

Programme Evaluation

Hidehiko Ichimura

Department of Economics

University College London

CeMMaP, IFS

1 Objectives of the course

1. Explain what the programme evaluation problem is.
2. Explain what the problems with the naive approach to programme evaluation are.
3. Review what the standard approaches to avoiding the problems are.
4. Review what the strength and weakness of each of the standard approaches are.
5. Explain how to implement each of the methods.

2 Tools used in this course

1. Idea of probability model.
2. Idea of conditional mean function and iterated expectation.
3. Idea of simulation.
4. Idea of linear regression model.

3 What is programme evaluation problem?

- i chooses $D_i = 1$ or 0
- Two potential outcomes Y_{0i} and Y_{1i} for any unit of observation i
- only one of them is observable for any unit of observation. That is, for any i we observe

$$Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i}.$$

- Let X_i denote the individual characteristics. We assume we observe (Y_i, D_i, X_i) .
- From this data we wish to study the impact of the program $Y_{1i} - Y_{0i}$.

Since we do not observe Y_{1i} and Y_{0i} simultaneously for anyone, $Y_{1i} - Y_{0i}$ is not known. There are many conditions studied in the literature that identify something about the distribution of $Y_{1i} - Y_{0i}$ from observations (Y_i, D_i, X_i) . This is the topic we discuss.

The approaches we discuss are

1. (Natural) Experiments
2. Matching approach
3. Difference in differences approach
4. Bloom's method
5. Exploiting Regression Discontinuity
6. Sample selection approach
7. IV: Constant treatment effect assumption or LATE
8. Bounds approach

Approaches 1–7 seeks to construct point estimates. Approach 8 gives up on a point estimate and attempt to clarify the informational content available within data.

4 Examples

The framework includes many interesting cases. The following are some examples:

- Training program on earnings or wages
- College attendance on labor force participation, earnings, or wages
- Union participation on earnings or wages
- Labor force participation on wages or earnings
- Number of children on labor force participation
- Whether to issue a stock or not on the value of firm.

5 Four parameters of interest

Since we do not observe Y_{1i} and Y_{0i} simultaneously for anyone, any parameter that depends on the joint distribution of Y_{1i} and Y_{0i} cannot be estimated. Thus the distribution of $Y_{1i} - Y_{0i}$ cannot be estimated, for example.

Four causal parameters considered in the literature:

1. $E(Y_1 - Y_0)$ or $E(Y_1 - Y_0|X)$,
2. $E(Y_1 - Y_0|D = 1)$ or $E(Y_1 - Y_0|D = 1, X)$,
3. $E(Y_1 - Y_0|D = 0)$ or $E(Y_1 - Y_0|D = 0, X)$
4. $E(Y_1 - Y_0|D(z) \neq D(z'))$ for some z and z' .

Note that

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

and that $E(Y_0 | D = 1)$ cannot be estimated.

The parameter $E(Y_1 - Y_0)$ has the same problem as well. Note that

$$\begin{aligned} E(Y_1) \\ = E(Y_1 | D = 1) \Pr(D = 1) + E(Y_1 | D = 0) \Pr(D = 0). \end{aligned}$$

6 What's wrong with $E(Y_1|D = 1) - E(Y_0|D = 0)$?

Bias if this is to estimate $E(Y_1|D = 1) - E(Y_0|D = 1)$:

$$\begin{aligned} & E(Y_1|D = 1) - E(Y_0|D = 0) \\ = & E(Y_1|D = 1) - E(Y_0|D = 1) \\ & + E(Y_0|D = 1) - E(Y_0|D = 0) \end{aligned}$$

Bias in more detail: Let $A = S_1 \setminus S_0$, $B = S_1 \cap S_0$,
 $C = S_0 \setminus S_1$

