Treatment Effects

Overview

- Preliminaries
- Control for Covariates
  - Selection on Observables
  - Regression Discontinuity
- Matching
- Difference in Difference
  - Difference in Difference
  - Triple Difference
- IV estimation for hidden bias
- Other approaches
  - Heckman’s MTE

Preliminaries: Dummy Variables

In econometrics, qualitative information is usually captured by defining a zero-one variable. It's called a dummy variable or a binary variable.

For example, $female_i = \begin{cases} 1 & \text{if individual } i \text{ is female} \\ 0 & \text{if individual } i \text{ is male} \end{cases}$.

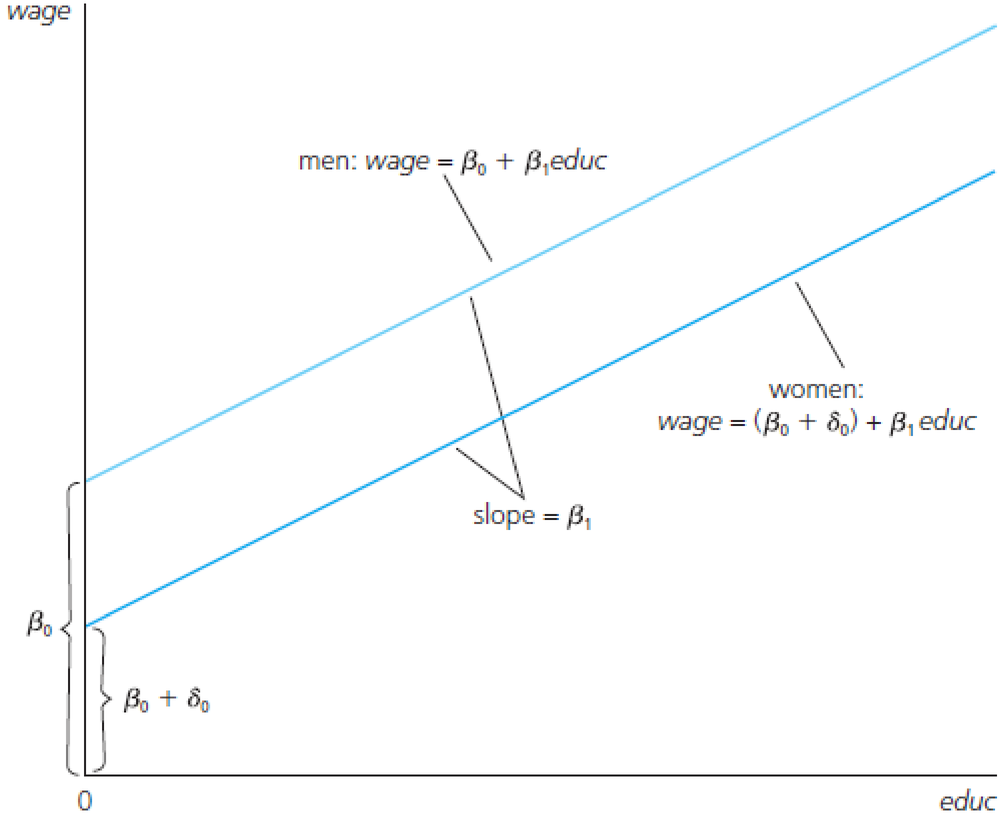
In regression, parameters of dummy variables have natural interpretations: differences between groups divided by dummy variables.

For more examples, dummy variables can be races, industries, regions, etc.

Thus, dummy variables are commonly used in policy analysis or program evaluations when the policy or program takes place with only one level, which is exactly the case we'll discuss.

Dummy Variables

This is an example in Wooldridge's textbook when the gender difference in wage in constant as $wage = \beta_0 + \delta_0 \text{fe}male + \beta_1 educ + u$.


wage
men: $wage = \beta_0 + \beta_1 educ$
women:
$wage = (\beta_0 + \delta_0) + \beta_1 educ$
slope = β_1
β_0
$\beta_0 + \delta_0$
0
educ

Dummy Variables

If there're male and female both married or single in study of wage, we can use three dummy variables in the regression.

$$\ln(wage) = \alpha_0 + \alpha_1 marrmale + \alpha_2 marrfem + \alpha_3 singfem + \cdots$$

Or, we can use interactions of dummy variables.

$$\ln(wage) = \beta_0 + \beta_1 female + \beta_2 marr + \beta_3 marr \cdot female + \cdots$$

In both examples, we regard the single male as a baseline and allow the "married effect" to differ between male and female.

α_1, α_2, α_3 are the differences between other groups and the baseline.

Dummy Variables

When the dependent variable is binary, we have a binary choice model.

Linear Probability Model: $E(Y|X) = X\beta$

Probit Model: $E(Y|X) = \Phi(X\beta)$ where $\Phi(\cdot)$ is the CDF of Normal Distribution.

Logit Model: $E(Y|X) = e^{X\beta}/(1 + e^{X\beta})$

Setup of Treatment Model

Rubin (1974), Holland (1986), Pearl (2000), Rosenbaum (2002):

One has two potential outcomes, Y_1 and Y_0.

Y_1 describes what would happen if one is treated.

Y_0 describes what would happen if one is not treated.

A binary variable *D* presents for the treatment.

So only one of Y_d is observed depending on D:

$$Y = D \cdot Y_1 + (1 - D) \cdot Y_0$$

D can be many interested variables, like medicine, education, job-training.

We’re interested in the causal relation between D and Y.

In other words, the difference between Y_1 and Y_0, so-called treatment effects.

Three No-Effect Concepts

Exact Same: $Y_1 = Y_0$

Exchangeability: $P(Y_0 \leq t_0, Y_1 \leq t_1) = P(Y_1 \leq t_0, Y_0 \leq t_1)$ for $\forall t_0, t_1$

Zero mean(or median): $E(Y_1) = E(Y_0)$ or $Med(Y_1) = Med(Y_0)$

Their relation:

Exact Same ⇒ Exchangeability ⇒ Zero mean/median

Definition of Treatment Effects

These treatment effects are defined by statistics:
mean treatment effects: $E(Y_1 - Y_0)$ or $E(Y_1 - Y_0|X)$
quantile treatment effects: $Q_{Y_1}(\tau|X) - Q_{Y_0}(\tau|X)$
distributional treatment effects: $F_{Y_1|X}(y) - F_{Y_0|X}(y)$

These mean treatment effects are defined by populations:
average treatment effects (ATE): $E(Y_1 - Y_0)$
average treatment effects on treated (ATT): $E(Y_1 - Y_0|D = 1)$
average treatment effects on untreated (ATUT): $E(Y_1 - Y_0|D = 0)$

ATE is useful when treatment has broad applicability.

ATT is mostly for those “focused” program. We don’t care the college return for one who doesn’t finish senior school.

Treatment Effects

For example, we want to know the return of college.

For college graduates who have D=1, we observe their wage Y_1.

And for those not entering college, we observe their wage Y_0.

So, we only observe one of Y_1 and Y_0 for an individual.

$$E(Y_1 - Y_0|D = 1) = E(Y|D = 1) - E(Y_0|D = 1)$$
$$E(Y_1 - Y_0|D = 0) = E(Y_1|D = 0) - E(Y|D = 0)$$

The key is how to estimate the **counterfactual** results.

Experiments

In an experiment of medicine:

(1) Individuals are randomly assigned to two groups.

(2) In one group, they take the medicine (treated group).

(3) In another group, they take no medicine (or just take placebo). It’s the untreated group (or control group).

(4) The difference between two groups are taken as the effect of medicine ($E(Y|D=1)-E(Y|D=0)$).

Under such randomization, the assignment of treatment is independent of everything (including all observed X and unobserved ε which may affect Y_d).

Experiments

Randomization Assumption: $Y_1 \perp D$ and $Y_0 \perp D$

$$E(Y_1|D=1) = E(Y_1|D=0) = E(Y_1)$$
$$E(Y_0|D=1) = E(Y_0|D=0) = E(Y_0)$$

$$E(Y|D=1) - E(Y|D=0) = E(Y_1|D=1) - E(Y_0|D=0)$$
$$= E(Y_1 - Y_0) = ATE = ATT = ATUT$$

Example: a randomized experiment for medicine (Rosner, 1995):

	N	age (SD)	edu (SD)	black men	black women
Treated	2365	71.6 (6.7)	11.7 (3.5)	4.9%	8.9%
Control	2371	71.5 (6.7)	11.7 (3.4)	4.3%	9.7%

Estimation

Under randomization assumption: treatment effects are estimated as,

$$\hat{\beta} = \frac{\sum D_i Y_i}{\sum D_i} - \frac{\sum(1 - D_i)Y_i}{\sum(1 - D_i)}$$

It's same to run OLS for the following equation:

$$Y_i = \alpha + \beta D_i + U_i$$

It's same to the difference between sample averages of two groups.

$$\hat{\beta} = \bar{Y}_{treated} - \bar{Y}_{control}$$

Internal and External Validity

Internal Validity: our conclusion truly represents the sample.

