フェスタ 国立大学2023　東京大学大学院数理科学研究科 Graduate School of Mathematical Sciences, THE UNIVERSITY OF TOKYO

ロジスティック回帰分析

ウェブ掲載用ハンドアウト

増田 弘毅

２０２３年度公開講座「統計と数学」
１１月２５日（土）

はじめに: データ解析の汎用的ホワイトボックス

- ソフトウェア開発
- 理論とは別の考察を要する
- プログラムコーディング

- データの特性・種類
- 抽出したい情報

- モデルの数理構造の構築
- 確率論，確率解析，確率過程論
- 統計学，統計的漸近理論

- モデルのあてはまりの良さ
- 変数選択問題
- 予測誤差

内容概観

- イントロ：<u>回帰分析</u>の概要
  - パラメトリックな<u>確率分布族</u>と説明変数
- ベルヌーイ分布と<u>(2値)ロジスティック回帰分析</u>
  - モデルの構造，推定方法，判別・予測
- 多項分布と<u>多項ロジスティック回帰分析</u>
  - カテゴリカルな応答変数への拡張

<u>回帰分析 (Regression analysis)</u>

- 「回帰」＝「元の状態へ戻ること」
- 統計では応答変数と説明変数の関係の統計分析全般をあらわす

説明変数 X → **応答変数 Y**

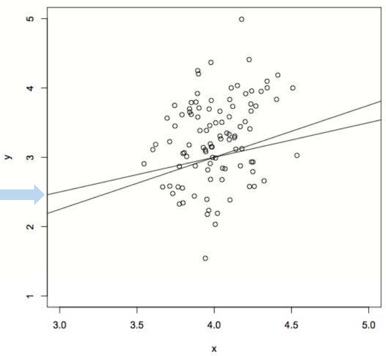
観測データ $(X_1, Y_1), \ldots, (X_n, Y_n)$ から確率構造を推測

フランシス・ゴルトン (1822-1911)

- 同別親子間の身長関係の傾向を調査
- 人類の身長の平均<u>回帰</u>傾向を示唆
- <u>**相関係数**</u>（線形関係）の概念を提唱

$$Y = aX + b + (\text{誤差})$$

応答変数が２値の場合？



要因のある２値反応データ

- 生活習慣病の<u>**発症**</u>確率の予測
- 疾患のリンパ節への<u>**転移**</u>確率の予測
- 化学成分量を要因として<u>**薬効有無**</u>の判断
- 試験をパスするにはどのくらい勉強すれば?

２値の確率分布を想定したい

$$x = (x_1, \ldots, x_p) \xrightarrow{\text{要因入力}} \text{変換} \xrightarrow{\text{出力}} y \in \{0, 1\}$$

ベルヌーイ (Bernoulli) 分布

- 表が出る確率が θ のコイン投げの結果の分布 $(0 < \theta < 1)$:

$$\text{確率} \begin{cases} P[Y = 1] = \theta \\ P[Y = 0] = 1 - \theta \end{cases} \Rightarrow \quad p(y; \theta) = \theta^y (1 - \theta)^{1-y}, \qquad y \in \{0, 1\}$$

  - P は Probability の頭文字.

- θ はさまざまな**要因** (説明変数) に依存して変わるはず

要因入力: $(\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_p})$

結果出力: $\boxed{\boldsymbol{y = 0}}$ **or** $\boxed{\boldsymbol{y = 1}}$

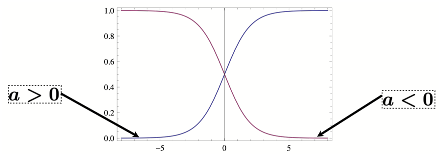
$$p(y; \theta(x)) = \theta(x)^y (1 - \theta(x))^{1-y}$$

確率のロジット変換

$$p = \frac{1}{1 + e^{-z}} \quad \overset{0<p<1,\ z\in\mathbb{R}}{\Longleftrightarrow} \quad z = \log\left(\frac{p}{1-p}\right)$$

- 各 $a \in \mathbb{R}$ に対し, $(-\infty, \infty)$ から $(0, 1)$ への関数

$$z \mapsto \frac{1}{1 + e^{-az}}$$



- 統計的パラメータ $\beta = (\beta_1, \ldots, \beta_p) \in \mathbb{R}^p$ の導入と変換:

$$\log\left(\frac{P[Y_i = 1]}{1 - P[Y_i = 1]}\right) = \beta^\top x_i := \langle \beta, x_i \rangle = \sum_{j=1}^{p} x_{ij}\beta_j$$

