Causal Inference Crash Course
Part 3: Inference

Julian Hsu

Causal Inference Series

1) Foundations
2) Defining Some ATE/ATET Causal Models
3) **ATE/ATET Inference, Asymptotic Theory, and Bootstrapping**
4) Best Practices: Outliers, Class Imbalance, Feature Selection, and Bad Control
5) Heterogeneous Treatment Effect Models and Inference
6) Difference-in-Difference Models for Panel Data
7) Regression Discontinuity Models
8) Arguable Validation

Overview

- This presentation will describe the "inference" in causal inference.
  A. Inference and consistency for OLS
  B. Challenge of applying asymptotic theory
  C. Bootstrapping is not a slow silver bullet

- We will only focus on inference for the ATE/ATET and not HTE. HTE incorporates additional inference challenges we will cover as part of HTE models.

Statistical inference overview

- Suppose we have a sample (X) and want to know whether its average is different from a given number, say zero.

$$X = (x_1, x_2, \dots, x_N) \ \ and \ \ X \sim F(\theta)$$

- We want to know whether a new sample from $F(\theta)$ would be different from zero on average.
- Our null hypothesis is that the average of X is zero.

Hypothesis testing and confidence intervals

- If we standardize the distribution of X, then we get a metric t that we know is distributed by a Student's t-distribution, which asymptotically approaches a normal distribution as the sample size increases

$$t = \frac{\bar{x}\ - 0}{se}, se = \frac{sample\ standard\ deviation}{\sqrt{n}}, \text{and } t \rightarrow^{d} N(0,1)$$

- This derivation relies on the Law of Large Numbers to that we can assume normality.
- This statistic tests our null hypothesis that $\bar{x} = 0$.
- This is useful because now we can model the variation in X if we drew more samples.
- We can now use this to form a confidence interval. A 95% confidence interval contains the range for 95% of future draws of X.

OLS statistical inference

- We can apply similar theories to do inference for an OLS regression
$$Y_i = \hat{\beta} X_i + \epsilon_i$$
- We previously showed that $\hat{\beta}$ will be unbiased. But how do we know the estimates are not driven by noise?
- Specifically, if we made another dataset, would we get the same value for $\hat{\beta}$?
- In other words, what is the distribution of $\hat{\beta}$?

Distribution of the OLS estimator

- We will use that $\hat{\beta}$ is consistent and converges to the true values.

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}(X'Y) \\ &= (X'X)^{-1}(X'(X\beta + \epsilon)) \\ &= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'\epsilon \\ &= \beta + (X'X)^{-1}X'\epsilon\end{aligned}$$

- How is this is distributed? We can then show that:

$$\sqrt{N}(\hat{\beta} - \beta) \rightarrow^{d} N(0, \Sigma)$$

- Where $\Sigma = \frac{1}{N}(X'X)^{-1}\frac{1}{N}(X'\epsilon\epsilon'X)\frac{1}{N}(X'X)^{-1}$
- If we gathered more data and recalculated $\hat{\beta}$ the distribution of those calculations would asymptotically converge to Σ
- This now tells us the joint distribution of $\hat{\beta}$. Now we can calculate confidence intervals.
  - See the Appendix for how to test hypothesis based on transformations of $\hat{\beta}$

Inference is not bias

- Confidence intervals are about whether we would get the same estimates a certain proportion of the time.
- A 95% confidence interval contains 95% of the possible estimates we would get from resampling the data.
- But $\hat{\beta}$ could be biased. $\hat{\beta}$ can consistently estimate a biased value.

$$\sqrt{N}(\hat{\beta} - \beta) \to^{d} N(\ bias\ , \Sigma)$$

- Therefore, $\hat{\beta}$ can be statistically significant and biased.

Inference is also not forecasting

- We interpret the confidence interval as what the estimate would be if we collected more (Y, X) data from $\mathrm{F}(\mathrm{y}|\mathrm{x}, \theta)$
- "More data" doesn't mean data from another context. For example, a confidence interval using data from $\mathrm{F}_{\mathrm{t}=1}(\mathrm{y}|\mathrm{x}, \theta)$ does not directly inform the results we would get from using data from $\mathrm{F}_{\mathrm{t}=2}(\mathrm{y}|\mathrm{x}, \theta)$
  - The confidence interval doesn't directly answer whether $\hat{\beta}$ would be the same if we collected data from next month.
- If the underlying data generating process changes over time, then we will have model misspecification biases.
- Model misspecification cause problems with inference.

