

一般化線形モデル・一般化線形混合モデルとベイズ化

樋口さぶろお <https://hig3.net>

龍谷大学大学院 先端理工学部研究科 数理・情報科学コース

計算科学特論 L08(2024-11-13 Wed)

最終更新: Time-stamp: "2024-11-13 Wed 14:19 JST hig"

今日の目標

- 一般化線形/混合モデルを説明できる
- これらのベイズ化, 階層ベイズモデルを説明できる
- 計算機を用いたベイズ推定の必要性を説明できる



L07-Q3

Quiz 解答: 確率変数の変数変換 (対数正規分布)

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \text{ である.}$$

$y = e^x$ より, $x \in \mathbb{R}$ に対して $y > 0$. $dy = e^x dx$ すなわち, $y^{-1} dy = dx$ より,

$$f_X(x) dx = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\log y - \mu)^2}{2\sigma^2}} y^{-1} dy.$$

よって,

$$f_Y(y) = \begin{cases} 0 & (y \leq 0), \\ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\log y - \mu)^2}{2\sigma^2}} y^{-1} & (0 < y). \end{cases}$$

ここまで来たよ

7 事後予測分布・多パラメタのベイズ推定・一般化線形モデル

8 一般化線形モデル・一般化線形混合モデルとベイズ化

- 一般化線形モデル・一般化線形混合モデル
- 一般化線形モデル・一般化線形混合モデルのベイズ化
- 計算機でベイズ推定

(復習) 一般化線形モデル (GLM)

Hoff ベイズ §10

ポアソン回帰モデル

Hoff ベイズ §10.1

植物の高さ x_i から、(今シーズンの) 種子の個数 Y_i を予測したい。
意味や $\lambda > 0$ を考えると、次が適切。

ポアソン回帰モデル

Hoff ベイズ §10.1

$$Y_i \sim \text{Po}(\lambda_i)$$

$$\log \lambda_i = \beta_1 + \beta_2 x_i$$

次の3つで指定される確率モデルを、**一般化線形モデル**
(GLM=Generalized Linear Model) **Hoff ベイズ §10.1** という。

- 確率分布 $\text{Po}(\lambda)$ パラメタを説明変数で説明する
- リンク関数 $g(\lambda) = \log(\lambda)$ 1 次関数に等置するための関数
- 線形予測子 $\beta_1 + \beta_2 x_i$

β_1, β_2 を最尤推定する。

L08-Q1

Quiz(線形モデルと最小二乗法)

一般化線形モデルとしての線形回帰モデルを考える.

$$Y_i \sim N(\mu_i, \sigma^2)$$

$$\mu_i = \beta_1 + \beta_2 x_i$$

- 確率分布 $N(\mu, \sigma^2)$
- リンク関数 $g(\mu) = \mu$ (恒等関数)
- 線形予測子 $\beta_1 + \beta_2 x_i$

測定値 $(x_1, y_1), \dots, (x_N, y_N)$ が与えられたとき,

- ① 対数尤度関数 $\log L(\beta_1, \beta_2)$ を書こう.
- ② 最尤推定における「対数尤度関数が最大」の条件が, 最小二乗法における「二乗誤差最小」の条件と等価であることを示そう.

L08-Q2

Quiz(ロジット関数とオッズ)

事象の確率を p とするとき、ロジット関数に現れる、事象と補事象の確率の比 $\frac{p}{1-p}$ はオッズ (odds) とよばれ、0 以上 $+\infty$ 以下の値をとる。

一方、ギャンブルにおいては、1.25 (decimal, Europe), 1 : 4 (fractional, UK), -400 (moneyline, US) などの表記をオッズとよぶことがある。

- ① decimal odds $d = 1.25$ という表記は、元本 1.00 をかけて、確率 p の事象が起きたとき、元本とあわせて 1.25 が払い戻される、利益 +0.25 が得られるという意味である（確率 $1 - p$ の補事象が起きたときは元本を含め払い戻されず、利益 -1.00 が得られる）。利益の期待値が 0 であるとき、decimal odds d と確率 p との関係を求めよう。
- ② fractional odds P/Q または $P : Q$ または $P-Q$ (P, Q は整数) という表記は、確率 p の事象が起きたときに利益 P を得るためには、元本 Q を賭けておく必要がある、という意味である。利益の期待値が 0 であるとき、fractional odds P/Q と確率 p との関係を求めよう。
- ③ moneyline odds $m = -400 < 0$ という表記は、確率 p の事象が起き

一般化線形混合モデル (GLMM)

Hoff ベイズ §11

植物の高さ x_i から、(今シーズンの) 種子の個数 Y_i を予測したい。
ただし、植物には高さだけでない個体差 r_i がある。
植物が成長して様々な x_i で測定されている。

- 確率分布 $\text{Po}(\lambda)$
- リンク関数 $g(\lambda) = \log(\lambda)$
- 線形予測子 $\beta_1 + \beta_2 x_i + r_i$

M=mixed 「混合」という語は何通りにも使われるが、ここでは、
固定効果 $\beta_1 + \beta_2 x_i$ と、

ランダム効果 r_i との混合という意味。

直接測定できないランダム効果は、例えば $r_i \sim N(0, s^2)$ のような分布にしたがうと考え、 r_i で積分してしまい、 β_1, β_2, s について最尤推定する。