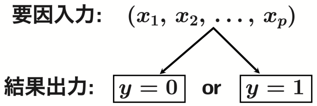
$$\Longleftrightarrow \quad P[Y_i = 1] = \left(1 + \exp(-\beta^\top x_i)\right)^{-1}$$

ロジスティック回帰モデル

- データセット $(x_1, y_1), \dots, (x_n, y_n)$:
  - 説明変数 $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})^\top \in \mathbb{R}^p$
  - 目的変数 $y_i \in \{0, 1\}$
- **未知パラメータ** $\beta = (\beta_1, \dots, \beta_p)$
- 目的変数 y_i の統計モデル:

$$p(y_i|x_i;\beta) = \pi(x_i;\beta)^{y_i}\{1-\pi(x_i;\beta)\}^{1-y_i},$$
$$\pi(x_i;\beta) = \frac{1}{1+\exp\left(-\sum_{j=1}^{p}\beta_j x_{ij}\right)}$$

- ベルヌーイ分布のパラメータ変動版



Cox, D.R. (1958) The regression analysis of binary sequences (with discussion). *J Roy Stat Soc B* **20**: 215–242.

最尤 (さいゆう) 法によるモデルの推定

$$p(y_i|x_i,\beta) = \pi(x_i;\beta)^{y_i}\{1-\pi(x_i;\beta)\}^{1-y_i}$$
$$\pi(x_i;\beta) = \frac{1}{1+\exp\left(-\sum_{j=1}^{p}\beta_j x_{ij}\right)}$$

- データの実現値 $\{(x_i, y_i)\}_{i=1}^n$ が生じた**対数確率 (対数尤度)**

$$\ell_n(\beta) = \log\left(\prod_{i=1}^n p(y_i|x_i,\beta)\right) = \sum_{i=1}^n \beta^\top x_i y_i - \sum_{i=1}^n \log\left\{1+\exp\left(\beta^\top x_i\right)\right\}$$

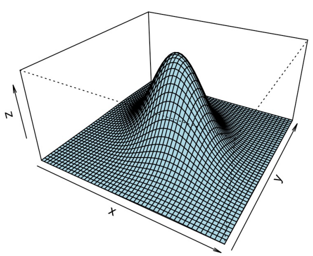
- **数値最適化**で ℓ_n を最大にする最尤推定量 $\hat{\beta}$ を求める: $\ell_n(\beta) \le \ell_n(\hat{\beta})$

最尤推定量の分布を近似できる ➔ 推定誤差の定量的評価

データが多いときの最尤推定量の基本性質

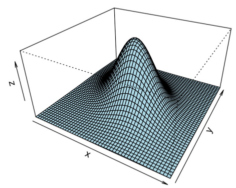
- **$\hat{\beta}$ の分布 $\underset{n\to\infty}{\approx}$ 真値 β_0 を平均とする正規分布**
- 2 次元の場合 (対称 $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix}$, $a_{11} > 0$, $\det(A) = a_{11}a_{22} - a_{12}^2 > 0$)

$$z = z(x, y)$$
$$\mapsto \frac{1}{2\pi\sqrt{\det(A)}}\exp\left(-\frac{1}{2}z^\top A^{-1} z\right)$$
$$= \frac{1}{2\pi\sqrt{\det(A)}}\exp\left\{-\frac{1}{2}\begin{pmatrix} x & y \end{pmatrix} A^{-1}\begin{pmatrix} x \\ y \end{pmatrix}\right\}$$

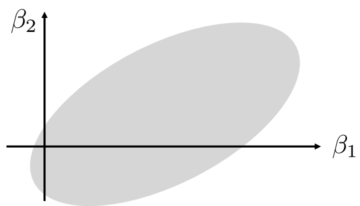


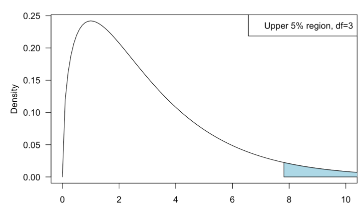
推定精度の評価法: 信頼集合の構成

- 例. データ $\{(x_i, y_i)\}_{i=1}^n \Rightarrow \beta_0 := (\beta_1, \beta_2)$ の 95%近似信頼領域



$$\underbrace{\left|(\hat{\beta}-\beta_0)^\top\left(-\frac{1}{n}\frac{\partial^2}{\partial\beta^2}\ell_n(\hat{\beta})\right)(\hat{\beta}-\beta_0)\right|}_{\underset{n\to\infty}{\approx}\chi^2(p)\ \text{(自由度 } p \text{ のカイ二乗分布)}} \le \text{“}p \text{ と } 95\% \text{ で決まる定数”}$$





