OPS の計算式における
出塁率と長打率の最適な価値配分
～一見手抜きだが便利な指標～

芝浦工業大学　数理科学研究会
BV17057　西脇 友哉

2019 年 1 月 9 日

目次

第Ⅰ部

導入

研究背景

昨今のプロ野球における**セイバーメトリクス**[*1]の浸透に伴い, 出塁率と長打率の和である **OPS** という打撃指標が定着しつつある. 互いに構造が異なる 2 つの指標を単純に足して得られた OPS は一見すると無意味なものに思えるが, この直感に反して相関係数が 0.9 を優に上回る程に得点との連動性が高い. この性質に私は興味を持ち, 更に「出塁率と長打率を等しい重み付けで総合化する」という計算式の発想が複数の変量を総合して主成分を求める**主成分分析**[*2]の原理に似ていると感じたことで, 「出塁率と長打率を主成分分析の手法に基づいて総合化するとどのような指標が得られるか」という疑問から研究の着手に至った.

先行研究の存在について確認したところ, セイバーメトリクスでは OPS から派生して既に NOI や GPA 等の指標が提唱されており, 更に OPS の計算式において出塁率と長打率を 1:1 の割合で評価する事に対する妥当性の議論もされているようであったが, 私が調べた限りでは OPS と主成分分析の間に関連性を見出した研究は他に見当たらなかった. これを根拠に, 本研究の新規性及び独創性を主張する.

前提

- 本研究は 2018 年 5 月 20 日に開催された第 22 回大宮祭で発表した内容の延長であり, 主な改善点として第 II 部の内容の修正と第 III 部の追加を行った. また, 研究に使用した NPB 公式戦のデータの範囲を, 「'12〜'17」(6 年間) から「'11〜'18」(8 年間) に拡大した.
- 紙面の都合上, NPB 所属球団の略称を以下の通りに表記する. ただし, 2011 年度以前の「横浜ベイスターズ」については, 「横」もしくは「B」と表記する.

パシフィック・リーグ			セントラル・リーグ		
西	L	埼玉西武ライオンズ	広	C	広島東洋カープ
ソ	H	福岡ソフトバンクホークス	ヤ	S	東京ヤクルトスワローズ
日	F	北海道日本ハムファイターズ	巨	G	読売ジャイアンツ
オ	Bs	オリックス・バファローズ	De	DB	横浜 DeNA ベイスターズ
ロ	M	千葉ロッテマリーンズ	中	D	中日ドラゴンズ
楽	E	東北楽天ゴールデンイーグルス	神	T	阪神タイガース

- 野球の指標のうち, 打率, 守備率, 勝率など, 割合を表すものを表記する際には 1 の位の「0」を省略することが一般的である. (例 : 打率 2 割 8 分 6 厘→ .286)
  本研究において, 過去の選手の成績を提示する際には慣例に従い「0」を省略して表記しているが, 計算式中において指標を用いる際には, 数値としての側面を意識し, 省略せずに表記することがある.
- 文章中の (p-q-r) は勝敗成績が「p 勝 q 敗 r 分」であることを表す.
- 文章中では年度を西暦の下 2 桁で表す. (例 : 2018 年→'18)
- 過去の試合結果をはじめ研究に必要なデータについては NPB (日本野球機構) の公式 HP から収集した.

[*1] 1970 年代に B.James により提唱された統計学的な野球の分析手法. 詳細は後述.
[*2] 多変量解析の手法の一つ. 手法の解説は後述.

第 1 章

用語解説

まず本研究に関係する公式記録と指標についての解説をここで行う．全て他の書籍や Web ページ上にも掲載されている内容であるが, 本研究は以下の記録や指標を前提知識としているため, この章を先に読むことが望ましい.

各項目にはそれぞれ規定打席到達者[*1]を対象とした成績と当てはまる人数の度数分布表を掲載した．控え選手の成績が反映されないため, 厳密にリーグの平均などを表すものではないが, 各指標に対するイメージを表から掴むことは出来るはずである.

1.1 公式記録

打率, 打点, 本塁打, 四球, 犠打などが打撃の公式記録に該当する．これらは従来から打者の評価に用いられてきた指標であり, 特に打率, 打点, 本塁打を総称して「打撃 3 部門」とよばれるなど, 長いプロ野球の歴史において確固たる地位を築いてきた．ここでは OPS の導出に用いられる出塁率と長打率の解説を行う．これらもまた公式記録に該当するが, 成績を評価する尺度としては打率等と比べあまり注目されていなかった．現在では, 後述のセイバーメトリクスの影響により重要性が高まっている.

1.1.1 出塁率 (OBP) On-base percentage

出塁率は文字通り「出塁する確率」を表すが, 「アウトにならない確率」と解釈をすることもできる．打率との最大の違いは, 以下の**計算式 2** が表すように四死球による出塁が考慮に入れられている点である.

計算式 1 (打率)**.**

$$\text{打率} = \text{安打} \div \text{打数}$$

計算式 2 (出塁率)**.**

$$\text{出塁率} = (\text{安打} + \text{四球} + \text{死球}) \div (\text{打数} + \text{四球} + \text{死球} + \text{犠飛})$$

野球で点を取る手段は「適時打」「本塁打」「犠飛」など複数存在するが, 全てに共通するのは「3 つ目のアウトを取られる前に走者を本塁へ返さなければならない」ということである．故に攻撃においてアウトにならないことは極めて重要であり, そこに四死球の価値が見出される．かつて出塁率は打率に比べて注目度が低かったが, 近年では打撃指標としての有用性が見直されている.

なお, 犠飛は打数に含まれないため打率の計算では分母から除外されているが, 出塁率の計算においては犠飛は分母に含まれる（凡退として扱われる）ため注意が必要である．四球と犠飛の数によっては出塁率が打率よりも低くなる例が存在しうる.

例 1.1. '00 の荒木雅博選手（中）が打率 .200, 出塁率 .167 を記録している．(打席数 12, 打数 10, 安打数 2, 四死球 0, 犠飛 2)

[*1] 規定打席：所属球団の試合数 ×3.1（小数点以下四捨五入）と定められている．つまり, ペナントレース全 143 試合消化時点での規定打席は 443 打席である.

表 1.1　規定到達者の出塁率 (OBP) の分布 ('11〜'18)

OBP		2018		2017		2016		2015		2014		2013		2012		2011	
以上	未満	パ	セ	パ	セ	パ	セ	パ	セ	パ	セ	パ	セ	パ	セ	パ	セ
.400		4	9	1	0	4	4	3	2	2	4	1	4	1	1	1	0
.380	.400	3	4	1	7	3	2	3	2	5	2	9	1	1	1	1	1
.360	.380	5	7	6	7	3	4	6	5	5	7	6	4	3	4	5	2
.340	.360	4	1	7	2	7	9	7	5	11	7	5	4	9	5	4	4
.320	.340	6	4	1	7	7	4	4	3	4	4	5	5	5	5	7	10
.300	.320	5	5	11	2	3	1	3	5	1	3	5	2	7	5	6	7
	.300	2	1	0	3	1	3	4	2	3	0	2	1	4	3	3	0
計		29	31	27	28	28	27	30	24	31	27	33	21	30	24	27	24

1.1.2　長打率 (SLG) Slugging percentage

長打率はその字面から「長打を打つ確率」と解釈されがちであるが, 長打率が 1.000 を超えない限りは内野安打でも長打率は上昇する. 実際は「1 打数につき稼いだ進塁の数」を意味する指標であり, 計算式 3 によって求められる.

