調査観察データの統計科学ゼミ-第 20-b 回-

Masamichi Ito

Osaka University Graduate School of Human Sciences
Adachi Lab M1

March 5, 2020

担当章

パラメトリックな回帰モデルの利用

Giula et al.(2006) は，図 7.1(p.194) の欠測部分をベイズモデルで補完する方法を提案した.

- D: すでに得られているデータ
- $\boldsymbol{\theta}$: モデルの母数
- $\boldsymbol{y}_A$: データ A の結果変数
- $\boldsymbol{y}_B$: データ B の結果変数

とすると，$\boldsymbol{y}_A, \boldsymbol{y}_B$ のベイズ事後予測分布は,

$$p(\boldsymbol{y}_A, \boldsymbol{y}_B|D) = \int\int p(\boldsymbol{y}_A|\boldsymbol{x}, \boldsymbol{\theta})p(\boldsymbol{y}_B|\boldsymbol{x}, \boldsymbol{\theta})p(\boldsymbol{x})p(\boldsymbol{\theta}|D)d\boldsymbol{x}d\boldsymbol{\theta} \tag{7.3}$$

と表現できることから，この分布から発生させた乱数で欠測値を補完して，シングルソースデータにする

パラメトリックな回帰モデルの利用 (続き)

(続き) ただ，(7.3) 式は，積分が含まれていることから，実際には MCMC を利用する.

- Giula et al.(2006) では，$p(\boldsymbol{y}_A|\boldsymbol{x},\boldsymbol{\theta}), p(\boldsymbol{y}_B|\boldsymbol{x},\boldsymbol{\theta})$ にパラメトリックな回帰モデルを仮定して解析してる
- $\boldsymbol{y}_A, \boldsymbol{y}_B$ が 2 値変数なら，$p(\boldsymbol{y}_A|\boldsymbol{x},\boldsymbol{\theta}), p(\boldsymbol{y}_B|\boldsymbol{x},\boldsymbol{\theta})$ をロジスティック回帰モデル (実際にはその変数分の積) とする

p.203 の例 7.1 読みましょう．ポイントは，この解析例が，マッチングの問題 (希薄化) が解決されている点.
しかし，

- 条件付き独立性の仮定が成立しているかどうかのチェックが必要.
- 条件付き独立性が近似的にせよ成立するためには，共通項目による変数 A(同じことだが，変数 B) の予測力が十分高い必要がある.
- ロジスティック回帰では予測力が不十分．→セミパラ回帰か，深層学習でも使うか？

カーネルマッチングによるデータ融合

データ融合でもセミパラ手法は有用
→カーネルマッチング(3.2 節) とディリクレ過程混合モデルを紹介

カーネルマッチングによるデータ融合

3.2 節の表記をちょっと変えるだけで利用可能である．「$z = 1$(データ A とされる群) において観測されない $\boldsymbol{y}_B$」を予測する式として，

$$\hat{y}_{B_i} = \frac{\sum_{j=1}^{N}(1 - z_j)K_{ij}y_{B_j}}{\sum_{j=1}^{N}(1 - z_j)K_{ij}}$$

が利用でき，「$z = 0$(データ B とされる群) において観測されない $\boldsymbol{y}_A$」の予測には，

$$\hat{y}_{B_i} = \frac{\sum_{j=1}^{N} z_j K_{ij} y_{A_j}}{\sum_{j=1}^{N} z_j K_{ij}}$$

が利用できる．

カーネルマッチングによるデータ融合

カーネルマッチングでは

- データ A の対象者の $\boldsymbol{y}_B$ の値を予測し代入する際にデータ B の対象者の $\boldsymbol{y}_B$ の値をカーネルによる重みで全て利用する
  $\leftarrow$ 通常のマッチングでは，「1 つまたは数個だけ」利用する．

- カーネルマッチングにおいて，特定の対象者の重み以外は 0 となる場合が通常のマッチング (カーネルマッチングが通常のマッチングを包含してる)
  $\rightarrow$ カーネルマッチングの方が優れた代入法．

ディリクレ過程混合モデル

セミパラなモデリングとして，

- ディリクレ過程混合モデル(Dirichlet process mixture model)
- ディリクレ過程事前分布を用いたセミパラベイズ

が最近の流行．なんでディリクレ過程が人気なの？

- 「全ての分布はある特定の分布の混合分布によって表現できる(Sethuraman, 1994)」$\rightarrow$ 混合モデル使おう
- しかし，通常の混合モデルでは事前に混合要素数を決めたり，事後的にモデル選択基準によって事後的に決定しなければならない...
- ディリクレ過程混合モデルでは，未知の混合要素数の混合分布によるモデリングを行うため，事前にモデルを選択しなくても良い!!

