ロジスティック回帰とその応用に関する研究

2012SE143 丸山拓也
指導教員：木村美善

1　はじめに

ロジスティック回帰の有効性をプロビットモデル, 極値モデルによる分析と比較することによって考察する. ロジスティック回帰については, 多項ロジスティック回帰と順序ロジスティック回帰についても分析を行う. 回帰式の有効性を調べるために, AIC, ピアソン残差統計量を用いる. なお, 解析にはフリーソフト「R」を用いた.

2　ロジスティック回帰

異なるレベルの刺激を与えるある実験において, 各刺激に対して反応した (=1) か否か (=0) の 2 値をとる確率変数を Y とする. 説明変数である刺激 $\mathbf{x}$ に対して反応する確率を p とすると, p =Pr$(Y=1|\mathbf{x})$ と表現できる. このときロジスティック回帰を用いて

$$p=\frac{\exp(\boldsymbol{\beta}^T\mathbf{x})}{1+\exp(\boldsymbol{\beta}^T\mathbf{x})} \tag{1}$$

と定義する. ただし, $\boldsymbol{\beta}$=$(\beta_1,\ldots,\beta_n)^T$, $\mathbf{x}=(x_1,\ldots,x_n)^T$. このとき,

$$\log\frac{p}{1-p}=\boldsymbol{\beta}^T\mathbf{x} \tag{2}$$

である. このロジット変換と呼ばれる変換により, 非線形関数であるロジスティック関数が線形回帰モデルに変換されパラメータ推定が可能になる.([1], [4] 参照).

2.1　オッズ比

p と $1-p$ の偏りを調べる量としてオッズ比が使われている. $\dfrac{p}{1-p}$ と表し, 2 つの結果の片方がもう片方よりも何倍起こりやすいかを意味する. ([1] 参照).

2.2　メディアン有効レベル

ある特定のレベルの反応を起こす刺激レベルを推定するときに用いられる. そのときにメディアン有効モデルを用いて, 推定された回帰曲線から反応する確率が 0.5 になる刺激量 $x_{0.5}$ は

$$\frac{\exp(\beta_0+\beta_1x_1)}{1+\exp(\beta_0+\beta_1x_1)}=\frac{1}{2}\Rightarrow x_{0.5}=-\frac{\beta_0}{\beta_1} \tag{3}$$

により推定される. ([1] 参照).

3　用量反応モデル

3.1　極値モデル

極値モデルは

$$\log[-\log(1-p)]=\beta_1+\beta_2x \tag{4}$$

で定義される. このモデルは p の値が 0.5 の近傍のときはロジスティックモデルに類似しているが, p が 0, 1 のときはズレが生じる. ([1] 参照).

3.2　プロビットモデル

このモデルでは, 正規分布を許容値分布として用いる.

$$p=\Phi(\frac{x-\mu}{\sigma}) \tag{5}$$

ただし, Φ は標準正規分布 $N(0,1)$ の分布関数である. このモデルは, $\beta_1=\dfrac{-\mu}{\sigma}$, $\beta_2=\dfrac{1}{\sigma}$ とおけば,

$$\Phi^{-1}(p)=\beta_1+\beta_2x \tag{6}$$

となるので, 連結関数を標準正規分布の逆累積分布関数とした一般化線形モデルである. ([1] 参照).

4　モデルの推定

4.1　ピアソン残差

今回はピアソン残差を用いて適合度を調査する. 共変量が m 個存在する場合は, m 個の残差が計算され, ピアソン残差は

$$X_k=\frac{y_k-n_k\hat{p}_k}{\sqrt{n_k\hat{p}_k(1-\hat{p_k})}},k=1,\ldots,m \tag{7}$$

で定義される. そして, 近似的に $X^2=\sum_{k=1}^{m}X_k^2\sim\chi^2(m-p)$ となる. ([2] 参照).

4.2　AIC

変数の選択を行う際の基準として, AIC(赤池情報量基準) を用いることがある.AIC は

AIC $=-2\times$(最大対数尤度)$+2\times$(パラメータの数)

と定義される. 最小の AIC をもつモデルが最良のモデルと考える. ([2] 参照).

5　多項ロジスティック回帰モデル

反応変数が 2 つ以上のカテゴリーを持ち, かつカテゴリー間に自然な順序がない場合にロジスティック回帰モデルを拡張する. まず, どれか 1 つのカテゴリーを基準カテゴリー (reference category) として選択する. ここでは第 1 カテゴリーとする. このとき, 他のカテゴリーに対するロジットを, $\mathrm{logit}(p_j)=\log\dfrac{p_j}{p_1}=\mathbf{x}_i^T\boldsymbol{\beta}_j,j=2,\ldots,J$ と定義する. この計 (J-1) 個のロジット式より, パラメータ β_j を推定すると

$$\hat{p}_1=\frac{1}{1+\sum_{j=2}^{J}\exp(\mathbf{x}^T\boldsymbol{\beta}_j)},\hat{p}_j=\frac{\exp(\mathbf{x}^T\boldsymbol{\beta}_j)}{1+\sum_{j=2}^{J}\exp(\mathbf{x}^T\boldsymbol{\beta}_j)}$$

となる. ([1] 参照).

