1 Introduction

被説明変数が離散的なデータ（質的データ）を扱う方法について 8 章では議論を深めている。

被説明変数 y_i と説明変数 x_i が結び付けられるモデルを作りたい。これは、回帰係数を $\beta = (\beta_0, \beta_1, \beta_2, ...)^T$、説明変数を $x = (1, x_{1i}, x_{2i}, ...)^T$ として、y_i と $x_i^T\beta$ とを結びつけるモデルを作ること。被説明変数が 1 になる確率、$P(y_i = 1)$ をモデル化したいときに、そのまま線形回帰モデルを当てはめると負になったり 1 を超えたりと問題が生じる。そこで、仮想的な因子

$$y_i^* = \boldsymbol{x}_i^T\boldsymbol{\beta} + \epsilon_i \tag{1}$$

を用いて[*1]

$$y_i = \begin{cases} 1 & (y_i^* \geq 0) \\ 0 & (y_i^* < 0) \end{cases} \tag{2}$$

と被説明変数を決定されるものとする。このとき、$-\epsilon_i$ の累積分布関数を $F(\cdot)$ とすると、被説明変数が 1 になる確率は

$$\begin{aligned} P(y_i = 1) &= P(y_i^* \geq 0) && (3)\\ &= P(\boldsymbol{x}_i^T\boldsymbol{\beta} + \epsilon_i \geq 0) && (4)\\ &= P(-\epsilon_i \leq \boldsymbol{x}_i^T\boldsymbol{\beta}) && (5)\\ &= F(\boldsymbol{x}_i^T\boldsymbol{\beta}) && (6) \end{aligned}$$

と表すことができる。誤差項（ϵ_i）がどんな確率分布に従うか（累積分布関数として何を使用するか）で、プロビットモデル（標準正規分布）とロジットモデル（ロジスティックス分布）が導入され、（教科書によると）学問領域によって異なるとのこと。

1.1 プロビットモデル

被説明変数が 1 になる確率を $P(y_i = 1) = p_i$ とする。標準正規分布の分布関数 $Phi(\cdot)$ を用いて

$$p_i = \Phi(\boldsymbol{x}_i^T\boldsymbol{\beta}) \tag{7}$$

とするモデルをプロビットモデルという。

1.2 ロジスティックモデル

ロジステック分布の分布関数を用いて

$$p_i = \frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i^T\boldsymbol{\beta})} \tag{8}$$

とするモデルをロジステックモデル（ロジステック回帰モデル）という。また、$\log(p_i/(1-p_i))$ をロジットもしくは対数オッズといい、

$$\log\frac{p_i}{1-p_i} = \boldsymbol{x}_i^T\boldsymbol{\beta} \tag{9}$$

と、線形の形式で表すことができるので扱いやすい。

[*1] 教科書 p.234 (8.2) 式、$Y_i^* = \beta_0 + \beta_1 X_i + \epsilon_i$

2　説明変数が 2 個以上の場合

本補助資料の 式 (6) らへんの議論をそのまま繰り返して、仮想因子は

$$y_i^* = \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i \tag{10}$$

であり、$P(y_i = 1) = p_i$ は

$$p_i = F(\boldsymbol{x}_i^T \boldsymbol{\beta}) \tag{11}$$

で与えられる。いまは二値分類で考えているので尤度関数は

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i} \tag{12}$$

$$= \prod_{y_i=0} (1 - F(\boldsymbol{x}_i^T \boldsymbol{\beta})) \cdot \prod_{y_i=1} F(\boldsymbol{x}_i^T \boldsymbol{\beta}) \tag{13}$$

であり、これを最大化することで β の最尤推定量 $\hat{\beta}$ を求めることができる。

2.1　式（8.15）〜

回帰係数 β に関する検定を行うために、最尤推定量 $\hat{\beta}$ の標本分布（漸近分布）を求める（教科書 p.130〜132 付近）[*2]。

対数尤度関数は

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{y_i \log p_i + (1 - y_i) \log(1 - p_i)\} \tag{14}$$

$$= \sum_{i=1}^{n} \{y_i \log F(\boldsymbol{x}_i^T \boldsymbol{\beta}) + (1 - y_i) \log(1 - F(\boldsymbol{x}_i^T \boldsymbol{\beta}))\} \tag{15}$$