ディリクレ過程混合モデルを利用したセミパラな離散変数の回帰分析モデル

従属変数が 2 値の場合，通常はロジスティック回帰モデルなどで，共変量との回帰モデルを表現するが，通常データに対する説明力は低い．そこで，

$$p(y=1|\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \frac{1}{1+\exp\{\boldsymbol{a}_k^t \boldsymbol{x} + b_k\}}$$

のように，ロジスティック回帰モデルの混合モデルを利用する．要素数 K に関しては，上限だけを決めておいて，データから推定する(有限ディリクレ過程混合モデル)

(図 7.4 を見ながら) 多様な回帰関数を表現可能であることがわかる．

ディリクレ過程混合モデル-まとめ-

$p(\boldsymbol{y}_A|\boldsymbol{x}), p(\boldsymbol{y}_B|\boldsymbol{x})$ を有限ディリクレ過程混合モデルを用いて表現し，ベイズ事後分布として，(7.3) を利用すれば，通常のロジスティック回帰分析よりも予測力が高く，ベイズ推定の枠組みで一貫した推論が可能となる．データ融合では，

- 条件付き独立性，ランダムな欠測の仮定を満たすべき！共変量の情報を最大限生かして予測を行いたい！
  $\rightarrow$ セミパラな回帰手法が望ましい
- 予測分布を構成したい
  $\rightarrow$ ベイズ統計学の枠組みを適用すべし

この 2 つの観点から，データ融合には，セミパラベイズのモデルを用いるのが適切

シングルソースデータが一部利用できる場合

データ A,B の両者が同時に測定されているデータ (データ C とする) がある場合，条件付き独立性は仮定しなくて良い
$\rightarrow$5 章の選択バイアスの特殊な状況と考えれば良い.

シングルソースデータが一部利用できる場合

- ランダムな欠測を仮定
  $\rightarrow$ $p(\boldsymbol{y}_A, \boldsymbol{y}_B|\boldsymbol{x})$ を正しくモデリングできれば，データ C のみで母数推定可能

- $p(\boldsymbol{y}_A, \boldsymbol{y}_B|\boldsymbol{x})$ ではなく，$p(\boldsymbol{y}_A, \boldsymbol{y}_B)$ に興味がある．$\boldsymbol{y}_A$ と $\boldsymbol{x}$,$\boldsymbol{y}_B$ と $\boldsymbol{x}$ の回帰関係を設定したくない
  $\rightarrow$ 重みつき M 推定や二重にロバストな推定

擬似パネル

擬似パネルデータ (pseudo panel data)

時点ごとに別の対象について行われた複数時点での調査データを合併して，シングルソースデータ化したもの．パネル調査のコストがかかる問題，脱落の問題への対症療法とみなせる．

擬似パネルの例

具体的にこれまで利用されてきたのは,

1. 属性で集計し，コーホートに分割し，同一のコーホートに所属する個人は等質とみなし，各コーホートの標本平均を観測値と考えて推定を行う.

2. 各時点の調査対象者をマッチングによって「同一対象」とみなす

3. 個人ごとの効果を表すパラメータに特定のモデルを仮定する.
(e.g. 個人ごとの固定効果が，性別や生年などの時間に依存しない属性と固定効果と相関がない変数によって説明されるとするモデル)

擬似パネルとその問題点

- 1 は，「コーホート内で個人が等質である」という仮定が強いことと，「コーホート数が統計解析でのサンプルサイズ」になることから，情報の損失が生じる
- 2 については，マッチングの問題点がそのまま生じる
- 3 は実際に擬似パネルデータを作成しているわけではないが，目的は擬似パネルデータの作成と同じ.
  - 「個人ごとの固定効果を説明する変数」が十分にないと，固定効果に関連する母数の推定にはバイアスが生じる
    - 共変量が多い場合は，セミパラ手法を利用することが望ましい.

実データによる性能比較

Just read these paragraphs by yourself!!!

References

- Giula, McCulloch, Rossi(2006).”A direct approach to data fusion”, *Journal of Marketing Research*, 43, p.73-83
- 岩崎 (2015) 統計解析スタンダード　統計的因果推論 ($\rightarrow$ おすすめ．読みやすい文章構成)
- 高井, 星野, 野間 (2016)「調査観察データ解析の実際
欠測データの統計科学-医学と社会科学への応用-」
- 星野 (2009)「調査観察データの統計科学-因果推論・選択バイアス・データ融合-」

Thank you for your attention!!

Any Questions?

メモに使ってね!!