回帰分析の理論とその応用に関する研究
— ロジスティック回帰を中心にして —

2009SE291 上嶌晃司

指導教員:木村美善

1 はじめに

重回帰分析の目的変数は量的変数であり，誤差項には正規性を仮定することが多い. しかし，実際の分析では，目的変数が事象発生の有無を表す質的変数 (カテゴリカルデータ) や事象の発生割合の場合も多く，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. なお，解析にはフリーソフト「R」を用いた.

2 回帰分析

2.1 モデルの定式化

重回帰分析のモデルは，n 個の観測値が与えられた場合，目的変数を $\boldsymbol{y}$, 説明変数を $\boldsymbol{x}_j$ $(j = 1, \ldots, k)$, $\boldsymbol{\varepsilon}$ を誤差項とすると回帰式は

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i \quad i = 1, \ldots, n$$

と表される. ただし，$\beta_0, \beta_1, \ldots, \beta_k$ は回帰係数を表す．以下，これをベクトルで表記する. 目的変数からなる $n \times 1$ 行列を $\boldsymbol{Y}$，定数項と説明変数からなる $n \times (k+1)$ 行列を $\boldsymbol{X}$，回帰係数からなる $(k+1) \times 1$ ベクトルを $\boldsymbol{\beta}$，誤差項の $n \times 1$ ベクトルを $\boldsymbol{\varepsilon}$ とすると

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

と表せる ([4], [5] 参照).

2.2 誤差項の仮定

重回帰モデルでは単回帰分析と同様に ε_i と y_i に以下の仮定を考える.

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

1 から 3 を満たすモデルを「線形回帰モデル」, 1 から 4 を満たすモデルを「線形正規回帰モデル」という ([4], [5] 参照).

3 ロジスティック回帰

3.1 ロジスティック回帰モデル

重回帰分析と同様に $g(\boldsymbol{x}_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$ のモデルを考える．ただし，$\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ik})^t$. 確率 $p(\boldsymbol{x}_i)$ は y_i の $\boldsymbol{x}_i$ に関する条件付き確率 $p(\boldsymbol{x}_i)$ であるので，このときロジスティック回帰モデルは

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

と表せる．回帰モデルの枠組において x_{ij} は説明変数，p は目的変数となり，各変数の定義域は

$$-\infty < x_{ij} < +\infty, \qquad 0 < p < 1$$

となる．この関係式はロジスティック反応関数と呼ばれる ([3], [6] 参照).

3.2 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 ([3], [6] 参照).

3.3 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 ([3] 参照).

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

3.4 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．$\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^t$, $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^t$, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})^t$ 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

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

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推定する ([6] 参照).

3.3 分離

目的変数が 2 値的な質的変数の場合のロジスティック回帰では，説明変数のある値以下では目的変数がすべて 0 で

その値を超えると全て 1 という場合に，係数が大きくなってしまい計算できず，最尤推定値が存在しない (逆の場合も同様)．このようなケースを分離という．分析においては，標本数が小さい，観測値のパターンごとの標本数が極端に異なる状況，つまりアンバランスなデザインの場合に起こる ([2], [6] 参照)．

4　分析

4.1　データ 1

2012 年度プロ野球のレギュラーシーズンにおける中日ドラゴンズの対セ・リーグチームの引き分けを除く 108 試合を対象に，得点，失点と勝敗の関係を調べるため，ロジスティック回帰分析を行った ([1] 参照，データは [7] より)．

