傾向スコア

北海道大学　経済学院　渡部元博

2023/05/15

目次

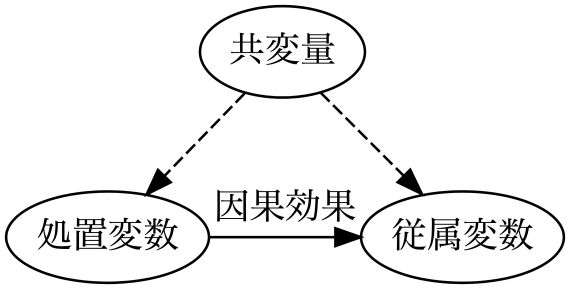
1　因果推論を行う前に確認すべき仮定

因果推論を実施する前に確認しておくべき仮定としては **SUTVA** がよく挙げられる (高橋 [2022,p31]) が、処置に一貫性があり、企業間の相互干渉がそもそも考えられない経済データにおいてはこの仮定が満たされない例はほとんど見られない。

ただし、もう一つの条件 **SITA**(強く無視できる割り当て条件) については、重要な仮定にもかかわらず、多くの応用研究で十分な検証がなされないまま、因果推論を実施している。この章では、共変量の投入による条件付き独立の確保及び SITA の定義と分析の際にチェックするべき事象を述べる。

1.1　共変量調整

因果推論を推定する際には、“潜在的な結果変数” と” 割り当て” いずれにも影響を与えらえる様々な共変量の影響を考える必要がある。経済分野では無作為割り当てを行うことができないので、共変量の影響をコントロールしなければ因果効果を測ることができない。



ここで、登場している 3 つの変数群である” 従属変数 (潜在的な結果変数)y_0, y_1“,” 処置変数 z“,” 共変量 x“全ての同時分布を考える。ここでは” 処置変数” は” 潜在的な結果変数” が欠測するかどうかについてのインディケータ変数であると考え、同時分布の分解を行う。

$$p(y_1, y_0, x, z) = p(z|y_1, y_0, x)p(y_1, y_0|x)p(x) \tag{1}$$

のように「潜在的な結果変数と共変量を条件づけた時の処置の分布」「共変量を条件づけた時の潜在的な結果変数の分布」および「共変量の分布」の 3 者の積で表現される。

このように同時分布を表現したとき、本来関心のある潜在的な結果変数の周辺分布、例えば $p(y_1)$ は

$$p(y_1) = \int p(z|y_1, y_0, x)p(y_1, y_0|x)p(x)dy_0dzdx$$

となり、この周辺分布は他の変数に依存しない。共変量の値に依存しない量を得るために共変量の分布について期待値を取ることを**共変量調整**と呼ぶ。

1.2 共変量を用いて因果効果を推定するための条件

処置が共変量に依存する場合を考えた場合、「処置はあくまで共変量にのみ依存し、結果変数には依存しない」と仮定する。この仮定は**強く無視できる割り当て (Strongly Ignorable Treatment Assignment)** 条件と呼ばれている。

これは、共変量 x の値を条件づける、つまり共変量の値が同じ対象だけで考えると処置群に割り当てられた時の潜在的な結果変数 y_1 と対照群に割り当てられた時の潜在的な結果変数 y_0 の同時分布が、処置変数 z と独立である

$$(y_1, y_0) \perp z|x$$

という条件である。これを同時分布の言葉で言い換えると

$$p(y_1, y_0, x, z) = p(z|x)p(y_1, y_0|x)p(x) \tag{2}$$

と表現できる。式 (1) と式 (2) より SITA は

$$p(z|y_1, y_0, x) = p(z|x)$$

つまり、どちらの群に割り当てられるかは共変量の値に依存し、従属変数による割り当てへの影響はあくまで「共変量と従属変数の関係」を通じてのみ間接的に存在している, という仮定であると言える。この条件については後の章で再度登場するのでそこで具体的なチェックの方法について整理する。

2　傾向スコアと伝統的な解析方法

傾向スコアの上位概念として**バランシングスコア**が挙げられる。このバランシングスコアを条件付けすることにより、共変量と処置が独立になるような「共変量の関数」である。傾向スコアはこのバランシングスコアの中でも最も粗い (1 次元に収めるため)。そして、傾向スコアの定義を再整理する。

傾向スコア　第 i 対象者の共変量の値を x_i, 処置変数の値を z_i とするとき、処置群へ割り当てられる確率 $\pi_i = p(z_i = 1|x_i)$ を第 i 対象者の傾向スコアという。

ここで、実際には各対象者の傾向スコアの真値はわからないので、データから推定する必要がある。推定においてはモデル設定が必要であるが、一般的にはプロビット回帰モデルやロジスティック回帰モデルが使用されることが多い。

2.1　傾向スコアを用いた解析方法

傾向スコアを用いて推定される因果効果は

$$E(y_1) - E(y_0) = E(y_1 - y_0) = E_e(E(y_1 - y_0|e)$$

であり、傾向スコアの定義を用いて、

$$E(y_1|e) = E(y_1|e, z = 1), \quad E(y_0|e) = E(y_0|e, z = 0)$$

結果として因果効果は

$$E(y_1) - E(y_0) = E_e[E(y_1|e, z = 1) - E(y_0|e, z = 0)] \tag{3}$$

ここから得られる有益な結論は観測されない潜在的な結果変数について考えなくても因果効果を推定することができるという点である。つまり、(3) 式の右辺において全ての期待値が観測可能なもので推定できているという点である。この結論から傾向スコアを用いたいくつかの解析の拡張を行うことができる。

傾向スコアを用いた調整法は全て二段階推定法であり、以下の二つのステップを踏む

1. **傾向スコアの推定・・・**処置変数 z を共変量 x によって説明するモデルを設定し、そのモデルの母数の推定を行う。母数の推定値を用いて、各対象者ごとに $z = 1$ に割り当てられる予測悪区立を計算し、これを傾向スコアの推定値とする。
2. **推定された傾向スコアを用いた調整・・・**上記で推定された傾向スコアを用いて、具体的な調整を行う方法として、傾向スコアの提唱者 Rosenbaum と因果推論の大家 Rubin は「マッチング」、「層別解析」、「共分散分析」の 3 つの方法を提案しており、これまではこの 3 つがよく用いられてきた。しかし、現在はこれらの改善点を克服した IPW の理論的研究がすすみ、利用例が増えてきている。IPW についてはのちに詳しく解説する。

傾向スコア解析の最大の利点は「従属変数と共変量の回帰モデルを必ずしも仮定する必要がない」という点である。この利点により、モデルの誤設定の可能性が低下し正しい推定を行うことができる。

2.2 Rosenbaum と Rubin が提唱した解析の問題点

これまで傾向スコアを用いた具体的な解析方法と利点について整理してきたが、以下のような欠点を抱えている。

1. マッチング・層別解析では因果効果の推定値を計算することはできるが、その標準誤差が正確には計算できない。
2. マッチング・層別解析ともに、各周辺期待値 $(E(y_1), E(y_0))$ の推定ができない
3. マッチングの基準や層の分け方に恣意性が残り、そこに理論的な保証がない
4. マッチングされない多くのデータが廃棄されることになる。

3 参考文献

- 高橋将宣 [2022]『統計的因果推論の理論と実践ー潜在的結果変数と欠測データ』共立出版
- 星野崇宏 [2009]『調査観察データの統計科学ー因果推論・選択バイアス・データ融合』岩波書店