計算式 3 (長打率)**.**

$$\text{長打率} = \text{塁打} \div \text{打数}$$

計算式 4 (塁打)**.**

$$\begin{aligned}\text{塁打} &= \text{単打} \times 1 + \text{二塁打} \times 2 + \text{三塁打} \times 3 + \text{本塁打} \times 4 \\ &= \text{安打} + \text{二塁打} + \text{三塁打} \times 2 + \text{本塁打} \times 3\end{aligned}$$

表 1.2　規定到達者の長打率 (SLG) の分布 ('11〜'18)

SLG		2018		2017		2016		2015		2014		2013		2012		2011	
以上	未満	パ	セ	パ	セ	パ	セ	パ	セ	パ	セ	パ	セ	パ	セ	パ	セ
.550		3	6	1	1	0	4	2	1	2	1	2	3	0	1	1	0
.500	.550	3	5	2	6	4	3	4	1	1	5	4	1	0	0	1	0
.450	.500	8	7	7	1	6	6	5	3	7	5	4	4	4	3	1	4
.400	.450	5	3	6	12	5	5	4	9	8	9	9	6	5	4	6	7
.350	.400	6	9	10	5	5	5	9	6	7	3	6	4	11	8	11	5
.300	.350	3	1	1	2	7	3	4	3	5	4	7	3	6	7	5	5
	.300	1	0	0	1	1	1	2	1	1	0	1	0	4	1	2	3
計		29	31	27	28	28	27	30	24	31	27	33	21	30	24	27	24

長打率が最大となるのは全打数で本塁打を打った場合なので最大値は 4.000 だが, NPB 歴代最高は .779（'13 W. バレンティン選手（ヤ)）である. 打率では安打はすべて等しい価値として計算されるが, 長打率の計算においては単打, 二塁打, 三塁打, 本塁打にそれぞれ異なる重み付けがなされるため, 出塁の頻度が低くても塁を稼げる（=長打力のある）打者を評価することができる.

例 1.2. '17 に鳥谷敬選手（神）は出塁率 .390, 長打率 .377 であったのに対し, 同年の鈴木誠也選手（広）は出塁率 .389, 長打率 .547 であった. この 2 選手の成績を比較すると, 出塁率はほぼ同じであるが, 長打率は鈴木選手が鳥谷選手を大きく上回っている. この事から, 鈴木選手は鳥谷選手よりも「塁を稼ぐ効率が良い」といえる.

1.2　セイバーメトリクス指標

セイバーメトリクス (SABR metrics) は, アメリカ野球学会の略称である「SABR」と測定法を意味する「metrics」を足した造語であり, 1980 年にビル・ジェームズ氏により提唱された統計学的見地による野球の分析手法である. 選手の成績から得点と失点への貢献度, ひいてはその選手に何勝分の価値があるかを計ることを主とし, その分析のために様々な指標を用いる. そして本研究の主題である「OPS」は打撃成績を評価するセイバーメトリクス指標の 1 つである.

以下の項目は OPS に関連のある指標及び勝率の予測に用いられる「ピタゴラス勝率」の解説となっている. 「IsoP」についても便宜上ここで扱っているが, 本研究において直接的な研究対象とはしていない. また, セイバーメトリクス指標は「打撃」「投球」「守備」「走塁」それぞれについて非常に多岐に渡って存在するものであり, この章で解説する指標はそれらの内のほんの一部に過ぎないという点は注意されたい.

1.2.1　OPS (On-base Plus Slugging)

OPS は, 前述の出塁率と長打率を同時に評価することができ, いわゆる「強打者」が高い数値を示す指標である. 理論上の最大値は 5.000（出塁率 : 1.000, 長打率 : 4.000）だが, NPB, MLB ともに例年リーグの平均は .700 程度である. OPS が提唱された背景には「打撃 3 部門」の指標としての欠陥があり, 特に打率に代わる評価基準として, 選手の年俸の査定や獲得選手の選考などに用いられるほどに定着した.

計算式 5 (OPS)**.**

$$\text{OPS} = \text{出塁率} + \text{長打率}$$

上記の式が示すように出塁率と長打率をそのまま足しただけという非常にシンプルな構造でありながら, 得点との連動性の高さに非常に優れている点が他の指標にはない長所である. 文献 [11] によると, NPB の過去 30 年分のシーズン成績によるチーム OPS と 1 試合の平均得点との重相関係数は $R^2 = 0.901$ であった. それに対し出塁率と平均得点では $R^2 = 0.7284$, 長打率と平均得点では $R^2 = 0.8182$ であり, ここから OPS による評価の精度の高さがうかがえる.

欠点としては走塁能力が考慮されていないことなどが挙げられる. この論文では解説を割愛するが, リーグや年度の異なる選手同士の OPS を比較するためにリーグ平均からの傑出度を表した OPS+ や, 評価項目を更に細分化した wOBA, wRC+ などといった指標が現在のセイバーメトリクスの主流とされている. しかしそれらは非常に計算式が複雑であることから, OPS の簡易指標としての有用性は非常に高いと言える.

表 1.3　規定到達者の長打率 (OPS) の分布 ('11〜'18)

OPS		2018		2017		2016		2015		2014		2013		2012		2011	
以上	未満	パ	セ	パ	セ	パ	セ	パ	セ	パ	セ	パ	セ	パ	セ	パ	セ
1.000		1	4	1	0	0	3	1	1	0	1	0	2	0	0	0	0
.900	1.000	4	6	1	3	1	1	2	1	3	3	4	1	0	1	1	0
.800	.900	9	10	5	9	10	9	7	4	8	10	9	7	4	4	3	3
.700	.800	11	7	16	12	9	10	10	12	13	9	10	7	13	7	12	10
.600	.700	2	3	4	3	8	3	9	5	7	4	9	4	9	10	11	9
	.600	2	1	0	1	0	1	1	1	0	0	1	0	4	2	0	2
計		29	31	27	28	28	27	30	24	31	27	33	21	30	24	27	24

1.2.2　NOI (New Offensive Index)

一部には「OPS は出塁率を過小評価している」という意見が存在し, NOI は OPS の見直し作業から考案された. Wikipedia によると, 考案者であるポール・デポデスタ氏は“過去のメジャーでのチームデータを基に,「出塁率と長打率の割合を 3 : 1 にすれば, 両者の和である NOI は OPS よりも実際のチームの総得点数との近似性がさらに高まる」”という推論を導き出した.

計算式 6 (NOI)**.**

$$\mathrm{NOI} = (\text{出塁率} + \text{長打率} \div 3) \times 1000$$

表 1.4　規定到達者の長打率 (NOI) の分布 ('11〜'18)

NOI		2018		2017		2016		2015		2014		2013		2012		2011	
以上	未満	パ	セ	パ	セ	パ	セ	パ	セ	パ	セ	パ	セ	パ	セ	パ	セ
600.00		1	5	1	0	1	4	1	1	0	1	0	3	0	1	0	0
550.00	600.00	5	6	1	4	1	0	4	1	4	4	4	1	0	0	2	0
500.00	550.00	7	8	7	11	11	8	6	7	11	10	12	6	5	4	4	3
450.00	500.00	12	7	10	8	10	11	11	8	8	8	9	7	13	9	9	13
400.00	450.00	2	4	8	4	5	3	5	6	6	4	7	4	8	9	11	7
	400.00	2	1	0	1	0	1	3	1	2	0	1	0	4	1	1	1
計		29	31	27	28	28	27	30	24	31	27	33	21	30	24	27	24

1.2.3　GPA (Gross Production Average)

GPA は, NOI と同様に OPS の改良版という位置づけで考案された打撃指標であり, この指標では“出塁率は約 80% 分長打率よりも価値がある”とされている. また, 計算式において最後に「÷4」という演算が行われているのは, 多くの野球ファンにとって馴染みのある「打率」に近い値を得られるようにするためであるとされている.

