Applied Methods for PhD

Estimations Strategies - Overview

Michael E. Kummer
Theoretical Slide Set 6, based on Stephen Kastoryano

NovaSBE, OTIM

METHODS FOR ESTIMATING TREATMENT EFFECTS

Treatment effect literature provides wide range of quite different estimators, many of which are regularly used in empirical work.

1. Field (or Social) experiments.
2. Regression (including factor models).
3. Matching.
4. Regression discontinuity.
5. Instrumental variable.
6. Control Functions.
7. Difference-in-difference.
8. Nonparametric bounds.
9. Timing-of-events.
10. Structural estimation (Roy-type models).

METHODS FOR ESTIMATING TREATMENT EFFECTS

Treatment effect literature provides wide range of quite different estimators, many of which are regularly used in empirical work.

1. Field (or Social) experiments.
2. Regression (including factor models).
3. Matching.
4. Regression discontinuity.
5. Instrumental variable.
6. Control Functions.
7. Difference-in-difference.
8. Nonparametric bounds.
9. Timing-of-events.
10. Structural estimation (Roy-type models).

POTENTIAL OUTCOMES MODEL

- Potential Outcome Model (aka. Rubin or Neyman-Rubin causal model) finds roots in Neyman (1923) Msc. thesis.
- Further groundwork in statistics by, Rubin (1974), Holland (1986 review).
- In economics, Heckman one of pioneering researchers on policy evaluation (often referring to Roy model).
- Let D_i be an indicator for receiving treatment ($\mathrm{D}_i = 1$) or not ($\mathrm{D}_i = 0$).

POTENTIAL OUTCOMES MODEL

- Potential Outcome Model (aka. Rubin or Neyman-Rubin causal model) finds roots in Neyman (1923) Msc. thesis.
- Further groundwork in statistics by, Rubin (1974), Holland (1986 review).
- In economics, Heckman one of pioneering researchers on policy evaluation (often referring to Roy model).
- Let D_i be an indicator for receiving treatment ($\mathrm{D}_i = 1$) or not ($\mathrm{D}_i = 0$).
- Each individual has two potential outcomes, $\mathrm{Y}_i^{\mathrm{D}=1}$ with treatment and $\mathrm{Y}_i^{\mathrm{D}=0}$ without treatment. We will also denote the realizations of these random variables as $y_i^{\mathrm{D}=1}$ and $y_i^{\mathrm{D}=0}$.
- The effect for each individual of participating in the treatment equals

$$\Delta_i = \mathrm{Y}_i^{\mathrm{D}=1} - \mathrm{Y}_i^{\mathrm{D}=0}$$

- Since only one of the random variables $\mathrm{Y}_i^{\mathrm{D}=1}$ and $\mathrm{Y}_i^{\mathrm{D}=0}$ can be observed, Δ_i will always be an unobserved random variable. The unobserved outcome is the *counterfactual outcome*.

ATE, ATET & ATENT

- *Parameter of interest* depends on the context of the study and the (sub-)population of interest. One often considered is Average Treatment Effect,

$$ATE = \mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=1} - \mathrm{Y}_i^{\mathrm{D}=0}] = \mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=1}] - \mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=0}]$$

- If the selection into treatment is not entirely random and some individuals are more likely to enter treatment then it may be preferable to focus on Average Treatment Effect on the Treated

ATE, ATET & ATENT

- *Parameter of interest* depends on the context of the study and the (sub-)population of interest. One often considered is Average Treatment Effect,

$$ATE = \mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=1} - \mathrm{Y}_i^{\mathrm{D}=0}] = \mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=1}] - \mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=0}]$$

- If the selection into treatment is not entirely random and some individuals are more likely to enter treatment then it may be preferable to focus on Average Treatment Effect on the Treated

$$ATET = \mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=1} - \mathrm{Y}_i^{\mathrm{D}=0} \,|\, \mathrm{D}_i = 1] = \mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=1} \,|\, \mathrm{D}_i = 1] - \mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=0} \,|\, \mathrm{D}_i = 1]$$

ATE, ATET & ATENT

- *Parameter of interest* depends on the context of the study and the (sub-)population of interest. One often considered is Average Treatment Effect,

$$ATE = \mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=1} - \mathrm{Y}_i^{\mathrm{D}=0}] = \mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=1}] - \mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=0}]$$

