2024 年 5 月 24 日 統計数理研究所 オープンハウス

共変量バランシングを組み込んだ二重に頑健な差分の差分法

馬場 崇充　**総合研究大学院大学 統計科学専攻 博士課程 (3年次編入) 5年**

1　はじめに

因果推論において，バイアスを除去したうえでaverage treatment effect on the treated (TET) を推定する方法として，傾向スコアを用いたsemiparametric difference-in-difference (SDID) が近年利用されている(Abadie 2005)．しかしながら既存の方法では傾向スコアのモデルが正しいことが仮定されていることが多く，モデルが誤特定されると推定量はバイアスを持つ．そこで本発表では，Imai and Ratkovic (2014)によるcovariate balancing propensity score (CBPS) や，二重頑健推定の考え方をSDIDと組み合わせることで，モデルの誤設定に頑健な方法を提案する．

2　モデルと仮定

$$y(t) = d_1 y_1(t) + d_0 y_0(t), \quad \Delta = d_1\Delta_1 + d_0\Delta_0$$

- 割り当て変数 d_0, d_1 ($\in \{0,1\}$かつ$d_0 + d_1 = 1$)
- $y_h(t)$: 割付hにおける時刻tの応答 ($h, t \in \{0,1\}$)
- $\Delta_h = y_h(1) - y_h(0)$ ($h \in \{0,1\}$)

推定対象: TET に$(g(\boldsymbol{x}) =)\boldsymbol{x}'\boldsymbol{\theta}$ のような線形モデルを仮定したときの $\boldsymbol{\theta}$ の推定と，処置群のデータを用いて$\boldsymbol{x}$について期待値をとることによるTETの推定．

$$g(\boldsymbol{x}) \equiv (\text{TET conditional on } \boldsymbol{x}) = \mathrm{E}[y_1(1) - y_0(1) \mid \boldsymbol{x}, D_1 = 1]$$

線形モデルにおける $\boldsymbol{\theta}$ の最適値の定義．

$$\boldsymbol{\theta}^* \equiv \operatorname*{argmin}_{\boldsymbol{\theta}} \mathrm{E}[\{\mathrm{E}[y_1(1) - y_0(1) \mid \boldsymbol{x}, d_1 = 1] - \boldsymbol{x}'\boldsymbol{\theta}\}^2 \mid d_1 = 1]$$

いくつかの仮定の下，傾向スコアを用いたセミパラメトリックなDID (SDID) は$\boldsymbol{\theta}^*$に収束する(Abadie 2005)．

$$\hat{\boldsymbol{\theta}} = \left\{\sum_{i=1}^n e_1(\boldsymbol{x}_i)\boldsymbol{x}_i\boldsymbol{x}_i'\right\}^{-1} \sum_{i=1}^n e_1(\boldsymbol{x}_i)\boldsymbol{x}_i\rho_i\Delta_i \tag{1}$$

このとき，$e_1(\boldsymbol{x}_i) = \mathrm{P}(D_1 = 1 \mid \boldsymbol{x}_i)$，$\rho_i \equiv D_{1i}/e_1(\boldsymbol{x}_i) - D_{0i}/e_0(\boldsymbol{x}_i)$．

仮定 1 (Conditional parallel trend).

$$\mathrm{E}[y_0(1) - y_0(0) \mid \boldsymbol{x}, d_1 = 1] = \mathrm{E}[y_0(1) - y_0(0) \mid \boldsymbol{x}, d_0 = 1]$$

仮定 2 (Common shocks). 介入の応答測定と，介入後の応答測定との間に，応答に影響を与えるような「別のイベント」が起きていない，もしくは起きているとしても二群に対して同じように作用している．

仮定 3 (Positivity).　$0 < P(d_1 = 1) < 1$ かつ $0 < \mathrm{P}(d_1 = 1 \mid \boldsymbol{x}) < 1$

3　共変量バランシングの組み込みによる二重頑健化

共変量バランシング(Imai and Ratkovic 2014)を組み込むことにより頑健な方法にする．具体的には，Ning et al. (2020) *Biometrika* を参考にして応答の経時的な変化量は共変量の線形モデルであることを想定し，

$$\mathrm{E}[\Delta_k \mid \boldsymbol{x}, d_k = 1] = \boldsymbol{x}'\boldsymbol{\beta}_k^*. \tag{2}$$

零行列をO，$\boldsymbol{d} = (d_0, d_1)'$として以下のモーメント条件を考える:

$$\boldsymbol{H}^{(1)}(\boldsymbol{d}, \boldsymbol{x}; \boldsymbol{\alpha}) = \mathrm{E}\left[e_1(\boldsymbol{x}; \boldsymbol{\alpha})\left\{\frac{d_1}{e_1(\boldsymbol{x}; \boldsymbol{\alpha})} - 1\right\}\boldsymbol{x}\boldsymbol{x}'\right] = \mathrm{O},$$
$$\boldsymbol{H}^{(0)}(\boldsymbol{d}, \boldsymbol{x}; \boldsymbol{\alpha}) = \mathrm{E}\left[e_1(\boldsymbol{x}; \boldsymbol{\alpha})\left\{\frac{d_0}{e_0(\boldsymbol{x}; \boldsymbol{\alpha})} - 1\right\}\boldsymbol{x}\boldsymbol{x}'\right] = \mathrm{O}.$$