計算式 7 (GPA)**.**

$$\mathrm{GPA} = (\text{出塁率} \times 1.8 + \text{長打率}) \div 4$$

表 1.5　規定到達者の長打率 (GPA) の分布 ('11〜'18)

GPA		2018		2017		2016		2015		2014		2013		2012		2011	
以上	未満	パ	セ	パ	セ	パ	セ	パ	セ	パ	セ	パ	セ	パ	セ	パ	セ
.300		6	11	2	4	2	4	4	2	3	4	4	4	0	1	1	0
.280	.300	1	6	4	6	5	7	4	2	3	7	8	3	3	2	3	3
.260	.280	9	4	5	6	9	5	7	8	11	8	6	4	3	5	3	1
.240	.260	9	6	11	8	5	7	7	6	7	4	7	6	13	6	7	10
.220	.240	2	3	5	2	5	3	5	4	5	4	6	3	4	7	11	7
.200	.220	1	0	0	1	2	0	2	2	2	0	1	1	5	3	2	3
	.200	1	1	0	1	0	1	1	0	0	0	1	0	2	0	0	0
計		29	31	27	28	28	27	30	24	31	27	33	21	30	24	27	24

1.2.4　IsoP (Isolated Power)

計算式 8 (IsoP)**.**

$$\begin{aligned}\text{IsoP} &= \text{長打率} - \text{打率} \\ &= (\text{二塁打} + \text{三塁打} \times 2 + \text{本塁打} \times 3) \div \text{打数}\end{aligned}$$

公式記録の長打率は重要度の高い指標であるが, その名の通りに純粋な長打力を表す訳ではない. IsoP は長打率から単打の要素を排除することで長打力の評価を可能にしている.

表 1.6　規定到達者の長打率 (IsoP) の分布 ('11〜'18)

IsoP		2018		2017		2016		2015		2014		2013		2012		2011	
以上	未満	パ	セ	パ	セ	パ	セ	パ	セ	パ	セ	パ	セ	パ	セ	パ	セ
.300		2	3	0	0	0	2	0	0	1	0	0	2	0	0	1	0
.250	.300	1	3	2	2	2	2	3	1	1	2	0	1	0	0	0	0
.200	.250	8	4	7	4	5	6	4	3	4	3	6	1	2	1	1	1
.150	.200	4	6	7	8	2	4	5	5	5	8	6	6	2	3	3	6
.100	.150	8	11	6	6	8	7	12	9	8	9	7	6	12	10	9	6
	.100	6	4	5	8	11	6	6	6	12	5	14	5	14	10	13	11
計		29	31	27	28	28	27	30	24	31	27	33	21	30	24	27	24

1.2.5　ピタゴラス勝率

ピタゴラス勝率とは, セイバーメトリクスの考案者であるビル・ジェームズ氏が提唱した式であり, 得点と失点から勝率を推定するために用いられる. 式の形がピタゴラスの定理（三平方の定理）と似ていることからピタゴラス勝率と命名されたと言われている.

計算式 9 (ピタゴラス勝率)**.**

$$\text{ピタゴラス勝率} = \frac{(\text{総得点})^2}{(\text{総得点})^2 + (\text{総失点})^2}$$

本来, ピタゴラス勝率は実際の勝率と比較することでチームの勝率が得点力や投手力に対して順当であるかを考察する[*2]という利用方法が一般的であるが, 本研究では第 III 部において実際の勝率をピタゴラス勝率で近似し, 得点数と失点数から勝率を予測するという目的でピタゴラス勝率を利用する.

[*2] 例として, あるチームの勝率がピタゴラス勝率よりも極端に低い場合, 「接戦での敗戦が多い」「継投策などの采配面に問題がある」などの原因が考えられる.

第 II 部

出塁率と長打率の価値配分

第 2 章

主成分分析による出塁率と長打率の総合化

2.1 方針

研究背景で述べた様に OPS の仕組みから着想し, 主成分分析により「出塁率と長打率を総合して評価する」ことを主目的とする. 1.2.1 で紹介した通り OPS は出塁率と長打率を 1 : 1 の割合で評価する指標であるが, 主成分は「観測データの持つ情報を最も良く表現できるように」生成された式であるため, 主成分分析によって OPS とは異なる比率で出塁率と長打率を評価する式が得られるはずである.

本研究では得られた主成分を数式であると同時に指標の一つとして扱う事とする. その上で,「OPS」「NOI」「GPA」「第 1 主成分得点」「第 2 主成分得点」という 5 指標それぞれについてチームの 1 試合平均得点との相関係数を計算し, どの指標が優れているかについて考察する. 主成分分析の手法については 2.2 で, 主成分得点については 2.5.1 でそれぞれ解説を行っている.

2.2 主成分分析の手順

主成分分析とは, 観測された多次元データのもつ情報をできるだけ失わないように（=データ間の違いを最もよく表現できるように）説明変数を合成し, 説明変数の線形結合で表される新たな変数を定義する方法である. 2 次元データをある軸上へ射影することを考えるとき, 分散を最大化する射影軸を**第 1 主成分**とよび, 第 1 主成分と直交するという条件の下で分散を最大化する射影軸を**第 2 主成分**とよぶ.

いま, 2 つの変数 $\boldsymbol{x} = (x_1, x_2)^T$ について観測された n 個の 2 次元データを

$$\boldsymbol{x}_1 = \begin{pmatrix} x_{11} \\ x_{12} \end{pmatrix}, \boldsymbol{x}_2 = \begin{pmatrix} x_{21} \\ x_{22} \end{pmatrix}, \cdots, \boldsymbol{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \end{pmatrix}, \cdots, \boldsymbol{x}_n = \begin{pmatrix} x_{n1} \\ x_{n2} \end{pmatrix}$$

とする. 次に, 説明変数の線形関数

$$y = w_1 x_1 + w_2 x_2 = \boldsymbol{w}^T \boldsymbol{x}$$

を考え, 主成分 y の分散 s_y^2 が最大になるような係数ベクトル $\boldsymbol{w} = (w_1, w_2)^T$ を求める. ただし,

$$\boldsymbol{w}^T \boldsymbol{w} = w_1^2 + w_2^2 = 1$$

という条件が生じるものとする.

公式 1 (分散の公式)**.**

$$\mathrm{Var}(ax+by) = a^2\mathrm{Var}(x) + b^2\mathrm{Var}(y) + 2ab\mathrm{Cov}(x,y) \quad (a,b:\text{const.})$$

証明.

$$\begin{aligned}
\mathrm{Var}(ax+by) &= \frac{1}{n}\sum_{i=1}^{n}\Big\{ax_i+by_i-\Big(\overline{ax+by}\Big)\Big\}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\big\{ax_i+by_i-(a\overline{x}+b\overline{y})\big\}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\big\{a(x_i-\overline{x})+b(y_i-\overline{y})\big\}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\big\{a^2(x_i-\overline{x})^2+b^2(y_i-\overline{y})^2+2ab(x_i-\overline{x})(y_i-\overline{y})\big\} \\
&= a^2\cdot\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2+b^2\cdot\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2+2ab\cdot\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}) \\
&= a^2\mathrm{Var}(x)+2ab\mathrm{Cov}(x,y)+b^2\mathrm{Var}(y)
\end{aligned}$$

□

この公式を利用して分散 s_y^2 を式で表すと,

$$\begin{aligned}
s_y^2 &= \mathrm{Var}(w_1x_1+w_2x_2) && (2.1)\\
&= w_1^2\cdot\mathrm{Var}(x_1)+w_2^2\cdot\mathrm{Var}(x_2)+2w_1w_2\cdot\mathrm{Cov}(x_1,x_2) && (2.2)
\end{aligned}$$