推定されたモデルに基づく判別法

- 入力 $x_0 = (x_{01}, \ldots, x_{0p})$ に対する**反応確率の予測**

$$\pi(x_0; \hat{\beta}) = \frac{1}{1 + \exp\left(-\sum_{j=1}^{p} \hat{\beta}_j x_{0j}\right)}$$

$$\begin{cases} \geq 0.5 & \Rightarrow \hat{Y}_0 = 1 \\ < 0.5 & \Rightarrow \hat{Y}_0 = 0 \end{cases}$$

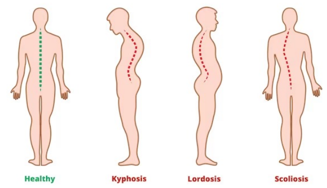
1	2
0.25	0.75

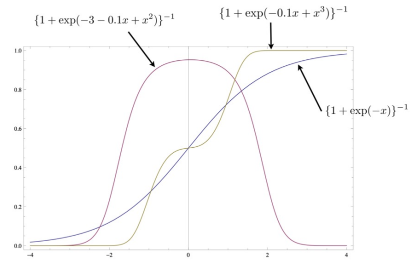
説明変数に関する非線形化

- 脊柱後弯症 (Kyphosis)
  - 発症頻度は生後月齢について**単調性がない**
- いつ施術すれば発症確率を抑制できる？

$$p(y_i | x_i, \beta') = \pi(x_i; \beta')^{y_i} \{1 - \pi(x_i; \beta')\}^{1-y_i}$$

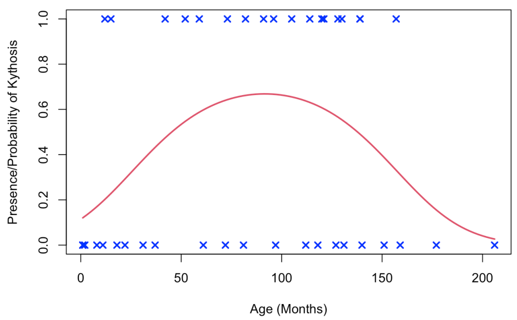
$$\pi(x_i; \beta') = \frac{1}{1 + \exp\left\{-\left(\beta_0 + \sum_{j=1}^{p} \beta_j b_j(x_i)\right)\right\}}$$





Hastie, T. and Tibshirani, R. (1990), Generalized Additive Models, *Chapman and Hall.*

- kyphosisデータ (R “gam” パッケージ)
  - 40人分データ
  - Ageの二次多項式でモデルを推定

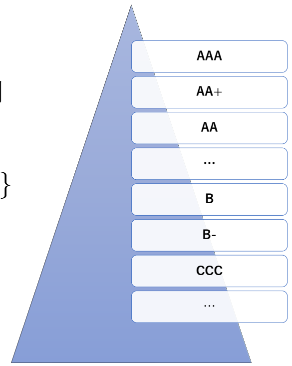


要因のあるカテゴリカル反応データ

- 例. 企業評価システム
  - 顧客のさまざまなデータを基に貸し倒れ確率を予測

$$x = (x_1, \ldots, x_p) \xrightarrow{\text{要因入力}} \text{変換} \xrightarrow{\text{出力}} y \in \{1, 2, \ldots, M\}$$

複数値をとる確率分布を想定する



多項 (Multinomial) 分布

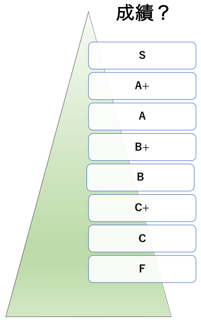
- $n, M \in \{1,2,\dots\}$, $\mathbf{p} = (\mathsf{p}_1, \dots, \mathsf{p}_M) \in (0,\infty)^M$, $\sum_{j=1}^M \mathsf{p}_j = 1$
- 変数 $Y = (Y_1, \dots, Y_M) \sim$ 多項分布 $\mathrm{Mult}(n, \mathbf{p})$

$$\iff P[Y_1 = y_1, \dots, Y_M = y_M] = \frac{n!}{y_1! \dots y_M!} \mathsf{p}_1^{y_1} \dots \mathsf{p}_M^{y_M},$$

$$y_j \in \{0, 1, \dots, n\}, \quad \sum_{j=1}^M y_j = n$$

- 多項定理: $(x_1 + \dots + x_n)^M = \sum_{k_1+\dots+k_n=M} \frac{M!}{k_1! \dots k_M!} x_1^{k_1} \dots x_n^{k_M}$

