units with $Z = 0$ and $Z = 1$?
  - What would be the difference if units with $Z = 0$ had actually had $Z = 1$/?

Causal Inference Doesn't *Need* Unfamiliar Statistics!

Especially with ranomization....

- Can rely on familiar inferential tools, like regression
- Main difference will be formalizing the underlying structure/assumptions to assess causal validity
  - What are the units?
  - What is the treatment?
  - What are the potential outcomes?
  - What is the *assignment mechanism*?
  - What is the role of covariates (blocking)?
- Much of the above is fairly trivial for randomized studies
- Eventually, we will consider non-randomized (observational) studies where the above are more critical

Regression Models in Randomized Experiments

- Common for the analysis of both experimental and observational data
- Four key features
  1. Models for the *observed* outcomes
  2. Model for the conditional mean only (not the full distribution)
  3. Average Treatment Effect will be a parameter of statistical model
  4. Whether the model accurately describes the conditional mean is immaterial for the large-sample unbiasedness of estimators for the average causal effect
     - I.e., Randomization guarantees unbiased estimation of the ATE even if the regression model is the "wrong" specification of the conditional mean of Y

Units, Covariates, Treatments, Outcomes

response $\qquad \in \{0,1\}$ "treatment"

$$Y_i = \alpha + \tau Z_i + \mathbf{X}_i\beta + \epsilon_i$$

units, $i = 1, \dots, n$

$$\underline{X_i} = (x_{i1}, x_{i2}, \dots, x_{ip})$$

"pre-treatment"

General guidance: Think of a single "treatment" as distinct from $\mathbf{X}$

Observed Y vs. Y^z

$$Y_i^{obs} = \alpha + \tau Z_i + \mathbf{X}_i\beta + \epsilon_i$$

$$E[Y|Z=1] = \alpha + \tau + X\beta = E[Y^1|Z=1]$$

$$E[Y|Z=0] = \alpha + X\beta = E[Y^0|Z=0]$$

$$E[Y|Z=1] - E[Y|Z=0] = \tau$$

Causal Inference as Prediction

- Special case of prediction
  - Predict what would have happened to the same unit(s) under different treatments
  - Vs. Predict what would happen for a unit with a particular set of features, treatments
- That is, try to predict the *other* potential outcome
  - A regression model may help to generate a "close substitute"

Table: Observed Data from the Hypothetical Dietary Experiment

Unit, i	Treatment Z_i	Potential Outcome, Y_i^c	Potential Outcome, Y_i^t	Observed Outcome, Y_i
Audrey	0	140	?	140
Anna	0	140	?	140
Bob	0	150	?	150
Bill	0	150	?	150
Caitlin	1	?	155	155
Cara	1	?	155	155
Dave	1	?	160	160
Doug	1	?	160	160

predict

Regression with No Covariates

and randomized treatment assignment

$$Y_i^{obs} = \alpha + \tau Z_i + \epsilon_i$$

- $\tau = E[Y^1 - Y^0]$
- $\hat{\tau}^{\text{ols}} = \bar{Y}_t^{\text{obs}} - \bar{Y}_c^{\text{obs}} = \hat{\tau}_{SATE} = \hat{\tau}_{PATE}$ randomiz.
- `stan_glm(y~ z)`
- Posterior distribution of τ dictates estimation of τ
  - Posterior median, sd, 95% interval, etc.

$$E[Y|Z=1] = \alpha + \tau = E[Y^1|Z=1] = E[Y^1]$$

$$-E[Y|Z=0] = \alpha = E[Y^0|Z=0] = E[Y^0]$$

$$\tau = E[Y^1] - E[Y^0]$$

Different Ways to Balance Covariates

Randomization is motivated in large part for its production of *covariate balance* across treatment groups

- Some methods can improve balance *by design*
  - Stratified randomized experiments /blocking
  - Paired randomized experiments

  (but balance is not *guaranteed* for any individual randomization)
- Could also attempt to correct covariate imbalance by *analysis*
  - Separate analysis within subgroups
  - Regression adjustment