Model misspecification also creates bias

- For example, the true model is: $Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon$
- But we instead estimate this model: $\mathrm{Y} = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2^2 + \eta$
- You have a misspecified model and so your estimate of β_1 will be different but can still be statistically significant.

Why can’t I just use LASSO and select features?

- Since LASSO selects features, we cannot do inference.
- LASSO coefficients are estimates using a penalty term for L1 regularization.
- Therefore, we cannot say that the coefficients from a LASSO regression are consistent and converge to the true coefficients.
- In other words, LASSO coefficients have two interpretations: the causal estimate of $\hat{\tau}$ and a bias towards zero to maximize prediction

Model misspecification in a regression adjustment model

- Recall the high-level model algorithm:
  - First, estimate the counterfactual control and treatment outcomes $\hat{Y}_0$ and $\hat{Y}_1$;
  - Then estimate ATE/ATET based on the differences between $\hat{Y}_0$ and $\hat{Y}_1$
- Ideally, $\hat{Y}_0$ and $\hat{Y}_1$ represent the true counterfactual outcomes. But if they are wrong, then the ATE/ATET estimate can still be wrong.
- But it can still be statistically significant.

How do we deal with model misspecification?

- Each model will generate some model misspecification bias
- The recommendation is to try do robustness checks. Try different model specifications, and they should provide similar results
  - Transforming features like squares
  - Linear and non-linear models
- The No Free Lunch Theorem (Wolpert and Macready, 1997) states that there is no model with universally superior performance, so relying on one model is guaranteed to eventually fail you

Review on what an estimate of β is

- $\hat{\beta} = \beta +$(Selection Bias) + (Model Misspecification Bias)
- Selection Bias is addressed by assuming we have satisfied the assumptions for a causal interpretation
- Model Misspecification Bias is addressed by robustness checks

Bootstrapping

- What happens if the estimator is consistent, but we cannot figure out how the estimator is distributed?
- Or, if we do not have a large enough sample size for asymptotic properties to kick in.
- Let's numerically calculate how the estimator is distributed.
- Recall that the distribution is interpreted as what the estimate would be if we redrew data.
- Bootstrapping assumes that the data we have X is sufficient to know what a redrawn dataset looks like.

Bootstrap setup

- $Y = \beta X + \epsilon$
- We want to get a bootstrap estimate for the variance of β, and we have pairs $(y_1, x_1), (y_2, x_2), \dots (y_N, x_N)$
- **Non-parametric bootstrap:**
    1. Resample N pairs from your sample with replacement S times
    2. For each bootstrap s, calculate β_s
    3. Use the variance of $\beta_1, \beta_2, \dots, \beta_S$ for the variance of β
- **Parametric bootstrap:**
    1. Calculate the joint distribution of $y|x \sim F(x, \theta)$
    2. Draw S pairs from $F(x, \theta)$, and do the same as 2. and 3. from the non-parametric bootstrap

You can bootstrap more than just variances

- For any given bootstrap s, you can calculate all sort of statistics from $Y_s = \beta_s X_s + \epsilon_s$
  - The p-value, standard error, confidence interval of β_s
  - Metrics of the regression like: F-statistic, R^2, or RMSE.
- As $S \rightarrow \infty$, the variance of bootstrap statistics approaches the truth.
- How many we do depends on the question we want to answer. More bootstraps gives us more precision.
- As a general practice, S should be large enough that the bootstrapped metric is stable enough.
  - Andrews and Buchinsky (2000); Cameron and Trivedi (2005) give us context dependent recommendations.

Final warning about bootstraps

- Bootstrapping only works if your estimator is consistent. An estimator is useless for inference if it is not consistent.
- For example, you can train an ML model to predict Y based on $X \in \mathbb{R}$ and $W = \{0,1\}$, then use $\hat{Y}(X, W = 1)$ and $\hat{Y}(X, W = 0)$. But you unless you can show that $\hat{Y}(X, W = 1) - \hat{Y}(X, W = 0)$ converges to the true treatment effect, then bootstrapping will not let you conduct proper inference.

