Causal Inference Crash Course: Arguable Validation for Cross-Sectional Models

Julian Hsu

Causal Inference Series

1) Foundations
2) Defining Some Causal Models
3) **Arguable Validation**
4) Inference, Asymptotic Theory, and Bootstrapping
5) Best Practices: Outliers, Class Imbalance, and Feature Selection
6) Heterogeneous Treatment Effect Models and Inference
7) [WiP] Models for Panel Data
8) [WiP] Regression Discontinuity Models

Overview

- Since panel models are relatively more straight-forward to test such as testing pre-treatment time parallel trends, this presentation focuses on arguable validation for cross-sectional models.
- We will also focus on standard propensity-based models, excluding approaches such as instrumental variable and regression discontinuity.
- We will cover some strategies:
  - Placebo tests
  - Coefficient stability following Oster (2019)
  - Feature balancing

What are we arguably validating?

- We are interested in estimating some treatment effect.
  - Ie, the impact of a selection change in a store's OPS.
- Recall the challenge is that we do not observe counterfactual outcomes – what the treatment observations' outcome would be if they were instead treated, or vice versa.
  - We don't know what the store's OPS would be if we did not change selection.
- We rely on assumptions like exogeneity, under which we can expect our model estimates the true treatment effect
- So we are trying to validate the assumptions

Actions you may take based on arguable validation

- Suppose you ran this OLS equation:

$$Y_i = \hat{\tau} W_i + \beta_1 X_i + \epsilon_i$$

for an outcome Y_i, treatment indicator W_i, and features X_i.

- What if you get an different estimate of $\hat{\tau}$ if you ran:

$Y_i = \hat{\tau} W_i + \beta_1 X_i^2 + \epsilon_i$; or

$Y_i = \hat{\tau} W_i + \beta_1 X_i + \beta_2 Z_i + \epsilon_i$?

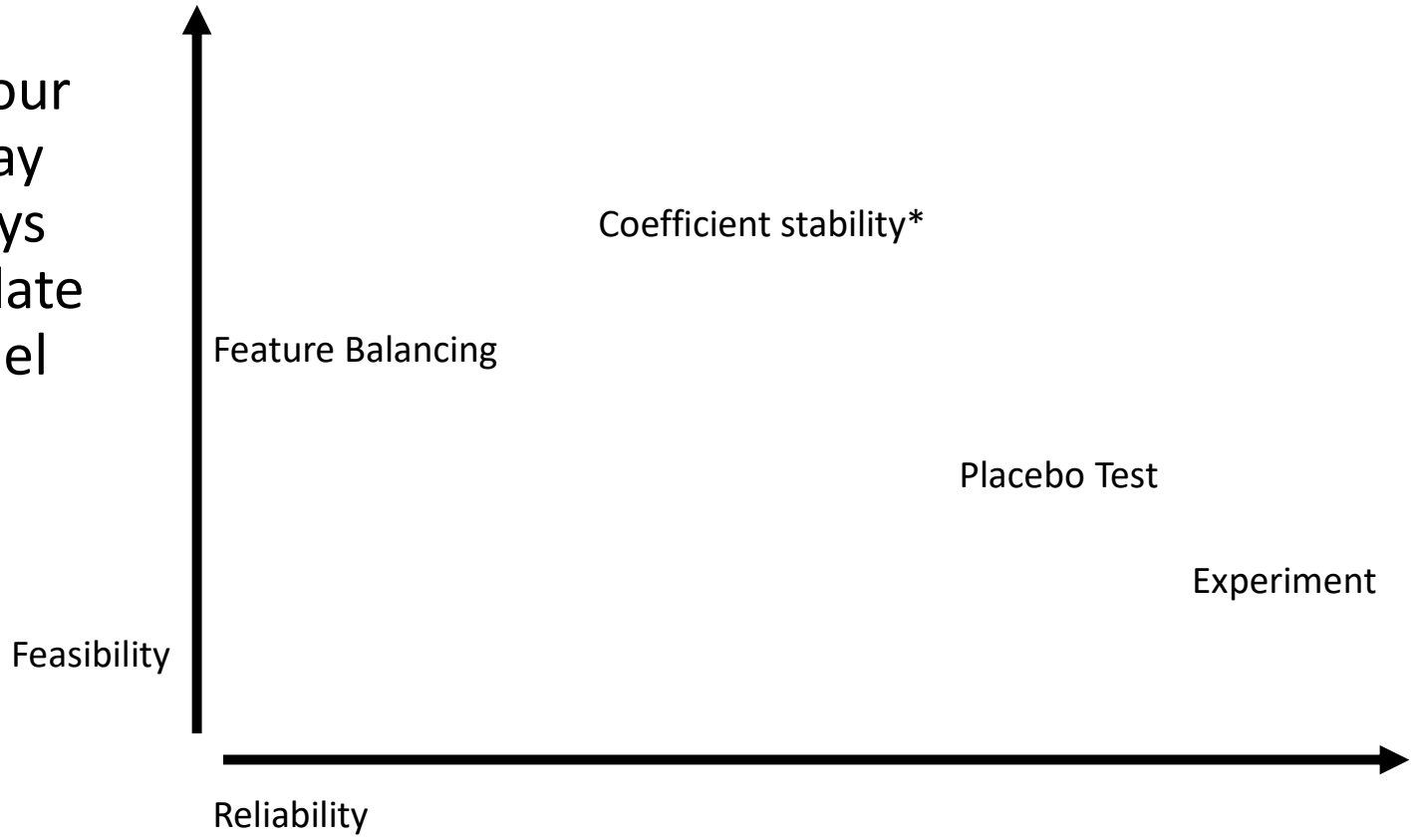
- Arguable validation helps you decide which model you should trust.