- If the selection into treatment is not entirely random and some individuals are more likely to enter treatment then it may be preferable to focus on Average Treatment Effect on the Treated

$$ATET = \mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=1} - \mathrm{Y}_i^{\mathrm{D}=0} \,|\, \mathrm{D}_i = 1] = \mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=1} \,|\, \mathrm{D}_i = 1] - \mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=0} \,|\, \mathrm{D}_i = 1]$$

- Notice that ATE can be decomposed into a weighted average effect on the treated and non-treated.

$$\begin{aligned} ATE = & (\mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=1} \,|\, \mathrm{D}_i = 1] - \mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=0} \,|\, \mathrm{D}_i = 1]) \cdot \Pr(\mathrm{D}_i = 1) \\ & + (\mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=1} \,|\, \mathrm{D}_i = 0] - \mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=0} \,|\, \mathrm{D}_i = 0]) \cdot \Pr(\mathrm{D}_i = 0) \\ = & ATET \cdot \Pr(\mathrm{D}_i = 1) + ATENT \cdot \Pr(\mathrm{D}_i = 0) \end{aligned}$$

Table of Contents

1. Overview/Intro
2. Experiment
3. Matching
4. Introduction RD
5. MM - Linear IV
6. DiD

SOCIAL EXPERIMENTS

- In field experiments, treatment assignment is randomized across individuals.

$$(\mathrm{Y}_i^{\mathrm{D}=1}, \mathrm{Y}_i^{\mathrm{D}=0}) \quad \perp \quad \mathrm{D}_i$$

- Treatment assignment is statistically independent of potential outcomes, which solves the problem of self-selection.

$$\begin{aligned} ATE = \mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=1}] - \mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=0}] &= \mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=1} \,|\, \mathrm{D}=1] - \mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=0} \,|\, \mathrm{D}=0] \\ &= \mathrm{E}[\mathrm{Y}_i \,|\, \mathrm{D}=1] - \mathrm{E}[\mathrm{Y}_i \,|\, \mathrm{D}=0] \end{aligned}$$

SOCIAL EXPERIMENTS

- In field experiments, treatment assignment is randomized across individuals.

$$(\mathrm{Y}_i^{\mathrm{D}=1}, \mathrm{Y}_i^{\mathrm{D}=0}) \quad \perp \quad \mathrm{D}_i$$

- Treatment assignment is statistically independent of potential outcomes, which solves the problem of self-selection.

$$\begin{aligned} ATE = \mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=1}] - \mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=0}] &= \mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=1} \,|\, \mathrm{D}=1] - \mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=0} \,|\, \mathrm{D}=0] \\ &= \mathrm{E}[\mathrm{Y}_i \,|\, \mathrm{D}=1] - \mathrm{E}[\mathrm{Y}_i \,|\, \mathrm{D}=0] \end{aligned}$$

- This is because treated and non-treated are random sub-samples of population,

$$\begin{aligned} \mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=1}] = \mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=1} \,|\, \mathrm{D}_i=1] = \mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=1} \,|\, \mathrm{D}_i=0] \\ \mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=0}] = \mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=0} \,|\, \mathrm{D}_i=1] = \mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=0} \,|\, \mathrm{D}_i=0] \end{aligned}$$

- This also implies $ATE = ATET = ATENT$.

DIFFERENCE-IN-MEANS ESTIMATOR

- With unconditional randomization from social experiment, $\mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=1} \mid \mathrm{D}_i = 1]$ and $\mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=0} \mid \mathrm{D}_i = 0]$ can be estimated by their sample means,

$$\mathrm{E}[\mathrm{Y}_i^{\widehat{\mathrm{D}=1} \mid \mathrm{D}_i} = 1] = \frac{\sum_{i=1}^{n} \mathrm{D}_i \, \mathrm{Y}_i}{\sum_{i=1}^{n} \mathrm{D}_i} \quad \text{and}$$

DIFFERENCE-IN-MEANS ESTIMATOR

- With unconditional randomization from social experiment, $\mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=1} \,|\, \mathrm{D}_i = 1]$ and $\mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=0} \,|\, \mathrm{D}_i = 0]$ can be estimated by their sample means,

$$\mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=\widehat{1 \,|\, \mathrm{D}_i} = 1] = \frac{\sum_{i=1}^n \mathrm{D}_i \, \mathrm{Y}_i}{\sum_{i=1}^n \mathrm{D}_i} \quad \text{and} \quad \mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=\widehat{0 \,|\, \mathrm{D}_i} = 0] = \frac{\sum_{i=1}^n (1-\mathrm{D}_i)\, \mathrm{Y}_i}{\sum_{i=1}^n (1-\mathrm{D}_i)}$$

with $\mathrm{D}_i = 0, 1$ and $\mathrm{Y}_i = (1-\mathrm{D}_i)\,\mathrm{Y}_i^{\mathrm{D}=0} + \mathrm{D}_i\,\mathrm{Y}_i^{\mathrm{D}=1}$ which are always observed.