Conclusion

- We have shown that statistical theorems are necessary to conduct inference for estimates
- Statistically significant estimates do not mean you have a causal estimate
  - Model misspecification biases
- Recommendations for understanding model misspecification biases and bootstrapping

Appendix Slides

Appendix Slides – Variance of Estimates

Using the variance

$$\sqrt{N}(\hat{\beta} - \beta) \to^{d} N(0, \Sigma)$$

- The diagonals $\sigma_{1,1}, \sigma_{2,2}, \ldots \sigma_{K,K}$ of Σ are the variance of $\hat{\beta}_1, \hat{\beta}_2, \ldots \hat{\beta}_K$. Then the standard error is $\mathrm{se_k} = \sqrt{\sigma_{k,k}}$. You then use the standard error to construct your confidence interval
- If you want to combine estimates, you need to use the covariance as well.
  - $Var(\hat{\beta}_1 + \hat{\beta}_2) = \sigma_{1,1} + \sigma_{2,2} + 2\sigma_{1,2}$
- If you want to know the variance of $g(\hat{\beta})$, then you need the Delta Method.
  - $\sqrt{N}\left(g(\hat{\beta}) - g(\beta)\right) \to^{d} N(0,\ \Sigma\,[g'(\beta)]^2)$
- Want to do both? See the next slide.

Standard errors from applying transformations of multiple parameters

- Standard errors from applying multiple transformations
  - https://www.stata.com/support/faqs/statistics/compute-standard-errors-with-margins/
- Another way this is used is to get the standard errors of a prediction, for example, $\hat{y} = \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$
  - https://stats.idre.ucla.edu/r/faq/how-can-i-estimate-the-standard-error-of-transformed-regression-parameters-in-r-using-the-delta-method/
  - Note that this is not the prediction interval which takes the error into account, only the confidence interval of the prediction.

Appendix Slides – You can’t do Inference with LASSO

Challenges to applying statistical inference

- High level note is that inference is about how the parameter is distributed, not about how well the prediction performs.

- We can see this if we were to use LASSO.

What about LASSO regressions

Model	Ordinary Least Squares (OLS)	Least Absolute Shrinkage and Selection Operator (LASSO)
Objective Function	$argmin_{\widehat{\beta},\hat{\tau}}\{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{\beta}X_i - \hat{\tau}T_i)^2\}$	$argmin_{\widehat{\beta},\hat{\tau}}\{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{\beta}X_i - \hat{\tau}T_i)^2\}$ $subject\ to\ \sum_{j=1}^{J}\lvert\widehat{\beta}_j\rvert + \lvert\hat{\tau}\rvert \leq C$

- LASSO regression coefficients are chosen to maximize prediction, subject to a constraint in the parameters.
- Intuitively, it assumes that coefficients are zero and there are penalties non-zero coefficients.
- Certainly, LASSO has better out-of-sample prediction. But can we use it for causal inference?

We cannot use LASSO for inference

- No, we can’t. Here is a technical and intuitive explanation.
- Technically, OLS identifies the causal estimate because of this moment condition you can get from solving the optimization problem:

$$E\left[(Y_i - \hat{\beta}X_i + \hat{\tau}T_i)\times T_i)\right] = 0$$

  But you can’t get this from a LASSO.
- Intuitively, a LASSO coefficient has two interpretations: the causal estimate of $\hat{\tau}$, and a feature selection of whether T_i is important to the prediction problem.
  - Then the unconfoundedness assumption may no longer hold.

Appendix Slides – Model Misspecification with Propensity Score Matching

Model misspecification in a propensity matching model

- High-level design for propensity score matching:
  - 1. Estimate a propensity score for all observations, $P(X_i)$
  - 2. Match treatment and control units in S groups with similar $P(X_i)$ values
  - 3. Find the differences within each $s \in S$ and aggregate them to estimate ATE/ATET
- Ideally, $P(X_i)$ represents the true propensity score. But if $P(X_i)$ is wrong, then the ATE/ATET estimate can still be wrong, but still be statistically significant.