Arguable validation of design

- You may draw the conclusion that across all modeling specifications, you still don't have a good estimate of $\hat{\tau}$! In this case, you need to either:
  - Choose a different design like difference-in-difference or regression discontinuity;
  - Deep dive your use case to find the natural experiment in your data; or
  - Understand the potential biases your design has.
- Given the breadth of options above, we will just focus on arguable validation methods

High Level

- Depending on your use case, you may use different ways to arguably validate your causal model



Placebo test

Easy to do, except when it isn’t.

Not a lot of wiggle room for pivoting.

The big idea

$$Y_i = \hat{\tau} W_i + \beta_1 X_i + \epsilon_i$$

- Are there outcomes for which we know the treatment effect ?
- A placebo outcome is one where we know the treatment effect is zero.

$$Y_i^{placebo} = \hat{\tau}^{placebo} W_i + \beta_1 X_i + \epsilon_i$$

- Placebo outcomes can help transform the causal validation to a traditional prediction one, because we have a “ground truth” value for $\hat{\tau}^{placebo}$.

Make a Covariate an Outcome

- A popular placebo is the outcome before the treatment: Y_i^{pre}

$$Y_i^{pre} = \hat{\tau}^{placebo} W_i + \alpha_1 X_i + \eta_i$$

- And we want to see whether $\hat{\tau}^{placebo}$ is statistically different from zero
- We could pick any features populated before the treatment, but the Y_i^{pre} is particularly good because it looks very similar to Y_i.

Feasibility – what if Y_i^{pre} is X_i?

- This approach sacrifices data. You will have one fewer control variable, or have one fewer pre-treatment period of data.
  - If Y_i^{pre} is the only thing you are controlling for, then you obviously can't use it as a placebo outcome.
- Sacrificing this data has risks if you think Y_i^{pre} is uniquely important for predicting Y_i and W_i.

$$Y_i = \hat{\tau} W_i + \beta_1 X_i + \beta_2 Y_i^{pre} + \epsilon_i$$
$$Y_i^{pre} = \hat{\tau}^{placebo} W_i + \beta_1 X_i + \epsilon_i$$

Drawbacks

- If you think that exogeneity does not hold once you no longer control for Y_i^{pre}, then you can't use it as a placebo outcome.
- You can test this by:

1. Seeing whether $\hat{\tau}$ varies over whether you control for Y_i^{pre} or not; or
2. You determine whether Y_i^{pre} is crucial for predicting Y_i or W_i.

- In my experience, past OPS is one of the best predictors of current OPS. The question is whether everything else you control for is can pick up the slack when you don't control for the past OPS.

Coefficient stability

Easy to do but requires some judgement.

Even if it fails, you have something to work with.

Setup

$$Y_i = \hat{\tau} W_i + \beta_1 X_i + \epsilon_i$$

- Variation in Y_i is explained by:

(a) observable control variables X_i;

(b) the treatment W_i; or

(c) unobserved variables ϵ_i.