DIFFERENCE-IN-MEANS ESTIMATOR

- With unconditional randomization from social experiment, $\mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=1} \,|\, \mathrm{D}_i = 1]$ and $\mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=0} \,|\, \mathrm{D}_i = 0]$ can be estimated by their sample means,

$$\mathrm{E}[\widehat{\mathrm{Y}_i^{\mathrm{D}=1} \,|\, \mathrm{D}_i} = 1] = \frac{\sum_{i=1}^{n} \mathrm{D}_i \, \mathrm{Y}_i}{\sum_{i=1}^{n} \mathrm{D}_i} \quad \text{and} \quad \mathrm{E}[\widehat{\mathrm{Y}_i^{\mathrm{D}=0} \,|\, \mathrm{D}_i} = 0] = \frac{\sum_{i=1}^{n} (1 - \mathrm{D}_i) \, \mathrm{Y}_i}{\sum_{i=1}^{n} (1 - \mathrm{D}_i)}$$

  with $\mathrm{D}_i = 0, 1$ and $\mathrm{Y}_i = (1 - \mathrm{D}_i) \, \mathrm{Y}_i^{\mathrm{D}=0} + \mathrm{D}_i \, \mathrm{Y}_i^{\mathrm{D}=1}$ which are always observed.

- The resulting estimator for the treatment effects is called the difference-in-means estimator,

$$\widehat{ATE} = \widehat{ATET} = \frac{\sum_{i=1}^{n} \mathrm{D}_i \, \mathrm{Y}_i}{\sum_{i=1}^{n} \mathrm{D}_i} - \frac{\sum_{i=1}^{n} (1 - \mathrm{D}_i) \, \mathrm{Y}_i}{\sum_{i=1}^{n} (1 - \mathrm{D}_i)}$$

- The difference-in-means estimator does not impose any structure on the model.

HETEROGENEOUS TREATMENT EFFECTS

- However, ATE may not be only parameter of interest if individuals respond differently to treatment.
- Individuals with different characteristics x_i have different treatment effects
- In such cases, we may want to evaluate treatment effects given a covariate level,

HETEROGENEOUS TREATMENT EFFECTS

- However, ATE may not be only parameter of interest if individuals respond differently to treatment.
- Individuals with different characteristics X_i have different treatment effects
- In such cases, we may want to evaluate treatment effects given a covariate level,

$$ATE(\boldsymbol{x}) = \mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=1} - \mathrm{Y}_i^{\mathrm{D}=0} | \mathrm{X}_i = \boldsymbol{x}] = \mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=1} | \mathrm{X}_i = \boldsymbol{x}] - \mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=0} | \mathrm{X}_i = \boldsymbol{x}]$$

- In case of a social experiments D_i is also independent of X_i, so $ATE(\boldsymbol{x}) = ATET(\boldsymbol{x}) = ATENT(\boldsymbol{x})$
- ...so why do we often see X_i included in field experiment regressions?

HETEROGENEOUS TREATMENT EFFECTS

- However, ATE may not be only parameter of interest if individuals respond differently to treatment.
- Individuals with different characteristics X_i have different treatment effects
- In such cases, we may want to evaluate treatment effects given a covariate level,

$$ATE(\boldsymbol{x}) = \mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=1} - \mathrm{Y}_i^{\mathrm{D}=0} | \mathrm{X}_i = \boldsymbol{x}] = \mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=1} | \mathrm{X}_i = \boldsymbol{x}] - \mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=0} | \mathrm{X}_i = \boldsymbol{x}]$$

- In case of a social experiments D_i is also independent of X_i, so $ATE(\boldsymbol{x}) = ATET(\boldsymbol{x}) = ATENT(\boldsymbol{x})$
- ...so why do we often see X_i included in field experiment regressions?
- If X_i discrete and low dimensional can stratify and apply the difference-in-means estimator.
- Difference-in-means becomes inefficient if X_i includes continuous variables or if the stratified samples become small.

