Quant II

Inference and Instrumental Variables

Ye Wang

March 04, 2020

Outline

- Inference
  - Asymptotics
  - Bootstrap
- Instrumental variables
  - Principal strata
  - Test for IV's validity

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?

Resampling algorithms

- Vanila bootstrap
- Wild bootstrap
- Cluster bootstrap
- Jackknife

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

```
## 95% CI from the percentile t-method: 2.314556 4.048762
```

```
## 95% CI from the percentile method: 2.317793 4.045422
```

```
## 95% CI from the Efron method: 2.290772 4.0184
```

Clustering

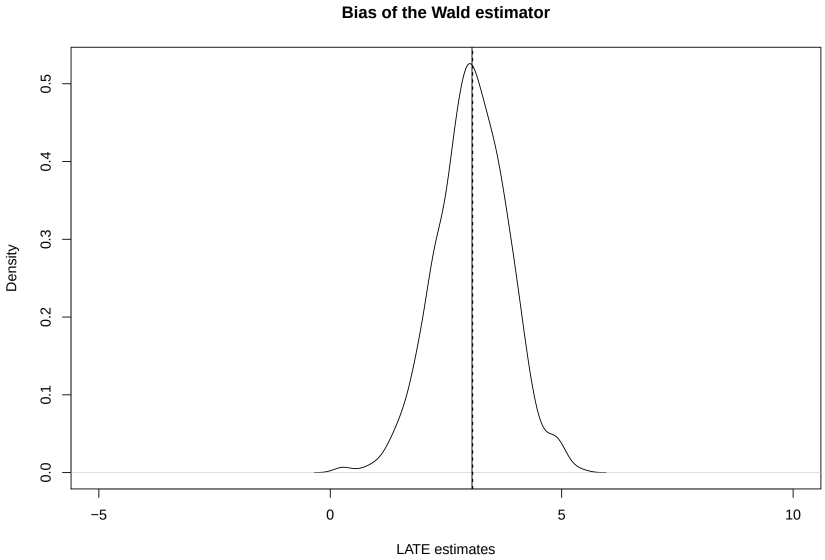
- Use the cluster SE only when you have clustering in sampling or clustering in design

```
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, clust
  rcse.cov <- dfc * sandwich(model, meat = crossprod(uj)/N)
  rcse.se <- coeftest(model, rcse.cov)
  return(list(rcse.cov, rcse.se))
}
```

LATE

- Principal strata: subsets defined by post-treatment variables
- Eg. ATT, LATE, etc.
- To estimate LATE requires four assumptions:
  - Unfoundedness
  - Exclusion restriction
  - Monotonicity (no defiers)
  - First stage

LATE



Bias of the Wald estimator
Density
0.0
0.1
0.2
0.3
0.4
0.5
−5
0
5
10
LATE estimates

LATE

```
## The share of always-takers is 20
##  and the estimate is 0.2

## The share of never-takers is 30
##  and the estimate is 0.28

## The share of compliers is 50
##  and the estimate is 0.52
```

A more complicated example

- Burde, Middleton, Samii and Wang, 2020
- An experiment in Afghanistan
- Y_i: grade in reading and mathematics of female students
- Z_i: access to community-based education (CBE)
- $D_i = 0$: no education; $D_i = 1$: attend government schools; $D_i = 2$: attend CBE

Principal strata

			If assigned to treatment ($z = 1$)		
			No school	Gov. School	CBE
			$D(1) = 0$	$D(1) = 1$	$D(1) = 2$
If assigned to control ($z = 0$)	No school	$D(0) = 0$	never taker	X	complier
	Gov. School	$D(0) = 1$	X	gov. adherent	substitutor
	CBE	$D(0) = 2$	X	X	always taker

Table 4: Principal strata.