ここまで来たよ

7 事後予測分布・多パラメタのベイズ推定・一般化線形モデル

8 一般化線形モデル・一般化線形混合モデルとベイズ化

- 一般化線形モデル・一般化線形混合モデル
- 一般化線形モデル・一般化線形混合モデルのベイズ化
- 計算機でベイズ推定

一般化線形モデル (GLM) のベイズ化

Hoff ベイズ §10

(パラメタ λ でなく) パラメタ β_1, β_2 を確率変数と考える
事後分布は 2 次元分布 $f(\beta_1, \beta_2 | y_1, \dots, y_N)$. (確率変数でないパラメタ $x_1, \dots, x_N$ にもよる)

一般化線形混合モデル (GLMM) のベイズ化

Hoff ベイズ §11

GLMM のベイズ化=階層ベイズ化

(パラメタ r_i だけでなく) パラメタ s を確率変数と考え, s の事前分布を考える.

階層事前分布 $f(r_i|s) \sim N(0, s^2)$.

s 超パラメター, ハイパーパラメター hyper-parameter

$f(s)$ 超事前分布, ハイパープライヤー hyper-prior

$y \leftarrow (\beta, r) \leftarrow s$ という確率変数の階層構造

階層モデル，マルチレベルモデル

Hoff ベイズ §8

植物の高さ x_i から，（今シーズンの）種子の個数 Y_i を予測したい．
ただし，植物には高さだけでない個体差 $r_i \sim N(0, s^2)$ がある．
さらに，畑 $j = 1, \dots, m$ があり，植物 i は $j(i)$ に植えられている．畑の
効果 $\bar{r}_{j(i)} \sim N(0, t^2)$ がある．

- 確率分布 $\text{Po}(\lambda)$ パラメタを説明変数で説明する
- リンク関数 $g(\lambda) = \log(\lambda)$ 1 次関数に等置するための関数
- 線形予測子 $\beta_1 + \beta_2 x_i + r_i + \bar{r}_{j(i)}$

個体と全体と中間に，グループ（畑）のレベル，階層があるという意味で，**マルチレベルモデル**，**階層モデル**という．
数学としては階層ベイズモデルの枠内に収まる．

L08-Q3

ベイズ推定

ベイズ統計学における，分布のパラメタのベイズ推定とは何か，

- 頻度論の点推定，区間推定と対比して，
- 事前分布，事後分布という用語を使って，

200 字程度の日本語で説明しよう．

ここまで来たよ

7 事後予測分布・多パラメタのベイズ推定・一般化線形モデル

8 一般化線形モデル・一般化線形混合モデルとベイズ化

- 一般化線形モデル・一般化線形混合モデル
- 一般化線形モデル・一般化線形混合モデルのベイズ化
- 計算機でベイズ推定

ベイズ推定はそう簡単にはできない…

Hoff ベイズ §4

ベイズ推定の値は、観測値 $y_1, \dots, y_N$ から定まる、パラメタ θ の次の分布の母ナントカ.

$$f_{\theta|Y}(\theta|y) = \frac{f_{Y|\theta}(y|\theta)f_{\theta}(\theta)}{\int_{-\infty}^{+\infty} f_{Y|\theta}(y|\theta)f_{\theta}(\theta)d\theta}$$

心配

- 知ってる分布になる？（共役事前分布ならなる）
- 知らない分布なら？
 - ▶ 分母の積分できる？
 - ▶ $E[\Theta|y] = \int_{-\infty}^{+\infty} \theta f_{\theta|Y}(\theta|y)d\theta$ とかを自分で積分しないといけない。
- 事前分布 $f_{\theta}(\theta)$ どう決める？

経験分布で代用

アイデア

母ナントカの計算が困難なら、 $f_{\theta|Y}(\theta|y)$ からサイズ n の標本（サンプル）を抽出して、その分布（**経験分布** (empirical distribution)）から推定すればいい.

モンテカルロ法 (MC=Monte-Carlo methods)

- 母期待値 $E[\Theta|y] \simeq$ 標本期待値 $\frac{1}{n}(\theta_1 + \dots + \theta_n)$
- 母比率 $\simeq$ 標本比率
- 母最頻値 $\simeq$ 経験分布のヒストグラムの最大階級値
- 母中央値 $\simeq$ 標本中央値

この n は、計算時間さえかければいくらでも大きくできる. (貴重な) 測定値 y_i の個数 N とは別.

- 知ってる分布になる場合, R や Python の `scipy.stats` で, `stats.dist.rvs(size=n)` で乱数を呼び出して取得できる.

心配タイプ1

- 知らない分布なら？

- ▶ サンプル抽出できる？ = 分布にしたがう乱数ある？ 書ける？
- ▶ 分母が計算できないなら？ 比率 $f_{\theta|Y}(\theta|y)/f_{\theta|Y}(\theta'|y)$ だけから乱数作れるといいな～

ひとつの解決策

マルコフ連鎖モンテカル口法 (MCMC=Markov Chain Monte-Carlo methods)

マルコフ連鎖は確率過程 (stochastic process) の1分類.

⇨ 来週以降

心配タイプ2

- 事前分布 $f_{\theta}(\theta)$ どうする？

いくつかの解決策

- 何か知見があって主観的に選ぶ分布
- 無情報事前分布，とても広い分布 例 $N(0, 100^2,)$, $U(-100, 100)$.
- 定数（全確率が発散する！）
- 変数変換しても変わらない ジュフリーズ分布