Regression with Covariates

(still with complete randomization)

$$Y_i^{obs} = \alpha + \tau Z_i + X_i\beta + \epsilon_i$$

- α, β regarded as "nuisance" parameters that do not get a causal interpretation
- We *do not* assume that the regression function is correctly specified
  - I.e., that the Y_i^{obs} is actually linear in X_i and Z_i
- $\hat{\tau}^{\text{ols}}$ is unbiased for τ_{SATE} and τ_{PATE}
- `stan_glm(y~ z + x1 + x2 + ...+ xp)`
- Posterior distribution of τ dictates estimation of τ
  - Posterior median, sd, 95% interval, etc.

Benefits of Adjusting for Pre-Treatment Covariates

Under randomization

$$Y_i^{obs} = \alpha + \tau Z_i + X_i \beta + \epsilon_i$$

- Including $\mathbf{X}$ does not change interpretation of τ
  - Not generally true under *nonrandomized* treatment
- Can adjust for *random* imbalances between groups
  - "Unlucky assignments"
- Can adjust for *systematic* imbalances between groups
  - Like blocking factors
- Ultimately bring the estimate $\hat{\tau}$ closer to the truth in any given sample/randomization
  - Less variability around truth $\Rightarrow$ smaller standard error of a given estimate

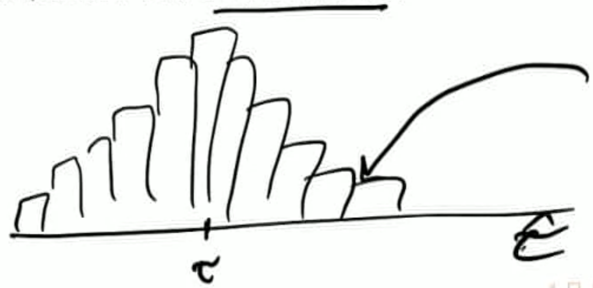
Benefits of Adjustment: Example

Effect of showing children an educational TV show

$n = 192$ elementary school classes
Z : New vs. standard educational TV program
Y : Reading ability measured at end of school year
X : Pre-test reading ability

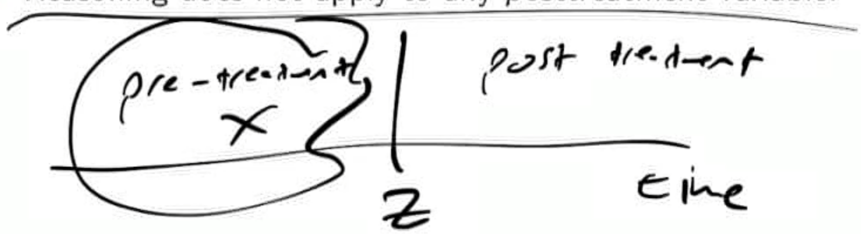
- Notice that pre-test reading is positively correlated with Y
- Notice that the average pre-test reading ability is Δ_x higher in the treatment group
  - "Unlucky" randomization
- (X positively correlated with Y) + ($\Delta_x > 0$) $\Rightarrow$ unadjusted comparison will overestimate τ

our estimate



Benefits of Adjusting for Pre-Treatment Covariates
Under randomization

- Same idea as the educational TV example applies to *any* pre-treatment covariate that may predict Y
- $\Rightarrow$ Always the chance to improve precision and bring $\hat{\tau}$ "closer to the truth" by adjusting for *predictors* of Y
- In *practice* we can't measure (or adjust) for *everything*
  - May need to prioritize which variable to include for adjustment
  - E.g., based on theory
  - E.g., based on some measure of imbalance and correlation with Y
- Reasoning does not apply to any *posttreatment* variable!



