調査観察データの統計科学ゼミ-第 7 回-

Masamichi Ito

Osaka University Graduate School of Human Sciences
Adachi Lab M1

December 5, 2019

担当章

差分の差 (DID) 推定量

- 差分の差 (difference in differences,DID)推定量
  - 「介入後の結果変数の差-介入前の結果変数の差」．純粋な政策効果の推定値として利用される．
  - 心理学や教育学では，不等価 2 群事前事後デザイン (nonequivalent group pretest-posttest design または，nonequivalent comparison group design) という名前で知られている．
  - 通常は，パネル調査 (panel survey) において，介入後の「処置群と対照群の差」と，介入前の「処置群と対照群の差」の差分を指す．
  - クロスセクションデータにも適用される
- パネル調査
  - 複数時点で同一の対象を繰り返し調査する調査方法
- クロスセクションデータ
  - 複数時点で，異なる対象者に対して，調査する調査方法

差分の差 (DID) 推定量

単純に考えると，

$$\begin{aligned}\text{対照群での平均の 2 時点間の差} &= \text{時間変化による効果}_0 + \text{誤差}_1 \\ \text{処置群での平均の 2 時点間の差} &= \text{時間変化による効果}_1 \\ &\quad + \text{介入プログラムの効果} + \text{誤差}_2\end{aligned}$$

であり，

$$\begin{aligned}\text{差分の差} &= \text{" 処置群での平均の 2 時点間の差"} \\ &\quad - \text{" 対照群での平均の 2 時点間の差"} \\ &= \text{介入プログラムの効果} + \text{誤差}_3\end{aligned}$$

と考えられるが，これでいいのか？

DID 続き

差分の差を正しく理解するために，2 つのグループのうち一方を政策の対象，他方を対象外とする状況を考え，政策の前後 (時点 $t = a, b\ a < b$) でデータが得られるとする．Rubin の因果モデルを拡張して，以下を定義する．

- y_a: 政策が実施される前での結果変数の値
- y_{1b}:「もし政策の対象となる集団に所属した場合における」政策実施後の結果変数の値
- y_{0b}:「もし製作の対象となる集団に所属しない場合における」政策実施後の結果変数の値
- z_t: t 時点でのグループへの所属のインディケータ．通常は時点間で変化しない ($z_a = z_b = z$) 場合を考える．
- δ: b 時点での測定値なら，$\delta = 1$，a 時点での測定値なら，$\delta = 0$ とするインディケータ

DID 続き (表を白板に書きます．)

DID 続き

b 時点で実際に得られる従属変数 y_b は

$$y_b = zy_{1b} + (1 - z)y_{0b} \tag{3.31}$$

と表せる．また，a 時点の結果変数も考慮した，観測される従属変数 y は

$$y = \delta y_b + (1 - \delta)y_a = \delta\{zy_{1b} + (1 - z)y_{0b}\} + (1 - \delta)y_a \tag{3.32}$$

となる．これらを用いて，差分の差を再度定義する．

DID 続き

差分の差 (DID)

$$
\begin{aligned}
DID &= \{E(y|z=1,\delta=1) - E(y|z=1,\delta=0)\} \\
&\quad - \{E(y|z=0,\delta=1) - E(y|z=0,\delta=0)\} \\
&= E(y_{1b} - y_a|z=1) - E(y_{0b} - y_a|z=0)
\end{aligned}
\tag{3.33}
$$

ここで,

- $E(y_{0b} - y_a|z=0)$: 結果変数の期待値の時間による変化
- $E(y_{1b} - y_a|z=1)$ は時間と介入両方の変化

であるため，この二つの差分が介入の純粋な効果になると考えられる.

DID を TET とみなすための条件

$t = b$ 時点での"処置群での" 因果効果 (TET, treatment effect for the treated) は

$$TET = E(y_{1b} - y_{0b}|z = 1)$$

と表現される．イメージとしては，「ある介入を行った後 ($t = b$ 時点) において，政策の対象となった場合の結果変数とならなかった場合の結果変数の期待値の差」.
※ 観測されない $E(y_{0b}|z = 1)$ を含んでいるため，何かしらの仮定を置かなければ推定できないことに注意
$\rightarrow$ ある条件のもとで，DID の推定量を TET の推定量として利用できる.

