調査観察データの統計科学ゼミ

第二回

Adachi Lab. M1 伊藤真道

October 12, 2019

大阪大学大学院人間科学研究科

担当した章

1. 「2.8 カーネル回帰モデルの利用とその問題点」
2. 「3.1 傾向スコアとは」

「2.8 カーネル回帰モデルの利用とその問題点」

2.8 カーネル回帰モデルの利用とその問題点

- 回帰関数の誤設計 → 大きなバイアスを生じる
  - → 回帰関数を用いないロバストな手法が欲すぃ…!

- 局所多項式回帰モデル (e.g. カーネル回帰モデル etc.)
  - ↑ の問題を避ける方法としてよく利用される.
  - カーネル回帰なら，(2.23) の $g(x|\beta_1), g(x|\beta_0)$ をカーネル回帰関数とすれば，共変量 x と結果変数 y の回帰関係を事前に決めなくても ok
  - さらに，バンド幅 h をいじると，線形〜完全フィットまで自由自在

2.8 カーネル回帰モデルの利用とその問題点

- ”ランダムな欠測”下でのカーネル回帰分析とこれを利用した因果効果の推定法
  - 共変量を所与とした時の Nadaraya-Watson 推定量は

$$\hat{E}[y|x] = \frac{\sum_{i=1}^{n} z_i K_h(x, x_i) y_{ij}}{\sum_{i=1}^{n} K_h(x, x_i) z_i}$$

    となる．但し，$z_{i1} = z_i, z_{i0} = 1 - z_i, (j = 1, 0)$ であり，$K_h(a, b)$ はバンド幅が h のカーネル関数である．
  - K_h は h が大きくなると，回帰関数がなめらかになる $\rightarrow h$ を平滑化パラメータとよぶ．
  - $K_h(a, b)$ として，ガウスカーネル

$$K_h(a, b) = (2\pi)^{-1/2} \exp\left\{-\frac{((a-b)/h)^2}{2}\right\}$$

    がよく利用される．

前項の続き

- $E[y_i]$ の推定量は，共変量を所与とした時の y の条件付き期待値 $\hat{E}[y|x]$ で欠測値を補間することにより，

$$\hat{E}[y_i] = \frac{1}{N}\sum_{i=1}^{n}\{z_{ij}y_{ij} + (1 - z_{ij})\hat{E}[y_j|x]\}$$

  と表現できる．
- 分散 $V(E[y_j])$ の推定量は，

$$\hat{V}(\hat{E}[y_j]) = \frac{1}{N}\sum_{l=1}^{n}\left[\frac{\sum_{i=1}^{n} K_h(x_l, x_i)z_{ij}y_{ij}^2}{h_j(x_i)\sum_{i=1}^{n} K_h(x_l, x_i)z_{ij}} + \frac{h_j(x_i) - 1}{h_j(x_i)}(\hat{E}[y|x_i])^2\right] - (\hat{E}[y_j])^2$$

  となる (式追えてない．すまねぇ…)．
  但し，$h_j(x) = \{\sum_{i=1}^{n} K_h(x, x_i)z_{ij}\}/\{\sum_{i=1}^{n} K_h(x, x_i)\}$ は，カーネル回帰による j 群への所属確率である．

前項の続き

- 推定量同士の共分散の推定値は，

$$\hat{Cov}(\hat{E}[y_1], \hat{E}[y_0]) = \frac{1}{N}\sum_{i=1}^{n} \hat{E}(y_1|x_i)\hat{E}(y_0|x_i) - \hat{E}(y_1)\hat{E}(y_0)$$

であることから，漸近的に，

$$\hat{E}(y_1) - \hat{E}(y_0) \sim N(E(y_1) - E(y_0), V(\hat{E}[y_1]) + V(\hat{E}[y_0]) - 2Cov(\hat{E}[y_1], \hat{E}[y_0]))$$

が成立する．これを利用して検定なども可能．

カーネル回帰の問題点

- バンド幅 h を変えると回帰関数の形状が大きく変わる (小さいとノイズに敏感，大きいとデータへの適合度が低下)
  - → 交差検証法などで h を決める
  - → モデル評価基準で h を決めることもできるが，分布を仮定しなければならない.
  - → 特に説明変数が多変量の場合のバンド幅の決め方で定番と呼べるものはない.