External Validity: our conclusion can be applied to other populations for prediction and so on.

Threats to Internal Validity:

- psychological effect: If one knows being treated, he may feel better. That’s why control group will take placebo in medicine experiments. But placebo isn’t available for many other treatments, like job-training.
- substitution effect: If one is told to sleep less or do less sport, he may make up for it in other ways.

Threats to External Validity:

- non-participation effect: The sample is different from the whole population. e.g., if data is collected on Internet, those not using Internet are excluded.

Violations to Randomization

In economics, our data is rarely from experiments.

Consider the case:

Half are Females with $Y_1 = 70\ and\ Y_0 = 75$. And 80% take the treatment.

Others are Males with $Y_1 = 50\ and\ Y_0 = 55$. And 20% take the treatment.

So $E(Y|D = 1) - E(Y|D = 0) = 66 - 59 = 7$. Positive.

0.8*70+0.2*50=66

Because of unbalance of observables (here is gender), we have overt bias. To solve it, we should control for covariates.

看得见的偏差

Violations to Randomization

Consider the case of college: 读大学的回报：70-50=20

Half are of high ability with $Y_1 = 70\ and\ Y_0 = 50$.

Others are of low ability with $Y_1 = 40\ and\ Y_0 = 30$. 大学回报：40-30=10

If tuition fee is 15, high ability goes to college while low ability not.

So $E(Y|D = 1) - E(Y|D = 0) = 70 - 30 = 40$. Much higher than those of high ability (20) or low ability (10). 显然，40高估了

Because of unbalance of unobservables (here is ability), we have hidden bias. To solve it, we may need some instruments.

Selection on Observables and Unobservables

Selection on observables (like gender):

(1) D is independent of Y_d, conditional on X ($f(\cdot)$ is density function).

$$f(Y_d|D,X) = f(Y_d|X)$$

D和Y0、Y1是独立的，所以可以不出现

(2) D is mean-independent of Y_d, conditional on X.

$$E(Y_d|D,X) = E(Y_d|X)$$

给定能力，就和potential outcome无关了

Only overt bias, no hidden bias.

observable

Selection on unobservables (like ability):

$$f(Y_d|D,X) \neq f(Y_d|X) \text{ and } f(Y_d|D,X,\varepsilon) = f(Y_d|X,\varepsilon)$$

$$\text{or } E(Y_d|D,X) \neq E(Y_d|X) \text{ and } E(Y_d|D,X,\varepsilon) = E(Y_d|X,\varepsilon)$$

Have hidden bias.

Control for Covariates

- Selection on observables
  - Cases
  - Identification
  - Treatments on different Populations
  - Two-stage method for Semi-linear Model
  - Efficient Estimators
  - Relation between Treatment and Regression

- Regression Discontinuity (RD)
  - sharp RD
  - fuzzy RD
  - Choices in RD

- Before-After (BA) DID

Control for covariates

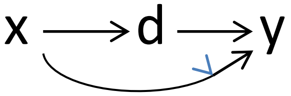
Failing to control covariates may result in biased estimators.

If $E(Y_0|D=0) \neq E(Y_0|D=1)$, then

$$E(Y|D=1) - E(Y|D=0) \neq E(Y_1 - Y_0|D=1).$$

There're three case when x may be controlled:

(1) must case:



不能控制的情况 (2) no case: d → y → x   d → x → y

(3) option case: d → y ← x (experiments)

Identification

反例：selection on y0，y1，工资低参加了training

Under mean-independence $E(Y_d|D,X) = E(Y_d|X)$,

$$
\begin{aligned}
ATT(X) &= E(Y_1 - Y_0|X, D = 1) \\
&= E(Y_1|X, D = 1) - E(Y_0|X, D = 1) \text{在什么位置上} \\
&= E(Y_1|X, D = 1) - E(Y_0|X, D = 0) \text{看，是一样的} \\
&= E(Y|X, D = 1) - E(Y|X, D = 0) \\
&= ATE(X) = ATUT(X)
\end{aligned}
$$

If ATE(X) is linear in X, we can run the following regression.

即可推出回归模型$Y_i = \alpha + X_i\beta + \delta_0 D_i + \delta_1 X_i D_i + U_i$

Then $ATE(X) = \delta_0 + \delta_1 X$.

Treatments on different populations

从右往左看来理解，承认是obsevable的，所以是相等的

Conditional TE	Assumption	Unconditional TE
ATE(X)	$E(Y_d \mid D, X) = E(Y_d \mid X)$	$\int ATE(X) \cdot f(X)\, dx$
ATT(X)	$E(Y_0 \mid D, X) = E(Y_0 \mid X)$	$\int ATT(X) \cdot f(X \mid D = 1)\, dx$
ATUT(X)	$E(Y_1 \mid D, X) = E(Y_1 \mid X)$	$\int ATUT(X) \cdot f(X \mid D = 0)\, dx$

Even if $ATE(X) = ATT(X) = ATUT(X)$,

there can be $ATE \neq ATT \neq ATUT$.

two-stage method for semi-linear model

非参数估计

In non-parametric estimation, there is a dimension problem.

In parametric estimation, there is a risk of misspecification.

Robins et al. (1992) suggest a way to relax the linear specification on X.

$$Y_i = \beta D_i + g(X_i) + U_i, \qquad E(U|D,X) = 0.$$

Conditional on X on both sides, we have

$$E(Y|X) = \beta E(D|X) + g(X)$$

Then

$$Y - E(Y|X) = \beta\big(D - E(D|X)\big) + U$$

two-stage method for semi-linear model

$$Y - E(Y|X) = \beta\big(D - E(D|X)\big) + U$$

Estimation:

- Estimate *E(Y|X)* and *E(D|X)*, e.g. $\mathrm{E}(D|X) = p_i(\alpha) = \Phi(x_i{}'\alpha)$ by probit.
- Regress *Y-E(Y|X)* on *D-E(D|X)* to estimate β.

stata command for the semi-linear model

$$Y - E(Y|X) = \beta\big(D - E(D|X)\big) + U$$

database: hprice3.dta

dependent variable: log house price

independent variables: log dist. from house to incin., log square footage of house, log square footage lot, number of rooms, number of bathrooms, house age, and log dist. to interstate.

remark: *linst* may have a nonparametric relation with *lprice*.

command:

```
semipar lprice ldist larea lland rooms bath age, nonpar(linst) ci
```

Efficient Estimators

Hahn (1998), Hirano et al. (2003) and Imbens (2004) provide efficient estimators for ATT and ATE.

$$\widehat{ATE} = \frac{1}{N}\sum\left[\frac{d_i y_i}{\pi_N(x_i)} - \frac{(1-d_i)y_i}{1-\pi_N(x_i)}\right]$$

即为yi的权重

$$\widehat{ATT} = \frac{1}{N_1}\sum\left[d_i y_i - \frac{\pi_N(x_i)(1-d_i)y_i}{1-\pi_N(x_i)}\right]$$

It’s a method of weighting. For example, in ATE, y_i is weighted by $\frac{1}{\pi_N(x_i)}$ if d_i=1, and $-\frac{1}{1-\pi_N(x_i)}$ otherwise.
However, the absolute weight may be extremely large for some data when π close to 0 or 1.

stata command for Efficient Estimators

database: cattaneo2.dta

dependent variable (Y): baby's birth weight (*bweight*)

treatment variable (D): whether the mother smokes (*mbsmoke*)

control variables (X): age, education, married, and a dummy for first baby

command:

inversed propensity score weight(ipw)

```
teffects ipw (bweight) (mbsmoke mage medu mmarried fbaby, probit)
```

probit是说明pi用什么方法来估计的

result: ATE is -230.9 grams, ATT is -219.6 grams.

an empirical example

β (t-value)	all observations	$0.001<\pi_N(x)<0.999$
ATUT	8.10 (2.07)	-0.01 (-0.01)
ATT	0.29 (15.79)	0.26 (14.88)
ATE	7.36 (2.07)	0.02 (0.03)

This is an empirical study about the job-training effect on *ln(unemployment duration)* with N=10285 and N_1=973.

The second column is the treatment effects estimated with all observations. The third column drops only 19 observations with $\pi_N(x)$ close to 0. However, *ATUT* and *ATE* changes dramatically.

The reason is that if estimated $\pi_N(x_i)$ close to 0, but D_i is 1, the weight is extremely large. The estimation is sensitive to these observations.