となる. w_1, w_2 を求めるためには, $Q(w_1,w_2) := s_y^2$ として,

$$Q(w_1,w_2) = w_1^2\cdot\mathrm{Var}(x_1)+w_2^2\cdot\mathrm{Var}(x_2)+2w_1w_2\cdot\mathrm{Cov}(x_1,x_2) \tag{2.3}$$

を最大にする最適化問題を解けば良い. そこで, ラグランジュの未定乗数法を用いて, ラグランジュ関数を次のように定義する.

$$\begin{aligned}
F(w_1,w_2) := &\quad Q(w_1,w_2)-\lambda(w_1^2+w_2^2-1) && (2.4)\\
= &\quad w_1^2\cdot\mathrm{Var}(x_1)+w_2^2\cdot\mathrm{Var}(x_2)+2w_1w_2\cdot\mathrm{Cov}(x_1,x_2)-\lambda w_1^2-\lambda w_2^2+\lambda && (2.5)
\end{aligned}$$

解はラグランジュ関数の停留点を求めることで得られるので, (2.5) を w_1, w_2 で偏微分して 0 とおくと

$$\begin{aligned}
\frac{\partial}{\partial w_1}F(w_1,w_2) &= 2(\mathrm{Var}(x_1)\cdot w_1+\mathrm{Cov}(x_1,x_2)\cdot w_2-\lambda w_1)=0\\
\frac{\partial}{\partial w_2}F(w_1,w_2) &= 2(\mathrm{Cov}(x_1,x_2)\cdot w_1+\mathrm{Var}(x_2)\cdot w_2-\lambda w_2)=0
\end{aligned}$$

したがって,

$$\begin{cases}\mathrm{Var}(x_1)\cdot w_1+\mathrm{Cov}(x_1,x_2)\cdot w_2 &= \lambda w_1\\ \mathrm{Cov}(x_1,x_2)\cdot w_1+\mathrm{Var}(x_2)\cdot w_2 &= \lambda w_2\end{cases} \tag{2.6}$$

が得られ, この式を行列で表現すると

$$\begin{pmatrix}\mathrm{Var}(x_1) & \mathrm{Cov}(x_1,x_2)\\ \mathrm{Cov}(x_1,x_2) & \mathrm{Var}(x_2)\end{pmatrix}\begin{pmatrix}w_1\\w_2\end{pmatrix}=\lambda\begin{pmatrix}w_1\\w_2\end{pmatrix} \tag{2.7}$$

となる. ここでの $\lambda, \boldsymbol{w}$ はそれぞれ以下に定義する**分散共分散行列** S の固有値 λ, 固有ベクトル $\boldsymbol{w}$ に相当し, 線形代数学の**固有値問題**でよくみられる $S\boldsymbol{w}=\lambda\boldsymbol{w}$ の形と対応している.

定義 2.1 (2×2 の分散共分散行列)**.**

$$S = \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix}, \quad s_{jk} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)(x_{ik} - \overline{x}_k), \quad j,k = 1,2.$$

$$s_{11} = \mathrm{Var}(x_1), \quad s_{12} = s_{21} = \mathrm{Cov}(x_1, x_2), \quad s_{22} = \mathrm{Var}(x_2)$$

より,

$$S = \begin{pmatrix} \mathrm{Var}(x_1) & \mathrm{Cov}(x_1, x_2) \\ \mathrm{Cov}(x_1, x_2) & \mathrm{Var}(x_2) \end{pmatrix}$$

となる. つまり,

$$\begin{pmatrix} x_1 \text{ の分散} & x_1 \text{ と } x_2 \text{ の共分散} \\ x_1 \text{ と } x_2 \text{ の共分散} & x_2 \text{ の分散} \end{pmatrix}$$

を表す.

S の固有値と固有ベクトルを求めることで, 最大固有値 λ_1 に対応する固有ベクトル $\boldsymbol{w}_1 = (w_{11}, w_{12})^T$ から式 (2.8) が示す第 1 主成分が得られ, 第 2 番目の固有値 λ_2 に対応する固有ベクトル $\boldsymbol{w}_2 = (w_{21}, w_{22})^T$ から式 (2.9) が示す第 2 主成分が得られる.

$$y_1 = w_{11}x_1 + w_{12}x_2 = \boldsymbol{w}_1^T\boldsymbol{x} \tag{2.8}$$

$$y_2 = w_{21}x_1 + w_{22}x_2 = \boldsymbol{w}_2^T\boldsymbol{x} \tag{2.9}$$

2.3　対象データ

1.2.1 では OPS を打者個人の成績として紹介したが, 個人の成績を対象とする場合得点との相関関係を調べることが難しくなるため, 以下では分析の対象をチームの成績とする.[*1] 具体的には, NPB の公式 HP に掲載されている'11〜'18 のペナントレースにおける各球団の打撃成績から「チームの出塁率」「チームの長打率」などを引用する. すなわち, データの個数は 12(球団) × 8(年) = 96(個) となる. 図 2.1 に出塁率と長打率の関係を表した. 観測データの具体的な数値については, 末尾の補足資料を参照されたい.

2.4　数値の代入と主成分の計算

以下, この節では特に断りがない限り x_1 **は出塁率**, x_2 **は長打率**を表すものとし, また $n = 96$ であるとする. 式 (2.2) に関して, 実際に行われたプロ野球の試合のデータから,

$$\begin{cases} \mathrm{Var}(x_1) & = 0.000185 \\ \mathrm{Var}(x_2) & = 0.000822 \\ \mathrm{Cov}(x_1, x_2) & = 0.000291 \end{cases} \tag{2.10}$$

が得られた. これらを代入し, 具体的に主成分を求める. 式 (2.7) に式 (2.10) を代入すると,

$$\begin{pmatrix} 0.000185 & 0.000291 \\ 0.000291 & 0.000822 \end{pmatrix}\begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \lambda\begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$$

[*1] 一応打撃・走塁成績による得点への貢献度を計る「RC (Runs Created)」という指標が存在するため不可能ではないが, RC は計算式が極めて複雑であり, その計算に労力をかけることが, 主目的である OPS の考察に支障をきたしかねないという結論に至った.

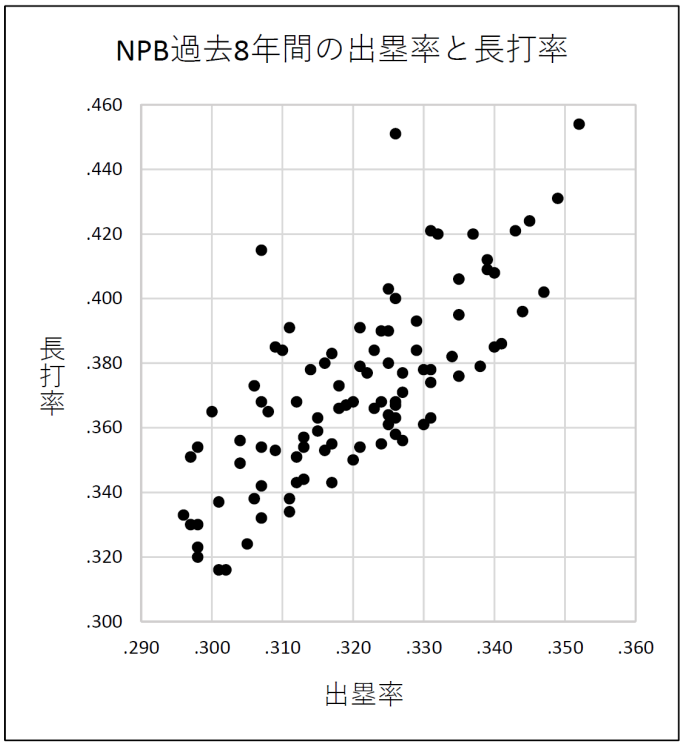

図 2.1　球団ごとの出塁率と長打率 (’11〜’18)

が得られる. $w_1^2 + w_2^2 = 1$ の条件の下, Mathematica を用いて固有値を求めると, 固有値は

$$\lambda = 0.00007208, 0.00093492$$

となった.