オリジナルの共変量バランシングと異なり，本稿の差分の差分法の設定では二次のモーメントをバランシングしていることに注意．行列の上三角部分を取り出す半ベクトル化を vech$(\cdot)$ と記したとき，$\boldsymbol{x}\boldsymbol{x}'$が対称行列であることから，

$$\boldsymbol{h}(\boldsymbol{d}, \boldsymbol{x}; \boldsymbol{\alpha}) \equiv (\mathrm{vech}\{\boldsymbol{H}_1(\boldsymbol{d}, \boldsymbol{x}; \boldsymbol{\alpha})\}', \mathrm{vech}\{\boldsymbol{H}_0(\boldsymbol{d}, \boldsymbol{x}; \boldsymbol{\alpha})\}')'$$

つまり，モーメント条件は

$$\mathrm{E}\{\boldsymbol{h}(\boldsymbol{d}, \boldsymbol{x}; \boldsymbol{\alpha})\} = \boldsymbol{0} \tag{3}$$

モーメント条件の次元(制約の数)がパラメータ$\boldsymbol{\alpha}$の次元より大きいover identifyな状況であるとき，この条件を満たすは$\boldsymbol{\alpha}$は必ずしも存在しない．しかし，$\mathrm{E}\{\boldsymbol{h}(\boldsymbol{d}, \boldsymbol{x}; \boldsymbol{\alpha})\}$をできるだけ$\boldsymbol{0}$に近づけるように$\boldsymbol{\alpha}$を推定すれば，得られる推定量は頑健性を示す．

定理. 二重に頑健な差分の差分法

モーメント条件(3)の経験版を満たすような$\hat{\boldsymbol{\alpha}}$を推定し，式(1)の傾向スコアを $e_1(\boldsymbol{x}; \hat{\boldsymbol{\alpha}})$ に置き換えることで得られる推定量$\hat{\boldsymbol{\theta}}^{\mathrm{DR}}$を考える．傾向スコアのモデル$e_1(\boldsymbol{x}; \boldsymbol{\alpha})$が正しい，あるいは応答の経時的な変化量は共変量の線形モデルであって(2)が成立するとき，この$\hat{\boldsymbol{\theta}}^{\mathrm{DR}}$は一致性を有し，$\boldsymbol{\theta}^*$に収束する．

4　数値実験

- 共変量 $\boldsymbol{x} = (x_1, x_2)'$，各$x_k \sim \mathrm{U}(0,2)$
- $y_0(0) \sim \mathrm{N}(0,1)$，$y_0(1) = y_0(0) + x_1 + \epsilon^{(0)}$，$\epsilon_0 \sim \mathrm{N}(0,1)$
- $y_1(0) \sim \mathrm{N}(0,1)$，$y_1(1) = y_1(0) + \beta x_1 + \epsilon^{(1)}$，$\epsilon^{(1)} \sim \mathrm{N}(0,1)$
- 割り当て $d_1 \sim \mathrm{logit}(-x_1 + \alpha x_2)$
- 傾向スコアの誤推定モデル: x_1のみの線形表現で書けると仮定
- 傾向スコアを提案手法で推定した場合と，ロジスティック回帰により推定した場合の因果効果の推定精度を比較する
- シミュレーション回数：3000回

				DR		MLE	
β	α	N	ATT	Est	95%CI	Est	95%CI
1.0	1.0	200	0.85	0.85	0.62, 1.10	0.71	0.44, 1.01
		600	0.85	0.85	0.72, 0.99	0.71	0.55, 0.87
	3.0	200	0.95	0.99	0.71, 1.28	0.50	-0.02, 0.93
		600	0.95	0.99	0.83, 1.15	0.50	0.24, 0.74
3.0	1.0	200	2.56	2.55	2.13, 2.96	2.30	1.82, 2.82
		600	2.56	2.55	2.33, 2.76	2.30	2.02, 2.58
	3.0	200	2.85	2.93	2.55, 3.33	2.28	1.70, 2.85
		600	2.85	2.93	2.71, 3.15	2.28	1.97, 2.59

5　結論

共変量バランシングを組み込んだ二重に頑健な差分の差分法を提案し，数値実験によりその頑健性を確認した．

参考文献

Abadie, A. (2005). Semiparametric difference-in-differences estimators, *The Review of Economic Studies*, **72**, 1–19.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score, *Journal of the Royal Statistical Society: Series B*, **76**, 243–263.