$$\begin{aligned}
& E(Y_0|D = 1) - E(Y_0|D = 0) \\
= & E\{E(Y_0|D = 1, X) | D = 1\} \\
& - E\{E(Y_0|D = 0, X)\} \\
= & E\{E(Y_0|D = 1, X) | D = 1, X \in A\} P(X \in A) \\
& + E\{E(Y_0|D = 1, X) | D = 1, X \in B\} P(X \in B) \\
& - E\{E(Y_0|D = 0, X) | D = 0, X \in C\} P(X \in C) \\
& - E\{E(Y_0|D = 0, X) | D = 0, X \in B\} P(X \in B) \\
= & E\{E(Y_0|D = 1, X) | D = 1, X \in A\} P(X \in A) \\
& - E\{E(Y_0|D = 0, X) | D = 0, X \in C\} P(X \in C) \\
& + \int_{X \in B} E(Y_0|D = 0, x) [f_X(x|D = 1) \\
& - f_X(x|D = 0)] dx \\
& + E\{E(Y_0|D = 1, X) \\
& - E(Y_0|D = 0, X) | D = 1, X \in B\} P(X \in B)
\end{aligned}$$

7 Job Training Programme Example

Heckman, Ichimura, Smith, Todd (1996) Fig 1, Table 1, Table 2.

8 Numerical Exercise

Model

$$Y_0 = 1 + X + X^2 + \varepsilon_0$$

$$Y_1 = 2 + X + X^2 + \varepsilon_1$$

$$D = 1(1 + X + u > 0)$$

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

9 Experiments

The ideal solution to this lack of data problem is to create the data we do not have.

9.1 Examples

- JTPA training program
- RAND health insurance study
- Tennessee STAR experiment

The identification condition is that the population under experiment is the same with the real situation we are interested in so that, denoting expectation under an experiment by E^* ,

$$E(Y_{1i} - Y_{0i} | D_i = 1, X_i) = E^*(Y_{1i} - Y_{0i} | D_i = 1, X_i).$$

9.2 Advantages and Disadvantages

Advantages:

- Some aspects of the probability model that generates data are under our control
 - One may be able to design an experiment that isolates the issue at hand
 - The same data are collected and individuals are placed in the same environment or differences are random between the two comparison group.
- A variation used to estimate the parameter of interest corresponds exactly to the population relationship which defines the parameter.

- “Doubly nonparametric” approach

distribution free and applicable under wide class of economic model.

Disadvantages

- Invariance of the parameter requires a separate analysis.
 - Extrapolation to other contexts may be difficult.
 - Assumptions justifying the procedure rarely explicit.
 - Costly and may not be timely
 - An experiment may not extend to reality
1. (a) Control groups may not comply with the experiments and may seek an alternative treatment or drop out.

- (b) Time horizon may be different so that participants may be different and behavior may be different for the same person.
- (c) Knowing that it is an experiment may change the behavior of participants.
- (d) The parameter we can recover under experiment may not correspond to the parameter we want to use in an application. For example, we may be able to recover an aspect of risk aversion parameter in an experiment. However, this parameter may not correspond to a parameter that plays an important role in say buying stocks because the levels of investment involved and the types of uncertainties a person faces may be rather different under experiment and reality. Another example is a difference between partial equilibrium results and a general equilibrium results.

- Not always available.

Many empirical questions are not covered.

10 Natural Experiment

Next best is to find something which is analogous to experiment.

10.1 Examples

- Use of twins to evaluate the impact of having additional child on labour force participation.
- Use of Mariel Boatlift to evaluate the impact of an increase in labour to wages and unemployment.

10.2 Advantages and Disadvantages

Analogous to experimental case but more restrictive as we cannot control most of the design.

10.3 Implementation of Natural Experimental approach

Regression analysis.

11 Matching

Identification condition is that for some variables X ,

$$E(Y_0|D = 1, X) = E(Y_0|D = 0, X)$$

holds. Note that when $D = 1 \{X'\theta + u \geq 0\}$, the condition requires that u and the residual term ε_0 are independent. Thus the condition does not allow sample selection on unobservables. In this sense the model is sometimes called selection on observables.

The condition can be used as follows to identify $E(Y_0|D = 1)$:

$$\begin{aligned} & E(Y_0|D = 1) \\ &= E[E(Y_0|D = 1, X) | D = 1] \\ &= E[E(Y_0|D = 0, X) | D = 1] \\ &= \int_{\Omega_1} E(Y_0|D = 0, X) f(X|D = 1) dX. \end{aligned}$$

Here Ω_1 denotes the support of $f(X|D = 1)$. Note that the expression requires that $E(Y_0|D = 0, X)$ to be defined over all points in Ω_1 . But that may not be.