Special Case: Gain Scores

Setting: One covariate is a "pre-treatment" or "pre-test" or "baseline" version of Y

Can create a "gain score", $g_i = Y_i - X_i$

$$g_i = \alpha + \tau Z_i + \varepsilon_i \longrightarrow \hat{\tau} = \bar{g}^t - \bar{g}^c$$

$$Y_i - X_i = \alpha + \tau Z_i + \varepsilon_i \longrightarrow Y_i = \alpha + \tau Z_i + X_i + \varepsilon_i$$

"β" $\equiv 1$

unnecessary assumption

Varying Treatment Effects and Interactions

General Consideration: The treatment effect may vary across units with different $\mathbf{X}$

$$\tau_{CATE|X} = E[Y^1 - Y^0 \mid X = x]$$

$$Y_i^{obs} = \alpha + \theta Z_i + X_i\beta + Z_i X_i \gamma + \epsilon_i$$

Treatment effect:

$$E[Y \mid Z=1] = \alpha + \theta + X\beta + X\gamma$$

$$E[Y \mid Z=0] = \alpha + X\beta$$

$$E[Y \mid Z=1] - E[Y \mid Z=0] = \theta + X\gamma$$

"treatment effect" a function of x

$$\hat{\tau}_{SATE} = \sum_{i=1}^{n} (\theta + \gamma x_i)$$

Tau hat should be multiplied by a factor of 1/n

Unbiasedness Under Randomization

Even when the model is "wrong"

$$Y_i = \alpha + \tau Z_i + X_i\beta + Z_i X_i \gamma + \epsilon_i$$

$$E[Y \mid Z=1] = \underbrace{\alpha + \tau + (\beta + \gamma)x}_{\text{"}X\beta_t\text{"}} = E[Y^1 \mid Z = \phi]$$

$$E[Y \mid Z=0]\ \underbrace{\alpha + \beta x}_{\text{"}X\beta_c\text{"}} = E[Y^0 \mid Z=0]$$

$$E[Y^{mis} \mid Z=1] = E[Y^c \mid Z=1] \overset{\text{randomization}}{=} E[Y^c \mid Z=0] = X\hat{\beta}_c$$

$$E[Y^{mis} \mid Z=0] = E[Y^t \mid Z=0] = E[Y^t \mid Z=1] = X\hat{\beta}_t$$

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} Z_i(\underset{Y^t}{Y_i^{obs}} - \underset{Y^c_{miss}}{X_i\hat{\beta}_c}) + (1-Z_i)(\underset{Y^t_{miss}}{X\hat{\beta}_c} - \underset{Y^c}{Y_i^{obs}})$$

$$\hat{\tau}_c = \frac{1}{n_c}\sum_{i:Z_i=0}(X_i\hat{\beta}_t - Y_i^{obs})$$

$$\hat{\tau}_c = \bar{X}_c\hat{\beta}_t - \bar{Y}_c^{obs} \left(-\bar{X}_t\hat{\beta}_t + \bar{Y}_t^{obs}\right)$$

$$= \bar{Y}_t^{obs} - Y_c^{obs} + \hat{\beta}_t(\underbrace{\bar{X}_c - \bar{X}_t}_{=0 \text{ on average}})$$

"balance"

$$\hat{\tau}_t = \frac{1}{n_t}\sum_{i:Z_i=1}(Y_i^{obs} - X_i\hat{\beta}_c)$$

$$= \bar{Y}_t^{obs} - \bar{X}_t\hat{\beta}_c + \bar{X}_c\hat{\beta}_c - \bar{Y}_c^{obs}$$

$$= \bar{Y}_t^{obs} - \bar{Y}_c^{obs} + \hat{\beta}_c(\underbrace{\bar{X}_c - \bar{X}_t}_{=0 \text{ on average}})$$

$$\hat{\tau} = \frac{n_c}{n}\hat{\tau}_c + \frac{n_t}{n}\hat{\tau}_t$$

Summary: Regression for Completely Randomized Experiments

- Easy to incorporate covariates
- Unbiased point estimates without covariates
- Consistent point estimates for the ATE with covariates, *even when model is not true*
  - Point estimates biased in finite samples
  - Works in large samples
- Model for *observed outcomes*
- Causal estimand (ATE) is a parameter (or parameters) in the statistical model