Segment 4: Observational Studies

Section 04: Balance and Overlap

Ways to Analyze Observational Studies

Under the assumption of conditional ignorability $Z \perp\!\!\!\perp Y^c, Y^t | X$

Two ideas that we have already seen:

1. Subclassification
   - Create strata of the sample based on $X = x$
   - $\tau_{CATE|X=x} = E[Y|Z = 1, X = x] - E[Y|Z = 0, X = x]$
   - $\tau_{SATE} = \sum \tau_{CATE|X=x} Pr(X = x)$
   - But what if there are many possible values of X?

2. Regression
   - $E[Y^z|\mathbf{X}] = \beta_0 + \mathbf{X}\beta + \tau Z$
   - $\hat{\tau}$ represents, in principle, a weighted average of $\hat{\tau}_{CATE|X=x}$
   - This is not fool proof...requires a bit more conditions that held by design in randomized studies but may not hold in observational studies

Balance and Overlap

Analysis with regression models (conditional on X):

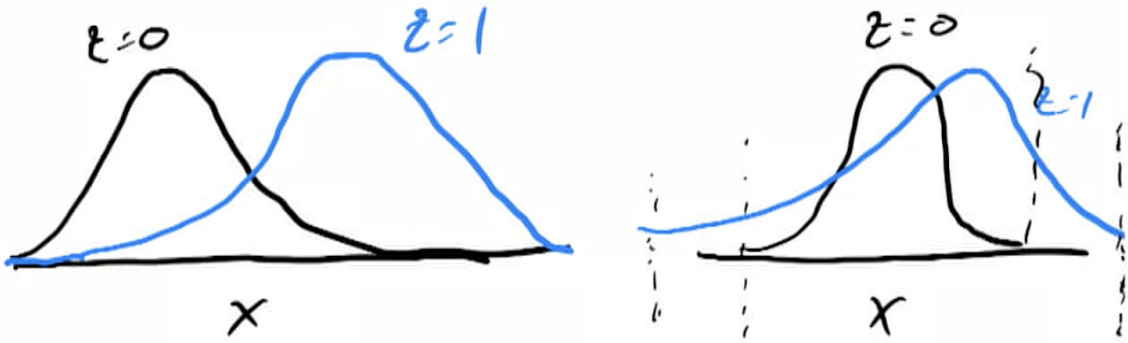
- In *randomized studies*, the regression model specification doesn't matter much
  - Unbiasedness even if the model is "wrong"
  - Potential precision gains
- In *observational studies*, the regression specification can lead us astray
  - Particularly if we have many X
- **Basic Idea:** Since balance (on X) is not guaranteed in an observation study, the model may fail at reliably predicting the "other" potential outcome
  - Because of *extrapolation*
- The dangerous circumstances can be described as conditions of *balance* and *overlap*

Balance

We know that the threat of *confounding* arises when a predictor of Y is also associated with Z

$\Rightarrow$ The distribution of $\mathbf{X}$ is different in the $Z = 0/1$ groups

- We know this as a lack of *balance*



Balance as a Threat to Regression Models

We have motivated regression models as one strategy to try to "recreate" balance

Key Property: More imbalance → regression model has to "work harder" → the more consequential the specification will be for estimating causal effects

$$Z=1: \; Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \theta + \text{error}_i$$
$$Z=0 \quad Y_i = \beta_0 + \beta_1 x_i + \beta_1 x_i^2 + \text{error}_i$$