Relation between Treatment and Regression

Y=Y0+(Y1-Y0)D,带入

Treatment Effects Model:

$$Y_1 = g_1(x) + U_1, Y_0 = g_0(x) + U_0$$
$$Y = D \cdot Y_1 + (1 - D) \cdot Y_0$$

Rewrite it as regression:

$$Y = D \cdot (g_1(x) - g_0(x) + U_1 - U_0) + g_0(x) + U_0$$

If $g_1(x) - g_0(x) = \delta$ and $U_1 = U_0$ (constant treatment effects), then

$$Y = D\delta + g_0(x) + U_0$$

So the regression form is more restrictive, and usually assuming constant treatment effects.

constant treatment时，都一样ATE

Homogeneity and Heterogeneity

If β$_i$ can be a random coefficient in

$$Y_i = \beta_i D_i + g(X_i) + U_i$$
$$Y_i = \beta D_i + g(X_i) + (U_i + (\beta_i - \beta)D_i)$$

Homogeneity: $\beta_i = \beta \Rightarrow$ traditional OLS constant effect, 所有人都一样

Heterogeneity: $\beta_i \perp D_i | X_i \Rightarrow \beta_i = \beta(X_i)$

Essential Heterogeneity: β_i not independent of D_i conditional on X_i

Regression Discontinuity

Hahn et al. (2001), Porter (2003), Imbens & Lemieux (2008).

The treatment *D* is (partially) determined by a threshold of *X*.

sharp RD: $D = 1\{X > \tau\}$

fuzzy RD: $\lim_{X \to \tau^+} E(D|X) \neq \lim_{X \to \tau^-} E(D|X)$

It’s still a “selection on observables”.

With threshold randomization, the individuals beside the threshold are analogous and comparable. It’s a quasi-experiment.

Example: vote share, score rank, area boundary.

Regression Discontinuity

sharp RD:

$Y = Y_1$ if $X > \tau$ and $Y = Y_0$ otherwise, $ATE = E(Y_1 - Y_0 | X = \tau)$

For example, $Y_d = \delta \cdot D + g(X) + U_d$.

If g is a continuous function and $\lim_{X \to \tau} E(U_1 | X) = \lim_{X \to \tau} E(U_0 | X)$,

then $\lim_{X \to \tau^+} E(Y|X) - \lim_{X \to \tau^-} E(Y|X) = \delta$.

sigema即为政策效应，Di为0，1

Estimation: $Y_i = \alpha + \beta_0 X_i + \beta_1 {X_i}^2 + \delta D_i + U_i$

Regression Discontinuity

Example of sharp RD:

In American college, students will get in trouble if their GPA<1.5.

$$D = 1[GPA < 1.5]$$

Students with GPA=1.51 and those with GPA=1.49 are similar in observables and unobservables.

The study also checks the distribution of GPA, and distributions of other observables conditional on GPA. They are all continuous at 1.5. So g(x) is continuous at x=1.5.

If it’s the case that some students have a connection with teacher, and can do something with their GPA expost, it violates the design because U (unobserved connection) is not continuous at the threshold.

Regression Discontinuity

fuzzy RD:

$$\lim_{X\to\tau^+} E(D|X) \neq \lim_{X\to\tau^-} E(D|X)$$

For example, $\mathrm{Y} = \delta D + g(X) + U$. Still, g is continuous.

$$\lim_{X\to\tau^+} E(Y|X) = \delta \cdot \lim_{X\to\tau^+} E(D|X) + g(\tau) + \lim_{X\to\tau^+} E(U|X)$$

$$\lim_{X\to\tau^-} E(Y|X) = \delta \cdot \lim_{X\to\tau^-} E(D|X) + g(\tau) + \lim_{X\to\tau^-} E(U|X)$$

1-2式

If $\lim_{X\to\tau^+} E(U|X) = \lim_{X\to\tau^-} E(U|X)$, then

$$\delta = \frac{\lim_{X\to\tau^+} E(Y|X) - \lim_{X\to\tau^-} E(Y|X)}{\lim_{X\to\tau^+} E(D|X) - \lim_{X\to\tau^-} E(D|X)}$$

比起sharp,

即多一个系数而已

Regression Discontinuity

Estimation of fuzzy RD:

$$\delta = \frac{\lim_{X \to \tau^+} E(Y|X) - \lim_{X \to \tau^-} E(Y|X)}{\lim_{X \to \tau^+} E(D|X) - \lim_{X \to \tau^-} E(D|X)}$$

$$Y = \alpha_1 + \beta_1 X + \cdots + \delta_1 \cdot 1[X > \tau] + U$$
$$D = \alpha_0 + \beta_0 X + \cdots + \delta_0 \cdot 1[X > \tau] + V$$

$$\hat{\delta} = \frac{\widehat{\delta_1}}{\widehat{\delta_0}}$$

Empirical Example

Van de Klaauw (2002) studies the effect of financial aid on college enrollment. The college has an ability index “x” for students and offers some financial aid depending on it.

$$aid = \gamma_1 \cdot 1[x \geq \tau_1] + \gamma_2 \cdot 1[x \geq \tau_2] + \gamma_3 \cdot 1[x \geq \tau_3]$$

The actual amount of aid differs from this function because other factors are also taken into consideration, which makes the RD fuzzy.

He estimates $\gamma_1, \gamma_2, \gamma_3$ as 1280, 1392 and 3145. By this RD, he estimates the effect of financial aid on college enrollment.

RD for financial aid effect on enrollment

	threshold 1	threshold 2	threshold 3
effect (SD)	0.010 (0.238)	0.040 (0.041)	**0.067 (0.029)**

Choices in RD

1. Bandwidth: how many data are near the threshold.

   If the threshold is GPA=1.5, we don't want to use GPA=4.0 which is quite different in characteristics. But if we only use 1.4<GPA<1.6, the sample size is too small.

2. Functional Form: linear, quadratic or higher polynomials.

   If the bandwidth is quite small, linear form is enough. If the bandwidth is large, we may want to use higher polynomials.

   With higher polynomials, the estimation is more robust, but less powerful.

3. Kernel Functions.

   Instead of functional form, we may use kernel functions for limitations.

stata command for sharp RD

database: votex.dta

dependent variable (Y): Log fed expenditure in district (*lne*)

treatment variable (D): whether the Democratic won the race

regression discontinuity: if Dem vote share higher than 0.5, they won; otherwise they lost. Here, $d = vote\ share - 0.5$. So discontinuity happens at d=0.

command:

```
rd lne d, graph mbw(100)
```

result: The treatment effect is -7.7%, not significant.

stata command for fuzzy RD

database: votex.dta

Remark: we use a simulated data where "*win*" isn't fully determined by vote share.

Simulated treatment: *ranwin*

Simulation command: gen byte ranwin=cond(uniform()<.1,1-win,win)

fuzzy RD command:

```
rd lne ranwin d, gr mbw(100)
```

result: The treatment effect is -8.4% as following:

$$\hat{\delta} = \frac{\lim_{X\to\tau^+} E(Y|X) - \lim_{X\to\tau^-} E(Y|X)}{\lim_{X\to\tau^+} E(D|X) - \lim_{X\to\tau^-} E(D|X)} = \frac{-7.7\%}{0.92} = -8.4\%$$

Empirical Example

Black (1999) studies how much parents are willing to pay for better education for children.

	school quality	bedrooms	bathrooms	bathrooms^2	building age
all houses	0.035 (0.004)	0.033 (0.004)	0.147 (0.014)	-0.013 (0.003)	-0.002 (0.0003)
boundary houses	0.016 (0.007)	0.038 (0.005)	0.143 (0.018)	-0.017 (0.004)	-0.002 (0.0003)

The table shows that results are similar for house characteristics, but the key effect is halved in the boundary house case, which controls the school district and neighborhood characteristics better than the all-house case does.

Before-After

Before-After is similar to RD where discontinuity takes place in time-dimension.

$$Y_t = \beta_t D_t + g(t) + U_t$$
$$D_t = 1[t \geq \tau]$$

Advantage: Use one's self as counterfactual results to control for individual effect.

Disadvantage: The treatment and its effect should take place so quickly that other covariates don't change during the period.

Matching

- Definition
- Estimation
- Implement
- Evaluation

Matching

$$ATT(X) = E(Y_1 - Y_0|X, D = 1) = E(Y|X, D = 1) - E(Y_0|X, D = 1)$$

Matching is to find observations in control group to imply $E(Y_0|X, D = 1)$.

For ATUT, matching is to find those in treated group to imply $E(Y_1|X, D = 0)$.

Still, we need “selection on observables”.

Observations in T group and C group with similar X are analogous.

$T=\{1,\ldots,i,\ldots,N_1\}$: set of treated group

$C=\{1,\ldots,j,\ldots,N_0\}$: set of control group

$C_i \subseteq C$: observations matched to $i \in T$

$T_j \subseteq T$: observations matched to $j \in C$

Matching

$$\widehat{ATT} = \frac{1}{N_{1u}} \sum_{i \in T_u} \left(y_i - \frac{\sum_{m \in C_i} y_{mi}}{|C_i|}\right)$$

$$T_u = \{i : C_i \neq \emptyset, i \in T\} \text{ and } N_{1u} = |T_u|$$

In other words, we use the mean of individuals in C_i matched to y_i as the counterfactual outcome for a treated individual $i \in T$.

0为匹配不到，1为匹配

In pair matching, $|C_i|$ = 0 or 1. $\widehat{ATT} = \frac{1}{N_{1u}} \sum_{i \in T_u} (y_i - y_{mi})$.