となる。ここで $\hat{\boldsymbol{\beta}}$ は一致推定量でありその漸近分布（$n \to \infty$ のときに従う分布）は

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N\left(0, \frac{1}{I_1(\theta)}\right) \tag{16}$$

であるので、フィッシャー情報量を計算する。$I_n(\theta) = -E[\partial^2 / \partial \theta^2 f(x, \theta)]$ を使いたいので、対数尤度関数の二階微分を計算する。

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left(\frac{y_i}{F(\boldsymbol{x}_i^T \boldsymbol{\beta})} - \frac{1 - y_i}{1 - F(\boldsymbol{x}_i^T \boldsymbol{\beta})} \right) f(\boldsymbol{x}_i^T \boldsymbol{\beta}) \boldsymbol{x_i} \tag{17}$$

$$\frac{\partial \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = \sum_{i=1}^{n} \left(-\frac{y_i}{F(\boldsymbol{x}_i^T \boldsymbol{\beta})^2} - \frac{1 - y_i}{1 - F(\boldsymbol{x}_i^T \boldsymbol{\beta})^2} \right) f(\boldsymbol{x}_i^T \boldsymbol{\beta})^2 \boldsymbol{x_i} \boldsymbol{x_i}^T$$

$$+ \sum_{i=1}^{n} \left(\frac{y_i}{F(\boldsymbol{x}_i^T \boldsymbol{\beta})} - \frac{1 - y_i}{1 - F(\boldsymbol{x}_i^T \boldsymbol{\beta})} \right) f'(\boldsymbol{x}_i^T \boldsymbol{\beta}) \boldsymbol{x_i} \boldsymbol{x_i}^T \tag{18}$$

[*2] 回帰係数に関する検定は教科書 p.59〜の議論で、最小二乗推定量 $\ell' \hat{\theta}$ が正規分布 $N(\ell' \theta, L' L \sigma^2)$ に従うということを利用していた。

ここで

$$E[y_i] = 1 \cdot F(\boldsymbol{x}_i^T\boldsymbol{\beta}) + 0 \cdot (1 - F(\boldsymbol{x}_i^T\boldsymbol{\beta})) = F(\boldsymbol{x}_i^T\boldsymbol{\beta}) \tag{19}$$

であることを用いると式 (18) の二項目の期待値はゼロになる。ゆえに

$$-E\left[\frac{\partial}{\partial\beta\partial\beta^T}\ell\right] = \sum_{i=1}^{n} \frac{f(\boldsymbol{x}_i^T\boldsymbol{\beta})^2\boldsymbol{x_i}\boldsymbol{x_i}^T}{F(\boldsymbol{x}_i^T\boldsymbol{\beta})(1 - F(\boldsymbol{x}_i^T\boldsymbol{\beta})} \tag{20}$$

となり、$I_1(\theta) = 1/nI_n(\theta)$ であることと、漸近分布が $n \to \infty$ であることを踏まえると、

$$I_1(\theta) = \lim_{n\to\infty} \sum_{i=1}^{n} \frac{f(\boldsymbol{x}_i^T\boldsymbol{\beta})^2\boldsymbol{x_i}\boldsymbol{x_i}^T}{F(\boldsymbol{x}_i^T\boldsymbol{\beta})(1 - F(\boldsymbol{x}_i^T\boldsymbol{\beta})} \tag{21}$$

が求まり、漸近分布が計算できる（教科書的には $\boldsymbol{A}$ を求めた）。

3　ベイズの定理

3.1　(9.1)〜(9.2) 式

ベイズの定理（教科書は分母を加法定理で変形した形を取っている）：

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} = \frac{P(X|Y)P(Y)}{\sum_Y P(Y|X)P(X)} \tag{22}$$

ここで少し見やすく（？）観測データを $\mathcal{D}$、パラメータを $\boldsymbol{w}$ とすると（Bishop（上）p.21)、

$$P(\boldsymbol{w}|\mathcal{D}) = \frac{P(\mathcal{D}|\boldsymbol{w})P(\boldsymbol{w})}{P(\mathcal{D})} \tag{23}$$

という形を取り、データ $\mathcal{D}$ を観測した事後に $\boldsymbol{w}$ に関する事後分布を評価することが可能になる。分母は左辺の事後分布の規格化の条件を満たすものであり、明らかなものとして省略してもよい。すると、ベイズの定理は言葉で書くと：

$$\text{事後確率} \propto \text{尤度} \times \text{事前確率} \tag{24}$$

となる。

p(D|w) という量は「データ集合 D に対する評価であって、パラメーターベクトル w の関数とみなせる。これを尤度関数と呼ぶ。これは、パラメーターベクトルを固定したときに観測されたデータ集合がどれくらい起こりやすいかを表している。」