2.4.1　第 1 主成分の導出

最大固有値 $\lambda_1 = 0.00093492$ に対する固有ベクトルは

$$\boldsymbol{w}_1 = \begin{pmatrix} 0.36176 \\ 0.93227 \end{pmatrix}$$

である. よって式 (2.8) より, 第 1 主成分は

$$y_1 = 0.36176x_1 + 0.93227x_2 \tag{2.11}$$

となる. 第 1 主成分の寄与率は,

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{0.00093492}{0.00093492 + 0.00007208} = 92.8(\%)$$

であり, ここから第 1 主成分がサンプルの様子を良く表現できているということが読み取れる.

2.4.2　第 2 主成分の導出

2 番目に大きい固有値 $\lambda_2 = 0.00007208$ に対する固有ベクトルは

$$\boldsymbol{w}_2 = \begin{pmatrix} -0.93227 \\ 0.36176 \end{pmatrix}$$

となる. よって式 (2.9) より, 第 2 主成分は

$$y_2 = -0.93227x_1 + 0.36176x_2 \tag{2.12}$$

となる．第 2 主成分の寄与率は,

$$\frac{\lambda_2}{\lambda_1+\lambda_2}=\frac{0.00007208}{0.00093492+0.00007208}=7.2(\%)$$

である．なお, 第 2 主成分の係数ベクトルは第 1 主成分の係数ベクトルと直交している．また, 求めた係数ベクトルは確かに $w_1^2+w_2^2=1$ を満たしている．

2.5　求めた主成分が持つ意味

2.5.1　主成分得点

第 1 主成分 y_1 及び第 2 主成分 y_2 の式に x_1, x_2 のデータの値を代入して得た具体的な数値をそれぞれ**第 1 主成分得点**, **第 2 主成分得点**とよぶ．以下, 第 1 主成分得点を p_1, 第 2 主成分得点を p_2 とする．

2.5.2　因子負荷量による判断

用語 1 (因子負荷量)**.** 因子負荷量とは, 因子分析において, 得られた共通因子が分析に用いた変数（観測変数）に与える影響の強さを表す値で, 観測変数と因子得点との相関係数に相当する．
-1 以上 1 以下の値をとり, 因子負荷量の絶対値が大きいほど, その共通因子と観測変数の間に（正または負の）強い相関があることを示し, 観測変数をよく説明する因子であると言える．(統計 WEB より)

主成分にどのような意味付けがなされるかは, 因子負荷量の大きさや符号から判断できる．因子負荷量が正 (負) ならば, その変数が増えると主成分の値は増える (減る)．本研究において観測変数は出塁率と長打率に相当し, 因子負荷量は表 2.1 のようになる．ここから, 第 1 主成分は出塁率と長打率を総合して評価する射影軸を表し, 第 2 主成分は各球団が出塁率と長打率のどちらを強みとしているかを表すものと解釈することができる.[*2]

表 2.1　因子負荷量

	出塁率	長打率
第 1 主成分	0.815	0.994
第 2 主成分	-0.580	0.106

[*2] 出塁率が増加すると p_2 が減少するため, p_2 が大きい球団は出塁率を強みとし, p_2 が小さい球団は長打率を強みとすることが分かる．

第 3 章

指標と得点の相関

3.1　相関係数

2.1 に基づき, 「OPS」「NOI」「GPA」「第 1 主成分得点」「第 2 主成分得点」の 5 指標それぞれについてチームの 1 試合平均得点との相関係数を計算したところ, 以下の表 3.1 の様になった. (比較対象として「打率」「出塁率」「長打率」についても相関係数を計算した.) この表から, OPS, NOI, GPA, 第 1 主成分得点とチーム平均得点との強い相関がみられたが, 第 2 主成分得点とチーム平均得点との間にはほとんど相関が無かったことがわかる.

表 3.1　チームの 1 試合あたりの平均得点との相関係数

指標	相関係数 r	指標	相関係数 r	主成分得点	相関係数 r
打率	0.817	OPS	0.968	p_1	0.954
出塁率	0.876	NOI	0.959	p_2	-0.170
長打率	0.931	GPA	0.968		

3.2　母相関係数の差の検定

3.2.1　2 種の相関係数を比較するために

前節の結果を基に, 得点との相関に優れた指標がどれであるかを判断したい. しかし, 表 3.1 が示す相関係数はプロ野球データ 8 年分という標本から計算した相関係数の推定値にすぎず, 各項目についての母相関係数は不明である. 標本相関係数の数値を単純比較することでは母集団の相関の強さの判断は出来ないとされているため, 母相関係数の差に関する仮説検定を行う必要がある. 用いる検定方法は次の通りである.

母相関係数の差の検定

標本サイズ n_A の 2 変数の観測値 $(x_i^A, y_i^A), i = 1, 2, \ldots, n_A$ から得られた相関係数を r_A, 標本サイズ n_B の 2 変数の観測値 $(x_i^B, y_i^B), i = 1, 2, \ldots, n_B$ から得られた相関係数を r_B とし, それぞれの母相関係数を ρ_A, ρ_B とするとき, 母相関係数の差の検定における帰無仮説 H_0 および対立仮説 H_1 は,

$$\text{帰無仮説 } H_0 : \rho_A = \rho_B$$
$$\text{対立仮説 } H_1 : \rho_A \neq \rho_B$$

である. また, 検定統計量 z_0 は,

$$z_0 = \frac{\dfrac{1}{2}\left\{\ln\left(\dfrac{1+r_A}{1-r_A}\right) - \ln\left(\dfrac{1+r_B}{1-r_B}\right)\right\}}{\sqrt{\dfrac{1}{n_A - 3} + \dfrac{1}{n_B - 3}}} \tag{3.1}$$

で与えられる. ここで ln は自然対数を表す. 検定統計量 z_0 は帰無仮説のもとで標準正規分布に従う.

以下の検定では, 検定統計量 z_0 の式に対し, 2.3 節より共通して $n_A = n_B = 96$ が代入されるものとする.

3.2.2　OPS vs. 打率

まず「OPS と得点の相関係数」と「打率と得点の相関係数」が統計的に有意な差を持つことを確かめたい. 本研究では OPS の方が打率よりも得点との相関に優れていることを前提として OPS を導入しているため, 今回は**片側検定**を用いる. 片側検定では, 検定統計量 z_0 に対して, 棄却限界値として $z(\alpha)$ を用いる. $z_0 > z(\alpha)$ ならば有意であり, そうでなければ有意でない. なお, 有意水準 $\alpha = 0.01$ とする.

ここでは, 式 (3.1) において

$$\begin{cases} x_i^A : \text{OPS} \\ y_i^A : \text{平均得点} \end{cases} \qquad \begin{cases} x_i^B : \text{打率} \\ y_i^B : \text{平均得点} \end{cases}$$

という対応関係をとり, この検定における帰無仮説 H_0 および対立仮説 H_1 は,

$$\text{帰無仮説 } H_0 : \rho_A = \rho_B$$
$$\text{対立仮説 } H_1 : \rho_A > \rho_B$$

である. 表 3.1 より $r_A = 0.968, r_B = 0.817$ を式 (3.1) に代入すると, $z_0 = 6.221$ が得られる. 棄却限界値 $z(0.01) = 2.326$ を用いると, $z_0 > z(0.01)$ より, 有意水準 $\alpha = 0.01$ で帰無仮説 H_0 が棄却される. よって, **OPS と得点の相関係数は打率と得点の相関係数よりも有意に高い数値をとる**.

