政治学方法論 I

第 9 回：ロジスティック回帰分析

矢内 勇生

神戸大学　法学部/法学研究科

2014 年 12 月 3 日

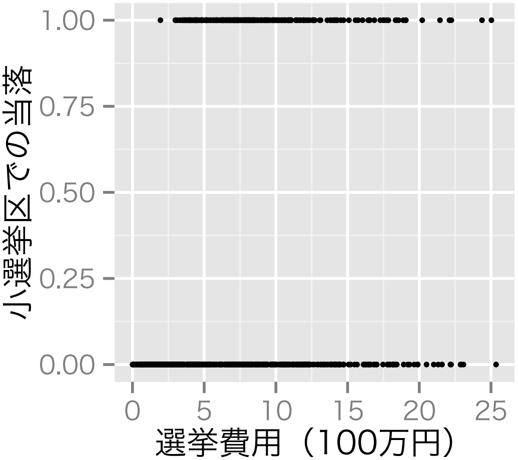
今日の内容

1. ロジスティック回帰とは？
   - イントロダクション
   - 説明変数が 1 つのロジスティック回帰分析
2. ロジスティック回帰を解釈する
   - 係数の意味
   - 統計的推論

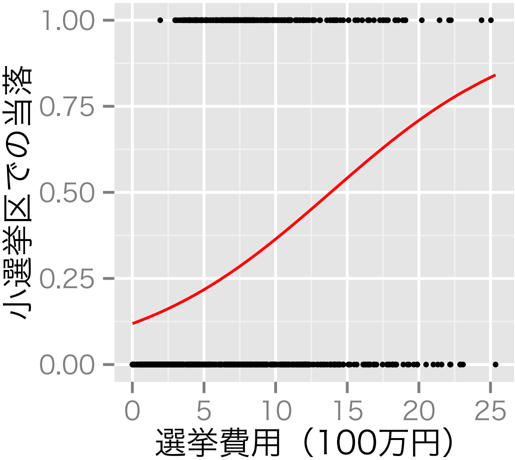
ロジスティック回帰の対象

- 応答変数が二値変数のとき
- 応答変数 $y_i = 0$ or 1 for all i
- 例
    - 内閣を支持するかしないか
    - 投票に参加するか棄権するか
    - 紛争・革命などが起きるか起きないか
    - etc.

二値変数の図示：当落 vs 選挙費用



ロジスティック曲線の当てはめ：当落 vs 選挙費用



ロジスティック回帰モデルの例

- 応答変数：小選挙における候補者の当落：y_i =0（落選） or 1（当選）
- 説明変数：候補者の選挙費用（100 万円）：expm
- このモデルのロジスティック回帰曲線：

$$\Pr(y_i = 1) = \mathrm{logit}^{-1}(-2.00 + 0.14 \cdot \mathrm{expm})$$

ロジスティック回帰モデル

- 応答変数が 0, 1：線形モデル「$X\beta$ + 誤差」はうまく当てはめられない
- 代わりに、応答変数 y が 1 をとる確率をモデル化する

$$\Pr(y_i = 1) = \mathrm{logit}^{-1}(X_i\beta)$$

- 仮定：成功確率 p_i が与えられれば、各 y_i は独立に決まる

$$y_i \sim \mathrm{Bernoulli}(p_i)$$

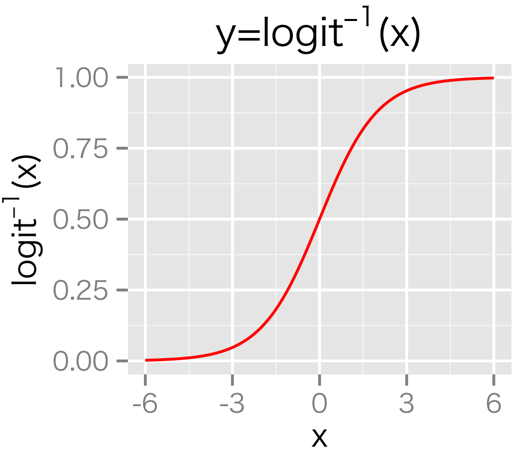
- $X\beta$ を線形予測子 (linear predictor) と呼ぶ