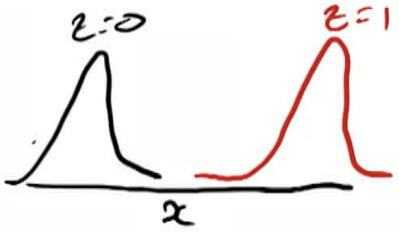
} true model

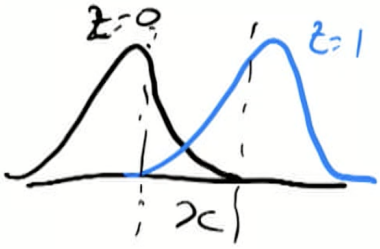
can show

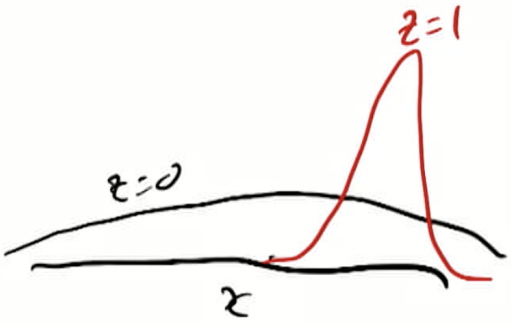
$$\theta = \bar{Y}_1 - \bar{Y}_0 \left[-\beta_1 (\bar{x}_1 - \bar{x}_0) - \beta_2 (\bar{x}_1^2 - \bar{x}_0^2) \right]$$

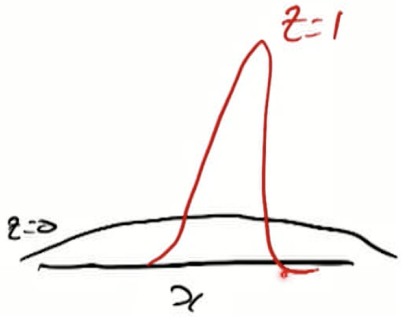
$$E[Y_i] = \beta_0 + \beta x + \theta Z$$

No Overlap: Extreme Lack of Balance









Lack of Overlap → Extrapolation

We know that what a regression model is doing implicitly is estimating:

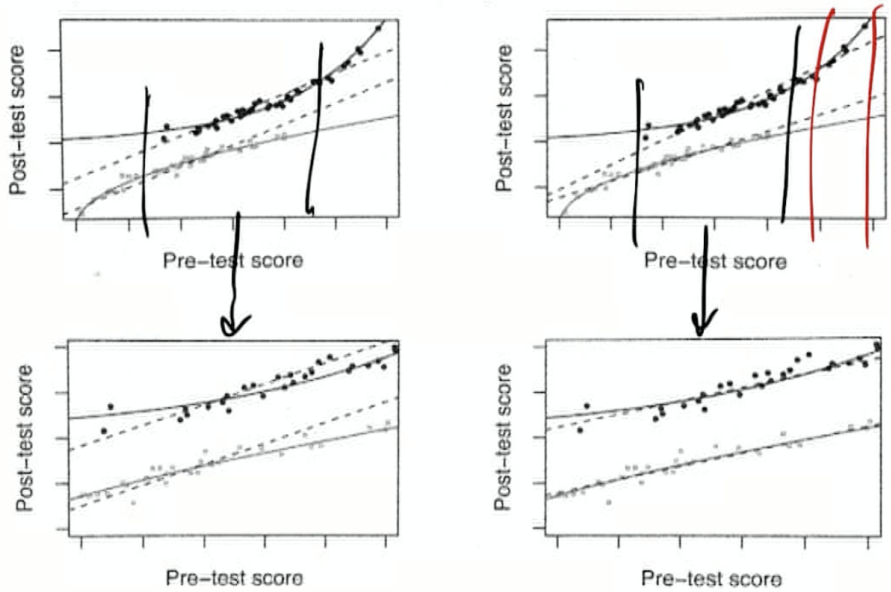
- How how average Y differs across levels of Z *among units with* $X = x$
- (which would be interpretable as a causal effect under certain assumptions)

But what happens when there are not *both* $Z = 0$ and $Z = 1$ units with $X = x$?

- The regression model doesn't "know" this, is agnostic with regard to overlap
- The parametric regression function will *extrapolate* an expected mean difference in $Y|Z = 0$ and $Y|Z = 1$ for $X = x$, *even when there is no such data in the sample*

Lack of Overlap → Extrapolation

Example from the Electric Company:



One Possible "Solution"

The dangers of lack of overlap amount to the dangers of the model extrapolation to areas of the distribution of $\mathbf{X}$ where there are no (or very few) treated or control units

One possible solution: Exclude those areas of $\mathbf{X}$ from the analysis!