- As we control for more observable things, the remaining unexplained variation must be (b) or (c)
- So the more features I control for, the closer I get to an unbiased estimate of $\hat{\tau}$

The big idea

- We need to look at how $\hat{\tau}$ and explained variation in Y_i change as we control for more things.
- For example, we can find that controlling for more things doesn't change $\hat{\tau}$. But if it doesn't change R^2 either, then we could still be missing something.
- We should be more confident in our estimate of $\hat{\tau}$ under Scenario 2 and Scenario 1, below:

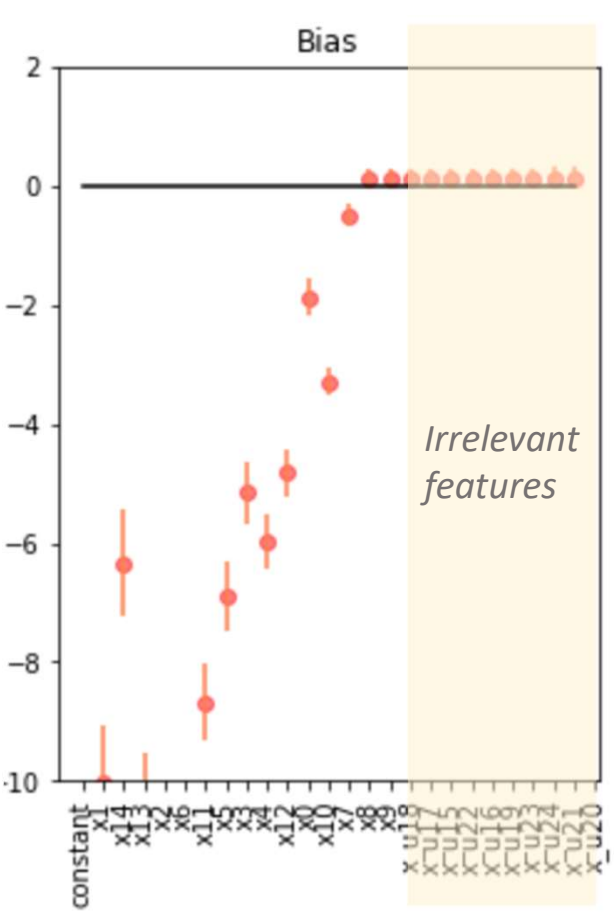
	Scenario 1		Scenario 2	
Control for … features	$\hat{\tau}$	R^2	$\hat{\tau}$	R^2
400	5	0.25	5	0.25
1000	5	0.27	5	0.90

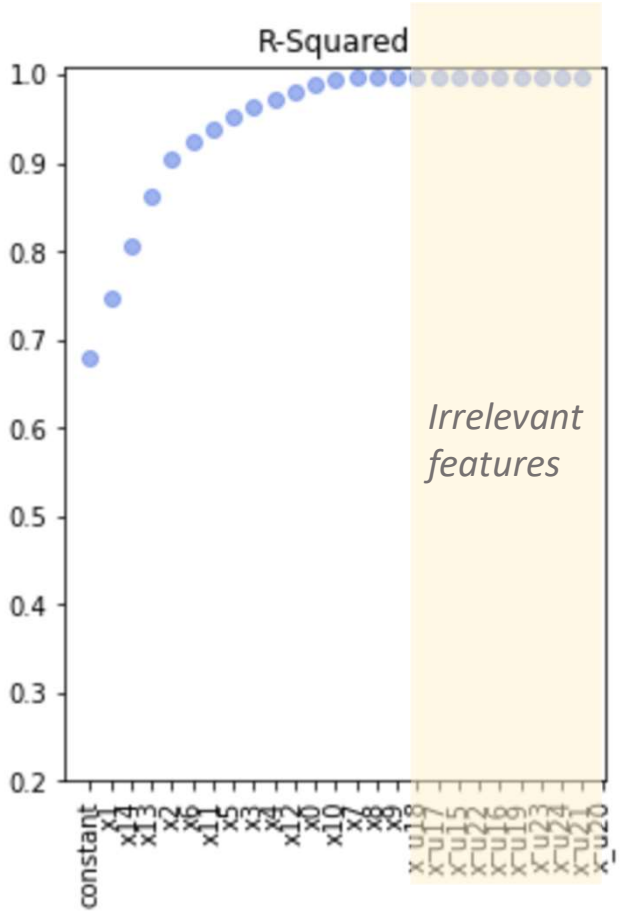
Coefficient stability test and interpretation