どのカテゴリーに属するかの要因を組み込む



- データ $(x_1, y_1), \dots, (x_n, y_n)$
  - 説明変数 $x_i = (x_{i1}, \dots, x_{ip}) \in \mathbb{R}^p$
  - 目的変数 (ラベル) $y_i = (y_{i1}, \dots, y_{iM}) \in \{0,1\}^M$, $\sum_{m=1}^M y_{im} = 1$
    - 目的変数 y_i が第 m 群に属する $\iff y_i = (0, \dots, 0, \underset{m}{1}, 0, \dots, 0)$
- 多項分布 $\mathrm{Mult}\left(1, \left(\frac{\exp(x_i^\top \beta_1')}{\sum_{j=1}^M \exp(x_i^\top \beta_j')}, \dots, \frac{\exp(x_i^\top \beta_M')}{\sum_{j=1}^M \exp(x_i^\top \beta_j')}\right)\right)$ で回帰?
- 多項分布 $\mathrm{Mult}\left(\left(\frac{\exp(x_i^\top \beta_1)}{1+\sum_{j=1}^{M-1} \exp(x_i^\top \beta_j)}, \dots, \frac{\exp(x_i^\top \beta_{M-1})}{1+\sum_{j=1}^{M-1} \exp(x_i^\top \beta_j)}, \frac{1}{1+\sum_{j=1}^{M-1} \exp(x_i^\top \beta_j)}\right)\right)$ で回帰

多項ロジスティック回帰モデルの推定

- データ $(x_1, y_1), \dots, (x_n, y_n)$
  - 説明変数 $x_i = (x_{i1}, \dots, x_{ip}) \in \mathbb{R}^p$
  - 目的変数 (ラベル) $y_i = (y_{i1}, \dots, y_{iM}) \in \{0,1\}^M$, $\sum_{m=1}^M y_{im} = 1$
    - 目的変数 y_i が第 m 群に属する $\iff y_i = (0, \dots, 0, \underset{m}{1}, 0, \dots, 0)$

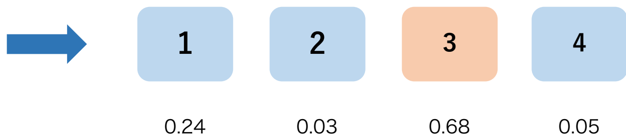
- "y_i が第 m 群に属する確率" $\pi_{im} = \pi_{im}(\beta) \overset{\text{モデリング}}{\longleftarrow} \frac{\exp(x_i^\top \beta_m)}{1 + \sum_{j=1}^{M-1} \exp(x_i^\top \beta_j)}, \qquad m = 1, \dots, M-1.$
- $P[Y_i = (y_1, \dots, y_M)] \leftarrow \pi_{i1}^{y_1} \pi_{i2}^{y_2} \dots \pi_{iM}^{y_M} =$ "$\pi_{i1}, \pi_{i2}, \dots \pi_{iM}$ のいずれか"
- 対数尤度は $\ell_n(\beta) = \ell_n(\beta_1, \dots, \beta_{M-1}) = \sum_{i=1}^n \sum_{m=1}^M d_{im} \log \pi_{im}(\beta)$
  - 統計量 $d_{im} := \begin{cases} 1 & (y_i \text{ の第 } m \text{ 成分が } 1), \\ 0 & (\text{それ以外}). \end{cases}$ $\Rightarrow$ 最尤推定量 $\hat{\beta}$: $\ell_n(\beta) \le \ell_n(\hat{\beta})$

多項ロジスティック回帰モデルによる多群判別

- 新たな実験点 x_0 に対して，目的変数 y_0 が m 群に属する**確率の予測値**が得られる:

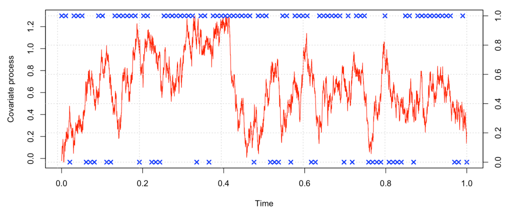
$$x_0 \mapsto \begin{cases} \dfrac{\exp(x_0^\top \hat{\beta}_m)}{1 + \sum_{j=1}^{M-1} \exp(x_0^\top \hat{\beta}_j)}, & 1 \le m \le M-1, \\ \dfrac{1}{1 + \sum_{j=1}^{M-1} \exp(x_0^\top \hat{\beta}_j)}, & m = M. \end{cases}$$

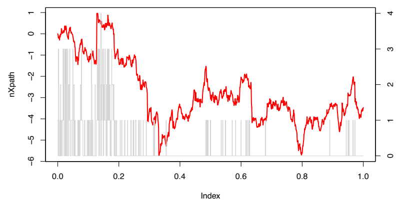
$\Rightarrow$ $\hat{\pi}_m(x_0)$ が最大となる m_0 に対して，"y_0 は m_0 に属する" と判断する (判別問題).



従属性カテゴリカル回帰モデル

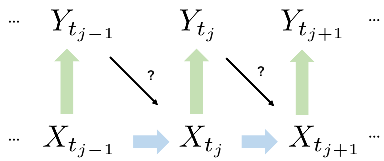
- 時間発展現象で説明される多値反応データ系列なども扱えるが,
- 全データの**同時分布**はだいぶ複雑になる.





Coxの部分尤度と統計的漸近理論

- **部分尤度** (Cox, 1975) の考え方に基づいた統一的な枠組みがある.



- イベント A が生じたときの B の条件つき分布: $P[B|A] := \dfrac{P[A\cap B]}{P[A]}$

$$\log \underbrace{f(\mathbf{x}_1^n, \mathbf{y}_1^n)}_{\text{同時分布}} = \log\left(\prod_{j=1}^n f(y_j|\mathbf{x}_1^j,\mathbf{y}_1^{j-1}) \times \prod_{j=1}^n f(x_j|\mathbf{x}_1^{j-1},\mathbf{y}_1^{j-1})\right)$$

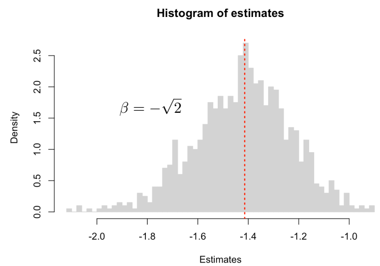
$$= \underbrace{\sum_{j=1}^n \log f(y_j|\mathbf{x}_1^j,\mathbf{y}_1^{j-1})}_{\text{部分尤度}} + \sum_{j=1}^n \log f(x_j|\mathbf{x}_1^{j-1},\mathbf{y}_1^{j-1})$$

$$\mathbf{z}_k^l := (z_k, z_{k+1}, \ldots, z_l)$$

Cox, D. R. (1975). Partial likelihood. *Biometrika*, *62*(2), 269-276.

オルンシュタイン-ウーレンベック過程

- $dX_t = (\mu - \lambda X_t)dt + \sigma dw_t$ （w は標準ブラウン運動)
- $\mathcal{L}(Y_{t_j}|X_{t_j}) = \text{Bernoulli}(\{1+\exp(-\beta X_{t_j})\}^{-1})$



おわりに：回帰モデリングの基本構造

$$f(y;\theta),\ \theta\in\Theta\subset\mathbb{R}^p \quad\Rightarrow\quad \{f(y;\theta(x,\beta)) : \beta\in\mathbb{R}^{p_\beta}\}$$
$$\Rightarrow\quad y\mapsto f\big(y;\theta(x_0,\hat\beta)\big)$$

- 未知パラメータに依存した確率分布 $f(y;\theta)$, $\theta\in\Theta\subset\mathbb{R}^p$
- θ を説明変数 x と別のパラメータ β の関数 $\theta(x,\beta)$ としてモデリングする.
- x に依存した**確率分布族 (統計モデル)** $\{f(y;\theta(x,\beta)) : \beta\in\mathbb{R}^{p_\beta}\}$ を作る.
- データ $D := \{(x_1,y_1),\ldots,(x_n,y_n)\}$ から β の推定量 $\hat\beta = \hat\beta(D)$ を構成する.
- 任意の入力 x_0 に対する出力 y の**予測分布** $y\mapsto f\big(y;\theta(x_0,\hat\beta)\big)$ を構成できる.