In weight matching, C_i=C. All observations in C play a role in counterfactual outcome. The weight is usually determined by kernel functions.

$$\widehat{ATT} = \frac{1}{N_1} \sum_{i \in T} (y_i - \sum_{j \in C} w_{ij} y_j) \text{ where } \sum_{j \in C} w_{ij} = 1.$$

Implement matching

Implementing matching is to choose many criteria on which to be matched.

Distance Measurement: we want to use y_j sharing a similar characteristics with y_i as counter-factual result. So their similarity is represented by their distance.

If the distance is $d(x_i, x_j)$, we have nearest-neighbor (NN) matching.

-If $d(x_i, x_j) = (x_i - x_j)'(x_i - x_j)$, it's called Euclidean Distance.

-If $d(x_i, x_j) = (x_i - x_j)'{\Sigma_N}^{-1}(x_i - x_j)$, it's called Mahalanobis Distance.

If the distance is $|p(x_i) - p(x_j)|$, we have Propensity Score Matching.

降低了维度

Caliper: $d(x_i, x_j)$ < *caliper or* $\left|p(x_i) - p\left(x_j\right)\right| < caliper$.

If caliper is too big, matching provides a poor counterfactual result.

If caliper is too small, there're few successful matchings.

Implement matching

Matching Number: within caliper, which are matched.

If we use all $j \in C$ with $d(x_i, x_j)$ < *caliper*, $|C_i|$ is random.

If we only use the nearest few observations with $d(x_i, x_j)$ < *caliper*, $|C_i|$ is fixed.

-If $|C_i|$ is fixed to 1 for a matched y_i, we have pair-matching.

-If $|C_i|$ is fixed to k(>1) for a matched y_i, we have multiple-matching

Stratification Matching: even if other characteristics are same, an observation with a different gender may provide a poor counter-factual result.

So, firstly we perform a stratification depending on x_1, and then match *i* and *j* by x_2 within the same strata. x_1 can be gender, age, education, etc.

Implement matching

If there are many x_i and few x_j somewhere, such x_j may be matched many times to different x_i. Thus, the estimation may be sensitive to such observations.

Greedy Matching: an observation is matched once at most.

Non-greedy Matching: an observation can be matched many times.

In algorithm, we have **sequential matching** and **non-sequential matching**. But pay attention that greedy sequential matching may have different results with different sequences.

stata command for nearest neighbor matching

database: cattaneo2.dta

dependent variable (Y): baby’s birth weight (*bweight*)

treatment variable (D): whether the mother smokes (*mbsmoke*)

control variables (X): age, education, married, and a dummy for first baby

command:

```
teffects nnmatch (bweight mage medu mmarried fbaby )
(mbsmoke)
```

result: ATE is -204.8 grams, ATT is -239.2 grams.

options: *ematch()* for straitification, *metric()* for distance metric

Evaluating Matching

In implementing matching, we choose many criteria. Thus, we want to see if the matching can be better by changing some criteria.

In evaluating, we consider three aspects:

1. how **close** the matching is: whether the matched ones is closed to observations, thus implying the counterfactual outcome.
2. how **balance** the matching is: whether the matched ones lie around observations evenly, thus the estimation is unbiased.
3. how much **information** we lose: how many observations are unmatched, thus absent from the estimation.

Thus, there is a trade-off between efficiency and information. If we choose a lower caliper, “distance & balance” becomes better, with a cost of more information loss.

Evaluating Matching

Distance

$$M_{|j|} = \frac{1}{|T_u|} \sum_{i \in T_u} |x_i - \frac{1}{|C_i|} \sum_{m \in C_i} x_{mj} |$$

Balance

$$M_{(j)} = \frac{1}{|T_u|} \sum_{i \in T_u} (x_i - \frac{1}{|C_i|} \sum_{m \in C_i} x_{mj})$$

Information

$$\% = \frac{|T_u|}{|T|}$$

Empirical Example

	No-strata	Multi-NG			Pair-NG	Pair-G	Pair-G
caliper	h=0.5	h<0.4	h=0.5	h=1	h=1	h=1	h=∞
ATT (t-value)	4.1%	12.5%	2.8%	2.7%	4.7% (0.77)	5.8% (0.84)	6.7% (1.21)
$x_{\|j\|}$ abs diff. (edu, exper)	0.393, 0.393, 0, … , 0	0.000, 0.000	0.437, 0.244	0.283, 0.415	0.333, 0.390	0.443, 0.375	0.902, 0.870
$x_{(j)}$ diff. (edu, exper)	-0.126, 0.126, 0, … , 0	0.000, 0.000	-0.186, 0.109	-0.130, 0.098	-0.124, 0.086	-0.193, 0.080	-0.252, 0.138
% of $\|T_u\|$	53.8%	37.1%	61.5%	73.4%	73.4%	61.5%	86.0%
average $\|C_i\|$	2.64	2.04	3.58	4.30	1	1	1

This is a study about effects of union membership on ln(wage).

We can see a trade-off between matching efficiency and information.

Empirical Example

From the above chart, we see some facts in this example.

Multi-matching shares the same $|T_u|$ and similar M (possibly larger) with pair matching, but uses more observations in the control group.

Greedy matching performs bad with a larger M and a lower % of $|T_u|$, but easier for calculation.

With a lower caliper h, we can have a lower M (especially for pair matching), with the cost of a lower % of $|T_u|$.

Propensity Score

Define $\pi(x) = E(D|X = x) \in (0,1)$, then given *π(X)*, the distribution of X is the same across two groups (treated group and control group).

proof: $P(D = 1, X \leq t|\pi(X) = \pi) = E(D \cdot 1[X \leq t]|\pi(X) = \pi)$

$= E(E(D|X) \cdot 1[X \leq t]|\pi(X) = \pi) = \pi \cdot P(X \leq t|\pi(X) = \pi)$

Thus, $P(X \leq t|D = 1, \pi(X) = \pi) = P(X \leq t|D = 0, \pi(X) = \pi)$.

π(X) is the minimum variable we need to control for (Rosenbaum and Rubin (1983a)).

$$D \perp (Y_1, Y_0) \mid X \;\Leftrightarrow D \perp (Y_1, Y_0) \mid \pi(X)$$

Propensity score matching

Advantage: reduction of dimension

Robin et al. (1992) find that estimated propensity scores perform better than true values in the second step.

an example of greedy pair matching with calipered propensity score

	caliper: 0.00001			**caliper: 0.001**		
ATT (t-value)	0.248 (5.94)			0.326 (11.38)		
% used treated	39%			91%		
	age	ex-firm	edu	age	ex-firm	edu
$M_{(\cdot)}$	0.113	-0.106	0.016	0.104	-0.079	-0.024
$M_{\lvert\cdot\rvert}$	1.997	1.622	0.702	2.086	1.694	0.835

stata command for propensity score matching

database: cattaneo2.dta

dependent variable (Y): baby's birth weight (*bweight*)

treatment variable (D): whether the mother smokes (*mbsmoke*)

control variables (X): age, education, married, and a dummy for first baby

command:

```
teffects psmatch (bweight) (mbsmoke mmarried mage fbaby medu, probit)
```

result: ATE is -201.1 grams, ATT is -232.0 grams.

options: *caliper()*

Difference in Difference

In matching, we use the control group $E(Y_0|D=0)$ to infer counterfactual results of the treated group $E(Y_0|D=1)$.

So, it’s biased with (unobserved) individual effect.

In Before-After (BA), we assume that the treatment happens so quickly that there’s no time effect.

So, it fails to control for time effect.

Now we use a combination of BA and matching.

Use a control group to deal with both individual and time effects.

Difference in Difference

$t \in \{0,1\}$: t=0 if before the treatment time; t=1 if after

$r \in \{0,1\}$: r=0 if in control group; r=1 if in treated group

$\tau = t \cdot r$ is the treatment. So τ=1 only if t=1 and r=1.

Like in experiments, only treated group after the treatment is treated.

$Y_{\tau t}$ is the potential outcome in time t when treatment status is τ.

So, $E(Y_{11} - Y_{01}|r = 1)$ is the interested ATT.

In data, there're four observed volumes (two groups in two times):
$E(Y_{11}|r = 1), E(Y_{00}|r = 1), E(Y_{01}|r = 0), E(Y_{00}|r = 0)$.

Difference in Difference

Panel Data without movers (e.g., schools in Beijing & Tianjin):

$$
\begin{aligned}
DID &= [E(Y|r=1,t=1) - E(Y|r=1,t=0)] \\
&\quad - [E(Y|r=0,t=1) - E(Y|r=0,t=0)] \\
&= E(Y_{11} - Y_{00}|r=1) - E(Y_{01} - Y_{00}|r=0) \\
&= E(Y_{11} - Y_{01}|r=1) + [E(Y_{01} - Y_{00}|r=1) - E(Y_{01} - Y_{00}|r=0)]
\end{aligned}
$$

Here, we use the control group to infer the time effect of treated group. Then we can figure out the treatment effects.