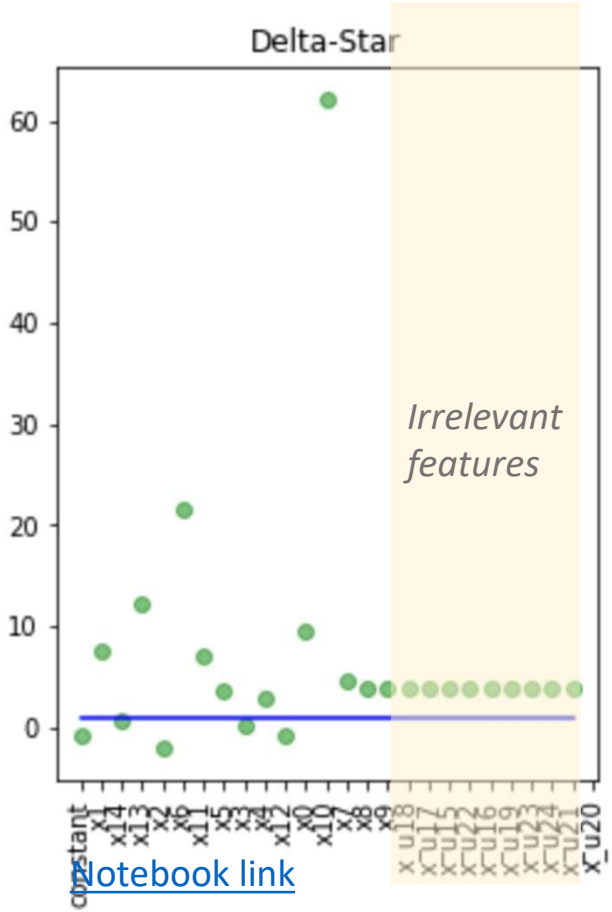
- "Given I can explain 99.9% of the variation Y_i after controlling for X_i and that my estimate of $\hat{\tau}$ is 10, how much selection bias is needed such that the incremental data needed for me to explain 100% of the variation changes $\hat{\tau}$ to 5?"
- Oster (2019) uses the omitted variable bias formula, the R^2 formula, and assumptions on what the selection bias might look like to answer this question.
- I won't go into this test statistic (shown below) for simplicity, but will show it in action in simulations

$$\delta^* = \frac{\left(\tilde{\beta}-\hat{\beta}\right)\left(\tilde{R}-\mathring{R}\right)\hat{\sigma}_y^2\hat{\tau_x} + \left(\tilde{\beta}-\hat{\beta}\right)\hat{\sigma}_X^2\hat{\tau}_x(\mathring{\beta}-\tilde{\beta})^2 + 2\left(\left(\tilde{\beta}-\hat{\beta}\right)\right)^2\left(\hat{\tau}_x(\mathring{\beta}-\tilde{\beta})\hat{\sigma}_X^2\right) + \left(\left(\tilde{\beta}-\hat{\beta}\right)\right)^3((\hat{\tau}_x\hat{\sigma}_X^2-\hat{\tau}_x^2))}{\left((R_{max}-\tilde{R})\hat{\sigma}_y^2(\mathring{\beta}-\tilde{\beta})\hat{\sigma}_X^2 + \left(\tilde{\beta}-\hat{\beta}\right)(R_{max}-\tilde{R})\hat{\sigma}_y^2(\hat{\sigma}_X^2-\hat{\tau}_x) + \left(\left(\tilde{\beta}-\hat{\beta}\right)\right)^2\left(\hat{\tau}_x(\mathring{\beta}-\tilde{\beta})\hat{\sigma}_X^2\right) + \left(\left(\tilde{\beta}-\hat{\beta}\right)\right)^3((\hat{\tau}_x\hat{\sigma}_X^2-\hat{\tau}_x^2))\right)}$$

Simulation results







Notebook link

Drawbacks on Interpretation

- Unlike the placebo test which you can either pass or fail based on whether the estimate on the placebo outcome is distinguishable from zero, the interpretation of coefficient stability is less clear.
- Coefficient stability embraces the fact that selection bias is unavoidable and tells us how likely it is the selection bias will change our results or conclusion.

Sample Interpretation

- As a user, we need to make decide these values about unobserved featurse that cause selection bias:
  - How much additional variation the unobserved features could potentially explain Y?;
  - The magnitude of the selection bias.

- *"We estimate a treatment effect of 10 and an $R^2 = 0.99$. We find δ^*=5, suggesting that unobserved factors need to be 5 times as important as the observed ones to explain the remaining 1% of the outcome's variation to bias our treatment effect by 15%."*

Feature balancing (aka covariate balancing)

Easiest to do,

But needs it’s own meta-validation.