LINEAR REGRESSION MODEL

- Alternatively, can specify a linear regression model (here only with one discrete covariate X_i), which can be estimated by OLS

$$\mathrm{Y}_i = \beta_0 + \beta_1 \mathrm{X}_i + \beta_2 \mathrm{D}_i + \beta_3 \mathrm{D}_i \mathrm{X}_i + u_i$$

- The linear regression model imposes stronger functional form assumption than difference in means.

LINEAR REGRESSION MODEL

- Alternatively, can specify a linear regression model (here only with one discrete covariate X_i), which can be estimated by OLS

$$\mathrm{Y}_i = \beta_0 + \beta_1 \mathrm{X}_i + \beta_2 \mathrm{D}_i + \beta_3 \mathrm{D}_i \mathrm{X}_i + u_i$$

- The linear regression model imposes stronger functional form assumption than difference in means.

- Since distribution of X_i is similar in the treatment and control group we have,

$$ATE(x) = \beta_2 + \beta_3 x \quad \text{and} \quad ATE = \beta_2 + \beta_3 \mathrm{E}[\mathrm{X}_i]$$

- But what happens if we no longer have a social experiment and the set of conditioning variables in the CSA can not be saturated in the model? Do the parameters still represent the *ATE*?

REGRESSION ESTIMATION OF TREATMENT EFFECTS

- Consider first a simple linear regression model,

$$\mathrm{Y}_i = \mathrm{X}_i' \boldsymbol{\beta} + \delta \, \mathrm{D}_i + \boldsymbol{u}_i$$

- δ captures treatment effect although it is not always clear in practice whether this is *ATE* or *ATET* or something else.

REGRESSION ESTIMATION OF TREATMENT EFFECTS

- Consider first a simple linear regression model,

$$\mathrm{Y}_i = \mathrm{X}_i'\boldsymbol{\beta} + \delta\,\mathrm{D}_i + u_i$$

- δ captures treatment effect although it is not always clear in practice whether this is *ATE* or *ATET* or something else.
- This regression is most often used when randomization of D_i is unconditional since in that case we know $ATE = ATET$.
- Sometimes also used within the context of factor models.
- This simple specification is unlikely to adequately capture correlations between covariates, treatment and unobservables.
- Errors are therefore unlikely to be balanced.

FREQUENTLY USED

- Real Credible Randomization is considered the Gold Standard
  - it is hard to get
  - Hence, NOT frequent.
- A very common application are Economic Programs in Development Evaluation
  - e.g.: teacher training in schools, microcredit programs
  - How done?: If you have 4oo schools in need, you select randomly from them.
- Other Nice examples:
  - Lottery Tickets
  - Vietnam Lottery (for war service)
  - Random Discounts

OUTLOOK - TWO STEP FITTED REGRESSION

- Regression approach with weaker specification assumption is to:

1. Estimate model for $\mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=1} \,|\, \mathrm{D}_i = 1, \mathrm{X}_i = \boldsymbol{x}]$ by linear regression $\mathrm{E}[\mathrm{Y}_i \,|\, \mathrm{D}_i = 1, \mathbf{x}_i]$. Using estimated coefficients generate predicted (fitted) values $\hat{\mathrm{Y}}_i^{\mathrm{D}=1}$ for all treated and non-treated individuals.

2. Estimate model for $\mathrm{E}[\mathrm{Y}_i^{\mathrm{D}=0} \,|\, \mathrm{D}_i = 0, \mathrm{X}_i = \boldsymbol{x}]$ by linear regression $\mathrm{E}[\mathrm{Y}_i \,|\, \mathrm{D}_i = 0, \mathbf{x}_i]$. Using estimated coefficients generate predicted (fitted) values $\hat{\mathrm{Y}}_i^{\mathrm{D}=0}$ for all treated and non-treated individuals.

