階層ベイス推定の概略

渡邊直樹

2022 年 12 月 10 日

はじめに

ベイズ推定とは，簡単にいうと，分析者の主観に基づいて設定された事前確率分布を観察されたデータとベイズルールを用いて改訂し，事後確率分布を求める「操作」といえる．母集団分布のパラメータを推定し，仮説を検定するタイプの統計学には馴染みがあったとしても，ベイズ推定はそのような統計学とはかなり趣が異なるだけでなく，計算機を用いた確率分布の計算を避けて通れないことがその利用を阻んできたことは否めない．個人で利用できる計算機（PC）の性能の飛躍的向上がベイズ統計学の利用を促進したことは間違いないだろう．それに伴ない，たとえば，20 年ほど前から医科学におけるデータ解析ではベイズ推定は頻繁に利用されるようになってきた．

しかし，ベイズ推定が急速に普及し始めた背景には，マーケティングなどの実務において階層ベイズ推定が利用されるようになってきたことがあるだろう．事前確率分布の設定が分析者の主観に依拠するとしても，データが観察されるたびにそれが改訂され，それを用いた予測が正確になっていく傾向があることも重要であり，たとえば，特定の属性を持つ個人のグループごとの推定と予測に対する需要の増加が階層ベイズ推定の普及を引き出したのかもしれない．そこでは，母集団全体の平均的な性質だけではなく，より詳細な分析結果を得ることができるからである．

以下では，階層ベイズ推定を実行するための技能習得を目的とした説明ではなく，そのモデリングの具体例を示し，そこでどのような計算が必要となるかを紹介する．それによって，このデータ解析技法の学習を始めるための準備としたい．

線型回帰分析への適用

ここでは多くの学習者にとって比較的馴染みのある線形回帰モデルに対して階層ベイズ推定の適用を考える．個人 i に関する m 個のデータを m 次のベクトル Y_i で表すとして，それに対して定数項と p 個の説明変数からなる線形回帰モデル

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + u_i \tag{1}$$

を設定しよう．ここで，Y_i は，たとえば，個人 i の m 個の商品に対する選好（willingness to pay）と考えると，各説明変数は対応する商品の属性，品質，価格などである．また，u_i は m 次の誤差項であり，平均 0，p 個の分散はすべて σ_i，共分散 0 の正規分布に従うとする．

この線形回帰モデルで説明変数 X_{ji} に対応するパラメータ β_j がすべての個人で共通とすると，標準的な回帰分析と同じであり，母集団の平均的な性質を推定することになる．階層ベイズ推定では，線形回帰モデルを

$$Y_i = \beta_{0i} + \beta_{1i} X_{1i} + \beta_{2i} X_{2i} + \cdots + \beta_{pi} X_{pi} + u_i \tag{2}$$

とし，たとえば，すべての個人からなる母集団の性質 h_j に加えて，個人の異質性を

$$\beta_{ji} = h_j + \delta_{ji} \tag{3}$$

としてモデルに取り込む．ここで，δ_{ji} は平均 0，分散は V_i の正規分布に従う確率変数である．また，すべての個人 i と任意の j と j' について，δ_{ji} と $\delta_{j'i}$ の共分散は 0 と単純化してもよい．

さらに，(3) ではなく，個人 i のある属性 z_i とその係数 α_i を導入して，

$$\beta_{ji} = h_j + \alpha_i z_i + \delta_{ji} \tag{4}$$

という形で個人の異質性を導入することもできる．個人 i の属性としては，年齢や性別などを思い浮かべると理解しやすいだろう．

階層ベイズ推定という名称は，(3) や (4) のように，個人 i のパラメータに潜在的な構造があることに由来する．ただ，これだけパラメータが多いと変数間の関係が非常に複雑で，(1) のようなモデルにおけるパラメータの推計は困難である．そこで，(2)，(3) または (4) のように事前確率分布を設定し，観察されたデータとベイズルールを用いてそれを改訂することで，事後分布を計算することで，個人 i の m 個の商品に対する選好に関する分析結果を得ようとすることになる．

(2)，(3) または (4) での事前確率分布の設定は分析者の主観によるものであり，それを設定した時点では客観性は担保されているわけではない．しかし，新しい商品やサービスを発売する際には，それらに関する購買データはそもそも存在しない．コンジョイント分析はそのような状況で実施される市場調査の方法の一つであり，しばしば階層ベイズ推定が適用される．

一般に確率分布の計算には積分が用いられるのだが，複雑な積分計算には「乱数」を用いた数値計算が用いられる．乱数自体は Excel でも発生させることができるので，実用を考えないならば，Excel でベイズ推定の原理を学習することは可能である．しかし，実務で用いるには，q 個の部品からなる機械の故障確率を扱う際に用いられるガンマ分布，経年劣化による故障確率を表現することに適したワイブル分布，個人の特異性を表現する際によく用いられるガンベル分布，滅多に発生しない現象を取り扱うことに適したポアソン分布や指数分布など，分析目的に合わせて，事前確率分布を設定する必要があるため，こうした確率分布に関する知識も実習前にある程度は学習しておく必要がある．こうした準備が必要なため，ベイズ統計の学習は初学者にとって敷居が高い．

今後の学習

(2)，(3) または (4) のような簡素な階層ベイズ推定であれば，R 単体で実行可能ではあるが，階層ベイズ推定を含むベイズ推定の十分な説明と練習用データセットが整備されている教科書は

「Stan と R でベイズ統計モデリング」，松浦健太郎 著，共立出版

である[1]．R 単体で階層ベイズ推定を行うには R のコマンドとパッケージに習熟する必要があるため，Stan というソフトウェアを用いることを推奨したい．この Stan を R で使えるようにしたものが RStan であり，R と同様に無料で配布されている．上記教科書は RStan の使用を前提として，ベイズ推定全般について説明している．乱数を用いた計算が収束しない場合の対処法も記述されている．簡単な階層ベイズ推定であれば，R 単体で実行可能ではあるが，EZR（easy R）にはベイズ推定のためのコマンドは用意されていない．

[1] EZR（easy R）にはベイズ推定のためのコマンドは用意されていない．

「ビジネスマンが一歩先を目指すベイズ統計学」，
朝野熙彦 編著，朝倉書店

でも，「Excel から RStan へステップアップ」というサブタイトルから分かるように，Excel の関数による乱数発生機能を使ってベイズ推定の原理に関する説明の後で，RStan によるいくつかの分析例が紹介されている．第 3 章の「やさしく読み解く階層ベイズ」では，R でコードを書き，Stan による乱数を用い計算の入門編になっている．ただ，本書全般に関して言えることだが，サンプルデータのサイズは小さく（20 人分），実務への応用を念頭においた実習には，もう少し詳細なデータがあると，なお良かったかもしれない．

「実践ベイズモデリング：解析方法と認知モデル」，
豊田秀樹 編著，朝倉書店

は，まったく初学者向きではないが，ある程度学習が進んだ人向けに階層ベイズ推定以外のベイズ推定の様々な技法を紹介しており，それらに対応する認知心理学に関する学術論文のサーベイも添付されている．

とりあえず，実習なしで，ベイズ推定の原理だけを学習するならば，

「完全独習 ベイズ統計学入門」，小島寛之 著，ダイヤモンド社

が良いかもしれない．この書籍の帯には「かけ算・わり算だけで理解できる!」と書かれている．ベイズ推定の原理をざっと理解し，どのように実務で使えそうかを考えるにはこうした書籍が適していると思われる．