- 次元の呪い
  - -説明変数が多い場合には，推定に必要なデータ量が指数的に増加すること
  - -次元数を d とすると，漸近的な MSE は $N^{-4/(d+4)}$ に比例し，次元が高いほど収束が遅くなる (p.57 図 2.7 を参照).

カーネル回帰の問題点-解決法編

- 説明変数ごとに，1 次元のカーネル関数を設定し，その積で多次元カーネル関数を表現する.
  - 説明変数間の相関や，交互作用項が導入できない.
- 正則化項の導入.
  - 正則化パラメータの決定や，正則化項の設計問題 (色々考えないといけない).
  - → 社会科学ではあまり利用されていない.
- 一般化加法モデルの利用.
  - 加法分離性 (additive separability) の仮定が成立することを示すのがむずい.
  - 一般化加法モデルとは，説明変数ごとに特定の関数を設定し，その和を全ての説明変数と結果変数の回帰関数とする (= 加法分離性) の仮定に基づくモデル

ノンパラとパラメトリックの融合

- ノンパラメトリックな手法 (ノンパラ)
  - (ここでは)「潜在的な結果変数と共変量の回帰関数」のモデル仮定を行わないという意味
  - カーネル回帰，一般化加法モデルはノンパラの一種
  - バンド幅 (カーネル回帰)，加法分離性の仮定の成立チェック (一般化加法モデル) など，問題点いっぱい

- パラメトリックな手法 (2.7 節，稲岡氏が紹介)
  - 分布の仮定や，回帰関数の形状間違うとバイアスすごいなど，問題点山盛り

- → 両者の中間的なモデル (セミパラメトリックモデル) を作れば，欠点を補い合い，いいとこ取りができるのでは?

セミパラメトリックモデル

- セミパラメトリックモデル
  - -研究者の関心のある部分だけにパラメトリックなモデルを仮定し，関心のない部分にはノンパラメトリックに解析を行う．
  - e.g. 説明変数 v で，潜在的な結果変数 y_1, y_0 を説明したいけど，共変量 x には興味がない
  - → 共変量 x に関しては，ノンパラメトリックに (= 回帰関係を仮定せずに)，説明変数 v と結果変数 y_1, y_0 に関しては，パラメトリックに (= 回帰関係を仮定して) 分析を行う．

次の章で説明する傾向スコアは，セミパラな手法の代表的なものである．

「3.1 傾向スコアとは」

3.1 傾向スコアとは

無作為割り当てが不可能な研究において因果効果を推定したい
→傾向スコア

- 傾向スコア (propensity score)
  - -複数の共変量を一つに集約すれば，その 1 変数の上で層別化などを行うことができ，マッチングや層別での問題が起こらないという考え方から生み出された概念.

- バランシングスコア (balancing score)
  - これで条件付けると，共変量と割り当てが独立になる，つまり，

$$x \perp\!\!\!\perp z | b(x) \tag{3.1}$$

    となるような「共変量の関数」.
  - 全てのバランシングスコアは，関数 g を使って，
    $p(z = 1|x) = g(b(x))$ とかけることが必要である.

バランシングスコア続き

- なぜなら，(3.1) が成立するためには，

$$
\begin{aligned}
p(z=1|b(x)) &= \int p(z=1|x,b(x))dx = E_{x|b(x)}[p(z=1|x)] \\
&= E_{x|b(x),p(z=1|x)}[p(z=1|x)] = p(z=1|x) \qquad (3.2) \\
&= p(z=1|x,b(x))
\end{aligned}
$$

の三つ目の等号が成立する条件として，$p(z=1|x)=g(b(x))$ が成立する必要がある．

伊藤の解釈

$E_{x|b(x)}[\cdot]$ は，$b(x)$ が与えられたもとでの x の分布の期待値である．ここで，もし，$p(z=1|x)=g(b(x))$，つまり，$p(z=1|x)$ が $b(x)$ の関数として表せるなら，

$$
b(x)\text{が既知} \Rightarrow g(b(x))=p(z=1|x)\text{も既知}
$$

となるはずである．

バランシングスコア続き

- -(3.2) から，「割り当て z を共変量 x で説明する」ことと，「割り当て z をバランシングスコア $b(x)$ で説明する」ことが同じことがわかる．