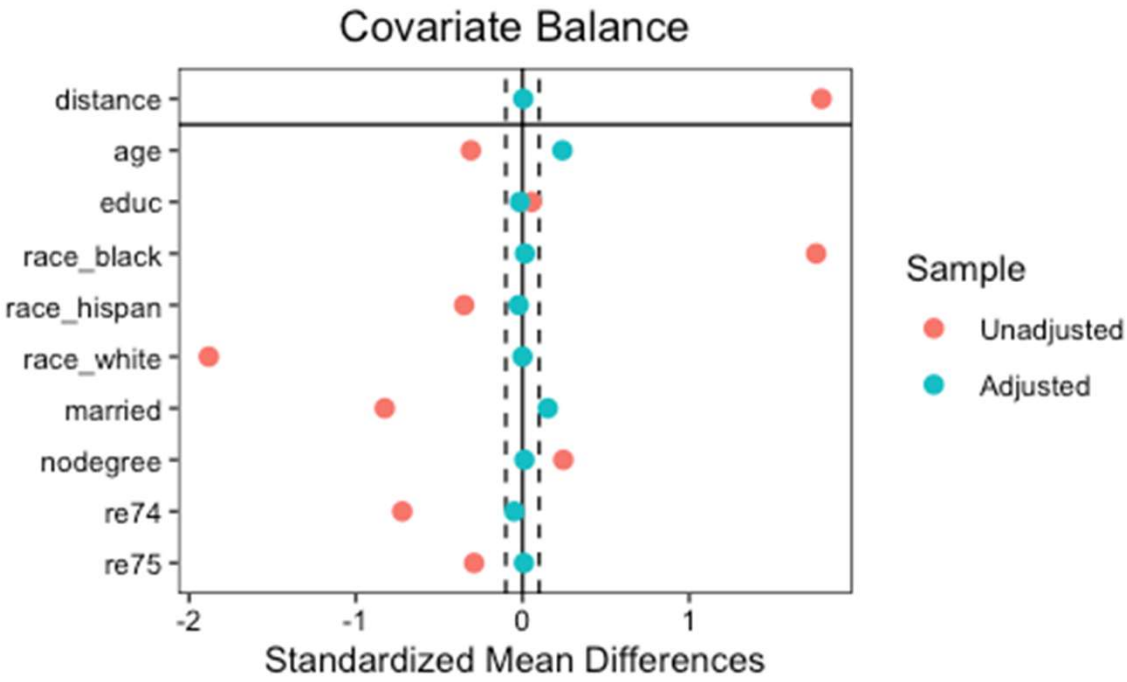
Setup

$$Y_i = \hat{\tau} W_i + \beta_1 X_i + \epsilon_i$$

- We can write the exogeneity condition as, "conditional on X_i, variation in Y_i is the treatment effect."
- This means that if it weren't for the treatment effect, the conditional difference between treatment and control is zero.
- This is the "synthetic twin" idea – we estimate $\hat{\tau}$ by comparing treatment and control groups that are otherwise similar.
- Propensity score matching does just this, by comparing groups that have the same likelihood of being treated.

The primitive to the big idea

- A common practice in propensity score matching is to test whether matched treatment and control pairs are similar in X_i.
- The matched/adjusted sample should be more similar in X_i than the raw/unadjusted sample.



From https://cran.r-project.org/web/packages/cobalt/vignettes/cobalt.html

The big idea

- We can do the same thing for any cross sectional model. This is because regression and weighting models do the same fundamental thing as matching.

$$x_i = \hat{\pi}_x W_i + \hat{\alpha}\, \hat{P}(X_i) + \epsilon_i$$

- Where x_i is an element of X_i, and $\hat{P}(X_i)$ is the estimated propensity score.
- $\hat{\pi}_x$ estimates the "adjusted" difference in X_i after matching.
- If all $\hat{\pi}_x$ estimates are zero, then we can argue the unconfoundedness assumption is valid.

Implementation

1. Estimate propensity score, $\hat{P}(X_i)$;
2. Estimate $\hat{\pi}_x$ for all your X_i;
3. Each time $\hat{\pi}_x$ cannot be distinguished from zero, you claim balance in x.
4. Failure to pass suggests controlling for additional features, or being specific in caveats for interpretation.

- You shouldn't expect each estimate $\hat{\pi}_x$ to be indistinguishable from zero because of the multiple hypothesis testing problem.
  - I like alpha-investing (following Foster (foster@) and Stein (robstine@) 2008) because you can prioritize what's most important to balance on.