3.2.3　第 1 主成分 vs. 打率

続いて, 「第 1 主成分と得点の相関係数」と「打率と得点の相関係数」が統計的に有意な差を持つという仮説に対し有意水準 $\alpha = 0.01$ の片側検定を行う. この検定における帰無仮説 H_0 および対立仮説 H_1 は 3.2.2 と同様であるが, 式 (3.1) においては

$$\begin{cases} x_i^A : \text{第 1 主成分} \\ y_i^A : \text{平均得点} \end{cases} \qquad \begin{cases} x_i^B : \text{打率} \\ y_i^B : \text{平均得点} \end{cases}$$

という対応関係をとるものとする. 表 3.1 より $r_A = 0.954, r_B = 0.817$ を式 (3.1) に代入すると, $z_0 = 4.997$ が得られる. 棄却限界値 $z(0.01) = 2.326$ を用いると, $z_0 > z(0.01)$ より, 有意水準 $\alpha = 0.01$ で帰無仮説 H_0 が棄却される. よって, **第 1 主成分と得点の相関係数は打率と得点の相関係数よりも有意に高い数値をとる**.

3.2.4　OPS vs. 第 1 主成分

次に,「OPS と得点の相関係数」と「第 1 主成分と得点の相関係数」が統計的に有意な差を持つかどうかを検定によって明らかにしたい. 表 3.1 から分かるように両者の標本相関係数の差は僅かであり, この段階では「どちらがより強い相関関係を持ちそうであるか」という指針を立てられないので, 今回は有意水準 $\alpha = 0.05$ として**両側検定**を行う.

両側検定では, 標準正規分布が平均 0 に対して対称分布であることを利用して, 検定統計量の絶対値 $|z_0|$ に対して, 棄却限界値として $z(\alpha/2)$ を用いる. $|z_0| > z(\alpha/2)$ ならば有意であり, そうでなければ有意でない.

ここでは, 式 (3.1) において

$$\begin{cases} x_i^A : \text{OPS} \\ y_i^A : \text{平均得点} \end{cases} \qquad \begin{cases} x_i^B : \text{第 1 主成分} \\ y_i^B : \text{平均得点} \end{cases}$$

という対応関係をとり, この検定における帰無仮説 H_0 および対立仮説 H_1 は,

$$\text{帰無仮説 } H_0 : \rho_A = \rho_B$$
$$\text{対立仮説 } H_1 : \rho_A \neq \rho_B$$

である. 表 3.1 より式 (3.1) に $r_A = 0.968, r_B = 0.954$ を代入すると, $z_0 = 1.292$ が得られる. 棄却限界値 $z(0.025) = 1.960$ を用いると, $|z_0| < z(0.025)$ より, 有意水準 $\alpha = 0.05$ で帰無仮説 H_0 が受容される. よって, **OPS と得点の相関係数は第 1 主成分と得点の相関係数と有意に異なる数値をとるとはいえない**[*1].

3.3　第 II 部 総括

3.3.1　主成分分析による研究結果

得られた主成分のうち, OPS のように「出塁率と長打率を総合的に評価するもの」として用いることができるのは第 1 主成分の方となった. では, 果たして第 1 主成分は指標として有用であると言えるだろうか.

まず, 3.2.3 より「第 1 主成分と得点の相関係数は打率と得点の相関係数よりも有意に高い数値をとる」ことが分かったため, 指標としての最低限の価値は保ったと言える. しかし, OPS と比較した際には, 3.2.2 より, 相関の強さに優劣をつけることは出来なかった. また, 第 1 主成分に複雑な係数が付いていることをある種の致命的な欠点であると見なすと, 出塁率, 長打率ともに係数が 1 である OPS に比べると**簡便性では明らかに劣っている**と判断できる.

これらの研究結果から, 万能ではないものの, OPS がいかに優れた指標であるかを再確認することができた.

3.3.2　OPS を用いた研究の発展性

第 II 部で OPS の利便性の高さについて触れたが, 次の第 III 部ではそれをより生かすべく, 分析への利用範囲を攻撃面からさらに拡大することを考える. OPS の考察をチーム勝率の予測へ展開させることで, 最終的には「2019 年度以降に贔屓の球団を優勝させるためには出塁率や長打率をどれ程上昇させれば良いか」という分析を可能にしたい. そこで 2 通りの分析方法を考案し, それぞれ「**手法 A**」「**手法 B**」と名付けた. 以下, 手法 A については第 4 章で, 手法 B については第 5 章で解説を行う.

[*1] つまり, $r_A = 0.968$ と $r_B = 0.954$ の差は統計的に意味を持たない.

第 III 部

出塁率・長打率向上目標の設定

第 4 章

手法 A ～勝率予測への展開～

3.3.2 小節で提示した 2 つの手法のうち, **手法 A** では, OPS から勝率の予測を行い, 勝率の変動に関する考察を攻撃面に帰着させるという方法をとる. 具体的には, 以下のような流れでの分析を考える.

1. **優勝ラインを設定する**（解説：4.1）
   まず優勝するために必要になる年間勝率（これを**優勝ライン**と呼ぶことにする）を定め, チーム成績の目標を立てる. そのためには過去のデータを利用して「優勝が見込める数値」と「優勝が見込めない数値」を判別する境界となる年間勝率の数値を求めればよい. そこで, **マハラノビス距離による 1 次元の判別分析**を行う.
2. **優勝ラインをピタゴラス勝率に近似させる**（解説：4.2）
   1 で設定した年間勝率の目標を**ピタゴラス勝率**[*1]に置き換える. ピタゴラス勝率とは得点と失点のみから勝率を予測する式であり, 真の勝率（年間勝率の実測値）と強い正の相関関係がある.
3. **ピタゴラス勝率から必要な得点数を求める**（解説：4.3）
   2 で得たピタゴラス勝率を達成するための得点数の目安を求める. その際に, 失点数を何らかの形で固定すること, すなわち変数の消去に相当する作業が必要となる. この段階からは得られたピタゴラス勝率を定数として扱うため, 固定した失点数の大小に応じて得点数も上下する. 例えば, 高い投手力や守備力を期待して失点数を低めの値で固定すれば得点数の条件は緩くなるが, 逆に投手力や守備力の不安を攻撃でカバーするようなチームを想定するのであれば, 得点数の条件は厳しくなる.
4. **目標の得点数を達成するために必要な OPS の目安を求める**（解説：4.4）
   OPS は得点と強い正の相関関係を持っていることから, **回帰直線**を求め, **単回帰分析**を基にチームの OPS の目標を設定する.
5. **目標の OPS を達成するために出塁率と長打率をどれ位上昇させるべきかを考察する**
   ある球団に対して, 目標とする OPS と今年度の OPS を比較し, その差を埋めるために出塁率や長打率をどれ位上昇させるべきか, また出塁率や長打率を上昇させるためには四球や長打などの各記録をどれ位上昇させるべきかなど, ケースバイケースで考察を行う.

4.1 優勝ラインの設定

4.1.1 リーグ優勝に求められる年間成績

リーグ優勝するためにはどれ程の成績が必要になるかを考える. チームの成績を評価する上で多くのプロ野球ファンにとってイメージしやすい指標は**貯金 · 借金**や**ゲーム差**などであるが, NPB のペナントレースの順位は**年間勝率**の序列によって決定されるため, ここでは成績の尺度に勝率を用いる.