3. Compute

$$ATE = \frac{1}{n}\sum_{i=1}^{n} \hat{\mathrm{Y}}_i^{\mathrm{D}=1} - \hat{\mathrm{Y}}_i^{\mathrm{D}=0}$$

$$ATET = \left(\frac{1}{n}\sum_{i=1}^{n} \mathrm{D}_i(\hat{\mathrm{Y}}_i^{\mathrm{D}=1} - \hat{\mathrm{Y}}_i^{\mathrm{D}=0})\right) / \frac{1}{n}\sum_{i=1}^{n} \mathrm{D}_i$$

4. Compute standard errors by bootstrapping this procedure.

Table of Contents

1 Overview/Intro

2 Experiment

3 Matching

4 Introduction RD

5 MM - Linear IV

6 DiD

MATCHING ESTIMATORS TO ACCOUNT FOR SELECTION ON OBSERVABLES

- Regression estimators can be sensitive to differences in the covariate distributions for treated and control units.
- If distribution of covariates for treated different than that for controls then fitted values can be sensitive to changes in specification.
- Matching also fits counterfactual outcomes but in a way less sensitive to specification.

MATCHING ESTIMATORS TO ACCOUNT FOR SELECTION ON OBSERVABLES

- Regression estimators can be sensitive to differences in the covariate distributions for treated and control units.
- If distribution of covariates for treated different than that for controls then fitted values can be sensitive to changes in specification.
- Matching also fits counterfactual outcomes but in a way less sensitive to specification.
- Idea is to find for each treated unit (a set of) 'similar' non-treated individuals (and vice-versa for non-treated).
- Assuming conditional independence holds, we can then estimate treatment effect by comparing outcomes for individuals with similar covariates.

MATCHING ESTIMATORS TO ACCOUNT FOR SELECTION ON OBSERVABLES

- Suppose: n_1 individuals observed to receive treatment and n_0 individuals without treatment

$$ATET = \frac{1}{n_1}\sum_{i=1}^{n_1}\left(\mathrm{Y}_i \cdot \mathrm{D}_i - \sum_{j=1}^{n_0} W(i,j)\cdot \mathrm{Y}_j \cdot (1-\mathrm{D}_j)\right)$$

- $W(i,j)$, where $\sum_{j=1}^{n_0} W(i,j) = 1$ for all i, weights non-treated individuals in such a way to construct the counterfactual for individual i in the treatment group.

MATCHING ESTIMATORS TO ACCOUNT FOR SELECTION ON OBSERVABLES

- Suppose: n_1 individuals observed to receive treatment and n_0 individuals without treatment

$$ATET = \frac{1}{n_1}\sum_{i=1}^{n_1}\left(\mathrm{Y}_i \cdot \mathrm{D}_i - \sum_{j=1}^{n_0} W(i,j)\cdot \mathrm{Y}_j \cdot (1-\mathrm{D}_j)\right)$$

- $W(i,j)$, where $\sum_{j=1}^{n_0} W(i,j) = 1$ for all i, weights non-treated individuals in such a way to construct the counterfactual for individual i in the treatment group.
- Many types of weighs to specify $W(i,j)$. Lead to different matching estimators: Nearest Neighbour Matching, Kernel Matching, one-to-one (with or without replacement), etc.
- If $W(i,j) = 1/n_0$, then we have the difference-in-means estimator.

OUTLOOK - Matching based on propensity score

- Finding identical individuals in the treatment and control group suffers from the curse of dimensionality.
- Exact matching on covariates is often not feasible.
- One commonly used approach to reduce dimensionality problem is to match based on propensity score.

OUTLOOK - Matching based on propensity score

- Finding identical individuals in the treatment and control group suffers from the curse of dimensionality.
- Exact matching on covariates is often not feasible.
- One commonly used approach to reduce dimensionality problem is to match based on propensity score.
- Propensity score is probability of entering treatment conditional on covariates:

$$p(\mathbf{x}_i) = Pr(\mathrm{D}_i = 1 | \mathrm{X}_i = \boldsymbol{x})$$

- Rosenbaum and Rubin (Biometrika, 1983) show that CIA for ATE and ATET imply

$$(\mathrm{Y}_i^{\mathrm{D}=1}, \mathrm{Y}_i^{\mathrm{D}=0}) \perp \mathrm{D}_i | p(\mathbf{x}_i) \qquad \text{and} \qquad \mathrm{Y}_i^{\mathrm{D}=0} \perp \mathrm{D}_i | p(\mathbf{x}_i)$$

- Instead of matching on all X_i, we can match on $p(\mathbf{x}_i)$.

OUTLOOK - Matching based on propensity score: estimation

1. Estimate binary model $Pr(\mathrm{D}_i = 1|\mathrm{X}_i = \boldsymbol{x})$ nonparametricaly or for example with Logit model.