Principal strata

- Denote the percentage of each strata as $\pi(D(0), D(1))$.
- $p(d, z) = \frac{\sum_{i=1}^{N} 1(D_i = d, Z_i = z)}{\sum_{i=1}^{N} 1(Z_i = z)}$.
- $\pi(2, 2) = p(2, 0)$
- $\pi(0, 0) = p(0, 1)$
- $\pi(1, 1) = p(1, 1)$
- $\pi(1, 2) = p(1, 0) - p(1, 1)$
- $\pi(0, 2) = p(0, 0) - p(0, 1)$

From LATE to ATE

- We can calculate the principal strata score using covariates.
- Although we do not know the strata each observation belongs to, it is still possible to fit a MLE to maximize the probability for the assignment $(\mathbf{D}, \mathbf{Z})$ to occur.
- Based on these scores we can weight LATE to get ATE.
- Aronow and Carnegie (2013); Ding and Lu (2016); Feller et al. (2018)

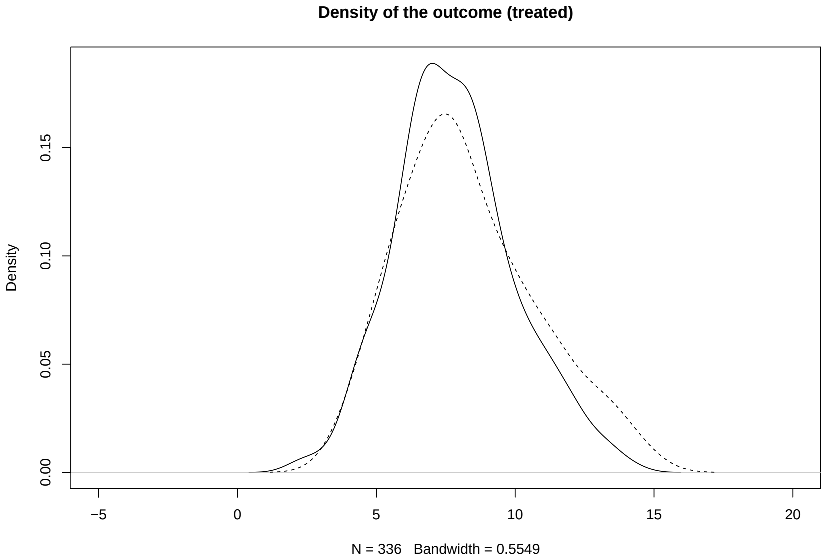
Many instrumental variables

- Conventional approach: GMM
  - Each IV gives a moment condition
  - Use GMM to combine all the conditions
- Modern approach: LASSO
  - The first stage is a prediction problem
  - Hence we can use all kinds of ML algorithms to fit the first stage
  - The concern is whether the approach leads to valid estimates

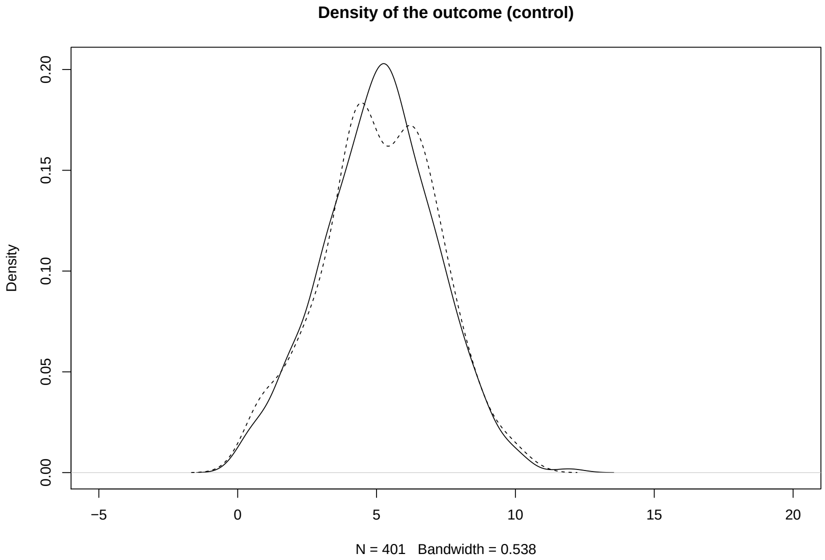
IV test

- A recently developed branch of the literature
- The exclusion restriction alone cannot be tested.
- But we can test its combination with the mototonicity assumption.
- $P(y, D = 1|Z = 1) > P(y, D = 1|Z = 0)$ and $P(y, D = 0|Z = 1) < P(y, D = 0|Z = 0)$

IV test


Density of the outcome (treated)
Density
0.00
0.05
0.10
0.15
−5
0
5
10
15
20
N = 336   Bandwidth = 0.5549

IV test



Density of the outcome (control)
Density
0.00
0.05
0.10
0.15
0.20
−5
0
5
10
15
20
N = 401 Bandwidth = 0.538