プロ野球における
勝率に関する統計的分析

2011SE069 池　侑弥
指導教員：白石高章

1　はじめに

本研究では，プロ野球の詳細データを解析していく．2011 年度と 2012 年度では統一球が使用されたため，この 2 年間を除いた 2006 年から 2014 年の 7 年間のデータを解析する. 最終的にこれらの数値的データに基づいてどの指標の組み合わせが最も勝率に影響したのか，またホームとビジターの勝率の差について統計的分析を用いて考察する．

2　データについて

[3] のウェブサイトに載せられている 2006 年から 2010 年，2014 年のプロ野球の詳細データを集めた．各年度のリーグ詳細成績から野球の基本である走攻守の「打率」,「防御率」,「盗塁」など，計 40 個の指標を用意した. 回帰分析の際には，従属変数（目的変数）に「勝率」をおき，その他の指標を説明変数において解析した. またロジスティック回帰分析を用いる為に引き分けを除き「勝ち」= 1,「負け」= 0 として表を作成した.

3　7 年分のホームとビジターの成績比較

表 1 と表 2 はホームとビジターにおける 7 年間の勝敗の合計数を表している. 両リーグともにホームの方が勝ち数が多くなった. 表 3 は中日のホームとビジターの勝敗の数を表している.

表 1.　ホームとビジターの勝敗数 (セ・リーグ)

戦績	勝ち	引分	負け	合計
ホーム	1619	57	1354	3030
ビジター	1329	66	1635	3030

表 2.　ホームとビジターの勝敗数 (パ・リーグ)

戦績	勝ち	引分	負け	合計
ホーム	1617	68	1315	3000
ビジター	1345	59	1596	3000

表 3.　ホームとビジターの勝敗数 (中日)

戦績	勝ち	引分	負け	合計
ホーム	295	10	200	505
ビジター	232	13	260	505

4　重回帰分析による考察

参考文献 [4] を用いて解析する. まずセ・リーグの戦術について解析する. 多重共線性に注意しながら変数増加法を用いて重回帰分析を行うと,「勝率」を従属変数におき「防御率」「OPS」「SP」「捕逸」を説明変数においたモデルが適切となった. 寄与率は 0.938，修正決定係数は 0.931(説明力 93.1 %) とかなり説明力が高く，すべての有意確率が約 0.000 であるので，このモデルは最適と言える. したがってセ・リーグは出塁率と長打率が重要で,「防御率」が低く救援陣の能力が高いチームが勝つ可能性が高いと言える.

次にパ・リーグの戦術について解析していく. セ・リーグと同様に重回帰分析を行うと,「勝率」を従属変数におき「防御率」「得点」「セーブ数」「内野安打率」を説明変数においたモデルが適切となった. 寄与率は 0.824，修正決定数は 0.805(説明力 80.5 %) と説明力が高く，すべての p 値が 0.05 以下となり，このモデルは最適と言える. したがってパ・リーグは得点するために泥臭い野球をし，救援陣が抑えに徹することができるチームが勝つ可能性が高いと言える.

5　主成分分析による考察

重回帰分析の際に選択された変数を使って相関係数行列を用いた主成分分析を行い，主成分得点の散布図を下図に表した. 番号について，$6n+1$, $(n = 0, 1, \cdots, 6)$ が 1 位のチーム，6 の倍数が 6 位のチームである. まずセ・リーグの分析の結果，第 1 主成分の寄与率が 42.6 %で第 3 主成分までの累積寄与率が 90.7 %となった. 第一主成分と各変数との間の主成分負荷量の数値は，防御率との間が 0.803，OPS との間が -0.322，SP との間が -0.891，捕逸との間が -0.402 となった. プロットの結果を考察すると，横軸が総合力を表していることがわかり，主成分負荷量の値から総合力は防御率と SP で決まると考えられる.

次にパ・リーグの分析の結果，第 1 主成分の寄与率が 42.2 %で第 3 主成分までの累積寄与率が 90.2 %となった. 第一主成分と各変数との間の主成分負荷量の数値は，防御率との間が 0.863，得点との間が 0.373，セーブ数との間が -0.780，内野安打率との間が -0.444 となった. プロットの結果を考察すると，横軸が総合力を表していることがわかり，主成分負荷量の値から総合力は防御率とセーブ数で決まると考えられる.