4　事前確率分布と事後確率分布

4.1　補助導入

他の教科書から、導入の補充：

> 事前分布 $p(\boldsymbol{w})$ は未知の母数に対して、観測値 $\mathcal{D}$ を得る以前に利用可能な情報を表すものとされる。...さらにベイズ法では観測値を得たあとで観測値の実現値を固定したもとでのパラメータの条件付き分布（事後分布 $p(\boldsymbol{w}|\mathcal{D})$）を考慮する。ベイズ法においてはこの事後分布を求めることが本質的であり、いっ

たん事後分布が求められれば推定、検定などの統計的推測問題の最適解が容易に求められる。... 一方で、ベイズ統計学の問題点は事前分布をどう設定するかという点である。

現代数理統計学（竹村）p.313

ということで、事前分布をどう設定するかがベイズ統計学のキモであるらしい。が、

ベイズ法における事前分布の選び方については、実際駅かつ広く受け入れられた標準的な方法がない。

現代数理統計学（竹村）p.313

とあるので、決定的な方法はないとのこと（数値計算などに便利であれば何でもいい？）。本教科書でも議論しているように、共役性という性質に基づいて事前分布を決定することで、解析的に便利な性質をもたせることができる。

4.2 (9.5)〜(9.6) 式

例 9.2 の尤度関数を計算してみると、n 回の試行中 x 回の成功を得る確率は

$$f(x|\theta) = {}_nC_x\theta^x(1-\theta)^{n-x} \tag{25}$$

であり、事前確率 $w(\theta)$ さえ決まれば事後確率が計算でき、それ以降の統計処理を行うことができる。事前分布については：

もし事前分布が θ と $(1-\theta)$ のべき乗に比例するように選ぶなら、事後分布は事前分布と尤度関数の積に比例するので、事前分布と同じ関数形式になる。なおこの性質は共役性と呼ばれ...

Bishop(上)p.69

以上のことから、ここではベータ分布を事前分布に選ぶとよい。ベータ分布は

$$w(\theta) = \mathrm{Beta}(\theta|\mathrm{a},\mathrm{b}) = \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1} \tag{26}$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \tag{27}$$

であり、事後確率の計算（9.3）式は

$$w'(\theta|z) \propto w(\theta)f(z|\theta) \tag{28}$$

$$\propto \theta^{a+x-1}(1-\theta)^{b+n-x-1} \tag{29}$$

となり、事後確率分布としても $Be(a+x, b+n-x)$ のベータ分布が出てくる。なのでこの計算では

$$a \to a+x, \quad b \to b+n-x \tag{30}$$

の変換をするだけでよいことになる[*3]。

一般に、尤度関数に対して事前分布 $w(\cdot)$ と事後分布 $w'(\cdot|\mathcal{D})$ が同一種類の分布に属するならば、単に統一分布俗内の変換を引き起こすだけであり、この分布族を f の自然共役分布の族という。

[*3] 事前分布から、この集合を観測したあとの事後分布を求めるには a の値を x だけ、b の値を $n-x$ だけ増やせば良いことがわかる。このことから事前分布のハイパーパラメータ a, b はそれぞれ成功した・失敗した回数の有効観測数として解釈できる（Bishop（上）p.70）

4.3 ex.9.3

開発中の新薬が偽薬と比較して有効となる確率を シータ として、まったく手探りでわからないとすれば事前確率分布として一様分布を選ぶことができる。

4.4 p.258

正規母集団 $N(\theta, \sigma^2)$ から抽出し大きさ n の標本を作成したとき、母分散 σ^2 が既知として尤度関数[*4]は

$$f(\mathcal{D}|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x_i-\theta)^2}{2\sigma^2}\right\} \tag{31}$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{\sum x_i^2 - 2\theta\sum x_i + \sum\theta^2}{2\sigma^2}\right\} \tag{32}$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{\sum x_i^2 - 2n\theta\bar{x} + n\theta^2}{2\sigma^2}\right\} \tag{33}$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{\sum(x_i-\bar{x})^2 + n(\bar{x}-\theta)^2}{2\sigma^2}\right\} \tag{34}$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{\sum(x_i-\bar{x})^2}{2\sigma^2}\right\} \cdot \exp\left\{-\frac{n(\bar{x}-\theta)^2}{2\sigma^2}\right\} \tag{35}$$