優勝するために最低限必要な年間勝率（=**優勝ライン**）を考えていくにあたり, 何の値を採用するべきかが問題になる. 平均値は 0.597, 中央値は 0.582 であるが, それらは最低ラインを計るものではないので優勝ラインに用いるべき

[*1] 1.2.5 を参照

ではない.[*2] かといって, この中の最低勝率である 0.539（'15 ヤクルト, 76-65-2）は優勝ラインとしては非常に心許ない.[*3] そこで, 優勝チーム（1 位）のグループ, 2 位チームのグループをそれぞれ作成し, **マハラノビス距離**（後述）による **1 次元の判別分析**の結果から優勝ラインを求めることにした.

また, 参考までに, 図 4.1 に各順位での年間勝率の分布を示した.

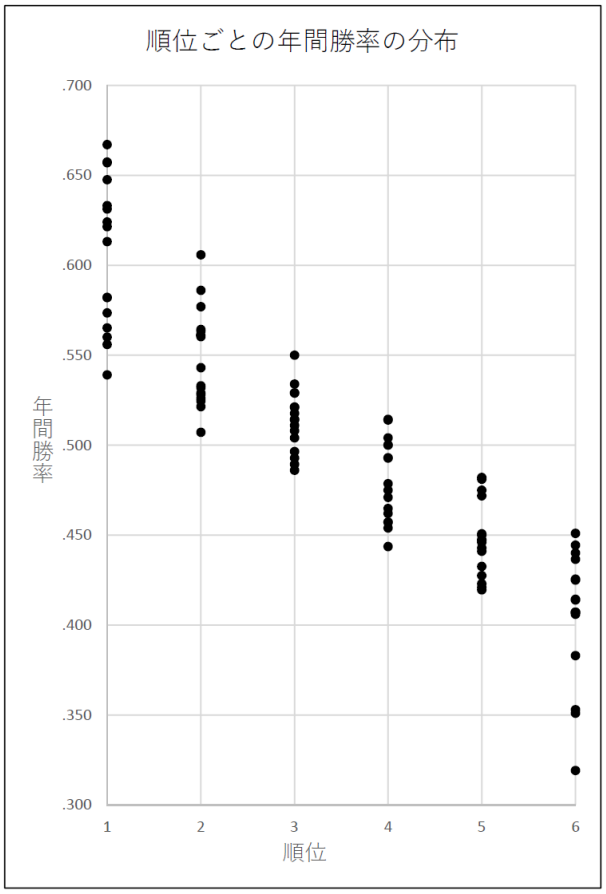

図 4.1　各順位での年間勝率の分布 ('11～'18)

4.1.2　マハラノビス距離の採用理由とグループ分け

以下はグループ分けの詳細と, ここで用いる記号の定義である.

$$\begin{cases} x: & \text{年間勝率（変数）} \\ \mu_k: & k \text{ 位群の平均} \\ \sigma_k: & k \text{ 位群の標準偏差} \\ D_k: & k \text{ 位群からのマハラノビス距離} \\ D_k^{'}: & k \text{ 位群からのユークリッド距離} \quad (k = 1, 2) \end{cases}$$

新たなサンプル（2019 年以降のデータ）が用意されたとき, そのサンプルは「1 位群」「2 位群」のうち, 群の平均 (μ) からの距離が近い群に分類される. 距離の算出にあたり, ユークリッド距離のようにグループのばらつきを考慮しない場合, 単純にサンプル (x) とグループの平均値 (μ) との差から距離が得られるため, $D_k^{'} := |x - \mu_k|$, すなわち $D_1^{'} = \mu_1 - x, D_2^{'} = x - \mu_2$ となる. しかし, 現実にはサンプルのばらつきを考慮した距離の調節が必要であり, 標準

[*2] 「入学試験に合格するために必要なのは合格最低点であって合格者平均点ではない」というものと似た論理である.

[*3] そもそもこの勝率で優勝出来たことはかなり稀な例であり, 直近 12 年間で年間勝率.539 を上回っていながら 2 位以下に終わったチームは延べ 16 チーム存在する. なお, 私個人の意見としては年間勝率が.539 であるからといって'15 ヤクルトの優勝の価値が損なわれるということは全く無く, むしろ混戦を極めたセ・リーグのペナントレースを制した強さを物語っている.

表 4.1　’07～’18 の 1 位, 2 位チームと年間勝率

年度		’18	’17	’16	’15	’14	’13	’12	’11	’10	’09	’08	’07
1 位群	パ	西	ソ	日	ソ	ソ	楽	日	ソ	ソ	日	西	日
		.624	.657	.621	.647	.565	.582	.556	.657	.547	.577	.543	.568
	セ	広	広	広	ヤ	巨	巨	巨	中	中	巨	巨	巨
		.582	.633	.631	.539	.573	.613	.667	.560	.560	.659	.596	.559
2 位群	パ	ソ	西	ソ	日	オ	西	西	日	西	楽	オ	ロ
		.577	.564	.606	.560	.563	.529	.533	.526	.545	.538	.524	.555
	セ	ヤ	神	巨	巨	神	神	中	ヤ	神	中	神	中
		.532	.561	.507	.528	.524	.521	.586	.543	.553	.566	.582	.549

偏差が異なるテストの成績の優劣を偏差だけでは比べられないのと同様に, ユークリッド距離では判別分析の手段として不十分である. そこで, 今回は

$$D_1 = \frac{\mu_1 - x}{\sigma_1}, \quad D_2 = \frac{x - \mu_2}{\sigma_2}$$

によって得られるマハラノビス距離を用いる.

4.1.3　優勝ラインの計算

優勝ラインは $D_1 = D_2$ となる勝率より求められるので,

$$\frac{\mu_1 - x}{\sigma_1} = \frac{x - \mu_2}{\sigma_2}$$

を満たす x を考える. これを解くと,

$$\begin{aligned} \sigma_2(\mu_1 - x) &= \sigma_1(x - \mu_2) \\ x(\sigma_1 + \sigma_2) &= \sigma_2\mu_1 + \sigma_1\mu_2 \end{aligned}$$

となる. 両辺を $\sigma_1 + \sigma_2 \neq 0$ で割り,

$$x = \frac{\sigma_2\mu_1 + \sigma_1\mu_2}{\sigma_1 + \sigma_2} \tag{4.1}$$

を得る. NPB の公式 HP から収集したデータを基に,

$$\begin{aligned} \mu_1 &= 0.597, \quad \sigma_1 = 0.042, \\ \mu_2 &= 0.549, \quad \sigma_2 = 0.024 \end{aligned}$$

を得られ, これらを式 (4.1) に代入することで.

$$x = 0.566$$

を得る. 故に (**.566**) を優勝ラインとして定める.

4.2　ピタゴラス勝率と実際の勝率の相関

4.2.1　無相関性の検定

図 4.2 からも直線的な関係がうかがえるが, 一般にピタゴラス勝率は実際の試合結果に基づく勝率との相関係数が高いといわれている. ’11～’18 の 8 年間について相関係数を計算したところ, $r = 0.920$ となり, 強い正の相関関係がみられた. しかし, 3.2 での議論と同様に, 標本標準偏差 r を求めただけでは母集団の相関関係の有無はわからない. そこで, 母相関係数 ρ が 0 であるか否かについて検定を行う.