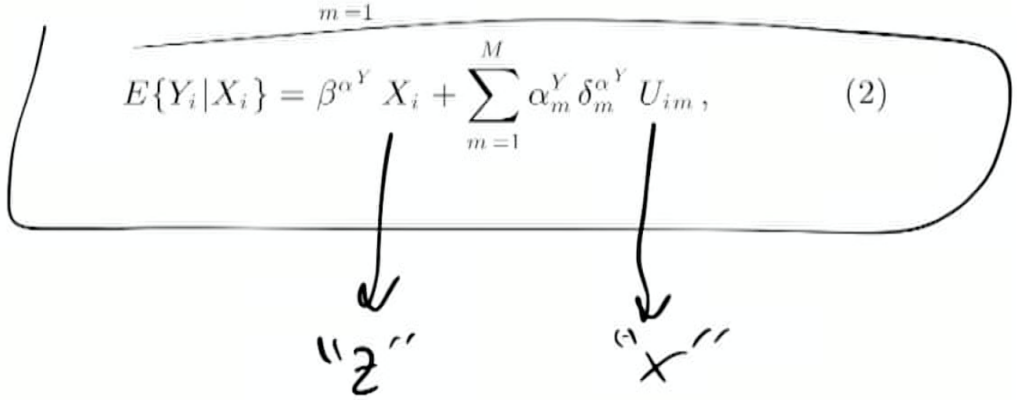
- Pruning, pre-processing, trimming
- **The good:** Make an inference for causal effect *only* in the areas where you are more likely to believe the model
- **The bad:** No longer making inference in the sample/population you started with
  - Those excluded due to lack of overlap may be different from those not excluded
  - Inference in the pruned/trimmed sample may not pertain to the entire population
  - "Changing the estimand"

Example: Bayesian Adjustment for Confounding

Underlying Models in WPD

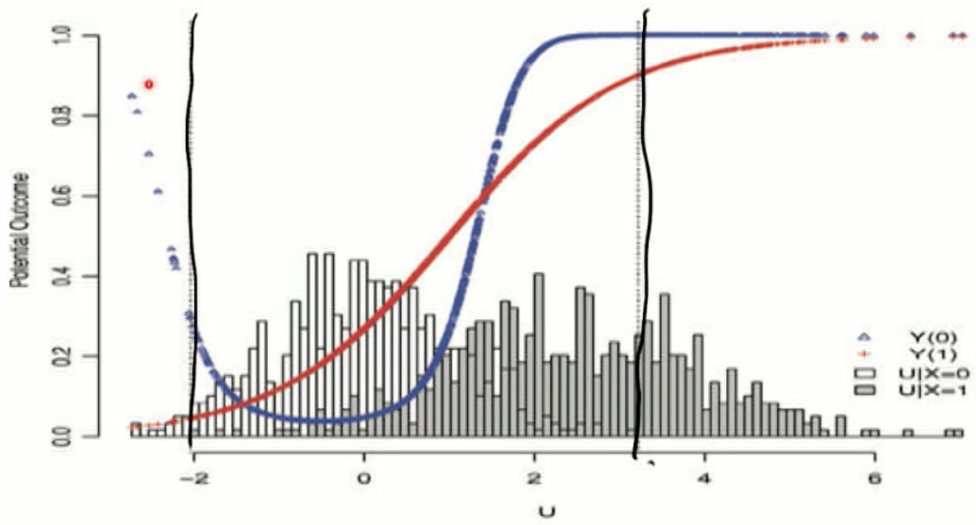
$$E\{X_i\} = \sum_{m=1}^{M} \alpha_m^X \delta_m^{\alpha^X} U_{im}\,, \qquad (1)$$

$$E\{Y_i|X_i\} = \beta^{\alpha^Y} X_i + \sum_{m=1}^{M} \alpha_m^Y \delta_m^{\alpha^Y} U_{im}\,, \qquad (2)$$