2. Next compute the $\hat{p}(\mathbf{x}_i)$ for all i.

3. Use nearest-neighbor matching, kernel-matching, etc. on $\hat{p}(\mathbf{x}_i)$ to match treated to an (weighted set of) individual(s).

4. In large samples, different estimators tend to be very similar.

5. After using propensity score matching, researchers often compare the distribution of X_i-variables in the treatment and constructed control group.

6. Check if X_i-variables are balanced, if not, trim for the group of treated which have overlapping non-treated (this will again affect the subpopulation you define as 'treated').

FREQUENT APPLICATIONS

- SOME matching is usually easy to do.
- The question is whether the matching is sufficient to address the endogeneity concern.
- Matching SHOULD be applied,
  - when enough observations available, but
  - treated are obviously different. (Selection on observables)
- Useful: Depending on Paper, it's something between
  - An improvement over your baseline specification ("Robustness Check")
  - Your identification strategy, if your matching is SUPER convincing...
- That said,
  - SOME Matching is usually better than no bother.
  - even if the condition you are satisfying is rather a necessary, not a sufficient one.

EXAMPLES

1. Twins
2. Android Apps

Table of Contents

1. Overview/Intro
2. Experiment
3. Matching
4. **Introduction RD**
5. MM - Linear IV
6. DiD

REGRESSION DISCONTINUITY DESIGN

- The credibility of your identification strategy relies on the credibility of your randomization.
- In the real world, treatment is rarely result of an arbitrary flip of a coin.
- But we can look for natural experiments or quasi-experiments which produce a credible randomization for (possibly a subset of) the population.

REGRESSION DISCONTINUITY DESIGN

- The credibility of your identification strategy relies on the credibility of your randomization.
- In the real world, treatment is rarely result of an arbitrary flip of a coin.
- But we can look for natural experiments or quasi-experiments which produce a credible randomization for (possibly a subset of) the population.
- Regression discontinuity (RD) design is an example of a quasi-experimental design in which the probability of receiving a treatment is a discontinuous function of one or more underlying variables.
- Regression discontinuity research designs exploit the fact that some administrative or organizational rule. rules are quite arbitrary.

REGRESSION DISCONTINUITY DESIGN

- The credibility of your identification strategy relies on the credibility of your randomization.
- In the real world, treatment is rarely result of an arbitrary flip of a coin.
- But we can look for natural experiments or quasi-experiments which produce a credible randomization for (possibly a subset of) the population.
- Regression discontinuity (RD) design is an example of a quasi-experimental design in which the probability of receiving a treatment is a discontinuous function of one or more underlying variables.
- Regression discontinuity research designs exploit the fact that some administrative or organizational rule. rules are quite arbitrary.
- These arbitrary rules provide good quasi-experiments when you compare people (or cities, firms, countries,...) who are just affected by the rule with people who are just not affected by the rule.
- Example: Van der Klaauw (2003) estimates effect of financial aid offers on student's decision to attend a college, exploiting discontinuity in administrative rule that relates aid to student's SAT score and the grade point average.

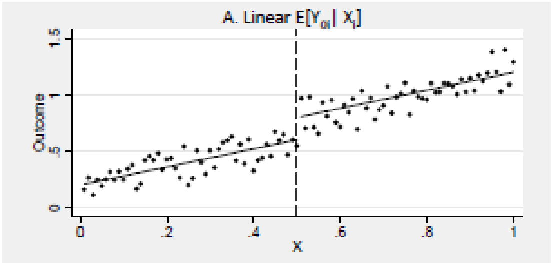
SHARP AND FUZZY RD DESIGN

There are two types of RD designs:

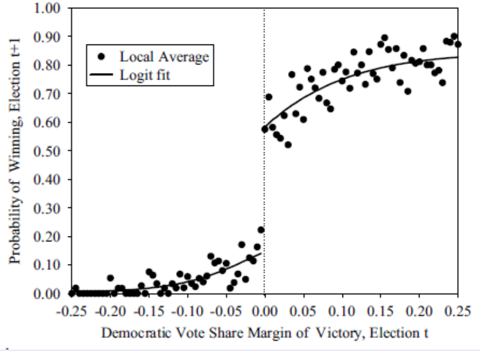
1. Sharp RD: treatment is a deterministic function of a covariate x_i.
2. Fuzzy RD: treatment is a probabilistic function of a covariate x_i (there is only partial compliance to the discontinuity).

