Quant II

Lab 5: Inference

Junlong Aaron Zhou

March 05, 2021

Outline

- Asymptotics
- Bootstrap
- Finite Sample

Inference

- The frequentist perspective: $\mathbf{W}_i = (Y_i, D_i, \mathbf{X}_i)$ is drawn from some unknown distribution f.

Inference

- The frequentist perspective: $\mathbf{W}_i = (Y_i, D_i, \mathbf{X}_i)$ is drawn from some unknown distribution f.
- The estimate θ is a functional of f: $\theta = \theta(f)$.

Inference

- The frequentist perspective: $\mathbf{W}_i = (Y_i, D_i, \mathbf{X}_i)$ is drawn from some unknown distribution f.
- The estimate θ is a functional of f: $\theta = \theta(f)$.
- Yet f is unknown.

Inference

- The frequentist perspective: $\mathbf{W}_i = (Y_i, D_i, \mathbf{X}_i)$ is drawn from some unknown distribution f.
- The estimate θ is a functional of f: $\theta = \theta(f)$.
- Yet f is unknown.
- We either focus on the limit of $\theta(f)$, θ_0, or use the empirical distribution function, $\hat{f}$, to approximate f.

Asymptotics

- How do we derive the asymptotic distribution of an estimator?

Asymptotics

- How do we derive the asymptotic distribution of an estimator?
- Cornerstone: central limit theorem
- The estimator converges to the normal distribution with the speed $\sqrt{N}$ when the error from each observation is relatively small and independent.

Asymptotics

- Most estimators in causal inference have the following form:

$$\hat{\tau} = \frac{1}{N} \sum_{i=1}^{N} \phi(\mathbf{W}_i; \beta)$$

where β represents nuisance parameters.

Asymptotics

- Most estimators in causal inference have the following form:

$$\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N} \phi(\mathbf{W}_i; \beta)$$

where β represents nuisance parameters.

- For example, in the Horvitz-Thompson estimator, $\phi(\mathbf{W}_i; \beta) = \frac{D_i Y_i}{p_i} - \frac{(1-D_i)Y_i}{1-p_i}$.
- $\phi(\mathbf{W}_i; \beta)$ is called the influence function of the estimator $\hat{\tau}$.

Asymptotics

- Most estimators in causal inference have the following form:

$$\hat{\tau} = \frac{1}{N} \sum_{i=1}^{N} \phi(\mathbf{W}_i; \beta)$$

  where β represents nuisance parameters.
- For example, in the Horvitz-Thompson estimator, $\phi(\mathbf{W}_i; \beta) = \frac{D_i Y_i}{p_i} - \frac{(1-D_i) Y_i}{1-p_i}$.
- $\phi(\mathbf{W}_i; \beta)$ is called the influence function of the estimator $\hat{\tau}$.
- When the value of β is known, $\hat{\tau} - \tau$ converges to the normal distribution under some regularity conditions.

Asymptotics

- When β is unknown, we have to take into account the uncertainty from its estimation.

Asymptotics

- When β is unknown, we have to take into account the uncertainty from its estimation.
- This part is more difficult.

Asymptotics

- When β is unknown, we have to take into account the uncertainty from its estimation.
- This part is more difficult.
- Usually we need the Delta method (Taylor expansion): $\phi(\mathbf{W}_i; \hat{\beta}) = \phi(\mathbf{W}_i; \beta) + \phi^{'}(\hat{\beta} - \beta)$.

Asymptotics

- When β is unknown, we have to take into account the uncertainty from its estimation.
- This part is more difficult.
- Usually we need the Delta method (Taylor expansion): $\phi(\mathbf{W}_i; \hat{\beta}) = \phi(\mathbf{W}_i; \beta) + \phi^{'}(\hat{\beta} - \beta)$.
- When the convergence rate of β is not too slow, the estimate is still asymptotically normal.

Asymptotics

- When β is unknown, we have to take into account the uncertainty from its estimation.
- This part is more difficult.
- Usually we need the Delta method (Taylor expansion): $\phi(\mathbf{W}_i; \hat{\beta}) = \phi(\mathbf{W}_i; \beta) + \phi^{'}(\hat{\beta} - \beta)$.
- When the convergence rate of β is not too slow, the estimate is still asymptotically normal.
- What if the parameter is infinite-dimensional?