Example: Bayesian Adjustment for Confounding

Data Simulated by GR



Example: Bayesian Adjustment for Confounding

GR Analyzes the simulated data a few ways

We assess the frequentist operating characteristics of three estimation procedures at $4 \times 4 \times 3$ different configurations: $\sigma^2 \in \{0.5, 1, 2, 4\} \times B \in \{0, 0.25, 0.5, 1, 2\} \times n \in \{600, 1200, 2400\}$. The first procedure is the simple difference in observed means. The second procedure uses a model, approximating WPD's procedure:

$$E(Y_i(X_i) \mid U_i, \theta) = \beta X_i + ns(U_i, 15), \tag{7}$$

where ns is the natural cubic spline with 15 *df*. The third

flexible function of U

Poor Performance!

Table 1

95% Interval coverage rate, bias and RMSE $n = 600$; cubic spline model (7)

σ^2	B	0	0.25	0.5	1	2
0.5	Coverage	0.05	0.04	0.01	0.00	0.00
	Abs. Bias	14.00	14.35	16.46	30.00	63.14
	RMSE	14.45	14.80	16.88	30.33	63.38
1	Coverage	0.93	0.80	0.28	0.00	0.00
	Abs. Bias	0.84	4.23	12.25	39.58	69.57
	RMSE	6.56	7.54	13.29	39.85	69.78
2	Coverage	0.11	0.25	0.13	0.00	0.00
	Abs. Bias	18.96	15.42	19.18	36.65	59.91
	RMSE	19.89	16.64	20.07	37.04	60.16
4	Coverage	0.00	0.00	0.00	0.01	0.05
	Abs. Bias	57.48	46.27	37.02	36.64	34.73
	RMSE	57.79	46.66	37.53	37.14	35.34

Absolute Bias and RMSE are in 10^{-3}.

Example: Bayesian Adjustment for Confounding

GR Analyzes the simulated data a few ways

The authors of the discussion (GR) then analyze the data with an approach where:

1. Units with non-overlapping U_i are removed
2. Remaining units are partitioned into subclasses with approximately balanced U
3. A regression model is fit to outcomes in each treatment group
4. Missing potential outcomes are predicted from the model (multiply imputed)
5. Treatment effect is estimated with mean and variance of the imputed estimates

Example: Bayesian Adjustment for Confounding

GR Analyzes the simulated data a few ways

Good Performance

Table 2

95% interval coverage rate, bias and RMSE $n = 600$; multiple imputation

σ^2	B	0	0.25	0.5	1	2
0.5	Coverage	0.97	0.96	0.94	0.96	0.98
	Abs. Bias	1.05	0.89	0.92	0.40	0.44
	RMSE	5.00	4.98	5.16	6.11	8.77
1	Coverage	0.94	0.96	0.95	0.96	0.96
	Abs. Bias	0.88	0.64	0.74	0.86	0.93
	RMSE	6.51	6.24	6.31	6.72	8.46
2	Coverage	0.93	0.95	0.96	0.93	0.95
	Abs. Bias	0.89	0.79	0.91	2.21	2.44
	RMSE	8.10	7.60	7.43	7.93	8.47
4	Coverage	0.95	0.96	0.96	0.94	0.92
	Abs. Bias	0.56	0.95	1.04	1.70	2.02
	RMSE	9.37	8.83	8.45	8.57	9.10

Absolute Bias and RMSE are in 10^{-3}.

Balance and Overlap in Observational Studies

Both issues relate to the ubiquitous notion of *confounding*

Confounding $\rightarrow$ imbalance $\rightarrow$ overlap

- Can have large implications for estimating causal effects with observational studies
- Can lead to results that are very sensitive to the particulars of a (regression) model specification
  - Which wasn't true in randomized studies
- Illustrations have been based on single X
- Same ideas apply to multiple $\mathbf{X}$
  - Need to consider balance/overlap among the *multivariate* distribution of $\mathbf{X}$
  - Can be very challenging
- One "solution": Change the estimand to only answer a question about the subset of units for which there is reasonable balance/overlap