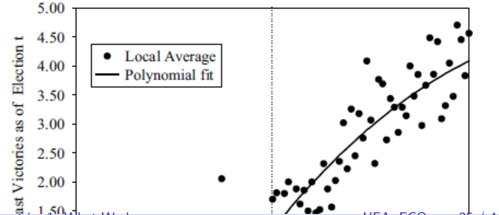
- x_i called *running variable* or *assignment variable* or *forcing variable*.

EXAMPLE LINEAR RD



EFFECT OF WINNING PREVIOUS ELECTION ON PROBABILITY OF WINNING CURRENT ELECTION





FREQUENT APPLICATIONS

- A credible RD is almost as good as an experiment.
  - A good RD, these days, is hard to find...
  - Hence, infrequent.
- Typical use is when a legislation has a rigorous cutoff
- Critical Questions: Ask yourself:
  - does the cutoff coincide with other important things?
    - A country border is not a good cutoff if the countries are quite different.
  - When including observations. How far can you afford to go away from the cutoff?
- → people will ask you these questions

Table of Contents

DiD

- Sometimes we have a shock, that does not affect “comparable” units of observation.
  - e.g. change in legislation in NY City, but not in Kansas
- BUT, if we get a before after too, we can account for the difference between them
- → Difference in Differences.
  1. Difference between Kansas and NY
  2. Difference before and after the regulation

DiD

	Kansas	NYC	Diff
before	2	1	*1.0*
after	3.2	1.3	*1.9*
Diff	1.2	0.3	**0.9**

- No longer assuming that the observations are counterfactuals (as in matching or regression)
- Works, if the **change over time** can be assumed to be equal absent treatment
- That's a higher level assumption, and sometimes it's even testable!

FREQUENT APPLICATIONS

- DiD is quite popular, assumptions are less restrictive and testable.
- You need
  - panel data
  - multiple observation per unit
  - some interesting economic “treatment”, that you cannot easily avoid
    - → regulation changes, new policies, etc.

→ that’s the big condition.

- if you can get that:
  - DiD is a nicely transparent method. This is a prime advantage over matching and IV
  - You can test the core assumption!
  - DiD strikes a nice balance between “doable” and transparent/“credible.”

Examples

Wikipedia 1
Wikipedia 2

Shown in class:
Feel free to check them out here, or ask me about them:
https://sites.google.com/site/kummermannheim/research

Table of Contents

1 Overview/Intro

2 Experiment

3 Matching

4 Introduction RD

5 MM - Linear IV

6 DiD

OUTLOOK - What is a method of moments estimator?

- Method of moments (MM) estimator solves sample moment conditions corresponding to population moment conditions (analogy principle).
- Example: if y_i is IID and $\mathrm{E}(y_i - \mu) = 0$ in the population
  - Use MM estimator $\hat{\mu}$ that solves corresponding sample moment condition $\frac{1}{n}\sum_{i=1}^{n}(y_i - \mu) = 0$.
  - This leads to sample mean $\hat{\mu} = \overline{y}$.
- General Method of Moments (GMM) allows more moment conditions than parameters.
- GMM encompasses all methods we have seen/shall see: OLS, GLS, IV, 2SLS and ML.

MM ESTIMATOR IN LINEAR MODEL

- We assume the linear structural model

$$y_i = \mathbf{x}_i'\beta + u_i$$

- And assume all x_{ik} are exogenous, so $\mathrm{E}(\mathbf{x}_i u_i) = \mathbf{0}$ holds and can be rewritten as,

$$\mathrm{E}(\mathbf{x}_i(y_i - \mathbf{x}_i'\beta)) = \mathbf{0}$$

- If $\mathrm{E}(\mathbf{x}_i\mathbf{x}_i')$ has full rank then above equation has unique solution

$$\beta = (\mathrm{E}(\mathbf{x}_i\mathbf{x}_i'))^{-1}\mathrm{E}(\mathbf{x}_i y_i)$$

- Method of Moments estimator replaces these two population moments by sample moments

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i(y_i - \mathbf{x}_i'\beta) = 0 \qquad \text{from which} \quad \begin{aligned}\hat{\beta}_{MM} &= \Big(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\Big)^{-1}\Big(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i y_i\Big)\\ &= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{y})\\ &= \hat{\beta}_{OLS}\end{aligned}$$

- So MM estimator of β is equivalent to OLS estimator of β.