DID を TET とみなすための条件

ある条件？具体的には，

$$E(y_{0b} - y_a | z = 1) = E(y_{0b} - y_a | z = 0) \tag{3.34}$$

つまり「もし政策の対象にならなかったときの経時変化が 2 つのグループ間で等しい」という条件が DID を TET とみなすための必要条件である．

なぜならば，

$$\begin{aligned} DID &= E(y_{1b} - y_a | z = 1) - E(y_{0b} - y_a | z = 0) \\ &= E(y_{1b} - y_a | z = 1) - E(y_{0b} - y_a | z = 1) \\ &\quad + E(y_{0b} - y_a | z = 1) - E(y_{0b} - y_a | z = 0) \end{aligned}$$

とかけるため，(3.34) が成立するならば，

$$DID = E(y_{1b} - y_a | z = 1) - E(y_{0b} - y_a | z = 1) = E(y_{1b} - y_{0b} | z = 1) = TET$$

となるからである．

DID を TET，TEU とみなすための条件

この条件は，無作為割り当て $E(y_a|z=1) = E(y_a|z=0) = E(y_a)$ という条件よりもゆるい！助かる！
何も介入しないければ結果変数が時間的変化を起こさない場合ならば，この条件は満たされる．

TEU と因果効果が一致するための条件は，

$$E(y_{1b} - y_a|z=1) = E(y_{1b} - y_a|z=0)$$

つまり，「もし政策を実行したときの経時変化が二つのグループ間で等しい」ことが条件となる．

DID の推定量

DID 自体は，単純に

$$\frac{1}{N_1}\sum_{i:z_i=1}^{N_1}(y_{bi}-y_{ai})-\frac{1}{N_2}\sum_{i:z_i=0}^{N_2}(y_{bi}-y_{a_i})$$

を用いて推定可能であるが，これが因果効果 (原文ママ，おそらく TET) として利用されている背景には，

$$E(y_{0b}-y_a|z=1)=E(y_{0b}-y_a|z=0) \tag{3.34}$$

が暗に仮定されている.

セミパラメトリックな差分の差推定

差分の差推定の仮定をもっと緩めたい．$\rightarrow$ 共変量 $\boldsymbol{x}$ の情報を用いることを考える．

$$TET = E(y_{1b} - y_{0b}|z = 1) = E_{\boldsymbol{x}}\left(E(y_{1b} - y_{0b}|z = 1, \boldsymbol{x})\right)$$

$$\begin{aligned} DID &= E_{\boldsymbol{x}}\left(E(y_{1b} - y_a|z = 1, \boldsymbol{x}) - E(y_{0b} - y_a|z = 0, \boldsymbol{x})\right) \\ &= E_{\boldsymbol{x}}\left(E(y_{1b} - y_a|z = 1, \boldsymbol{x}) - E(y_{0b} - y_a|z = 1, \boldsymbol{x})\right. \\ &\quad \left. + E(y_{0b} - y_a|z = 1, \boldsymbol{x}) - E(y_{0b} - y_a|z = 0, \boldsymbol{x})\right) \end{aligned} \tag{3.35}$$

から，DID と TET が一致するための条件は，

$$E(y_{0b} - y_a|z = 1, \boldsymbol{x}) = E(y_{0b} - y_a|z = 0, \boldsymbol{x}) \tag{3.36}$$

つまり，共変量を共通にしたときに，「政策の対象にならなかったときの経時変化が 2 つのグループ間で等しい」という条件である．
これは，(3.34) よりもゆるい．

セミパラメトリックな差分の差推定

ゆるいんだけども，実際には，

- $E(y_{1b}|z=1,\boldsymbol{x}), E(y_{0b}|z=0,\boldsymbol{x}), E(y_a|z=1,\boldsymbol{x}), E(y_a|z=0,\boldsymbol{x})$ の回帰関数を正しく設計する必要がある.
- (3.36) の条件が成立するためには十分な数の共変量を利用する必要がある.

$\rightarrow$ 共変量が多い場合に回帰関数を設計する必要がない方法 ($=$ セミパラメトリックな手法) を開発すべき

セミパラメトリックな差分の差推定

- Abadie(2005) のセミパラメトリックな DID 推定法
  -回帰関数を設定する代わりに傾向スコア $e = E(z = 1|\boldsymbol{x})$ を用いる.
  すると，TET は,

$$
\begin{aligned}
E(y_{1b} - y_{0b}|z = 1) &= \int E(y_{1b} - y_{0b}|z = 1, \boldsymbol{x}) p(\boldsymbol{x}|z = 1) d\boldsymbol{x} \\
&= E_{\boldsymbol{x}} \left[E(\rho(y_b - y_a)|\boldsymbol{x}) \frac{p(z = 1|\boldsymbol{x})}{p(z = 1)} \right] \\
&= E\left(\frac{y_b - y_a}{p(z = 1)} \frac{z - e}{1 - e} \right) \qquad (3.37)
\end{aligned}
$$

  ただし,

$$
\rho = \frac{z - e}{e(1 - e)}
$$

  である.

