効果検証入門　第二章

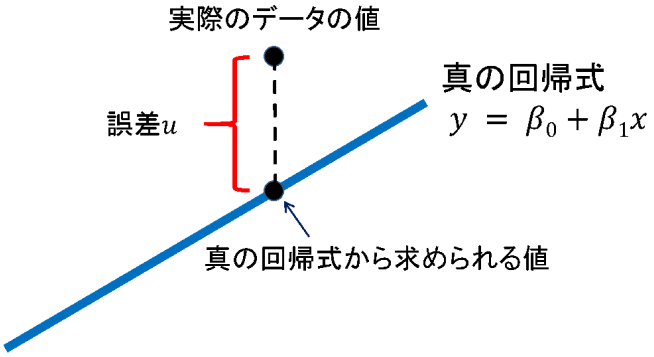
効果検証入門　第二章
介入効果を測るための回帰分析

2.1 回帰分析の導入

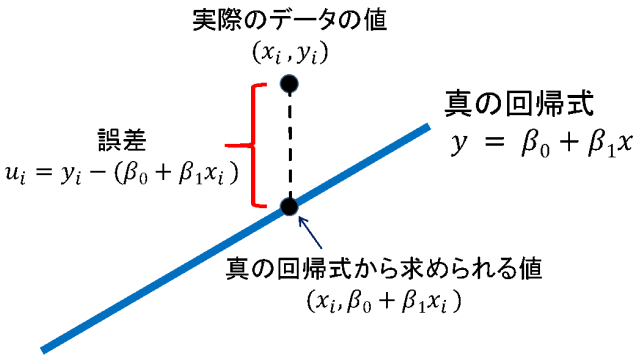
セレクションバイアスが存在するときにその影響を取り除くことができる最も基本的な方法が回帰分析

2.1.1 回帰分析

回帰分析のイメージをつかむために単回帰分析をとりあげる



B_0：切片。X_iが0のときのY_iの値を示す
B_1：線の傾き。X_iが1増加したときのYの期待値の増加分を示す



誤差u_iを小さくするため、誤差の二乗を最小にするように回帰分析のパラメータをデータから推定する操作のことを最小二乗法（OLS）とよぶ

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \{y_i - (\beta_0 + \beta_1 x_i)\}^2$$

最小二乗法により推定されたとは「偏回帰係数」とよばれる。
これらは実際のデータから算出された推定値であり、真の回帰式におけるととは異なることから$\hat{B}_0, \hat{B}_1$と表す

パラメータ（$\hat{B}$）の数値は母集団上での回帰分析で得られるパラメータBに対する推定値となっている。そのため、母集団の一部として得られたデータにおいて行われる回帰分析の特性を知ることができる
母集団において、YとYの条件付き期待値と回帰式の関係性は以下に表される

$$Y = E[Y|X] + u = B_0 + B_1 X + u$$

2.1.2　効果分析のための回帰分析

効果分析のための回帰分析では、データに存在するセレクションバイアスの影響をなるべく減らして、施策を行った場合と行わなかった場合の期待値の差分を推定する

登場する3つの変数

- 被説明変数（Y）：介入による効果を確認したい変数
- 介入変数（Z）：施策の有無を表す変数
- 共変量（Z）：セレクションバイアスを発生させていると分析者が想定する変数

複数種類の変数をモデルに含む回帰分析を重回帰分析とよぶ

$$E = [Y|X, Z] = B_0 + B_1 X + B_2 Z$$

$$Y = E[Y|X, Z] + u = B_0 + B_1 + B_2 Z + u$$

2.1.3 回帰分析による効果の推定

回帰分析によって条件付き期待値を近似した場合、介入の有無ごとの条件付き期待値は以下のようになる

$$E[Y|X, Z = 1] = B_0 + B_1 X + B_2 X^2 + B_3 1$$

$$E[Y|X, Z = 0] = B_0 + B_1 X + B_2 X^2 + B_3 0$$

介入の効果はこれらの期待値の差分であることから、回帰分析による効果の推定ではこれらの差分をだすと、

$$E[Y|X, Z = 1] - E[Y|X, Z = 0] = B_3$$

回帰分析において、興味のあるパラメータはB_3のみであることがわかる

2.1.4 回帰分析における有意差検定

母集団上のB_3が0であるという可能性を検証するために有意差検定が必要

2.1.5 Rによるメールマーケティングデータの分析〜2.2.3 RによるOVBの確認

notebook参照

2.2.4 OVBが与えてくれる情報

交絡（こうらく）因子：目的変数と介入変数の両方に関係がある因子のこと

モデルに含まれていない変数が目的変数と介入因子の両方に関係する場合、回帰分析から得られたZに関数効果の推定にはOVBが含まれる
また、分析結果が致命的な問題を抱えているのかは、変数と目的変数、介入変数との相関の強さに依存する

2.2.5 CIA

CIAとはモデルに含まれていない変数によるOVBが0になるような状態のこと

また、共変量の値が同一のサンプルにおいて、介入変数Zがランダムに割り振られているのに等しい状態ともいえる

2.2.6 変数の選び方とモデルの評価

CIAを満たすための二つの問題

- バイアスの評価ができない
  OVAが0になるまでの残りのバイアスの大きさを示してくれない
- 必要な共変量がデータにはない
  手持ちのデータにふくまれる変数だけでバイアスが十分に減らせない可能性がある

手持ちのデータには含まれない変数がセレクションバイアスを引き起こしているか評価するためのSensitivity Analysisという方法がある

重要だと分析者が認識でいている共変量以外の共変量をモデルから抜くことで、効果の推定値が大きく変動しないかを確認する

2.2.7 Post treatment bias〜2.3.6 分析のまとめ

notebook参照

2.4 回帰分析に対する様々な議論

- 効果検証においては、予測能力の高い低いは関係ない
    - 例：RCTだと介入変数のみで回帰分析するため予測能力は低いが、効果検証としては有用
- 介入と効果が非線形な関係をもつ場合、線形回帰以外の形も検討する
    - 「介入効果を比率で表したい」「介入と効果の関係が比率で扱われるべきである」場合は対数を利用する
- 多重共線性に注意
    - 介入変数と相関が強すぎる共変量を入れると、推定効果の分散が大きくなりすぎて、検定結果が意味のないものになる
    - 介入変数に関係ない多重共線性は、何も影響しないため考慮する必要なし
- 説明変数の間で相互作用がある場合は、推定されるパラメータにズレが生じる