Under the **assumption**: $E(Y_{01} - Y_{00}|r=1) = E(Y_{01} - Y_{00}|r=0)$, which means **the control group shares the same time trend with treated group**, DID is the interested ATT ($E(Y_{11} - Y_{01}|r=1)$).

Empirical Example

Card (1990) studies the effect of immigration on unemployment of Miami.

Atlanta, Houston, Los Angeles and Tampa are chosen as control group.

	1979	1981	1981-1979 (SD)
Miami	8.3	9.6	1.3 (2.5)
Control	10.3	12.6	2.3 (1.2)
DID		-1.0 (2.8)	

With a BA design, the treatment effected is 1.3. But two years are too long to assume no time trend. After controlling for time effect, DID gives an estimation of -1.0 which has a different sign from BA.

Difference in Difference

Panel Data with movers: for example, an household can be observed in both periods. But there exist some movers between two groups (from Beijing to Tianjin and vice versa)

Cross-sectional Data: a house can only be observed when it's sold in this period. So most houses only appear once in data.

Thus, $(r = 1, t = 1)$ and $(r = 1, t = 0)$ are different groups: they have nearly no overlap in cross-sectional data, and they only overlap on non-movers in panel data.

Like that in compliers, we define $r = r_1 \cdot t + r_0 \cdot (1 - t)$, where r_1 and r_0 are one's potential belongings in period 1 and period 0. For non-movers, r_1=r_0=r.

Difference in Difference

Panel Data with movers or Cross-sectional Data:

Kang et al. (2005) assumptions:

(i) t is mean-independent of Y_{11}, Y_{01} and Y_{00} given r_1 and r_0.

(ii) $E(Y_{0c}|r_1 = d) = E(Y_{0c}|r_0 = d)$ for any $c, d \in \{0,1\}$.

Assumption (i) is for same time trend as that in non-movers.

Assumption (ii) means that the treated group in period 1 and period 0 are identical in expectation of Y_{0c}. So the movers are randomly selected, and don't bring any bias for DID (no self-selection for move or observed).

In the panel data example: If people who like the policy move to Beijing. In the cross-section example of house: If after the policy, people are less likely to sell houses with a high unobserved ε. Then these violate assumption (ii).

Difference in Difference

Panel Data with movers or Cross-sectional Data:

$$
\begin{aligned}
DID = {} & [E(Y|r=1, t=1) - E(Y|r=1, t=0)] \\
& - [E(Y|r=0, t=1) - E(Y|r=0, t=0)] \\
= {} & E(Y_{11}|r_1 = 1, t=1) - E(Y_{00}|r_0 = 1, t=0) \\
& - E(Y_{01}|r_1 = 0, t=1) + E(Y_{00}|r_0 = 0, t=0) \\
= {} & E(Y_{11}|r_1 = 1) - E(Y_{00}|r_0 = 1) - E(Y_{01}|r_1 = 0) \\
& + E(Y_{00}|r_0 = 0) \\
= {} & [E(Y_{11}|r_1 = 1) - E(Y_{01}|r_1 = 1)] \\
& + [E(Y_{01}|r_1 = 1) - E(Y_{01}|r_1 = 0)] + [E(Y_{00}|r_0 = 0) \\
& - E(Y_{00}|r_0 = 1)] \\
= {} & E(Y_{11}|r_1 = 1) - E(Y_{01}|r_1 = 1) = ATT
\end{aligned}
$$

Estimation of DID

$$\widehat{DID} = \left[\bar{Y}_{t_1,treated} - \bar{Y}_{t_0,treated}\right] - \left[\bar{Y}_{t_1,control} - \bar{Y}_{t_0,control}\right]$$

Linear form: $y = x'\beta + \beta_t t + \beta_r r + \delta\tau + u$; by OLS $\Rightarrow \hat{\delta}$

$\hat{\delta} = \widehat{DID}$. OLS or fixed effects can be used to estimate DID.

Empirical Example

This is a study about a Seoul’s (a city in Korea) labor policy on working rate during 2000 to 2001.

	r=0, t=1	r=1, t=1	sum
r=0, t=0	2341	132	2473
r=1, t=0	146	515	661
sum	2487	647	3134

As shown in the chart, there’re 278 movers (9% of whole).

132+146=278 movers

If we only use non-movers’ data, the estimation is -4.9% (-1.57).

If we use all data, the estimation is -5.8% (-1.90).

Triple difference

Consider we want to study the effects of a policy applied to college students in Beijing. So, college students in Beijing are treated group.

Now, we want to find a control group. If we use college students in Tianjin as control group, there’s a bias of time-varying region effects (a policy applied to all students in Beijing). If we use junior college students in Beijing as control group, there’s a bias of time-varying college effects (a policy applied to college students everywhere).

Triple difference provides a way to use both as control groups to eliminate such biases.

Triple difference

$r: region.$ $r = 1$ if in Beijing.

$q: qualification.$ $q = 1$ for college students.

$t: time.$ $t = 1$ if after the policy.

$\tau: treatment.$ Only applied to the qualified individuals in region 1 in period 1. So $\tau = r \cdot q \cdot t$

$y = x'\beta + \beta_t t + \beta_r r + \beta_g \overset{q}{g} + \beta_{rt} r \cdot t + \beta_{qt} q \cdot t + \delta\tau + u$+Brq

β_{rt}: time-varying region effect

β_{qt}: time-varying qualification effect

δ: treatment effect

By OLS or triple difference, we can attain the treatment effect.

Empirical Example

Gruber (1994) studies the effect of mandated maternity benefits on log wage.

Because of the policy, it's costly to hire women of child-bearing age. So it may lower the wage of these people.

The states adopt this policy is r=1 (e.g., New York), while some other states are control group (r=0).

Married women of child-bearing age are qualified people (q=1), while males are unqualified people (q=0).

With TD, we control for other policies which may affect all people in New York.

Empirical Example

		t=0	t=1	time-difference
women	New York	1.547	1.513	-0.034
	other states	1.369	1.397	0.028
	region diff.	0.178	1.116	
	DID for women			-0.062 (0.022)
men	New York	1.759	1.748	-0.011
	other states	1.630	1.627	-0.003
	region diff.	0.129	0.121	
	DID for men			-0.008 (0.014)
	TD		-0.054 (0.026)	

The DID estimation for women is -6.2%.

The TD estimation is -5.4%.

Selection on Unobservables

In above chapters, we talk about selection on unobservables.

However, if the treated group and the control group are different in unobserved variables ε which affect both the treatment D and the response Y_d, there exists hidden bias (or selection on unobservables).

Selection on unobservables:

$$f(Y_d|D,X) \neq f(Y_d|X) \text{ and } f(Y_d|D,X,\varepsilon) = f(Y_d|X,\varepsilon)$$

In consume preferences, it often occurs that people choose the better one, so $D_i = 1[Y_{1i} > Y_{0i}]$. So the treated group tends to have a large U_1 and a lower U_0 (U_d is the error term of Y_d). This is a violation of selection on observables.

Check for hidden bias

It's possible to check coherence with multiple responses.

Thun et al. (1997) studies the effect of drinking habits on death rates with a large data set of N=490,000.

	0	<1	1	2-3	>3
cirrhosis, alcoholism	5.0 (0.9)	4.3 (0.9)	7.7 (1.9)	10.4 (1.9)	23.9 (4.5)
injuries, external causes	22.7 (1.9)	25.5 (2.2)	17.7 (2.8)	18.9 (2.7)	17.1 (4.0)

The cirrhosis and alcoholism row shows that the death rate increases as more alcohol is consumed.

But with extra response variable (injuries), we can detect hidden bias due to the unobservables consisting of alertness and carefulness.

traditional IV estimation

Under homogeneity, if we have a exogenous instrument Z:

$$Y = X\beta + \delta D + U \text{ where } E(U|X, Z) = 0$$

Rewrite the regression as:

$$Y = X\beta + \delta \cdot E(D|X, Z) + [Y - E(Y|X, Z)]$$

Now the new error term $[Y - E(Y|X, Z)]$ is mean-independent of X, Z.
So we can have a 2SLS or Wald Estimator (when Z is binary):

$$\hat{\delta} = \frac{\mathrm{Cov}(Y, Z|X)}{Cov(D, Z|X)} = \frac{E(Y|X, Z = 1) - E(Y|X, Z = 0)}{E(D|X, Z = 1) - E(D|X, Z = 0)}$$

IV estimation

Under essential heterogeneity: Consider the return of college education, it differs across individuals. Let ε_i to be ability, and β_i for treatment effects. It's reasonable to think that ε_i and β_i are correlated. Moreover, ability may affect whether you go to college (the treatment).

$$Y_i = \beta_i D_i + g(X_i) + U_i$$
$$Y_i = \beta D_i + g(X_i) + (U_i + (\beta_i - \beta)D_i)$$

Because D is in the error term, any Z correlated with D is also correlated with the error term. The traditional IV estimation doesn't work.