第一項はデータから決まる量（かつここでは母分散既知）なので、(9.2b) の計算の際に分母の積分から飛び出て、分母・分子で打ち消し合う。以上から (9.3) 式は

$$w'(\theta|z) \propto w(\theta) \exp\left\{-\frac{n(\bar{x}-\theta)^2}{2\sigma^2}\right\} \tag{36}$$

であり、事前分布も正規分布にとると計算の都合上便利で、

$$w(\theta) = \frac{1}{\sqrt{2\pi}\tau} \exp\left\{-\frac{(\theta-\lambda)^2}{2\tau^2}\right\} \tag{37}$$

とすれば、事後分布は

$$w'(\theta|z) \propto \exp\left\{-\frac{n(\bar{x}-\theta)^2}{2\sigma^2}\right\} \exp\left\{-\frac{(\theta-\lambda)^2}{2\tau^2}\right\} \tag{38}$$

$$\tag{39}$$

[*4] 母分散未知の場合のは逆ガンマ関数が共役事前分布となる。

指数部分だけを取り出すと（θ で平方完成）

$$-\frac{1}{2}\left(\frac{n(\bar{x}-\theta)^2}{\sigma^2}+\frac{(\theta-\lambda)^2}{\tau^2}\right) \tag{40}$$

$$=-\frac{1}{2}\left(\frac{n(\bar{x}-\theta)^2\tau^2+(\theta-\lambda)^2\sigma^2}{\sigma^2\tau^2}\right) \tag{41}$$

$$=-\frac{1}{2}\left(\frac{(n\tau^2+\sigma^2)\theta^2-2(n\bar{x}\tau^2+\lambda\sigma^2)\theta+n\bar{x}^2\tau^2+\lambda^2\sigma^2}{\sigma^2\tau^2}\right) \tag{42}$$

$$=-\frac{1}{2}\left(\frac{(n\tau^2+\sigma^2)\left(\theta-\frac{(n\bar{x}\tau^2+\lambda\sigma^2)}{n\tau^2+\sigma^2}\right)^2-\ldots+n\bar{x}^2\tau^2+\lambda^2\sigma^2}{\sigma^2\tau^2}\right) \tag{43}$$

定数部分（θ を含まない項）はおいておいて

$$-\frac{1}{2}\left(\frac{\left(\theta-\frac{n\bar{x}\tau^2+\lambda\sigma^2}{n\tau^2+\sigma^2}\right)^2}{\sigma^2\tau^2/(n\tau^2+\sigma^2)}\right) \tag{44}$$

ということで、事後分布として

$$\text{平均} \ : \ \frac{n\bar{x}\tau^2+\lambda\sigma^2}{n\tau^2+\sigma^2}=\frac{\lambda/\tau^2+n\bar{x}/\sigma^2}{1/\tau^2+n/\sigma^2} \tag{45}$$

$$\text{分散} \ : \ \frac{\sigma^2\tau^2}{n\tau^2+\sigma^2} \tag{46}$$

の正規分布が出てくる。

5 Appendix

5.1 フィッシャー情報量

$X = (X_1, ..., X_n)$ の同時確率密度関数を $f_n(\boldsymbol{x}, \theta)$ で表すときに、フィッシャー情報量は次式で定義される：

$$I_n(\theta) = E_\theta \left[\left(\frac{\partial}{\partial \theta} \log f_n(x, \theta) \right)^2 \right] = -E_\theta \left[\frac{\partial^2}{\partial \theta^2} \log f_n(\boldsymbol{x}, \theta) \right] \tag{47}$$

$I_n(\theta)$ は n 個のデータのフィッシャー情報量を表す。

5.2 クラメールラオの下限

不偏推定量の分散の下限値を与える不等式で、フィッシャー情報量を用いて表される

$$V[\hat{\theta}] \geq \frac{1}{I_n(\theta)} \tag{48}$$

5.3 漸近有効性について

母数の真の値を $\boldsymbol{\theta}$、最尤推定量を $\hat{\boldsymbol{\theta}}$ と表す。このとき $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ は $N(0, I(\boldsymbol{\theta})^{-1})$ に分布収束する。

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N\left(0, \frac{1}{I(\theta)}\right) \tag{49}$$