Asymptotics

- When β is unknown, we have to take into account the uncertainty from its estimation.
- This part is more difficult.
- Usually we need the Delta method (Taylor expansion): $\phi(\mathbf{W}_i;\hat{\beta}) = \phi(\mathbf{W}_i;\beta) + \phi^{'}(\hat{\beta} - \beta)$.
- When the convergence rate of β is not too slow, the estimate is still asymptotically normal.
- What if the parameter is infinite-dimensional?
- What if the influence function is not smooth?

Asymptotics

- Or we can do honest inference.

Asymptotics

- Or we can do honest inference.
- Split the sample, use one subsample to estimate the nuisance parameters and the other to estimate the causal effect.

Asymptotics

- Or we can do honest inference.
- Split the sample, use one subsample to estimate the nuisance parameters and the other to estimate the causal effect.
- Honest inference is popular in techniques based on machine learning.

Asymptotics

- Or we can do honest inference.
- Split the sample, use one subsample to estimate the nuisance parameters and the other to estimate the causal effect.
- Honest inference is popular in techniques based on machine learning.
- e.g. Ratkovic (2019)

Asymptotics

- There are other techniques for deriving the asymptotic distribution.
  - Moment generating function
  - Stein's method

Asymptotics

- There are other techniques for deriving the asymptotic distribution.
  - Moment generating function
  - Stein's method
- It is more difficult to derive the asymptotic distribution when the dimension is high.
- Now the number of variables increases at the same speed as the number of observations.
- For example, the empirical covariance no longer converges.

Asymptotics

- The roadmap if you are interested...

Asymptotics

- The roadmap if you are interested. . .
- Van der Vaart: Asymptotic Statistics (1998)
- Van der Vaart and Wellner: Weak Convergence and Empirical Processes (1996)
- Newey and McFadden: Large sample estimation and hypothesis testing (1994)
- Wainwright: High-Dimensional Statistics: A Non-Asymptotic Viewpoint (2016)

Bootstrap

- We estimate the distribution f with the estimated $\hat{f}$, a parameter with infinite dimension.

Bootstrap

- We estimate the distribution f with the estimated $\hat{f}$, a parameter with infinite dimension.
- The confidence interval is actually a “plug-in” estimator, but we plug in a function rather than a value.

Bootstrap

- We estimate the distribution f with the estimated $\hat{f}$, a parameter with infinite dimension.
- The confidence interval is actually a "plug-in" estimator, but we plug in a function rather than a value.
- There are different ways of plugging in the $\hat{f}$.
- How to resample?
- How to calculate the confidence interval?

Bootstrap: Motivation

- Where does it come from?
- $X \sim F(X|\theta) \to \{x_i\}_{i=1}^n$
- $\hat{\theta} = h(x_1, \dots, x_n)$
- Different samples from $F(X|\theta)$ produce different $\hat{\theta}$s that estimate the true θ
- Ideally: take different $\{x_i\}_{i=1}^n$ from $F(X|\theta)$, compute $\hat{\theta}$ for each of them, and get $\hat{\sigma}^2_{\hat{\theta}}$
- But we often have just one sample...
- So, we simulate samples! Parametrically or non-parametrically

Resampling algorithms

- Vanilla bootstrap
- Wild bootstrap
- Cluster bootstrap
- Jackknife

Bootstrap: nonparametric

- Instead of $X \sim F(X|\theta)$, assume $X \sim \hat{F}(X|\theta)$
- Our previous sample becomes the new population from which we sample
- Algorithm:
  - Choose B, number of pseudo-samples
  - Sample $\{x_1^{(1)}, \ldots, x_n^{(1)}\}, \ldots, \{x_1^{(B)}, \ldots, x_n^{(B)}\}$
  - Compute $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)}$
- $\hat{\sigma}^{*2} = \frac{1}{B-1}\sum_{j=1}^{B}(\hat{\theta}^{(j)} - \bar{\hat{\theta}})^2$, where $\bar{\hat{\theta}} = \frac{1}{B}\sum_{j=1}^{B}\hat{\theta}^{(j)}$
- $(1-\alpha)\%$ CI: cut off $\frac{\alpha}{2}\%$ smallest and largest $\hat{\theta}^{(j)}$ values