MM ESTIMATOR IN IV MODEL

- We assume the linear structural model

$$y_i = \mathbf{x}_i'\beta + u_i$$

- And assume NOT all elements of $\mathbf{x}_i$ are exogenous.
- In particular, $\mathrm{E}(x_{ik}u_i) = 0$ for $k = 1, \ldots, L$ but $\mathrm{E}(x_{ik}u_i) \neq 0$ for $k = L+1, \ldots, K$ so $\mathbf{x}_i$ has $K - L$ endogenous variables.
- For notational convenience we rewrite, $\mathbf{x}_i = \begin{bmatrix} \mathbf{x}_{i1} \\ \mathbf{x}_{i2} \end{bmatrix}$, with $\mathbf{x}_{i1}$ the vector of exogenous variables and $\mathbf{x}_{i2}$ the vector of endogenous variables.

OUTLOOK - IV estimator

- In IV model we have a $K-L$ size vector $\mathbf{z}_{i1}$ of additional variables z_{ij} which are exogenous $\mathrm{E}(z_{ij}u_i)=0$, $j=1,\ldots,K-L$.
- Taking $\mathbf{z}_i$ as the vector of exogenous variables we can write moment conditions,

$$\mathrm{E}(\mathbf{z}_i(y_i-\mathbf{x}_i'\beta))=\mathbf{0} \qquad \mathbf{z}_i=\begin{bmatrix}\mathbf{x}_{i1}\\ \mathbf{z}_{i1}\end{bmatrix}$$

- If $\mathrm{E}(\mathbf{z}_i\mathbf{x}_i')$ has full rank then above equation has unique solution

$$\beta=(\mathrm{E}(\mathbf{z}_i\mathbf{x}_i'))^{-1}\mathrm{E}(\mathbf{z}_iy_i)$$

- Method of Moments estimator replaces these population moments by sample moments

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i(y_i-\mathbf{x}_i'\beta)=0 \qquad \text{from which} \quad \hat{\beta}_{IV}=\Big(\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i\mathbf{x}_i'\Big)^{-1}\Big(\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_iy_i\Big)$$
$$=(\mathbf{Z}'\mathbf{X})^{-1}(\mathbf{Z}'\mathbf{y})$$

- So MM estimator of β is the IV estimator of β.

TWO STAGE ESTIMATION OF IV

- We can also obtain IV estimate from two stage estimation

1. Regress all variable(s) x_{ik} on the instrument $\mathbf{z}_i$ using OLS.

$$x_{ik} = \mathbf{z}_i'\gamma + v_{ik}$$

   Calculate the predicted (fitted) values $\hat{x}_{ik}$ of x_{ik}.

2. Use the predicted values (instead of the actual values) $\hat{x}_{ik}$ from the first regression as the explanatory variable in the structural equation, and estimate using OLS.

$$y_i = \hat{x}_{i1}\beta_1 + \ldots + \hat{x}_{iK}\beta_K + u_i$$

- Resulting estimate of the coefficient on predicted $\hat{x}_{ik}$ are the IV estimate of β_k.
- Interpret this as ‘purging’ endogenous variable of the correlation with the error.

IV ASSUMPTIONS

- Recall the instrument $\mathbf{z}_{ij}$ for endogenous variable x_{ik} must satisfy two properties:
  1. Exogeneity IV.1: $\mathrm{Cov}(z_{ij}, u_{ik}) = 0$
  2. Relevance IV.2: $\mathrm{Cov}(z_{ij}, x_{ik}) \neq 0$
- IV.1 is necessary otherwise moment equations do not hold.
- IV.2 is also necessary otherwise rank condition would not hold.
- To see this write in simple case of one endogenous variable x_{iK} and one instrument z_{i1},

$$x_{iK} = x_{i1}\gamma_1 + \ldots + x_{iK-1}\gamma_{K-1} + z_{i1}\gamma_K + v_{iK}$$

- If $\gamma_K = 0$, that is $\mathrm{Cov}(z_{i1}, x_{iK}) = 0$, then $\mathrm{E}(\mathbf{z}_i\mathbf{x}_i')$ will not have full rank.

FREQUENT APPLICATIONS

- A **really** convincing IV is not easier to find than a good RD.
- Like with matching SOME IV is usually relatively easy.
- The Question will be:
  - Can you defend your exclusion restriction? (IV.1)
  - → They will ask you, you cannot test it.
  - This is also the reason for IV's popularity.
- That said, like with matching: SOME IV is better than no IV