By definition, $E(Y_0|D = 0, X)$ measurable with respect to the distribution of X given $D = 0$, but not necessarily with respect to X given $D = 1$. The last equality follows when the support of $f(X|D = 1)$ and that of $f(X|D = 0)$ coincide. This is equivalent to saying that

$$0 < \Pr(D = 1|X) < 1.$$

at any point of the support of X .

To see this note that

$$\Pr(D = 1|X) f(X) = f(X|D = 1) P(D = 1)$$

and that

$$\Pr(D = 0|X) f(X) = f(X|D = 0) P(D = 0)$$

When the support of the two are the same

$$0 < \Pr(D = 1|X) < 1$$

holds and vice versa.

As this condition may be too stringent, we may not be able to recover $E(Y_0|D = 1)$. In general, what we can identify is $E(Y_0|D = 1, X \in S)$ where S is a subset of the common support of $f(X|D = 1)$ and that of $f(X|D = 0)$.

To see this

$$\begin{aligned} & E(Y_0|D = 1, X \in S) \\ &= E[E(Y_0|D = 1, X) | D = 1, X \in S] \\ &= E[E(Y_0|D = 0, X) | D = 1, X \in S] \\ &= \int_S E(Y_0|D = 0, X) f(X|D = 1, X \in S) dX. \end{aligned}$$

On S , both $E(Y_0|D = 0, X)$ and $f(X|D = 1, X \in S)$ are well defined.

To carry this out in estimation, one needs to estimate the high dimensional nonparametric estimation.

Let $P(X) = \Pr(D = 1|X)$. Rosenbaum and Rubin observed that when

$$E(Y_0|D = 1, X) = E(Y_0|D = 0, X)$$

holds it is also true that

$$E(Y_0|D = 1, P(X)) = E(Y_0|D = 0, P(X))$$

whenever $P(X) > 0$.

To see this note that,

$$\begin{aligned} & E(Y_0|D = 1, P(X)) \\ = & \frac{E(Y_0D|P(X))}{E(D|P(X))} \\ = & \frac{E[E(Y_0D|X)|P(X)]}{P(X)} \\ = & \frac{E[E(Y_0|X)P(X)|P(X)]}{P(X)} \\ = & E(Y_0|P(X)). \end{aligned}$$

This means that estimation can be carried out using the estimation of the conditional mean function on $P(X)$. The probability is sometimes called propensity score.

11.1 Examples

- Heckman, Ichimura, Todd 1997 REStud
- Dearden, Ferri, Meghir 2000
- Blundell, Costa Dias, Meghir 2001

12 Difference in differences

We assume that panel data are available and that we have data on people switching the treatment status over time. So now for the switchers, we do observe outcomes that correspond to two treatment statuses, albeit at different periods.

The time effect is controlled using nonswitchers.

Maintained assumption is that the time effect does not vary across switchers and stayers.

It is still the case that we do not observe Y_{0it} and Y_{1it} simultaneously in the same period t for anyone.

- $m_{1t}(x) = E(Y_{1t}|X = x)$

- $m_{0t}(x) = E(Y_{0t}|X = x)$

- Switchers:

$$\begin{aligned} & E(Y_{1t} - Y_{0t-1} | \text{switchers}, X = x) \\ = & m_{1t}(x) - m_{0t-1}(x) \\ & E(\varepsilon_{1t+1} - \varepsilon_{0t} | \text{switchers}, X = x) \end{aligned}$$

- Stayers:

$$\begin{aligned} & E(Y_{0t} - Y_{0t-1} | \text{stayers}, X = x) \\ = & m_{0t}(x) - m_{0t-1}(x) \\ & E(\varepsilon_{0t+1} - \varepsilon_{0t} | \text{stayers}, X = x) \end{aligned}$$

- Assume unobserved trend for switchers and stayers are the same.
- Previous assumption was that levels are the same given X .

