回帰分析の理論の研究
―ロジスティック回帰を中心に―

2010SE045 平野佑馬
指導教員：木村美善

1　はじめに

3 年次に学んだ重回帰分析の基本は，目的変数が量的変数の場合であり，誤差項には正規性を仮定することが多かった．しかし，実際問題の分析では，目的変数が質的変数や事象の確率，割合の場合が多く，この場合は誤差項に正規性は仮定されない．ロジスティック回帰はこのようなデータを中心にロジット変換や 2 項分布に従う確率分布を用いて，一般化線形モデルで分析することができる．本研究の目的はロジスティック回帰分析を中心に回帰分析の理論と応用について研究することである．またモデルの適合度やモデル（変数）の有意性を考慮して回帰分析を学ぶ．なお，解析にはフリーソフト「R」を用いた．

2　回帰分析

2.1　回帰モデルについて

目的変数を Y，p 個の説明変数を $X_1,\cdots,X_p$ とすると，Y と $X_1,\cdots,X_p$ の関係は，回帰モデルにより次のように近似される．

$$Y = f(X_1,\cdots,X_p) + \varepsilon$$

関数 $f(X_1,\cdots,X_p)$ は Y と $X_1,\cdots,X_p$ の関係を表し，ε は近似によって生じる確率誤差を示している．回帰分析は関数 f を求める手法であり，特に f が $X_1,\cdots,X_p$ の一次式で表される次式のような線形回帰モデルを仮定したものが一般に広く用いられている．

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i, \quad i = 1,\cdots,n$$

回帰分析の主な目的は，未知な値である回帰係数 $\beta_0,\beta_1,\cdots,\beta_p$ を推測することであり，係数を確定し回帰式を求めることによって，説明変数の変化が結果に対してどの程度の影響を及ぼすかを予測することができる．上式のモデルを，行列とベクトルで

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

と書き換えることができる．ただし，$\boldsymbol{Y}$ は目的変数からなる $n\times 1$ ベクトル，$\boldsymbol{X}$ は定数項と説明変数からなる $n\times(p+1)$ 行列，$\boldsymbol{\beta}$ は回帰係数からなる $(p+1)$ ベクトル，$\boldsymbol{\varepsilon}$ は誤差からなる $n\times 1$ ベクトル ([3] 参照)．

2.2　誤差項の諸仮定

重回帰モデルでは，以下の性質を仮定する．

仮定 1　$E(\varepsilon_i) = 0, \quad i = 1,2,\cdots,n$　(不偏性)
仮定 2　$\mathrm{var}(\varepsilon_i) = \sigma^2, \quad i = 1,2,\cdots,n$　(等分散性)
仮定 3　$\mathrm{cov}(\varepsilon_i,\varepsilon_j) = 0, \quad i \neq j$　(無相関性)
仮定 4　$\varepsilon_i \sim N(0,\sigma^2\boldsymbol{I}), \quad i = 1,2,\cdots,n$　(正規性)

仮定 1〜3 を満たすモデルを「線形回帰モデル」，仮定 1〜4 を満たすモデルを「線形正規回帰モデル」という．

この線形回帰モデルにおいては，目的変数 y の分布の変化を適切に捉える説明変数の組を求めることが最も重要な問題であり，変数選択の問題といわれる ([4] 参照)．

3　変数選択について

3.1　情報量規準 AIC

実際のデータを分析する際の有効なモデル選択規準が，情報量規準 AIC である．情報量基準 AIC は

$$\mathrm{AIC} = -2(\text{最大対数尤度}) + 2(\text{モデルの自由パラメータ数})$$

で与えられる．

最大対数尤度とはそのモデルでの対数尤度の最大値であり，自由パラメータとは，そのモデルの含む未知母数の個数である．モデルの尤度関数が適切に定義されているかぎり，どんな統計的問題にも適用可能である．しかしその反面，モデルの分布型を特定化しなくてはならない，という欠点もある ([2]，[4] 参照)．

4　ロジスティック回帰

4.1　ロジスティック回帰モデル

重回帰分析と同様に $g(\boldsymbol{x}_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ip} + \varepsilon_i$（ロジット）のモデルを考える．ただし，$\boldsymbol{x}_i = (x_{i1},\cdots,x_{ip})^t$．$p(\boldsymbol{x}_i) = P(Y_i = 1|\boldsymbol{x}_i)$ を $\boldsymbol{x}_i$ のもとでの $Y = 1$ となる条件付き確率とすると，ロジスティック回帰モデルは

$$p(\boldsymbol{x}_i) = \frac{\exp(g(\boldsymbol{x}_i))}{1+\exp(g(\boldsymbol{x}_i))} = \frac{1}{1+\exp(-g(\boldsymbol{x}_i))}$$

である ([1]，[4]，[5] 参照)．

4.2　対数オッズ (ロジット)

2 水準の場合，p を確率として $p/(1-p)$ をオッズという．これは片方がもう片方の何倍起こりやすいかを意味する．この対数をとった $\log p/(1-p) = \log p - \log(1-p)$ を対数オッズという．オッズの対数をとることをロジット変換という ([1] 参照)．

4.3 メディアン有効レベル

ロジスティック回帰モデルは反応のレベルを調べることができる．特に反応の半分のレベルはメディアン有効レベルと呼ばれる．説明変数が 1 つのとき，メディアン有効レベルは回帰曲線から x を逆推定することができる ([1] 参照).