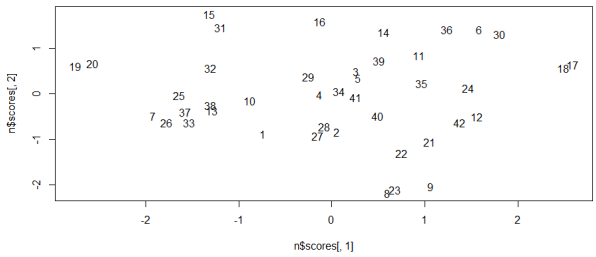

図 1　セ・リーグの主成分得点の散布図

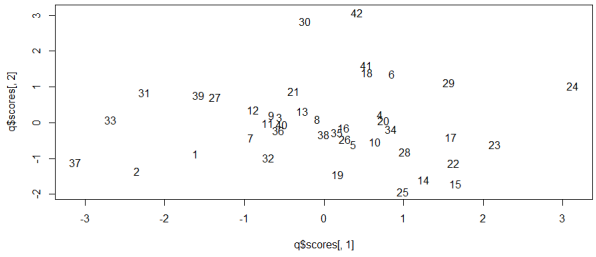

図 2　パ・リーグの主成分得点の散布図

6　ロジスティック回帰曲線による得点と失点についての考察

安藤 [1] では球団別の得点と失点の関係を調べていたので，今回は各年度の 1 位と 6 位の得点と失点の関係を調べ，ロジスティック回帰モデルに当てはめた. そして特徴のある結果となった一部を下図に載せた. 実線が 1 位のチームで，点線のチームが 6 位のチームである.

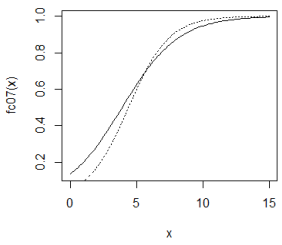

図 3　セ・リーグの 2007 年度の得点 (横軸) と勝率 (縦軸)

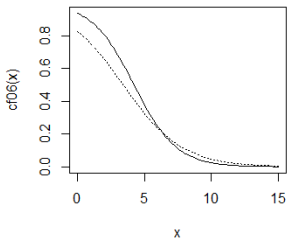

図 4　セ・リーグの 2006 年度の失点 (横軸) と勝率 (縦軸)

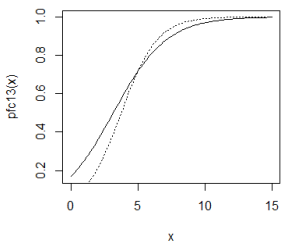

図 5　パ・リーグの 2013 年度の得点 (横軸) と勝率 (縦軸)

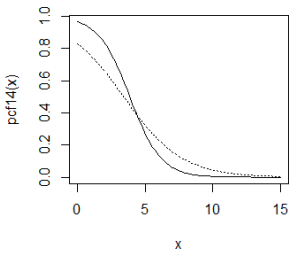

図 6　パ・リーグの 2014 年度の失点 (横軸) と勝率 (縦軸)

図 3 と図 5 により得点に関しては高い得点での勝率が重要ではないことがわかる. むしろ，少ない得点で勝つことのほうが重要で，打ち勝つ必要はないことがわかる. 次に図 4 と図 6 により失点に関しては，少ない失点での勝率が高いほうが重要で，守りで勝てるチームが強いと言える.

7　ホームとビジターの勝率の差の検定

最後に参考文献 [2] を用いて表 3 について解析する.2 標本比率モデルの漸近的な検定により，ホームの方が有利かどうかを検定する.ホームでの勝率を p_1，ビジターでの勝率を p_2とおく.帰無仮説 $H_0 : p_1 = p_2$ vs.対立仮説 $H_2 : p_1 \neq p_2$の検定を考える.また検定統計量 T は，次のようにおける.

$$T \equiv \frac{2\{\arcsin(\sqrt{\hat{p_1}}) - \arcsin(\sqrt{\hat{p_2}})\}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$n_1 = 295$，$n_2 = 232$，$\hat{p_1} = 0.764$，$\hat{p_2} = 0.678$ を上式に代入すると，$T = 3.979$ が求められる.

水準 $\alpha = 0.01$ で検定すると，$z(\alpha/2) = 2.576$ なので $T > z(\alpha/2)$ となり帰無仮説 H_0は棄却される. また信頼係数 0.99 の $\arcsin(\sqrt{p_1}) - \arcsin(\sqrt{p_2})$ に関する漸近的な信頼区間は，

$$0.044 < \arcsin(\sqrt{p_1}) - \arcsin(\sqrt{p_2}) < 0.206$$

したがって，

$$\begin{aligned} & 0 < \arcsin(\sqrt{p_1}) - \arcsin(\sqrt{p_2}) \\ \iff \quad & \arcsin(\sqrt{p_1}) > \arcsin(\sqrt{p_2}) \\ \iff \quad & p_1 > p_2 \end{aligned}$$

よって，ホームでの勝率の方が高くなることがわかり，有利であることが導かれる.

8　おわりに

パ・リーグとセ・リーグどちらも投手力が重要であることがわかった. 大量失点してしまった場合には，敗戦処理投手をうまく使い有力投手を温存し，勝率の高いホームの試合に焦点を合わせ，確実に勝利を積み重ねる必要がある. 144 試合という長期決戦であるためにメリハリをつけた試合運びが重要であるとわかった.

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』2010 年度南山大学情報理工学部情報システム数理学科卒業論文.

[2] 白石高章:『統計科学の基礎』. 日本評論社，東京.

[3] プロ野球ヌルデータ置き場　http://lcom.sakura.ne.jp/NulData/，2014 年 12 月参照

[4] 中村永友:『R で学ぶデータサイエンス 2　多次元データ解析法』. 共立出版，東京.