12.1 Examples

- Ashenfelter and Krueger 1994 AER
- Chay and Greenstone 2001
- Panel fixed effect models

13 Bloom's Method and Regression Discontinuity

13.1 Examples

- Use of a test-score to give “merit scholar” status
- Use of Maimonides rule to determine class size

13.2 Three assumptions:

1. Probability distribution of D depends on z and it changes discontinuously at $z = z_0$. So that

$$\Pr(D = 1|Z = z_0^-) \neq \Pr(D = 1|Z = z_0^+).$$

2. Treatment effect varies continuously with respect to z .
3. Treatment effect and Treatment are independent near z_0 .

$$\begin{aligned} & E(Y_1 - Y_0|Z = z_0) \\ &= \frac{E(Y|Z = z_0^+) - E(Y|Z = z_0^-)}{\Pr(D = 1|Z = z_0^+) - \Pr(D = 1|Z = z_0^-)}. \end{aligned}$$

13.3 Implementation of Regression Discontinuity

How do we estimate $\Pr(D = 1|Z = z_0^+)$

or $\Pr(D = 1|Z = z_0^-)$, etc?

14 Sample selection model

Here the idea is to exploit the variable that shifts the participation decision but not outcomes.

Can be viewed as a generalization of the Bloom's method.

Let it be Z_i . The assumptions to identify $E(Y_0|D = 1, X)$ are

$$E(Y_0|X, Z) = E(Y_0|X)$$

For each X there exist Z
such that $\Pr(D = 1|X, Z) = 0$.

Under these conditions, for each X by choosing those

Z that correspond to $\Pr(D = 1|X, Z) = 0$, we have

$$\begin{aligned} & E(Y_0|X) \\ &= E(Y_0|X, Z) \\ &= E(Y_0|D = 0, X, Z) \end{aligned}$$

and that using

$$\begin{aligned} & E(Y_0|X) \\ &= E(Y_0|D = 1, X) \Pr(D = 1|X) \\ & \quad + E(Y_0|D = 0, X) \Pr(D = 0|X) \end{aligned}$$

or

$$\begin{aligned}
& E(Y_0|D = 1, X) \\
= & \frac{E(Y_0|X) - E(Y_0|D = 0, X) \Pr(D = 0|X)}{\Pr(D = 1|X)} \\
= & E(Y_0|D = 0, X) \\
& + \frac{E(Y_0|D = 0, X, Z) - E(Y_0|D = 0, X)}{\Pr(D = 1|X)}.
\end{aligned}$$

Analogously, assumptions to identify $E(Y_1|D = 0, X)$ are

$$E(Y_1|X, Z) = E(Y_1|X)$$

For each X there exist Z
such that $\Pr(D = 1|X, Z) = 1$.

Clearly the Z does not have to be the same for the two

cases. The strategy of identifying using observations at the extreme points is sometimes called identification at infinity, although the terminology is somewhat misleading. It is based on the case where

$$\Pr(D = 1|X, Z) = \Phi(X'\theta + Z'\gamma)$$

where Φ is the standard normal CDF. In this case in order for the condition to hold, literally some Z need to diverge to infinity to achieve identification.

Note that the assumption allows random treatment effect and very general selection structure.

14.1 Example

- Schafgans JAE 1998

15 Instrumental Variable Method

Basic assumption is

$$Y_{1i} - Y_{0i} = \Delta(X_i) + U_i$$

where U_i satisfies

$$E(U_i | X_i, D_i) = 0.$$

That is the treatment impact differs across observational units only via observables X_i and a random variable not systematically related with X_i or D_i .

When this holds

$$\begin{aligned} Y_i &= Y_{0i} + D_i (Y_{1i} - Y_{0i}) \\ &= Y_{0i} + D_i \Delta (X_i) + D_i U_i \end{aligned}$$

so that when we write $Y_{0i} = m_0 (X_i) + \varepsilon_{0i}$, where $m_0 (x) = E (Y_{0i} | X_i = x)$,

$$Y_i = m_0 (X_i) + D_i \Delta (X_i) + \varepsilon_{0i} + D_i U_i.$$

Note that by the assumed property of U_i ,

$$E(D_i U_i | X_i, D_i) = 0.$$

Also by the definition of $m_0(x)$, $E(\varepsilon_{0i} | X_i) = 0$. If in addition ε_{0i} and D_i are uncorrelated so that

$$E(\varepsilon_{0i} | D_i = 0, X_i = x) = E(\varepsilon_{0i} | D_i = 1, X_i = x) = 0,$$

then the difference of the nonparametric regressions of Y_i on X_i and $D_i = 1$ and $D_i = 0$ estimates $\Delta(X_i)$.