ロジスティック関数 (logistic [inverse logit] function)

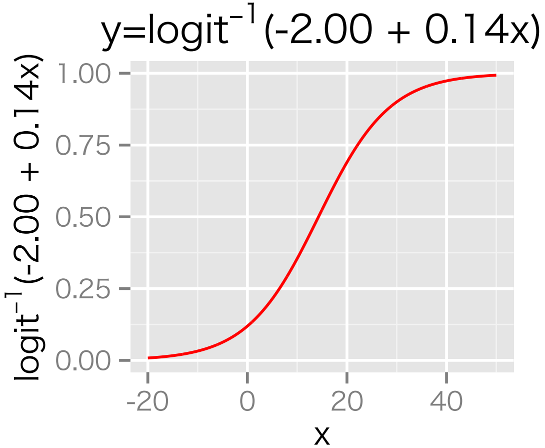
$$\mathrm{logit}^{-1}(x) = \frac{e^x}{1 + e^x} = \frac{1}{1 + \exp(-x)}$$

- 連続変数 $x \in (-\infty, \infty)$ を (0, 1) に写す関数
- 確率を扱うのに適している
- ロジット関数の逆関数なので、logit^{-1} と書く

ロジスティック曲線の例（1）



ロジスティック曲線の例（2）



ロジスティック関数とロジット関数（1）

ロジット関数 (logit function)

- ロジスティック関数の逆関数
- 連続変数 $z \in (0,1)$ を $(-\infty, \infty)$ に写す

$$x_i = \mathrm{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$$

ロジスティック関数とロジット関数（2）

応答変数と線形予測子の関係：2 つの表現法

1. ロジスティック（逆ロジット）関数による表現

$$\Pr(y_i = 1) = \mathrm{logit}^{-1}(X_i\beta) = \frac{1}{1 + \exp(-X_i\beta)}$$

2. ロジット関数による表現

$$\begin{aligned} \Pr(y_i = 1) &= p_i \\ \mathrm{logit}(p_i) &= \log\left(\frac{p_i}{1 - p_i}\right) = X_i\beta \end{aligned}$$

どちらの表現も重要！

ロジスティック「曲線」

- ロジスティック曲線は直線ではない：効果が一定ではない
- x の 1 単位変化に対応する応答変数の成功確率の変化量は一定ではない：x の値によって成功確率の変化量が変わる
  - $\text{logit}(0.5) = 0$, $\text{logit}(0.6) = 0.4$：logit 上の 0.4 単位の変化は、確率上では 50%から 60%への 10 ポイント変化
  - $\text{logit}(0.07) = -2.6$, $\text{logit}(0.1) = -2.2$：logit 上の 0.4 単位の変化は、確率上で 7%から 10%への 3 ポイント変化
- 説明変数の一定変化に対する応答変数の成功確率の変化量：x の値が最小値または最大値に近いときに小さく、中心付近の値のときに大きい

平均値付近のデータを評価する（1）

$$\Pr(\text{当選}) = \text{logit}^{-1}(-2.00 + 0.14 \cdot \text{expm})$$

- ロジスティック曲線の非線形性：どこで評価するかによって効果量が異なる
- 平均値で成功確率の値を評価する
  - $\text{mean}(\text{exam}) = 6.12$
  - $\text{logit}^{-1}(-2.00 + 0.14 \cdot 6.12) = 0.24$