セミパラメトリックな差分の差推定

$$\rho = \begin{cases} \frac{1}{e} & (z=1) \\ -\frac{1}{1-e} & (z=0) \end{cases}$$

であることを考慮すると，(3.37) の 1 行目からの 2 行目への式変形は，

$$\begin{aligned} & E(\rho(y_b - y_a)|\boldsymbol{x}) \\ & = E(\rho(y_b - y_a)|z=1,\boldsymbol{x}) \times e + E(\rho(y_b - y_a)|z=0,\boldsymbol{x}) \times (1-e) \\ & = E(y_b - y_a|z=1,\boldsymbol{x}) - E(y_b - y_a|z=0,\boldsymbol{x}) \\ (& = E(y_b - y_a|z=1,\boldsymbol{x}) - E(y_b - y_a|z=1,\boldsymbol{x}) = E(y_{1b} - y_{0b}|z=1,\boldsymbol{x})) \end{aligned}$$

であることを利用している．

セミパラメトリックな差分の差推定

以上から，パネル調査データを利用する際には，TET を傾向スコアを用いて重み付けした下記の推定量

$$\frac{1}{N}\sum_{i=1}^{N}\frac{y_{bi}-y_{ai}}{p(z=1)}\frac{z_i-e_i}{1-e_i} \ or \ \frac{\sum_{i=1}^{N}(y_{bi}-y_{ai})\frac{z_i-e_i}{1-e_i}}{\sum_{i=1}^{N}e_i}$$

を用いて推定すれば良い.

- 実際には，傾向スコアの推定値を用いて上記の推定量を計算するが，その一致性と漸近正規性は持つ (Abadie, 2005)
- この推定量は形自体は差分の差という単純な形ではない
- しかし，パネル調査データを利用することで，(3.36) というゆるい条件のもとで TET を推定することが可能.
  (通常は，強く無視できる割り当て，平均での独立性条件が成立しなければならない)

クロスセクションデータを利用した差分の差推定量

差分の差推定の問題は，介入前後の 2 時点で同一の調査対象に対して測定するパネル調査を行う必要あるということである．
しかし....

- 一般にパネル調査では対象を追跡するのが困難
- 正確に追跡を行うためにはコストがかかる．
- 2 時点目で調査を受けない” 脱落” が多数起こりうる

→ そこで，クロスセクションデータの繰り返しによって，TET を推定することが望まれるケースもある．
クロスセクションデータを利用する場合は，介入前のデータに含まれる対象者についても「政策介入の対象となるかならないか」を知ることができる場合がある．

クロスセクションデータを利用した差分の差推定量

(テキスト p.109 の例を読む)

クロスセクションデータを利用した差分の差推定量

先ほどの例は，様々な前提条件を仮定しないと適切な解析とは言えない．
$\rightarrow$ これを理解するために，以下の表記を利用して，潜在的な結果変数を用いてモデルの記述を行う．

- δ: b 時点の調査対象であれば 1，a 時点での調査対象であれば 0 とするインディケータ
- z_a, z_b: a, b 時点のそれぞれで，処置群か，対照群かを示すインディケータ

これらを用いると，結果変数は，

$$y = \delta\{z_b y_{1b} + (1 - z_b) y_{0b}\} + (1 - \delta) y_a \tag{3.39}$$

と表すことができる．a 時点では介入の効果はないので．z_a が $0, 1$ のどちらでも y_a が得られることに注意．

白板に p.100 の図を書きます

クロスセクションデータを利用した差分の差推定量

単純な差分の差は,

$$
\begin{aligned}
DID &= E(y|\delta = 1, z_b = 1) - E(y|\delta = 0, z_b = 1) \\
&\quad - \{E(y|\delta = 1, z_b = 0) - E(y|\delta = 0, z_b = 0)\} \\
&= E(y_{1b}|\delta = 1, z_b = 1) - E(y_a|\delta = 0, z_b = 1) \\
&\quad - \{E(y_{0b}|\delta = 1, z_b = 0) - E(y_a|\delta = 0, z_b = 0)\} \qquad (3.40)
\end{aligned}
$$

の推定量

$$
\begin{aligned}
&\frac{\sum_{i=1}^N \delta_i z_{bi} y_i}{\sum_{i=1}^N \delta_i z_{bi}} \frac{\sum_{i=1}^N (1-\delta_i) z_{ai} y_i}{\sum_{i=1}^N (1-\delta_i) z_{ai}} \\
&- \left\{ \frac{\sum_{i=1}^N \delta_i (1-z_{bi}) y_i}{\sum_{i=1}^N \delta_i (1-z_{bi})} \frac{\sum_{i=1}^N (1-\delta_i)(1-z_{ai}) y_i}{\sum_{i=1}^N (1-\delta_i)(1-z_{ai})} \right\}
\end{aligned}
$$

は何も仮定を置かない場合，TET の不偏推定量ではない.