However in many cases, the assumption that the participation decision be uncorrelated with the unobserved error term in no program situation may be too restrictive.

In this case if we can find an instrumental variable Z_i that satisfies $E(\varepsilon_{0i}|X_i, Z_i) = 0$ we can still estimate $\Delta(X_i)$ provided we strengthen the assumption about U_i to

$$E(U_i|X_i, Z_i, D_i) = 0.$$

$$\begin{aligned} & E(Y_i|X_i, Z_i) \\ &= m_0(X_i) + \Pr(D_i = 1|X_i, Z_i) \Delta(X_i). \end{aligned}$$

Thus

$$= \frac{\Delta(X_i) \text{Cov}(E(Y_i|X_i, Z_i), \Pr(D_i = 1|X_i, Z_i) | X_i)}{\text{Var}(\Pr(D_i = 1|X_i, Z_i) | X_i)}.$$

Thus if there is a conditional variation in Z_i given X_i , by exploiting this variation and the assumption that Z_i influences D_i one can estimate $\Delta(X_i)$.

15.1 Limitations

1. The model implies

$$\begin{aligned} E(Y_1 - Y_0|X) &= E(Y_1 - Y_0|D = 1, X) \\ &= E(Y_1 - Y_0|D = 0, X) \end{aligned}$$

as they are all equal to $\Delta(X)$. That is, the parameters that have very different meanings are the same under the identifying assumption and hence the estimation method which is based on this assumption may be questionable.

2. Model does not allow heterogeneous impact through unobserved variables that is systematically related with D_i or X_i or Z_i , the instrumental variable.

Imbens and Angrist (1994) gave an alternative causal interpretation to the IV estimate when the instrumental variable is a binary variable.

Assuming that the IV Z_i only affects D_i we have for $Z_i = z$

$$Y_i(z) = Y_{0i} + D_i(z)(Y_{1i} - Y_{0i})$$

so that

$$Y_i(z) - Y_i(z') = (D_i(z) - D_i(z'))(Y_{1i} - Y_{0i}).$$

Assume that $D_i(z) \geq D_i(z')$ always (monotonicity).

This implies

$$\begin{aligned} & E \left[Y_i(z) - Y_i(z') \right] \\ &= E \left[Y_{1i} - Y_{0i} \mid D_i(z) \neq D_i(z') \right] \Pr \left(D_i(z) \neq D_i(z') \right) \end{aligned}$$

so that

$$\begin{aligned} & E \left[Y_{1i} - Y_{0i} \mid D_i(z) \neq D_i(z') \right] \\ &= \frac{E \left[Y_i(z) - Y_i(z') \right]}{\Pr \left(D_i(z) \neq D_i(z') \right)}. \end{aligned}$$

The denominator can be estimated, under the monotonicity assumption, by $\Pr(D_i(z) = 1) - \Pr(D_i(z') = 1)$.

15.2 Advantage

The interpretation does not depend on the constant treatment effect assumption. The result shows that the IV method estimates a causal parameter allowing for unobserved heterogeneity.

15.3 Limitations

Two of the limitations of this approach are common to all the observational methods: the IV (Z_i) needs to be independent with outcomes given X_i and needs to shift $\Pr(D_i = 1|Z_i, X_i)$ given X_i . Another limitation is its interpretability as an interesting parameter and how the interpretation generalizes when the IV takes on more than two values. The latter two limitations are resolved in Ichimura and Taber (2003) by explicitly considering a policy impact parameter and by studying an alternative estimator to IV.

16 Bounds

Manski examined bounds on $E(Y_0|X)$. Note that

$$\begin{aligned} & E(Y_0|X) \\ = & E(Y_0|D = 1, X) P(X) \\ & + E(Y_0|D = 0, X) [1 - P(X)]. \end{aligned}$$

By putting bounds on Y_0 given X is

$$Y_{0L}(X) \leq Y_0|X \leq Y_{0H}(X).$$

Then

$$\begin{aligned} & Y_{0L}(X) P(X) \\ & + E(Y_0|D = 0, X) [1 - P(X)] \\ \leq & E(Y_0|X) \\ \leq & Y_{0H}(X) P(X) \\ & + E(Y_0|D = 0, X) [1 - P(X)]. \end{aligned}$$

If $P(X)$ is close to zero, the bound is tight.