平均値付近のデータを評価する（2）

$$\Pr(\text{当選}) = \mathrm{logit}^{-1}(-2.00 + 0.14 \cdot \mathrm{expm})$$

- 平均値付近で説明変数が成功確率に与える影響を評価する
  - mean(exam) = 6.12：expm=7 と expm=6 を比較する
  - $\mathrm{logit}^{-1}(-2.00 + 0.14 \cdot 7) - \mathrm{logit}^{-1}(-2.00 + 0.14 \cdot 6) = 0.026$
  - 選挙費用を（平均値付近で）1 単位増やすと、当選確率が約 2.6 ポイント上昇する

平均値付近のデータを評価する（3）

$$\Pr(\text{当選}) = \mathrm{logit}^{-1}(-2.00 + 0.14 \cdot \mathrm{expm})$$

- 解析的に評価すると、

$$\frac{d}{dx}\mathrm{logit}^{-1}(-2 + 0.14x) = 0.14\frac{\exp(-2 + 0.14x)}{[1 + \exp(-2 + 0.14x)]^2}$$

- $\bar{x} = 6.12$ を代入すると、

$$\frac{d}{dx}\mathrm{logit}^{-1}(-2 + 0.14 \cdot 6.12) = 0.026$$

係数を「4 で割る」

- ロジスティック曲線の傾き：$X\beta = \alpha + \beta x = 0$ のときに最大
- $\mathrm{logit}^{-1}(0) = 0.5$
- このときの傾きは、

$$\frac{d}{dx}\mathrm{logit}^{-1}(0) = \beta\frac{e^0}{(1+e^0)^2} = \frac{\beta}{4}$$

- 効果の最大値は、係数を 4 で割った値！
- 当選確率の場合：係数は 0.14： $0.14/4 = 0.035 \rightarrow$ 選挙費用が 1 単位の増加は、当選確率を「最大で」3.5 ポイント上げる

オッズ比 (odds ratios)

- 成功確率が p, 失敗確率が $1-p$ のとき、成功のオッズ：

$$\frac{p}{1-p}$$

  - $p = 0.5 \rightarrow$ オッズは 1
  - $p = 1/3 \rightarrow$ オッズは 0.5
- オッズ比：2 つのオッズの比

$$\frac{p_1}{1-p_1} \Big/ \frac{p_2}{1-p_2}$$

- オッズ比を使うメリット：上限値がない

オッズ比でロジスティック回帰を解釈する

- ロジスティック回帰のオッズ：

$$\begin{aligned}\frac{\Pr(y=1|x)}{\Pr(y=0|x)} &= \left(\frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}\right) \Big/ \left(1-\frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}\right)\\ &= \exp(\alpha+\beta x)\end{aligned}$$

- 両辺の自然対数をとると

$$\log\left(\frac{\Pr(y=1|x)}{\Pr(y=0|x)}\right) = \alpha+\beta x = \mathrm{logit}(\alpha+\beta x)$$

- 対数スケールでは、x が 1 単位変化するとオッズが β だけ変化する
- 元のスケールでは、$\exp(\beta)$ の変化
- 例：$\exp(0.14) = 1.15 \rightarrow x$ が 1 単位変化すると、当選確率のオッズは 1.15 倍になる

係数の推定値と標準誤差

- ロジスティック回帰の（1つの）目的：線形予測子内の β を推定する
- 推定法：最尤法 (maximum likelihood method)

[来週の内容]

- β の点推定値 $\hat{\beta}$ から $\pm 2\mathrm{se}$ の範囲にある値はデータと整合的
- 衆院選当落の例：$\hat{\beta} = 0.14$, $\mathrm{se} = 0.01 \rightarrow$ このデータによると、β は $[0.14 \pm 2 \cdot 0.01] = [0.12, 0.16]$ にありそう

統計的有意性

- $\hat{\beta}$ が 0 から 2se 以上離れている：統計的に有意な効果
- 衆院選当落の例：$\hat{\beta} = 0.14$ は統計的に有意で符号は正：選挙費用の増加が当選確率を上昇させるという効果が統計的に認められる
- 切片の有意性は検討しない：興味がないから
- 相互作用の解釈は慎重に

Rで分析してみよう！

来週の内容

最尤法