Bootstrap: parametric

- Plug $\hat{\theta}$ into $F(X|\theta)$
- Simulate $X \sim F(X|\hat{\theta})$
- Algorithm:
  - Choose B, number of pseudo-samples
  - Sample $\{x_1^{(1)}, \ldots, x_n^{(1)}\}, \ldots, \{x_1^{(B)}, \ldots, x_n^{(B)}\}$ from $F(X|\hat{\theta})$
  - Compute $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)}$
- $\hat{\sigma}^{*2} = \frac{1}{B-1} \sum_{j=1}^{B} (\hat{\theta}^{(j)} - \bar{\hat{\theta}})^2$, where $\bar{\hat{\theta}} = \frac{1}{B} \sum_{j=1}^{B} \hat{\theta}^{(j)}$

Construct the confidence interval

- The percentile t-method: $\frac{\hat{\theta}-\hat{\theta}^*}{\hat{\delta}^*}$
- The percentile method: $\hat{\theta} - \hat{\theta}^*$
- The Efron method: $\hat{\theta}^*$

```
## 95% CI from the percentile t-method: 2.4103 4.426186

## 95% CI from the percentile method: 2.434548 4.403184

## 95% CI from the Efron method: 2.486494 4.45513
```

Finite Sample

- We might only have small sample (non-asymptotic)
- Example? Small number of cluster, small N in panel, etc.

Finite Sample

- We might only have small sample (non-asymptotic)
- Example? Small number of cluster, small N in panel, etc.
- t distribution
- t-statistics is so called "pivotal-statistics", i.e. not depending on nuisance estimator.
- deal with finite sample problem better.

Finite Sample

- Bell-McCaffrey Solution moves the t-distribution to a χ^2-distribution, and find the degree of freedom adjustment for the t distribution.
- Similar idea with cluster.

Comprision of Confidence Set

1.11 Definition. C_n *is a* **finite sample** $1-\alpha$ **confidence set** *if*

$$\inf_{F \in \mathfrak{F}} \mathbb{P}_F(\theta \in C_n) \geq 1-\alpha \quad \text{for all } n. \tag{1.12}$$

C_n *is a* **uniform asymptotic** $1-\alpha$ **confidence set** *if*

$$\liminf_{n \to \infty} \inf_{F \in \mathfrak{F}} \mathbb{P}_F(\theta \in C_n) \geq 1-\alpha. \tag{1.13}$$

C_n *is a* **pointwise asymptotic** $1-\alpha$ **confidence set** *if,*

$$\text{for every } F \in \mathfrak{F}, \quad \liminf_{n \to \infty} \mathbb{P}_F(\theta \in C_n) \geq 1-\alpha. \tag{1.14}$$

Clustering

- Use the cluster SE only when you have clustering in sampling or clustering in design
- Easy to implement using `clubSandwich`

```
robust.se <- function(model, cluster){
  require(sandwich)
  require(lmtest)
  M <- length(unique(cluster))
  N <- length(cluster)
  K <- model$rank
  dfc <- (M/(M - 1)) * ((N - 1)/(N - K))
  uj <- apply(estfun(model), 2, function(x) tapply(x, cluster,
  rcse.cov <- dfc * sandwich(model, meat = crossprod(uj)/N)
  rcse.se <- coeftest(model, rcse.cov)
  return(list(rcse.cov, rcse.se))
}
```

Clustering

- Example: Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. "How much should we trust differences-in-differences estimates?." The Quarterly journal of economics 119.1 (2004): 249-275.
- It claims serial correlation in panel data should be taken into account.

Clustering

- Example: Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. "How much should we trust differences-in-differences estimates?." The Quarterly journal of economics 119.1 (2004): 249-275.
- It claims serial correlation in panel data should be taken into account.
- However, it doesn't solve problems like:
- AR(1) process: $Y_{it} = \alpha Y_{i,t-1} + \beta X$
- Interference: $Y_i(\overrightarrow{D}) = Y_i(d_1, d_2, \ldots)$

Take-aways

- Know your data structure and sample.
- Deal with cluster carefully.
- Cluster doesn't solve every problems.
- Correct bootstrap works (most of time).