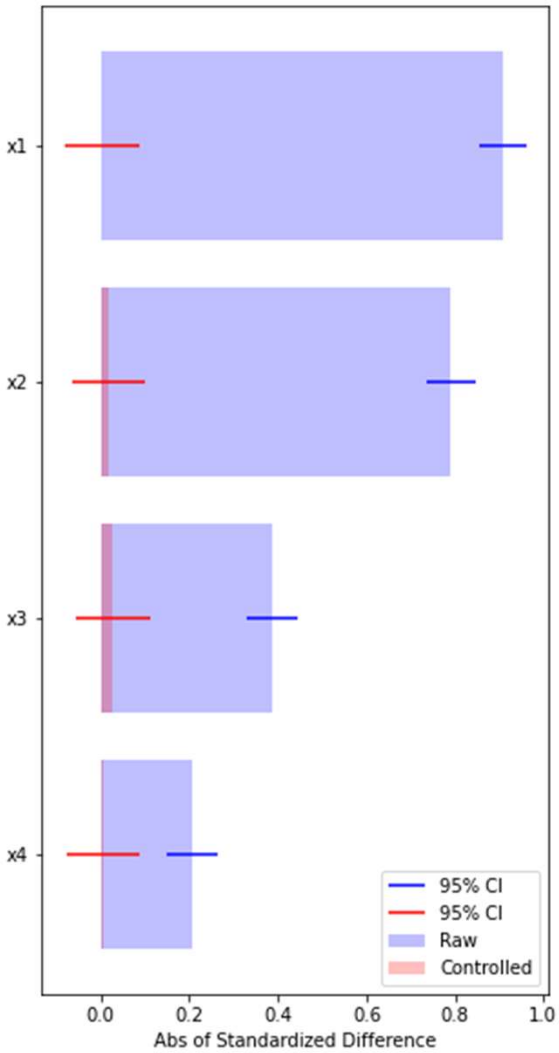
Drawbacks

- This approach requires you to be sure you have a good propensity score, $\hat{P}(X_i)$.

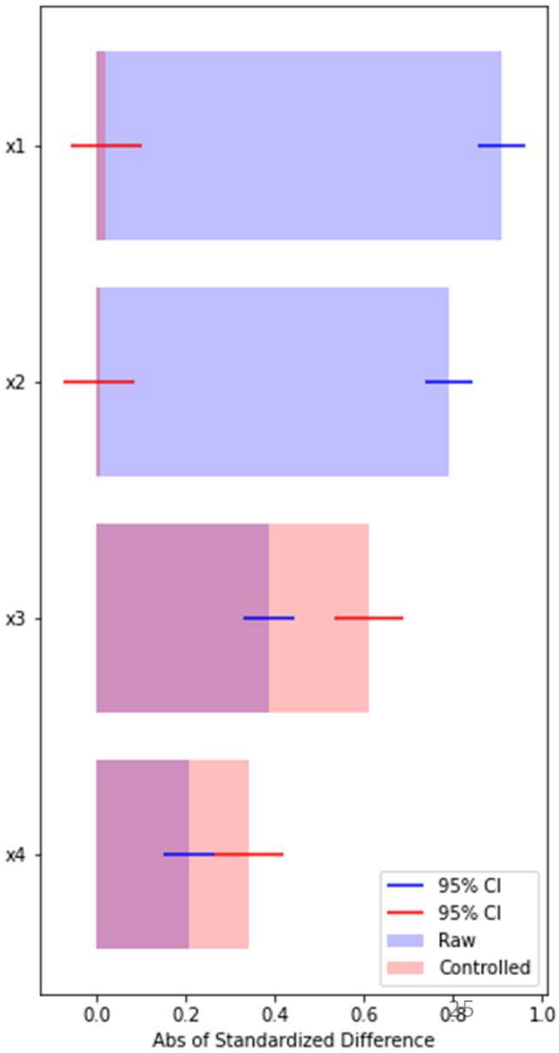
$$x_i = \hat{\pi}_x W_i + \hat{\alpha}\, \hat{P}(X_i) + \epsilon_i$$

- Notice that $\hat{P}(X_i)$ is estimated, so your confidence interval for $\hat{\pi}_x$ will be too small because you are not taking into account measurement error.
- You need to bootstrap estimates of $\hat{P}(X_i)$ to get the correct confidence interval for $\hat{\pi}_x$.
- Therefore, you need to do two loops: (1) bootstrap $\hat{P}(X_i)$, and (2) over $x \in X_i$.

