政治学方法論 I
ロジスティック（ロジット）回帰（2）

矢内 勇生

神戸大学　法学部/法学研究科

2014 年 12 月 17 日

今日の内容

1. 最尤法によるロジスティック回帰
   - ロジスティック回帰の例
   - 数値計算
2. ロジスティック回帰の当てはまりの評価
   - 当てはまりを評価する

問題の設定

例 1：小選挙区の当落と過去の当選回数（架空のデータ）

過去の当選回数は、小選挙区での当落に影響した？ どの程度影響した？

- 応答変数 y 過去の当選回数別の当選者数
- 説明変数 t (terms)：0 以上の整数

$\rightarrow$ ロジスティック回帰を当てはめる

変数の確認

過去の当選回数 (t_i)	人数 (n_i)	当選者数 (y_i)
0	3	1
1	2	1
2	1	0
3	2	1
4	3	2
5	3	2
6	0	0
7	1	1
合計	15	8

ロジスティック回帰

- この問題をロジスティック回帰として定式化：

$$
\begin{aligned}
p_i &= \Pr(y_i|n_i,\pi_i) = \binom{n_i}{y_i}\pi_i^{y_i}(1-\pi_i)^{n_i-y_i} \\
\pi_i &= \frac{\exp(\beta_1+\beta_2 t_i)}{1+\exp(\beta_1+\beta_2 t_i)} \\
Y_i &\sim \mathrm{Bin}(n_i,\pi_i)
\end{aligned}
$$

- π_i ベルヌーイ試行の成功確率
- Y_i は互いに独立だとする
- 推定する母数：β_1 と β_2

尤度関数の特定

- $\binom{n_i}{y_i} = a_i$ とおく
- 観測値 i に関する尤度関数

$$\begin{aligned} L_i(\boldsymbol{\beta}) &= p_i = a_i \pi_i^{t_i}(1-\pi_i)^{n_i-t_i} \\ &= a_i \left(\frac{\exp(\beta_1+\beta_2 x_i)}{1+\exp(\beta_1+\beta_2 x_i)} \right)^{y_i} \left(\frac{1}{1+\exp(\beta_1+\beta_2 x_i)} \right)^{n_i-y_i} \end{aligned}$$

- y_i が互いに独立だとすると、全体の尤度関数は、

$$\begin{aligned} L(\boldsymbol{\beta}) &= \prod_{i=1}^{n} L_i(\boldsymbol{\beta}) \\ &= \prod_{i=1}^{n} a_i \left(\frac{\exp(\beta_1+\beta_2 x_i)}{1+\exp(\beta_1+\beta_2 x_i)} \right)^{y_i} \left(\frac{1}{1+\exp(\beta_1+\beta_2 x_i)} \right)^{n_i-y_i} \end{aligned}$$

対数尤度関数の特定

- 全体の対数尤度関数（定数項は省略）は、

$$
\begin{aligned}
\log L(\boldsymbol{\beta}) &= \log \prod_{i=1}^{n} L_i(\boldsymbol{\beta}) \\
&= \sum_{i=1}^{n} \log \left(\frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)} \right)^{y_i} \left(\frac{1}{1 + \exp(\beta_1 + \beta_2 x_i)} \right)^{n_i - y_i} \\
&= \sum_{i=1}^{n} \log \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i}
\end{aligned}
$$

- ここから先は、R で計算する

問題の設定

例 2：小選挙区の当落と選挙費用（架空のデータ）

選挙費用（100 万円単位で測定）は、小選挙区での当落に影響した？ どの程度影響した？

- 応答変数 r (response, result)：当選なら 1、落選なら 0
- 説明変数 x (expenditure)：0 以上の連続値（測定単位=100 万円）

$\rightarrow$ ロジスティック回帰を当てはめる

ロジスティック回帰

- この問題をロジスティック回帰として定式化：

$$
\begin{aligned}
\pi_i &= \Pr(r_i = 1) = \frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)} \\
R_i &\sim \mathrm{Bern}(\pi_i)
\end{aligned}
$$

- π_i ベルヌーイ試行の成功確率
- r_i, $(i = 1, 2, \ldots, n)$ は互いに独立だとする
- 推定する母数：β_1 と β_2

尤度関数の特定

- 観測値 i に関する尤度関数

$$\begin{aligned} L_i(\boldsymbol{\beta}) &= \Pr(r_i|\boldsymbol{\beta}, \mathbf{x}) \\ &= \pi_i^{r_i}(1-\pi_i)^{1-r_i} \\ &= \left(\frac{\exp(\beta_1+\beta_2 x_i)}{1+\exp(\beta_1+\beta_2 x_i)}\right)^{r_i}\left(\frac{1}{1+\exp(\beta_1+\beta_2 x_i)}\right)^{1-r_i} \end{aligned}$$

- r_i が互いに独立だとすると、全体の尤度関数は、

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} L_i(\boldsymbol{\beta})$$

- $\boldsymbol{\beta} = [\beta_1, \beta_2]^T$
- $\mathbf{x} = [x_1, \ldots, x_n]^T$

対数尤度関数の特定

- 全体の対数尤度関数は、

$$
\begin{aligned}
\log L(\boldsymbol{\beta}) &= \log \prod_{i=1}^{n} L_i(\boldsymbol{\beta}) \\
&= \sum_{i=1}^{n} \log \left(\frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)} \right)^{r_i} \left(\frac{1}{1 + \exp(\beta_1 + \beta_2 x_i)} \right)^{1-r_i} \\
&= \sum_{i=1}^{n} \log \pi_i^{r_i} (1 - \pi_i)^{1-r_i}
\end{aligned}
$$

- ここから先は、R で計算する

最大値の求め方

- 理想：尤度関数を推定したい母数で微分して、最大値を求める
- 問題：微分した後、簡単に解が求められるとは限らない
- 数値計算 (numerical methods, computation) で最大値を「探す」
  - 二分法
  - 勾配法
  - ニュートン法（ニュートン・ラフソン法）
  - etc.

的中率の計算

- ロジスティック回帰の予測：各観測値が 1 になる「確率」が予測できる
- 本当に知りたいのは、結果が 1 になるか 0 になるか
- 確率を使い、1 になるか 0 になるかを予測する
    1. 確率がある数値を超えたら 1、そうでなければ 0
    2. シミュレーション
- 予測した結果が観測した結果と同じになる割合を計算する
- この割合を当てはまりの良さの指標にする
- 基準点は 0.5（当てずっぽうでも半分は当たるから）

ROC 曲線

- ROC (receiver operating characteristic, 受信者操作特性) 曲線
- 縦軸に「真陽性」の割合 (感度 [sensitivity])、横軸に「偽陽性」の割合 ($1-$ 特異度 [specificity]) をとる
- $\pi > c$ のとき予測値 1、$\pi \leq c$ のとき予測値 0 とする
- c を 1 から 0 まで変化させ、曲線を描く
- 応答変数が完全にランダム：曲線は 45 度線になるはず
- ROC 曲線が左上にあるほど、予測精度の高いモデル
- ROC の下側の面積 (AUC) が大きいほど「良い」モデル

赤池情報量基準 (AIC)

- Akaike Information Criterion (AIC)

$$\mathrm{AIC} = -2\log L(\hat{\theta}) + 2k$$

- k は自由な母数（パラメタ）の数
- AIC が小さいほど「良い」モデル
    - 対数尤度の最大値が大きいほど良い
    - 母数の数が少ないほど良い