Section 12 : Instrumental Variables

Andrew Bertoli

20 November 2013

Roadmap

1. Instrumental Variables
2. Homework

Instrumental Variables

Motivation

We frequently have exogenous variation for treatments that we don't really care about, like weather or natural disasters.

IV designs try to exploit this variation to make inferences about treatments that we do care about.

Instrumental Variables

Notation

Z: Instrument

T: Treatment

X: Coviariates

Y: Outcome

Note: None of these variables have to be dichotomous.

Instrumental Variables

Assumptions

1. SUTVA
2. As-if randomness of Z
3. Exclusion Restriction (Z only affects Y through T)
4. Non-zero causal effect of Z on T
5. Z either increases or decreases T for all units (no defiers)

Instrumental Variables

Question: How do these assumptions compare to the assumptions in a natural experiment?

Answer: 1 and 2 are the same, while 3, 4, and 5 are additional assumptions.

Instrumental Variables

In the one-way non-compliance example, the instrument is the intention to treat. Our estimator for the effect of T on Y (for compliers) is

$$\frac{\hat{Y}_{Z=1} - \hat{Y}_{Z=0}}{\hat{\alpha}}$$

where $\hat{Y}_{Z=1} - \hat{Y}_{Z=0}$ is the Intention to Treat estimator and $\hat{\alpha}$ is the estimated proportion of compliers.

Instrumental Variables

Assumptions

1. SUTVA (?)
2. As-if randomness of Z ✓
3. Exclusion Restriction (Z only affects Y through T)
4. Non-zero causal effect of Z on T
5. Z either increases or decreases T for all units (no defiers)

Instrumental Variables

Assumption (3) holds if being offered the treatment does not change any unit's potential outcomes.

Instrumental Variables

Assumptions

1. SUTVA (?)
2. As-if randomness of Z ✓
3. Exclusion Restriction (Z only affects Y through T) (?)
4. Non-zero causal effect of Z on T ✓
5. Z either increases or decreases T for all units (no defiers)

Instrumental Variables

Assumption (5) holds since no one who was not offered the treatment got it.

Instrumental Variables

Assumptions

1. SUTVA (?)
2. As-if randomness of Z ✓
3. Exclusion Restriction (Z only affects Y through T) (?)
4. Non-zero causal effect of Z on T ✓
5. Z either increases or decreases T for all units (no defiers) ✓

Instrumental Variables

These assumptions are usually much harder to believe with other types of instrumental variable studies.

Instrumental Variables

Example

Question: Does walking reduce stress levels?

Problem: The amount that people walk is not random.

Instrumental Variables

Instrument: Segway decides to give away free vehicles randomly at Cal.

Instrumental Variables

How could the assumptions be violated

1. SUTVA
2. As-if randomness of Z
3. Exclusion Restriction (Z only affects Y through T)
4. Non-zero causal effect of Z on T
5. Z either increases or decreases T for all units (no defiers)

Instrumental Variables

How could the assumptions be violated

1. Student's without Segways borrow them from their friends
2. The giveaway is not random
3. Getting a free Segway makes some people happier, decreasing their stress
4. No one who gets a Segway uses it
5. No one walks more if they get a Segway

Instrumental Variables

Set-up

There are 30,000 students at Cal. None originally have a Segway.

Segway offers to give 10,000 of them a free Segway.

About 5% of students are compliers

We have data on every students walking habits and stress levels after the giveaway.

Instrumental Variables

```
> head(data)
  Stress.Level Walk.Hours Offered.Segway Complier
1    0.9710777   2.028637              0       NA
2    0.5376507   1.938973              0       NA
3    0.9078909   4.658039              0       NA
4    0.6899065   3.287785              1        0
5    0.8000950   1.308127              1        0
6    0.6386251   4.648876              0       NA
```

Instrumental Variables

Question: What is the ITT (Intention to Treat) Estimator?

Answer: The estimated effect of being offered a Segway on stress levels.

Instrumental Variables

```
> t.test(Stress.Level~Offered.Segway,data=data)

  Welch Two Sample t-test

data:  Stress.Level by Offered.Segway
t = 1.3351, df = 20051.32, p-value = 0.1819
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.001104341  0.005822629
sample estimates:
mean in group 0 mean in group 1
      0.7495060       0.7471468
```

Instrumental Variables

Question: What is the Wald Estimator?