Simulation

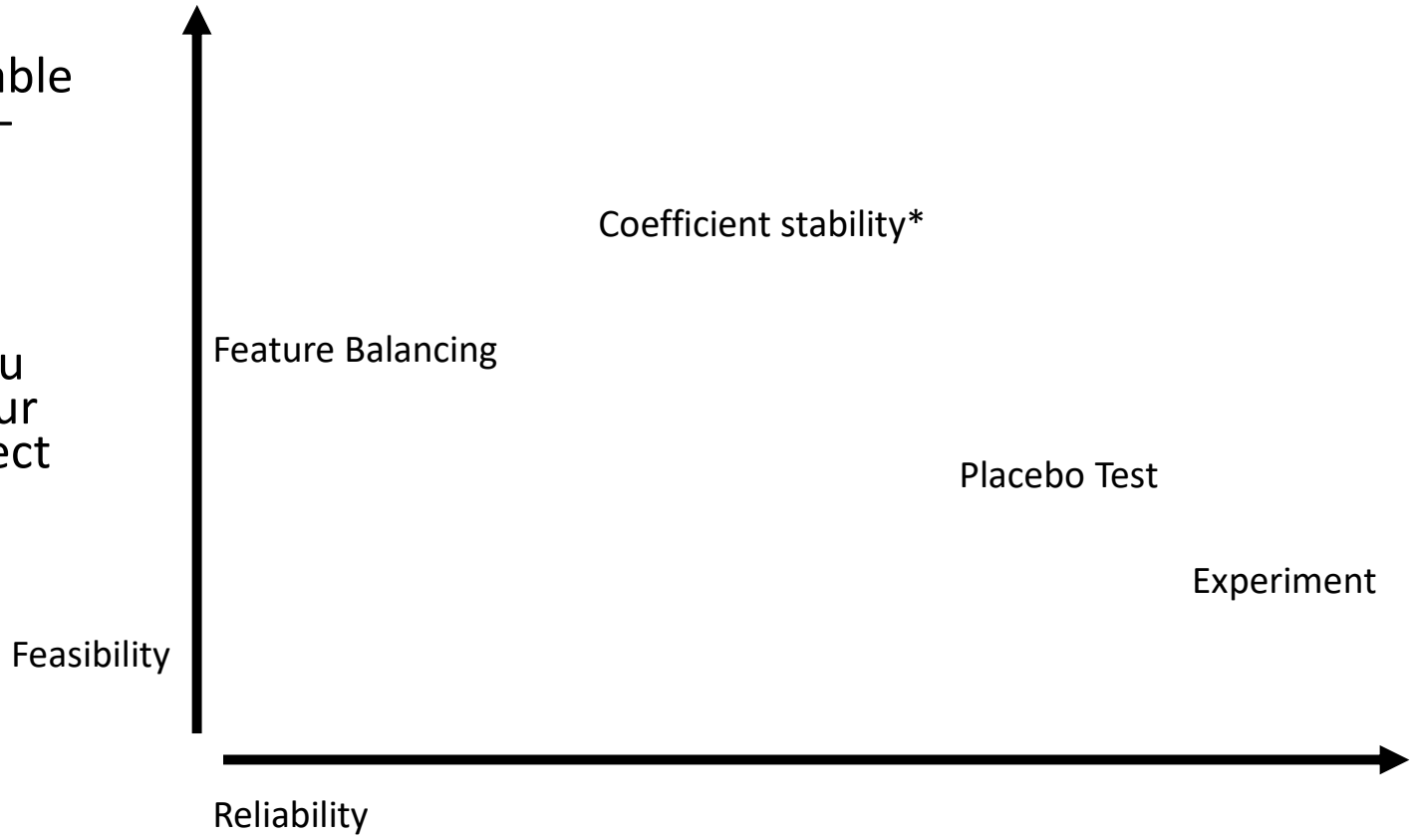


- When exogeneity condition is true on the left hand side, the controlled differences are smaller.
- When it is not true, feature balancing calls out which features we are imbalanced in
- [Notebook link]



Conclusion

- Four types of arguable validation for cross-sectional causal models
- They all require interpretation of results and help you decide whether your assumption is correct



Causal Inference Series

1) Foundations
2) Defining Some Causal Models
3) **Arguable Validation**
4) Inference, Asymptotic Theory, and Bootstrapping
5) Best Practices: Outliers, Class Imbalance, and Feature Selection
6) Heterogeneous Treatment Effect Models and Inference
7) [WiP] Models for Panel Data
8) [WiP] Regression Discontinuity Models

Literature Review of Related Papers

- Placebo Tests
  - Imbens, Wooldridge. (2009). “Recent Developments in the Econometrics of Program Evaluation.”link
  - Imbens. (2003). “Matching Methods in Practice: Three Examples.” link
- Coefficient Stability
  - Oster. (2019). “Unobservable Selection and Coefficient Stability: Theory and Evidence.” link;
  - Imbens. (2003). “Sensitivity to Exogeneity Assumptions in Program Evaluation.” link
- Feature Balancing
  - Imai, Ratkovic. (2014). “Covariate balancing propensity score.” link
  - Sant’Anna, Song, Xu. (2018). “Covariate Distribution Balance via Propensity Scores.” link
  - Athey, Imbens, Wager. (2016). “Approximate Residual Balancing: De-Biased Inference of Average Treatment Effects in High Dimensions.” link
  - Ben-Miachel, Feller, Hirschberg, Zubizaretta. (2021).” The Balancing Act in Causal Inference.” link

Appendix Slides

Conducting an experiment

The hard work is at the setup,

But you may not get you what you want

Big idea: create the randomization you wish you had

$$Y_i = \hat{\tau}^{obs} W_i + \beta_1 X_i + \epsilon_i$$

- When we cannot run an experiment, we can use observational data. We want to control for X_i to deal with selection bias.

$$Y_i = \hat{\tau}^{exp} W_i + \zeta_i$$

- If we could randomly assign W_i, then we will have no selection bias by design. We want to know whether $\hat{\tau}^{obs} = \hat{\tau}^{exp}$?