IV estimation

Model with a binary instrument Z:

$$Y = D \cdot Y_1 + (1 - D) \cdot Y_0$$
$$D = Z \cdot D_1 + (1 - Z) \cdot D_0$$

Consider that Z is randomly assigned (e.g., scholarship), D_z is one's potential choice for college education if he receives Z=z.

Angrist, Imbens (1995) defines four types:

(1) **always takers** with D_1=D_0=1, who always take the treatment whatever Z is.

(2) **never takers** with D_1=D_0=0, who never take the treatment.

(3) **compliers** with D_1=1 D_0=0, who comply what Z is (D=Z).

(4) **defiers** with D_1=0 D_0=1, the opposite of compliers.

IV estimation

Table 1	Z=0	Z=1
D=0	never takers, compliers	never takers, defiers
D=1	defiers, always takers	compliers, always takers

Table 2	Z=0	Z=1
D=0	never takers, compliers	never takers
D=1	always takers	compliers, always takers

In Table 1 is the four defined types. But we can't identify their proportions with data.

Table 2 is the case without defiers. In this case, we can identify the proportion of each type.

$$P(always\ takers) = P(D = 1|Z = 0)$$

$$P(never\ takers) = P(D = 0|Z = 1)$$

IV estimation with heterogeneity

Abadie et al. (2002) Assumptions:

- **INDEPENDENCE**: (Y_1, Y_0, D_1, D_0) is jointly independence of Z given X.
- NONTRIVIAL ASSUMPTION: $P(Z = 1|X) \in (0,1)$.
- FIRST-STAGE: $E(D_1|X) \neq E(D_0|X)$. (existence of compliers)
- **MONOTONICITY**: $P(D_1 \geq D_0 \; |X) = 1$. (no defiers)

Theorem 1: Wald estimator is the average treatment effects of compliers: $\frac{E(Y|X,Z=1)-E(Y|X,Z=0)}{E(D|X,Z=1)-E(D|X,Z=0)} = E(Y_1 - Y_0|X, D_1 > D_0)$

Theorem 2: $\frac{E(\kappa \cdot h(Y,D,X)|X)}{P(D_1>D_0|X)} = E(h(Y,D,X)|X, D_1 > D_0)$

IV estimation with heterogeneity

Theorem 1: $\frac{E(Y|X,Z=1)-E(Y|X,Z=0)}{E(Y|X,Z=1)-E(Y|X,Z=0)} = E(Y_1 - Y_0|X, D_1 > D_0)$

Proof: (X is always given)

$$E(Y|Z=1)$$
$$= E(Y_1|D_1 = D_0 = 1, Z = 1) \cdot P(D_1 = D_0 = 1|Z = 1)$$
$$+ E(Y_1|D_1 > D_0, Z = 1) \cdot P(D_1 > D_0|Z = 1)$$
$$+ E(Y_0|D_1 = D_0 = 0, Z = 1) \cdot P(D_1 = D_0 = 0|Z = 1)$$

By INDEPENDENCE,

$$E(Y|Z=1)$$
$$= E(Y_1|D_1 = D_0 = 1) \cdot P(D_1 = D_0 = 1) + E(Y_1|D_1 > D_0\ )$$
$$\cdot P(D_1 > D_0) + E(Y_0|D_1 = D_0 = 0) \cdot P(D_1 = D_0 = 0)$$

$$E(Y|Z=0)$$
$$= E(Y_1|D_1 = D_0 = 1) \cdot P(D_1 = D_0 = 1) + E(Y_0|D_1 > D_0\ )$$
$$\cdot P(D_1 > D_0) + E(Y_0|D_1 = D_0 = 0) \cdot P(D_1 = D_0 = 0)$$

IV estimation with heterogeneity

$$\begin{aligned}
&E(Y|Z=1)\\
&\qquad = E(Y_1|D_1=D_0=1)\cdot P(D_1=D_0=1) + E(Y_1|D_1>D_0)\\
&\qquad \cdot P(D_1>D_0) + E(Y_0|D_1=D_0=0)\cdot P(D_1=D_0=0)\\
&E(Y|Z=0)\\
&\qquad = E(Y_1|D_1=D_0=1)\cdot P(D_1=D_0=1) + E(Y_0|D_1>D_0)\\
&\qquad \cdot P(D_1>D_0) + E(Y_0|D_1=D_0=0)\cdot P(D_1=D_0=0)
\end{aligned}$$

$$E(Y|Z=1)-E(Y|Z=0) = E(Y_1-Y_0|D_1>D_0)\cdot P(D_1>D_0)$$

Similarly,

$$E(D|Z=1)-E(D|Z=0) = P(D_1=1)-P(D_0=1) = P(D_1>D_0)$$

Thus,

$$\frac{E(Y|X,Z=1)-E(Y|X,Z=0)}{E(Y|X,Z=1)-E(Y|X,Z=0)} = E(Y_1-Y_0|X,D_1>D_0)$$

Local Average Treatment Effects

$E(Y_1 - Y_0|X, D_1 > D_0)$ is the ATE only for compliers. So it's called LATE.

For always takers, they always take the treatment. So we only have information of Y_1|always. Without further assumptions, we have no idea about Y_0|always (counter-factual). So we cannot estimate their treatment effects. Same reason for never takers.

However, LATE has a shortcoming. Because the definition of compliers depends on Z, LATE also depends on Z. With different choices of Z, we can have different LATEs for different groups of compliers. For example, to study return of college, we have many IVs, like parents' education, tuition fee, distance from home, number of siblings.

This problem is partly solved in Marginal Treatment Effects.

IV estimation with heterogeneity

Theorem 2: $\frac{E(\kappa \cdot h(Y,D,X)|X)}{P(D_1 > D_0|X)} = E(h(Y,D,X)|X, D_1 > D_0)$ where

$$\kappa = 1 - \frac{D \cdot (1-Z)}{1-P(Z=1)} - \frac{(1-D) \cdot Z}{P(Z=1)}$$

Because $P(D_1 > D_0|X) = E(D|X, Z = 1) - E(D|X, Z = 0)$, the left hand side can be estimated.

$$E(Y_1 - Y_0|X, D_1 > D_0) = E\left(\frac{D \cdot Y}{P(Z = 1|X)} - \frac{(1 - D) \cdot Y}{P(Z = 0|X)}\middle| X, D_1 > D_0\right)$$

Define $h(Y, D, X) = \frac{D \cdot Y}{P(Z=1|X)} - \frac{(1-D) \cdot Y}{P(Z=0|X)}$, we can solve ATE.

IV estimation with heterogeneity

Theorem 2: $\frac{E(\kappa \cdot h(Y,D,X)|X)}{P(D_1 > D_0|X)} = E(h(Y,D,X)|X, D_1 > D_0)$ where

$$\kappa = 1 - \frac{D \cdot (1-Z)}{1-P(Z=1)} - \frac{(1-D) \cdot Z}{P(Z=1)}$$

Because

$$Q_\tau(Y_1|X, D_1 > D_0) = arg \min_{g_1(x)} E(\rho_\tau(Y - g_1(x)) \cdot D|X, D_1 > D_0)$$

$$Q_\tau(Y_0|X, D_1 > D_0) = arg \min_{g_0(x)} E(\rho_\tau(Y - g_0(x)) \cdot (1-D)|X, D_1 > D_0)$$

Define $h_1(Y,D,X) = \rho_\tau(Y - g_1(x)) \cdot D$ and $h_0(Y,D,X) = \rho_\tau(Y - g_0(x)) \cdot (1-D)$, we can solve Quantile Treatment Effects.

stata command for IV quantile treatment effects with a binary instrument

database: card.dta

dependent variable (Y): lwage

treatment variable (D): college

instrumental variables (Z): nearc4

control variables (X): black, experience, mother’s education and region

remark: the compliers are those who go to college if and only if living near a college.

command:

```
ivqte lwage (college=nearc4), quantile(0.5) variance dummy(black) continuous(exper motheduc) unordered(region) aai
```

result: college return is 69%.

Empirical Example

Abadie, Angrist, Imbens (2003) studies the effect of a job-training program on wage. So D=1 if taking the program.

They use the qualification of this program as a binary instrument Z. Z=1 if qualified for the program.

Without qualification, one can’t take the program. So D_0=0. In other words, there’re no “always takers” or “defiers”.

Remark: in general case, MONOTONICITY can’t be tested.

By assignment of Z, they divide the whole sample into “treatment” and “control”.

If Z is randomly assigned, “treatment” and “control” group should be balanced in all aspects.