$$\frac{\exp(\beta_0+\beta_1 x_1)}{1+\exp(\beta_0+\beta_1 x_1)}=\frac{1}{2}\Longrightarrow x_{0.5}=-\frac{\beta_0}{\beta_1}$$

4.4 尤度

$\boldsymbol{\beta}$ の推定のために，独立な n 個の標本を収集したとする．$\boldsymbol{Y}=(Y_1,Y_2,\cdots,Y_n)^t$，$\boldsymbol{y}=(y_1,y_2,\cdots,y_n)^t$，$\boldsymbol{x}_i=(x_{i1},x_{i2},\cdots,x_{ip})^t$ とする一般モデルのもとで，確率変数 $\boldsymbol{Y}$ の観測値が $\boldsymbol{y}$ となる確率は

$$\begin{aligned}L(\boldsymbol{\beta})&=\prod_{i=1}^{n}p(\boldsymbol{x}_i)\\&=\prod_{i=1}^{n}\left(\frac{\exp(g(\boldsymbol{x}_i))}{1+\exp(g(\boldsymbol{x}_i))}\right)\end{aligned}$$

であり，この $L(\boldsymbol{\beta})$ を尤度または尤度関数という．$L(\boldsymbol{\beta})$ を最大にする最尤推定値は，対数をとり対数尤度を求め，加法形式にした上で Newton-Raphson 法などのアルゴリズムにより反復計算をして推定する ([5] 参照).

4.5 逸脱度 (デビアンス)

最尤推定に基づくロジスティック回帰モデルでは次に示す尤度の比（尤度比）の対数の -2 倍の量

$$D=-2\log\left(\frac{\text{モデルの尤度}}{\text{完全にフィットしたモデルの尤度}}\right)$$

を利用する．これは逸脱度 (デビアンス) と呼ばれ，モデルの適合度を総合的に要約して評価する尤度比検定統計量である．逸脱度を重回帰モデルで計算すると残差平方和 SSE に一致する．逸脱度は現在のモデルが正しいという仮説のもとで漸近的に χ^2 分布の上側 $100\alpha\%$ 点より小さければ，有意水準 α で適合度が良くないと判断する根拠が乏しくなる ([5] 参照).

4.6 Nagelkerke の R^2

モデルの当てはまりについては，3.1 節の情報量規準 AIC の他，Nagelkerke の R^2 という値が使われることがある．これは線型回帰の場合に使われる自由度調整済重相関係数の二乗（いわゆる決定係数）を一般化したものである．

$$R^2=\frac{1-(\hat{L_0}/\hat{L})^{2/n}}{1-\hat{L_0}^{2/n}}=\frac{1-\exp((D-D_{null})/n)}{1-\exp(-D_{null}/n)}$$

L は尤度（L_0 は帰無仮説の下での尤度）を示し，D は逸脱度（線型回帰における残差平方和のようなもの），D_{null} は帰無仮説の下での逸脱度である．n はサンプルサイズである．決定係数と同じく 0 から 1 の間の値をとり，モデルがデータのどれくらいの割合を説明しているかを表す指標である ([6] 参照).

5 分析

5.1 データ

Springfield の Baystate 医療センターの 189 人の出生についてのデータであり，低体重出生とそのリスク因子の関連を調べた．目的変数を「低体重出生の有無」，説明変数を「年齢」,「最終月経時体重（ポンド)」,「人種」,「喫煙の有無」,「非熟練労働経験数」,「高血圧の既往」,「子宮神経過敏の有無」,「妊娠の最初の 3 ヶ月の受診回数」,「児の出生時体重 (g)」とし，AIC が最小となる変数を選択した．データは「R」の MASS ライブラリより引用した．

5.2 分析結果

表 1 変数選択後の分析結果 (ロジスティック)

	Estimate	Std.Error	z-value	Pr(> |z|)
(Intercept)	−0.087	0.952	−0.091	0.928
最終月経時体重	−0.016	0.007	−2.320	0.020
黒人	1.326	0.522	2.539	0.011
他の有色人種	0.897	0.434	2.068	0.039
喫煙あり	0.939	0.399	2.354	0.019
非熟練労働経験数	0.503	0.341	1.475	0.140
高血圧既往あり	1.855	0.695	2.669	0.008
子宮神経過敏あり	0.786	0.456	1.721	0.085

表 2 Baystate 医療センターにおける低体重出生リスクのロジスティック回帰分析結果

		95% 信頼区間	
独立変数	オッズ比	下限	上限
人種（白人）			
黒人	3.765	1.355	10.68
他の有色人種	2.452	1.062	5.878
喫煙あり（なし）	2.557	1.185	5.710
高血圧既住あり（なし）	6.392	1.693	27.3
子宮神経過敏あり (なし)	2.194	0.888	5.388

Nagelkerke の R^2 : 0.223，AIC: 217.99，D_{null} : 234.67(自由度 188)，D : 201.99(自由度 181)

6 おわりに

本研究では回帰分析を中心に，ロジスティック回帰分析の理論について理解を深めることができた．

参考文献

[1] 粕谷英一：一般化線形モデル，共立出版，2012.

[2] 小西貞則・北川源四郎：情報量基準，朝倉書店，2004.

[3] 宮川公男：基本統計学，有斐閣出版，1977.

[4] 佐和隆光：回帰分析，朝倉書店，1979.

[5] 丹後俊郎・山岡和枝・高木晴良：ロジスティック回帰分析 — SAS を利用した統計解析の実際—，朝倉書店，1996.

[6] http://www.iic.tuis.ac.jp/edoc/journal/ron/r8-1-2/r8-1-2.pdf