クロスセクションデータを利用した差分の差推定量

以下の特別な仮定が成立しているなら，(3.40) 式にこれを代入することで DID=TET が示せる

- 2 時点間で調査対象者は等質である

$$
\begin{aligned}
E(y_{1b}|z_b, \delta = 1) &= E(y_{1b}|z_b, \delta = 0) = E(y_{1b}|z_b) \\
E(y_{0b}|z_b, \delta = 1) &= E(y_{0b}|z_b, \delta = 0) = E(y_{0b}|z_b) \\
E(y_a|z_a, \delta = 1) &= E(y_a|z_a, \delta = 0)
\end{aligned}
$$

  赤字の部分は欠測.

- 介入しなかった場合の結果変数の変化が b 時点における処置群と対照群で等しい

$$
E(y_{0b} - y_a|z_b = 1) = E(y_{0b} - y_a|z_b = 0)
$$

- a 時点での結果変数の平均は 2 つのデータの処置群で共通であり，対照群でも共通

$$
E(y_a|z_b) = E(y_a|z_a)
$$

クロスセクションデータを利用した差分の差推定量

- 反復されたクロスセクションデータ (repeated cross-section data)
-2 つのデータが等質であるという仮定を置いたもとでクロスセクションデータを利用すること.
- 2 番目の条件はパネル調査 (調査対象者が同じ)(3.34) と基本的に同じ. 加えて,「2 時点間での等質性」「a 時点での 2 つのグループでの等質性」を仮定している.
- 2 時点間での介入の対象となる集団に違いがない場合は $z_b = z_a$ とと仮定して良い. この場合 3 番目の条件は常に成立する.
- しかし, この仮定が成立することをデータから示すのは困難

続き，共変量の利用

調査対象者について様々な共変量が利用できる場合には，先ほどの特別な仮定の期待値を $\boldsymbol{x}$ を所与とする場合に変更した条件下で，

$$\begin{aligned} DID = & E_{\boldsymbol{x}}\left[E(y_{1b}|\delta=1, z_b=1, \boldsymbol{x}) - E(y_a|\delta=0, z_b=1, \boldsymbol{x})\right. \\ & \left. - \{E(y_{0b}|\delta=1, z_b=0, \boldsymbol{x}) - E(y_a|\delta=0, z_b=0, \boldsymbol{x})\}\right] \end{aligned} \tag{3.41}$$

が TET と一致する.

- 調査対象者についての様々な情報を所与とする場合には，3 つの特別な条件はより成立しやすくなる.
- しかし，共変量に関しての回帰関数の設計が必要
  $\rightarrow$ 回帰関数を指定しないセミパラな推定量が欲しい！

Abadie(2005) のアプローチ

Abadie(2005) は$z_a = z_b = z$という条件下で，パネル調査での推定量 $((3.37))$ と同様に傾向スコアを用いて，

$$TET = E(y_{1b} - y_{0b}|z = 1) = E_M \left[\frac{e}{p(z = 1)} \phi \times y \right] \tag{3.42}$$

と表せることを示した．ただしここで，

$$\phi = \frac{\delta - \lambda}{\lambda(1 - \lambda)} \frac{z - e}{e(1 - e)}$$

であり，E_M は

$$p(y, z, \delta, \boldsymbol{x}) = \lambda\delta p(y_b = y, z, \boldsymbol{x}) + (1 - \lambda)(1 - \delta)p(y_a = y, z, \boldsymbol{x})$$

に関する期待値．

続き

(3.42) から，2 時点のクロスセクションデータを利用した TET の推定量は，

$$\frac{1}{N}\sum_{i=1}^{N}\left[\frac{e_i}{p(z=1)}\frac{\delta_i-\lambda}{\lambda(1-\lambda)}\frac{z_i-e_i}{e_i(1-e_i)}y_i\right]$$

もしくは，

$$\frac{\sum_{i=1}^{N} e_i \frac{\delta_i-\lambda}{\lambda(1-\lambda)}\frac{z_i-e_i}{e_i(1-e_i)}y_i}{\sum_{i=1}^{N} e_i}$$

であり，これは漸近正規性を持つ，一致推定量である (Abadie,2005).

続き

まとめ

以上のように，差分の差推定の議論は，

- 研究デザインを洗練させる
- データを二度取得する努力 で強く無視可能な割り当て条件よりも，ゆるい仮定のもとで，因果効果の推論が可能であることを示している.

→研究デザインと統計手法両方を洗練させた方法論 のさらなる発展の方向性を示唆!!

References

- 高井, 星野, 野間 (2016)「調査観察データ解析の実際 欠測データの統計科学-医学と社会科学への応用-」
- 星野 (2009)「調査観察データの統計科学-因果推論・選択バイアス・データ融合-」

Thank you for your attention!!

Any Questions?

付録経験尤度法

後日！すみま千円！

メモに使ってね!!