Empirical Example

说明是随机的分配的，支持IV

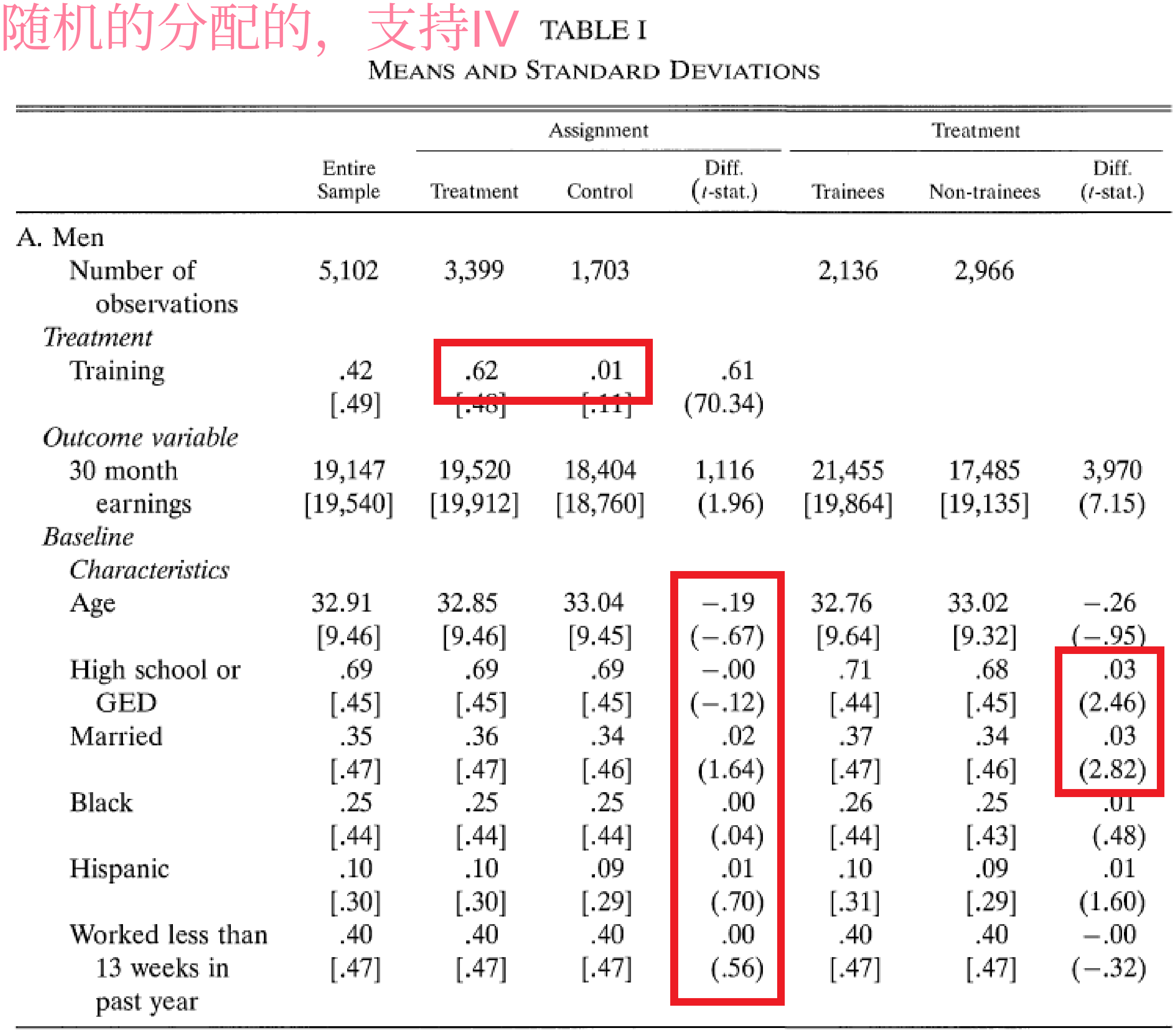


TABLE I

MEANS AND STANDARD DEVIATIONS

	Entire Sample	Assignment: Treatment	Assignment: Control	Assignment: Diff. (*t*-stat.)	Treatment: Trainees	Treatment: Non-trainees	Treatment: Diff. (*t*-stat.)
A. Men							
Number of observations	5,102	3,399	1,703		2,136	2,966	
Treatment							
Training	.42 [.49]	.62 [.48]	.01 [.11]	.61 (70.34)			
Outcome variable							
30 month earnings	19,147 [19,540]	19,520 [19,912]	18,404 [18,760]	1,116 (1.96)	21,455 [19,864]	17,485 [19,135]	3,970 (7.15)
Baseline Characteristics							
Age	32.91 [9.46]	32.85 [9.46]	33.04 [9.45]	−.19 (−.67)	32.76 [9.64]	33.02 [9.32]	−.26 (−.95)
High school or GED	.69 [.45]	.69 [.45]	.69 [.45]	−.00 (−.12)	.71 [.44]	.68 [.45]	.03 (2.46)
Married	.35 [.47]	.36 [.47]	.34 [.46]	.02 (1.64)	.37 [.47]	.34 [.46]	.03 (2.82)
Black	.25 [.44]	.25 [.44]	.25 [.44]	.00 (.04)	.26 [.44]	.25 [.43]	.01 (.48)
Hispanic	.10 [.30]	.10 [.30]	.09 [.29]	.01 (.70)	.10 [.31]	.09 [.29]	.01 (1.60)
Worked less than 13 weeks in past year	.40 [.47]	.40 [.47]	.40 [.47]	.00 (.56)	.40 [.47]	.40 [.47]	−.00 (−.32)

Empirical Example

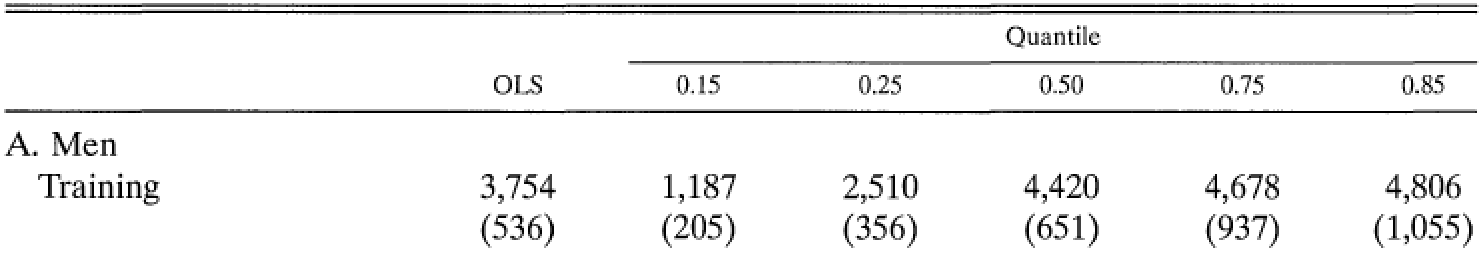
From the above table, we see some facts:

a) 62% of treatment takes the program and nearly none of control takes the program. So in this case, 62% are compliers and 38% are never takers.

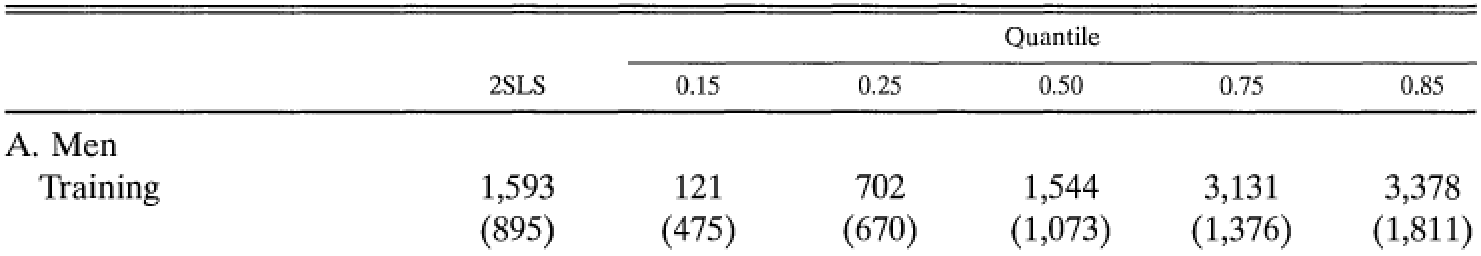
b) education and married are unbalanced between D=1 and D=0. This is a signal of hidden bias, like resulted from ability.

c) observed characteristics (age, education, etc.) are balanced between Z=1 and Z=0. The instrument (qualification) is randomly assigned to individuals.

Empirical Example



	OLS	Quantile				
		0.15	0.25	0.50	0.75	0.85
A. Men						
Training	3,754	1,187	2,510	4,420	4,678	4,806
	(536)	(205)	(356)	(651)	(937)	(1,055)



	2SLS	Quantile				
		0.15	0.25	0.50	0.75	0.85
A. Men						
Training	1,593	121	702	1,544	3,131	3,378
	(895)	(475)	(670)	(1,073)	(1,376)	(1,811)

OLS and Quantiles without IV are upward biased.

Quantile Treatment Effects vary much, and provide more information than mean treatment effects.

Marginal Treatment Effect

- Model and Definition
- Relation between MTE and other TEs
- Misspecification of P(Z)
- Violation of Monotonicity
  - omitting some IV
  - random coefficients

Marginal Treatment Effect

$$Y = Y_1 \cdot D + Y_0 \cdot (1 - D) \text{ and } D = 1[G(Z) > V]$$

Consider D is employment. It sounds reasonable to assume that you receive a job if the offered wage is larger than your reserved wage. So $D = 1[w_o > w_r]$.