- ” 強く無視できる割り当て”(= 割り当てはあくまで共変量のみに依存する) が成立しているとすると，

$$
\begin{aligned}
p(z=1|y_1,y_0,b(x)) &= \int p(z=1,x|y_1,y_0,b(x))dx \\
&= \int p(z=1|y_1,y_0,x,b(x))p(x|y_1,y_0,b(x))dx \\
&= E_{x|y_1,y_0,b(x)}[p(z=1|y_1,y_0,x)] \\
&= E_{x|y_1,y_0,b(x)}[p(z=1|x)] \\
&= p(z=1|x)
\end{aligned}
$$

強く無視できる割り当て条件下でのバランシングスコア

- これを (3.2) と合わせると，

$$p(z|y_1, y_0, b(\boldsymbol{x})) = p(z|b(\boldsymbol{x}))$$

であることがわかる．これはつまり，

$$\begin{aligned}
p(z|y_1, y_0, b(\boldsymbol{x})) &= \frac{p(z, y_1, y_0, b(\boldsymbol{x}))}{p(y_1, y_0, b(\boldsymbol{x}))} \\
&= \frac{p(y_1, y_0|z, b(\boldsymbol{x}))p(z, b(\boldsymbol{x}))}{p(y_1, y_0|b(\boldsymbol{x}))p(b(\boldsymbol{x}))} \\
&= \frac{p(y_1, y_0|z, b(\boldsymbol{x}))}{p(y_1, y_0|b(\boldsymbol{x}))} p(z|b(\boldsymbol{x})) = p(z|b(\boldsymbol{x})) \\
&\Leftrightarrow \frac{p(y_1, y_0|z, b(\boldsymbol{x}))}{p(y_1, y_0|b(\boldsymbol{x}))} = 1 \\
&\Leftrightarrow p(y_1, y_0|z, b(\boldsymbol{x})) = p(y_1, y_0|b(\boldsymbol{x}))
\end{aligned}$$

つまり，BS を条件づければ，$(y_1, y_0), z$ は独立，

$$(y_1, y_0) \perp\!\!\!\perp z|b(\boldsymbol{x})$$

となる．(さっきは，共変量 $\boldsymbol{x}$ と割り当て z であったことに注意)

傾向スコア再訪

Def. 傾向スコア

第 i 対象者の共変量の値を x_i，割り当てを z_i とするとき，群 1 へ割り当てられる確率

$$e_i = p(z_i = 1|\boldsymbol{x}_i)\ (0 \leq e_i \leq 1)$$

を第 i 対象者の傾向スコアという．

- -実際には，真値がわからない → データから推定
- 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を用いて推定.
- 例えばロジスティック回帰モデルなら，定数 1 を含めた x_i を用いて，

$$p(z_i = 1|\boldsymbol{x}_i) = e_i = \frac{1}{1 + \exp(-\boldsymbol{\alpha}^t \boldsymbol{x}_i)} \tag{3.4}$$

傾向スコア再訪

- この時，割り当てに関する尤度は，

$$\prod_{i=1}^{N} \left(\frac{1}{1+\exp(-\boldsymbol{\alpha}^t \mathbf{x}_i)} \right)^{z_i} \left(1 - \frac{1}{1+\exp(-\boldsymbol{\alpha}^t \mathbf{x}_i)} \right)^{1-z_i} \tag{3.5}$$

  となる．

- これを最大化する MLE$\hat{\boldsymbol{\alpha}}$ を用いることで，i 番目の対象者の傾向スコアは，

$$\hat{e}_i = \left(\frac{1}{1+\exp(-\hat{\boldsymbol{\alpha}}^t \mathbf{x}_i)} \right)^{z_i}$$

  と表される．

- -モデルを仮定しないノンパラで，傾向スコアを推定することもある．
  - -しかし，次元の呪いが存在するため，一般には，傾向スコアの推定にはパラメトリックな手法が用いられる．

メモに使ってね！

Questions?