4.2　分析結果 1

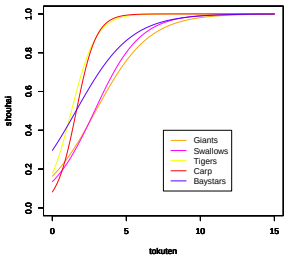

図 1　得点と勝率の関係

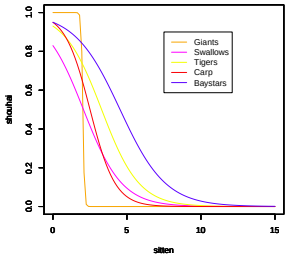

図 2　失点と勝率の関係

中日の勝率が 5 割に達するための中日の得点は，巨人とヤクルトが相手の場合，約 3 点である．しかし阪神，広島，横浜が相手の場合，1 点台の得点であることがわかった．また，対巨人の失点に関する分析では，疑似完全分離が起こった．これは境界である 2 点を除いて勝敗が完全に分離されたためであった．対巨人以外では，得点と失点に関して対広島の回帰係数が大きいことから，対広島は 1 点の試合に対する勝敗への影響が大きい試合であることがわかった．

4.3　データ 2

次に同試合を対象に，初回から 3 回終了時 (序盤)，4 回から 6 回終了時 (中盤)，7 回から試合終了時 (終盤) の各結果を説明変数とし，それぞれ分析を行った．尚，説明変数には「 単打」，「 二塁打」，「 三塁打」，「 本塁打」，「 盗塁」，「 併殺」，「 犠打」，「 犠飛」，「 四死球」，「 被安打」，「 被本塁打」，「 与四死球」，「 失策」，「 敵失策」，「 ホーム」を用い， AIC が最小となる説明変数を選択した ([1] 参照，データは [7] より)．

4.4　分析結果 2

序盤は単打，盗塁でチャンスを作ることが勝敗に影響するが，中盤以降それらがあまり影響しないことが言える．これから，序盤で試合の流れが決まり，先制点を取り逃げ切ることが勝利に繋がると言える．中盤は序盤と同様に本

表 1　分析結果

	序盤		中盤		終盤	
	係数	P 値	係数	P 値	係数	P 値
(切片)	1.150	0.023	0.273	0.529	0.375	0.106
単打	0.326	0.024	–	–	–	–
本塁打	1.583	0.036	1.708	0.035	–	–
盗塁	1.178	0.082	–	–	–	–
犠打	–	–	0.880	0.137	–	–
犠飛	–	–	1.709	0.179	–	–
四死球	–	–	0.363	0.105	0.542	0.016
被安打	−0.450	0.003	−0.197	0.114	–	–
被本塁打	−1.781	0.016	−1.136	0.037	−2.441	0.004
与四死球	−0.740	0.005	–	–	−1.044	0.0004
ホーム	–	–	–	–	0.962	0.044

塁打，被本塁打の影響が大きく，序盤に比べて安打の勝敗への影響は小さくなるが，四死球の影響が大きくなる．また，中盤の犠打，犠飛は勝利に繋がると言える．終盤は本塁打，四死球でしかほとんど勝ちに繋がらず，ホームでの試合が影響することから，安打や本塁打で打ち勝つというよりは四死球で粘り，試合の流れを引き寄せる，そして，ナゴヤドームの大きさや特徴である高いマウンドに慣れた中継ぎや抑え投手陣で逃げ切るという野球であると言える．終盤の中日は本塁打やヒットをあまり打てず，打っても勝利にそれほど繋がらないことが言える．また四死球，与四死球は多くの場面で単打よりも効果的であった．これは投手が崩れるところから得点できるためだと考えられる．

5　おわりに

統一球が導入されて「先制して最後まで守り勝つ」という風潮があるが，分析からも終盤の逆転が困難であることが言えた．安打，本塁打が期待できない状況の中，四球が試合の多くの場面で有効的であった．

本研究でロジスティック回帰分析や関連したその他の理論について理解を深めることができて良かった．

参考文献

[1] 安藤道太: 2010 年度プロ野球球団別の統計的分析，南山大学数理情報学部数理科学科卒業論文，2011.

[2] 粕谷英一: 一般化線形モデル, 共立出版，2012.

[3] 中村永友: 多次元データ解析法, 共立出版，2009.

[4] Rencher A.C and Schaalje G.B: *Linear Models in Statistics*，John Wiley & Sons, Inc，2007.

[5] 佐和隆光: 回帰分析，朝倉書店，1979.

[6] 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.

[7] nikkansports.com http://www.nikkansports.com/.