$w_o(X)$ is the offered wage from your employer, affected by your characteristics (edu, exper, etc.) and market environment.

$w_r=w_r(X,Z_0)+V=w_r(Z)+V$ is your reserved wage, affected by your observed characteristics X and some unobserved preference (Z_0 and V). For example, birth season may reflect some of your personality. V is your preference for work (someone may be work holism).

Here V is assumed to be continuous and independent of Z.

Then we can rewrite $w_0(X) > w_r(Z) + V$ as $G(Z) > V$.

Marginal Treatment Effect

$$Y = Y_1 \cdot D + Y_0 \cdot (1 - D) \text{ and } D = 1[G(Z) > V]$$

$$Y_1 = \mu_1(X, U_1) \text{ and } Y_0 = \mu_0(X, U_0)$$

Heckman et al. (2006) assumptions:

1. (U_0 , U_1 , V) are independent of Z conditional on X (INDEPENDENCE)
2. The distribution of G(Z) conditional on X is nondegenerate (RANK)
3. The distribution of V is continuous.

Vytlacil (2002) proves that such selection structure is equivalent to the assumptions of LATE.

Marginal Treatment Effect

$Y = Y_1 \cdot D + Y_0 \cdot (1 - D)$ and $D = 1[G(Z) > V]$

$G(Z) > V \Leftrightarrow F_V\big(G(Z)\big) > F_V(V)$

$F_V(\cdot)$ is the distribution function of V.

So $F_V(V) \sim Unif(0,1)$

Rewrite it as $D = 1[P(Z) > U_D]$

Because of independence, $U_D | Z \sim U(0,1)$.

So $E(D|Z) = P(Z)$ (propensity score).

Define $MTE(x, u_D) = E(Y_1 - Y_0 | X = x, U_D = u_D)$.

Relation between MTE and LATE

$$MTE(x, u_D) = E(Y_1 - Y_0 | X = x, U_D = u_D)$$

If Z is binary, because $D = 1[P(Z) > U_D]$,
$D = 1$ if $Z = 1 \Leftrightarrow P(Z = 1) > U_D$
$D = 0$ if $Z = 0 \Leftrightarrow P(Z = 0) \leq U_D$.
So compliers are those with $P(Z = 1) > U_D \geq P(Z = 0)$,
always takers are those with $U_D \geq P(Z = 1)$,
and never takers are those with $P(Z = 0) > U_D$.

$$LATE(x, p, p') = E(Y_1 - Y_0 | X = x, p > U_D > p')$$
$$= \frac{1}{p - p'} \int_{p\prime}^{p} MTE(x, u_D)\, du_D$$

Relation between MTE and other TEs

Given X, Z is a random variable independent of U_D.

So, P(Z) is also a random variable.

$f(p|x)$ is its conditional PDF.

$$ATE(x) = \int_0^1 MTE(x, u_D)\, du_D$$

$$ATT(x) = \frac{1}{E(P|X = x)} \int_0^1 MTE(x, u_D) \int_{u_D}^1 f(p|x)\, dp\, du_D$$

(ATT are for those with $D = 1 \Leftrightarrow P(Z) > U_D$)

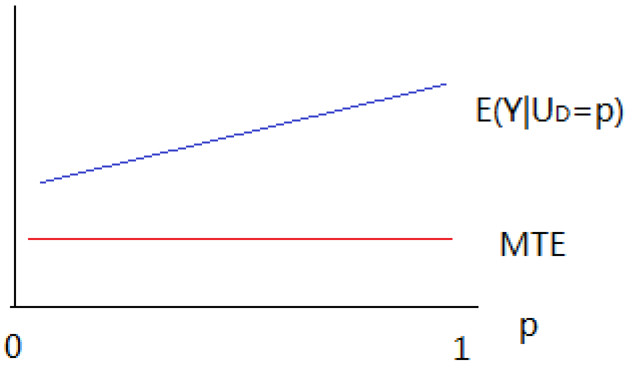
Identification of MTE

Conditional on X,

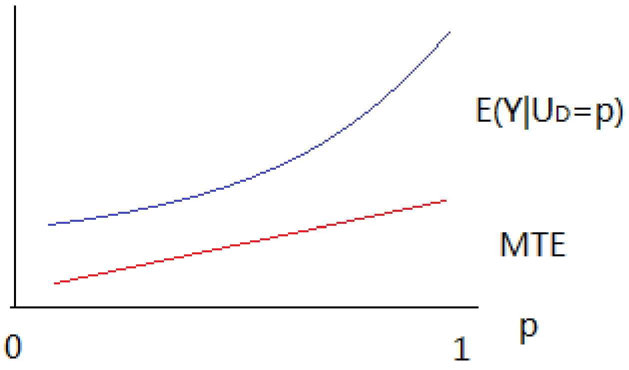
$$\begin{aligned} E(Y|P(Z)=p) &= E(Y_0 + (Y_1 - Y_0)\cdot D|P(Z)=p) \\ &= E(Y_0) + P(D=1|P=p)\cdot E(Y_1 - Y_0|D=1, P=p) \\ &= E(Y_0) + p\cdot E(Y_1 - Y_0|p > U_D) \\ &= E(Y_0) + \int_0^p MTE(u_D)du_D \end{aligned}$$

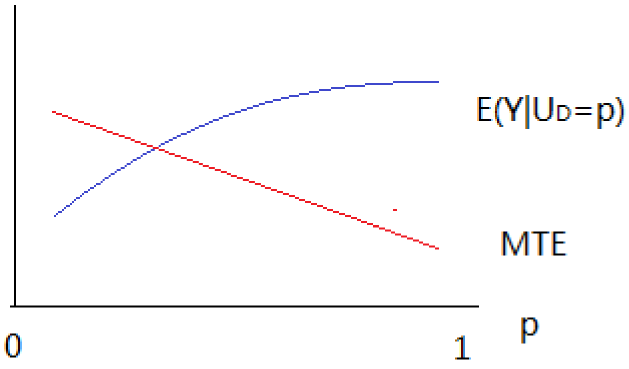
$$\therefore\ \frac{\partial E(Y|X=x, P(Z)=p)}{\partial p} = MTE(x,p)$$

Some patterns



constant treatment effects





Misspecification of P(Z)

In empirical study, we may run probit for E(D|Z).

But the estimated J(Z) may not be the true P(Z).

$$\widehat{LATE} = \frac{Cov(Y,Z)}{Cov(D,Z)} = \frac{\int_0^1 w_{IV} \cdot MTE(x, u_D)\, du_D}{\int_0^1 w_{IV}\, du_D}$$

where $w_{IV} = E(J(Z) - E(J(Z))|P(Z) > u_D) \cdot Pr(P(Z) > u_D)$

If $J(Z) = P(Z), E(J(Z)|P(Z) > u_D)$ is weakly increasing in u_D. *So* w_{IV} is positive.

If $J(Z) \neq P(Z), w_{IV}$ may be negative somewhere.

This may be one reason why IV estimation sometimes gives a wrong sign.

Violation of Monotonicity

Monotonicity requires that $P(Z_i) > P(Z_j)$ for all $Z_i = z$ and $Z_j = z'$.

Omitting some instruments:

Assume $P(Z) = 0.5 + 0.1Z_1 - 0.2Z_2$ and $(Z_1, Z_2) \in \{(0,0), (1,0), (1,1)\}$.

Now we only use Z_1 as instrument, and omit Z_2.

$P(Z_1 = 0) = 0.5$ and $P(Z_1 = 1) = 0.6 \; or \; 0.4$.

Whatever you order $P(Z_1 = 0)$ and $P(Z_1 = 1)$, there're some defiers.

Violation of Monotonicity

Random-Coefficients:

Consider parents' education as binary IVs. So $Z_1, Z_2 \in \{0,1\}$.

There're two kind of families. For father-families, $P(Z) = 0.2 + 0.4Z_1 + 0.2Z_2$. And for mother-families, $P(Z) = 0.2 + 0.2Z_1 + 0.4Z_2$.

In other words, β is a random coefficient in $P(Z) = Z\beta$.

As a result, $P(Z_1 = 1, Z_2 = 0) = 0.6 \; or \; 0.4$. Same to $P(Z_1 = 0, Z_2 = 1)$.

Still, whatever you order $P(Z_1 = 1, Z_2 = 0)$ and $P(Z_1 = 0, Z_2 = 1)$, there exist some defiers.

stata command for marginal treatment effects

database: margte_dgps (a simulated one)

dependent variable (Y): *lwage*

treatment variable (D): *enroll*

instrumental variables (Z): distance to college (*distCol*)

control variables (X): experience and mother’s education

command:

```
margte lwage exp exp2 momsEdu, treatment(enroll momsEdu distCol) first
```

result: average college return is 33%. MTEs vary from -60% to 120%.