6　順序ロジスティック回帰

もし, 反応変数カテゴリー間に自然な順序が存在すれば, それをモデルに取り組むことが可能である. 概念的にはある状態を測る連続的な潜在変数 z が存在するとき, 潜在変数の代わりであるような J 個のカテゴリーを設け, そのいずれかに振り分けることがしばしば行われる. しかし, これは変数 z に区分点 $C_1, \ldots, C_{J-1}$ を設け, それに従って z の値を計測することに相当する. ([1] 参照).

6.1　比例オッズモデル

パラメータベクトル $\boldsymbol{\beta}_j = (\beta_{0j}, \beta_{1j}, \ldots, \beta_{p-1,j})^T$ において, 切片項 β_{0j} だけが j に依存し, それ以外は j に依存しないような場合,

$$\log \frac{P(z > C_j)}{P(z \leq C_j)} = \beta_{0j} + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1} \qquad (8)$$

と簡略化される. ([1] 参照).

7　多項ロジスティック回帰の解析例

自動車装備に対する嗜好調査として, McFadden らは, 自動車ドライバーに対し, 安全性や自動車装備への嗜好に関する聞き取り調査を実地した ([3] 参照). 総数 300 人に対する調査データにロジスティックモデル

$$\log\left(\frac{p_j}{p_1}\right) = \beta_{0j} + \beta_{1j} x_1 + \beta_{2j} x_2 + \beta_{3j} x_3, j = 2, 3 \quad (9)$$

を当てはめる. 今回は 2 つの基準カテゴリーにより解析を行った. 1 つ目の基準は [1] と同じように,「女性」,「年齢 18-23 歳」を説明変数の基準として行い, 2 つ目は,「男性」,「年齢 > 40」を説明変数の基準カテゴリーとして,

$$x_1 = \begin{cases} 1 & \text{女性} \\ 0 & \text{男性} \end{cases}, x_2 = \begin{cases} 1 & \text{18-24} \\ 0 & \text{それ以外} \end{cases}, x_3 = \begin{cases} 1 & \text{24-40} \\ 0 & \text{それ以外} \end{cases}$$

とした.

表 1 ロジステックモデルを当てはめた結果

パラメータ β	推定値 b(標準誤差)	オッズ比 e^b
β_{02} 定数	0.609(0.365)	
β_{12} 女性	0.388(0.301)	1.47
β_{22}18-24 歳	− 1.588(0.403)	0.20
β_{32}24-40 歳	− 0.459(0.423)	0.63
β_{03} 定数	1.065(0.350)	
β_{13} 女性	0.813(0.321)	2.26
β_{23}18-24 歳	− 2.917(0.423)	0.05
β_{33}24-40 歳	− 1.439(0.416)	0.24
AIC=596.702, X^2=3.926		

1 つ目の解析結果は [1] の解析結果と一致した. 2 つ目の解析結果は表 1 である. 2 つの解析結果を総合的に考察すると, オッズ比により, 年齢が増加するとエアコンおよびパワーステアリングの重要性が高まり, 女性のほうが重要性が高まることがわかる. この結果は [1] で書かれている考察と一致する. そしてピアソン残差の平方和は $X^2 = 3.926$ となり, $\chi^2_4(0.05) = 9.488$ と比較したときこのモデルはよく当てはまっていることがわかる.

8　順序ロジスティック回帰の解析例

多項ロジスティックと同じデータを順序を取り込んで解析する. 比例オッズモデル

$$\log\left(\frac{p_2 + p_3}{p_1}\right) = \beta_{01} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \qquad (10)$$

$$\log\left(\frac{p_1}{p_2 + p_3}\right) = \beta_{02} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \qquad (11)$$

を当てはめる. ([1] 参照). ただし, x_1,x_2,x_3 は前節と同じである. 当てはめた結果が表 2 である.

表 2 各モデルを当てはめた結果

回帰モデル	AIC	X^2
ロジスティック	591.296	4.564
プロビット	591.844	5.168
極値	592.885	6.075

順序ロジスティックモデル（3 つのモデル) は $\chi^2_7(0.05) = 14.067$ より小さい. そして, 多項ロジスティックと比較をすると, どのモデルも χ^2 の当てはまりはよいが, AIC には差が生じた. その結果極値モデルが一番当てはまりが悪いとわかる. このデータの場合, 順序ロジスティックモデルのほうが変数が減るので, 順序を取り込んだロジスティックモデルのほうが良いと考えられる.

9　おわりに

本研究はロジスティック回帰の様々な問題について研究することができ, ロジスティックモデルの有効性を理解できた. さらに, 統計学で重要なピアソン残差, 推定値そして AIC についても学ぶことができた. 大学院でも統計学を専攻していくので, この研究を今後の大学院の勉学に活かしていきたい.

参考文献

[1] Dobson, A.J. : An Introduction To Generalized Linear Models 2nd edition , Chapman & Hall/CRC, 2002(田中豊 訳:一般化線形モデル入門, 共立出版，東京，2008.)

[2] 柏谷英一：一般化線形モデル, 共立出版, 東京，2012.

[3] McFadden, M., Powers, J., Brown, W. and Walker, M. : Vehicle and driver attributes affecting distance from the steering wheel in motor vehicles. Human Factors, 42, 676-668, 2000

[4] 中村永友：多次元データ解析法, 共立出版, 東京，2009.