Experimentation for validation

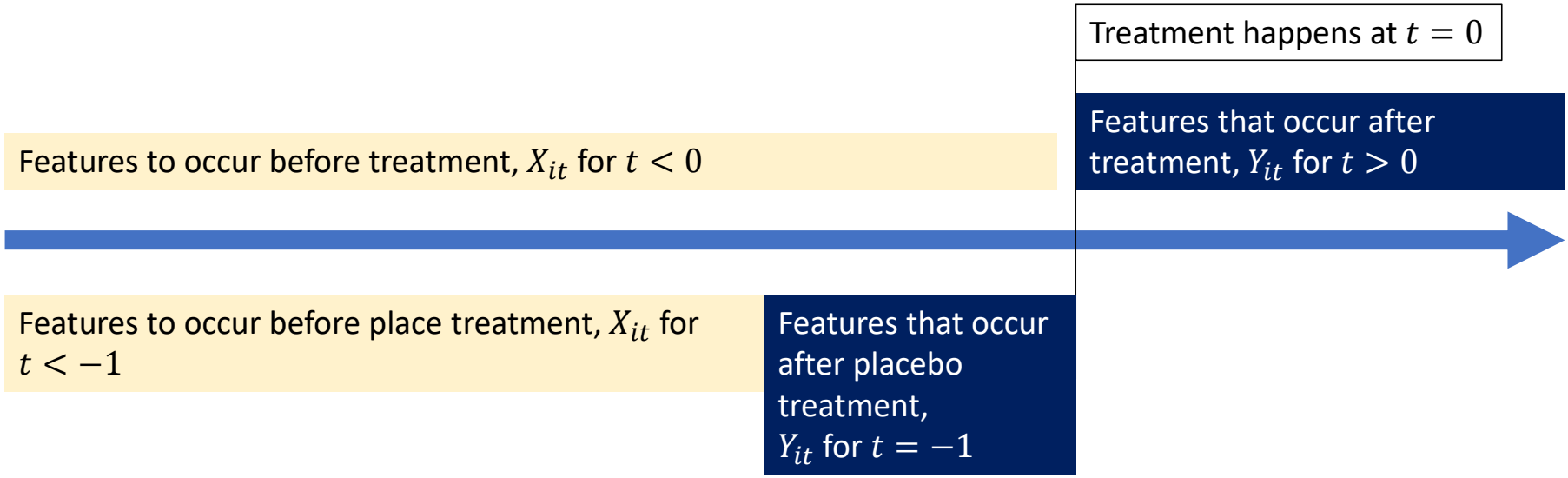
- Conducting an experiment randomly assigning W_i could validate our estimate of $\hat{\tau}$.
- Observational and experimental analysis can be substitutes or complements:

Substitutes	Complements
1. Do the experiment instead of observational analysis.	1. Use observational analysis to know where or how to conduct the experiment. 2. Experiment is designed to validate some of the observational results.

Drawbacks

- You may not be able to randomly assign W_i.
  - For example, assigning random prices or purposely delaying delivery
- Experiments may be underpowered
- Different experimental and observational samples may lead to incorrectly concluding $\hat{\tau}^{obs}$ is wrong.
  - For example, different types of customers, different time periods.

Placebo Type 2: Shift your Entire Dataset Back



- You always use these controls and these outcomes. In this placebo approach, you pretend the treatment happened before it actually did.

Placebo tests are not A/A tests

$$Y_i = \hat{\tau} W_i + \beta_1 X_i + \epsilon_i$$

- Instead of changing Y_i to Y_i^{pre}, why do a hold-out approach of all control observations and we randomly assign W_i?
- By randomly assigning W_i, we don't have any selection biases to correct.

Optimize for feature balance

- The feature balancing described before has separate estimation steps for estimating the propensity score and the balancing test.
- Why not incorporate feature balancing as an objective to estimating the propensity score?
- Explored in a few papers:
- Sant'Anna, Song, Xu. https://arxiv.org/abs/1810.01370
- Athey, Imbens, Wager. https://arxiv.org/pdf/1604.07125