

TMLE の理論と応用

横浜市立大学 データサイエンス学部

田栗 正隆

1. 要旨

ランダム化を伴わない観察研究のデータから薬剤の因果効果を推定したい場合、治療群間が比較可能でないことによる交絡の調整がデータ解析上の問題となる。交絡調整をする方法は、結果変数に対する回帰モデルに基づく方法や傾向スコアに基づく方法など、いくつか提案されているが、本発表では近年注目されている TMLE (Targeted Maximum Likelihood Estimation) の理論と応用について紹介する。

2. TMLE の手順

A を 2 値の治療変数、 W を潜在的な交絡因子の組、 Y を結果変数とする。疫学でよく知られている標準化の一般化である G-formula に基づけば、興味のある平均因果効果に対する統計パラメータは以下のように表すことができる。

$$\Psi(P_0) = E_0[E_0(Y|A=1, W) - E_0(Y|A=0, W)]$$

ここで P_0 は n 個の i.i.d. 確率変数として観察される $O = (W, A, Y)$ の従う確率分布である。標的パラメータである $\Psi(P_0)$ は W の周辺分布 $Q_{W,0}$ と条件付き期待値 $\bar{Q}_0(A, W) = E_0(Y|A, W)$ を通してのみ P_0 に依存するので、 $\Psi(Q_0)$ と書くこともできる。ただし $Q_0 = (Q_{W,0}, \bar{Q}_0)$ である。

TMLE の推定手順は、以下のようなものである。

手順 1: アウトカム回帰 $\bar{Q}_0(A, W) = E_0(Y|A, W)$ の推定

手順 2: 傾向スコアモデル $g_0(a, W) = P_0(A=a|W)$ の推定

手順 3: $H(A, W)$ (clever covariate) の推定

$$H(A, W) = \frac{A}{P_0(A|W)} - \frac{1-A}{P_0(A|W)}$$

手順 4: $\bar{Q}_0(A, W)$ の推定量のアップデート

$$\bar{Q}_0(A, W) = \bar{Q}_n^0 + \varepsilon H_n^0$$

という回帰モデルを $\bar{Q}_0(A, W)$ の初期推定量 $\bar{Q}_n^0$ をオフセット、 $H(A, W)$ の初期推定量 H_n^0 を説明変数、 ε を未知パラメータとして最尤法によって当てはめる。

これにより、 $\bar{Q}_0(A, W)$ の推定結果を繰り返し更新する。

手順 5: $\Psi(Q_0)$ に最終的な $\bar{Q}_0(A, W)$ の推定量を代入し、TMLE 推定量を得る。

参考文献

Van der Laan, M. J., and Rose, S. (2018). *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*. Springer.