Answer: ITT Estimator/ Estimated Effect of Instrument on Treatement

$$\frac{E[Y_i|Z_i=1]-E[Y_i|Z_i=0]}{E[T_i|Z_i=1]-E[T_i|Z_i=0]}$$

Note: For one-way non-compliance in an experiment, this equals

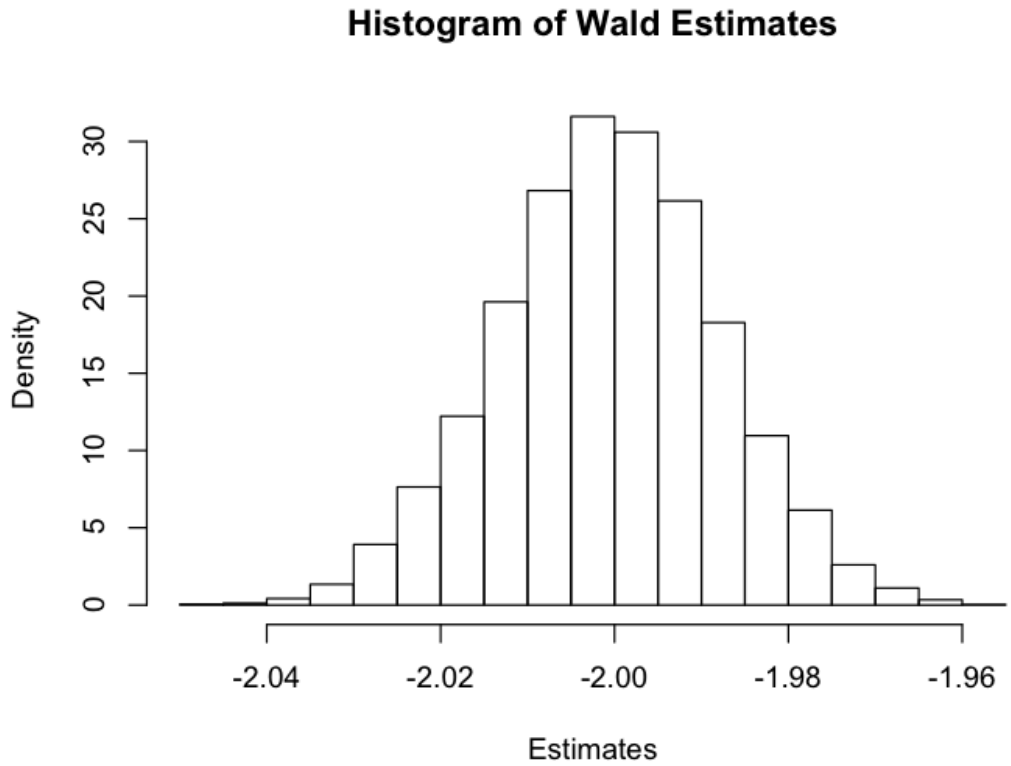
$$\frac{\bar{Y}^T - \bar{Y}^C}{\hat{\alpha}}$$

Instrumental Variables

```
> with(data,
+
+ mean(Stress.Level[Offered.Segway==1])-mean(Stress.Level[Offered.Segway==1])
+ /
+ mean(Walk.Hours[Offered.Segway==1])-mean(Walk.Hours[Offered.Segway==1])
+
+ )
[1] -1.981156
```

Instrumental Variables

Using Monte Carlo simulation, we have



Question: What was the true treatment effect of walking on stress levels?

Answer: 0

Instrumental Variables

One Solution: Use Permutation Inference

Homework

Problem 2: Say you have a sample of n draws of some random variable, and you estimate the median of this random variable by taking the median of your sample. You then decide to estimate the standard error of your estimator by taking the average difference between the medians of many bootstrapped samples and the median of your original sample. Prove whether this method will give you an unbiased estimate of the standard error of the median estimator? What will happen as $n \to \infty$?

Homework

Problem 3a: Using your own functions (not the "Matching" library), select some important covariates and find matches for the treated units using Mahalanobis distance and an estimated propensity score. Do not look at the outcomes. Simply match a control unit to every treated unit. For each matching method, construct a balance plot showing the p-values for the covariates before and after matching.

Homework

Problem 3b: Now use GenMatch to find matches for all the treated units. Write a loss function that prioritizes getting balance on age. Make a balance plot showing the p-values before and after matching.

Homework

Problem 3c: Now write another loss function that ensures that the balance with GenMatch will not be worse on any covariate than it was under either of your two matching schemes in Part (a). Construct another balance plot showing the p-values for this approach.

Homework

Problem 4

Create a simulation where you have a population, a treatment of interest, and a non-random sample of the population that you can run an experiment on. Please do not create arbitrary variables. Have a substantive story to go along with your data. Make sure that you have at least 10 covariates for each unit. Make the probability of being selected into the experimental sample

$$P(I_i = 1) = max(0, \lambda_i / max(\lambda))$$

where $\lambda_i = \beta_1 \cdot X_{i1} + \beta_2 \cdot X_{i2} + ...\beta_n \cdot X_{in}$, the $\beta_1, \beta_2, ..., \beta_n$ are parameters that you set, and the $X_{i1}, X_{i2}, ..., X_{in}$ are some of the covariates for unit i. This formulation will guarantee that $0 \leq P(I_i = 1) \leq 1$.

Homework

For the units outside your sample, compute the probability of treatment assignment as a function of some of the covariates. Make the outcomes for each unit a function of treatment and some of the covariates. Also calculate the values for the population treated units had they not been treated and find the true ATE for these units.

Homework

When setting this simulation up, make sure that you have some covariates that only affect the probability of being selected into the experimental sample, some covariates that only affect the likelihood of treatment assignment, and some covariates that only affect the outcome. Obviously, you also want some covariates that affect all three, and others that affect two of the three.

Homework

Once you have constructed the data, use maximum entropy weighting to estimate the effect of the treatment of interest on the subset of units that were treated but not in your experiment (the population treated units).

Homework

a) Approximately how bad are the bias and variance of your estimator of the PATT if you control for only half of the covariates? You can determine this by recording the difference between the true PATT and the estimated PATT on every run and looking at the distribution of these differences.

Homework

b) What happens to the bias and variance of the PATT estimator when you drop control variables that only affect the likelihood of being in the sample? What about if you drop control variables that only affect the likelihood of treatment assignment in the population? What about control variables that only affect the outcome?