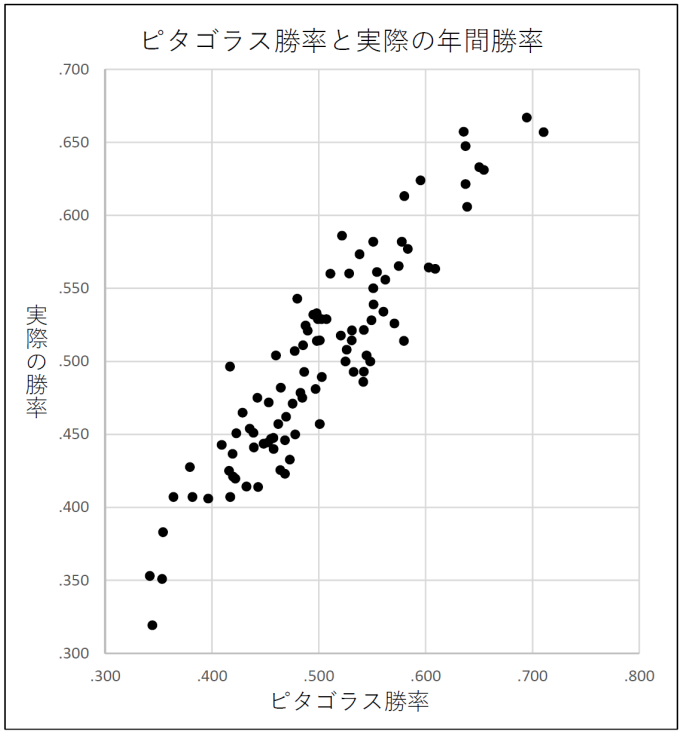

図 4.2　ピタゴラス勝率と実際の年間勝率の散布図 ('11～'18)

無相関性の検定

標本サイズ n の 2 変数の観測値 $(x_i, y_i), i = 1, 2, \ldots, n$ が与えられたとき, 無相関性の検定では x と y の間の母相関係数 ρ が 0 であるか否かが検定される. 帰無仮説 H_0 および対立仮説 H_1 は,

$$\text{帰無仮説 } H_0 : \rho = 0$$
$$\text{対立仮説 } H_1 : \rho \neq 0$$

であり, 検定統計量 t_0 は,

$$t_0 = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}} \tag{4.2}$$

で与えられる. ここで r_{xy} は x と y の標本相関係数である. 検定統計量 t_0 は帰無仮説のもとで自由度 $n-2$ の t 分布に従う.

無相関性の検定を用いて, ピタゴラス勝率と実際の勝率の相関関係を評価する. この検定では, 検定統計量 t_0 と棄却限界値[*4] $t_{n-2}(\alpha/2)$ を比較する. $|t_0| > t_{n-2}(\alpha/2)$ ならば帰無仮説を棄却し, $|t_0| \leq t_{n-2}(\alpha/2)$ ならば帰無仮説を受容する.

$r_{xy} = 0.920$ より, 検定統計量 t_0 は, 式 (4.2) を用いて

$$t_0 = \frac{0.920\sqrt{96-2}}{\sqrt{1-0.920^2}} = 22.822$$

となる. 検定統計量 t_0 は, 帰無仮説 H_0 のもとで, 自由度 $96-2=94$ の t 分布に従う. 自由度 94 の t 分布の上側 2.5 パーセント点 $t_{94}(0.025)$ は, $t_{94}(0.025) = 1.986$ である.[*5] $t_0 > t_{94}(0.025)$ なので, 有意水準 $\alpha = 0.05$ で帰無仮説 H_0 が棄却される. 従って, **ピタゴラス勝率と実際の勝率は無相関ではない**.

[*4] ここでの棄却限界値は自由度 $n-2$ の t 分布における上側 $100 \cdot (\alpha/2)$ パーセント点である.

[*5] 今回は Excel を用いて棄却限界値を求めたが, 一般に自由度が大体 30 以上であれば t 分布を標準正規分布で精度良く近似できると言われている.

4.2.2　相関係数の推定

文献 [9] によれば, “無相関性の検定では, 標本サイズが大きい場合に, 相関係数が小さくても有意になるおそれがある”とされているため, 母相関係数の区間推定を行う.

母相関係数の $100\cdot(1-\alpha)\%$ 信頼区間

標本サイズ n の 2 変数の観測値 $(x_i, y_i), i = 1, 2, \ldots, n$ が与えられたときの相関係数を r_{xy} とする. このとき, $100\cdot(1-\alpha)\%$ 信頼区間は

$$\left[\frac{\exp(2a)-1}{\exp(2a)+1},\quad \frac{\exp(2b)-1}{\exp(2b)+1}\right]$$

で与えられる. ここで

$$a = \frac{1}{2}\ln\left(\frac{1+r_{xy}}{1-r_{xy}}\right) - \frac{1}{\sqrt{n-3}}\cdot z(\alpha/2), \tag{4.3}$$

$$b = \frac{1}{2}\ln\left(\frac{1+r_{xy}}{1-r_{xy}}\right) + \frac{1}{\sqrt{n-3}}\cdot z(\alpha/2) \tag{4.4}$$

であり, $z(\alpha/2)$ は, 標準正規分布の上側 $100\cdot(\alpha/2)$ パーセント点である.

これを基に, ピタゴラス勝率と実際の勝率の相関係数の 95% 信頼区間を求める. $r_{xy} = 0.920$ なので, 式 (4.3), (4.4) より

$$a = \frac{1}{2}\ln\left(\frac{1+0.920}{1-0.920}\right) - \frac{1}{\sqrt{96-3}}\times 1.96 = 1.388,$$

$$b = \frac{1}{2}\ln\left(\frac{1+0.920}{1-0.920}\right) + \frac{1}{\sqrt{96-3}}\times 1.96 = 1.795$$

が得られる. 故に, 母相関係数に対する 95% 信頼区間は,

$$\text{下側信頼限界}: \frac{\exp(2\times 1.388)-1}{\exp(2\times 1.388)+1} = 0.883$$

$$\text{上側信頼限界}: \frac{\exp(2\times 1.795)-1}{\exp(2\times 1.795)+1} = 0.946$$

より, $[0.883, 0.946]$ となる.

4.3　失点数の固定

1.2.5 で紹介した計算式 9 の左辺に優勝ラインの 0.566 を代入すると,

$$0.566 = \frac{(\text{総得点})^2}{(\text{総得点})^2 + (\text{総失点})^2} \tag{4.5}$$

という関係式が得られる. ここから総失点の数値を固定することで, 「年間勝率 .566 を達成するために必要となる総得点数」を求めたい. 左辺を先に 0.566 で固定したことから, 式 (4.5) において総得点と総失点は**トレードオフ**[*6]の関係にある. それを踏まえて, 総失点について厳しさの異なるいくつかの基準を定める.

総失点の平均を μ, 標準偏差を σ とすると, データより $\mu = 550.99, \sigma = 76.07$ が得られた. (参考までに, 1 位のチームの失点数の平均値は, 499.00 であった.) これを基に, 式 (4.5) を変形して得た式

$$(\text{総得点}) = (\text{総失点}) \times \sqrt{\frac{0.566}{1-0.566}} \tag{4.6}$$

に代入計算を行うことで, 一例として表 4.2 のような基準を設定することが出来る. 式 (4.6) から求めた「年間勝率 .566 を達成するために必要となる総得点数」を, **要求得点**（造語）と呼ぶことにする.

[*6] 一方を追求すれば他方を犠牲にせざるを得ないという状態. 今回のケースでは総得点と総失点の条件を両方同時に甘く設定することが出来ない状態にある.

表 4.2　失点数の基準の例とそれに対応した得点数の条件

失点		得点	
平均 + 偏差	数値	要求得点	条件の厳しさ
μ	550.99	**629.23**	比較的厳しい
$\mu - 0.5\sigma$	512.95	**585.79**	中間
$\mu - \sigma$	474.92	**542.35**	比較的緩い

例として,「打ち勝つ野球」で優勝を成し遂げた球団の代表格として挙げられる'18 の西武のように, 失点数の多いチームの場合はより多い得点が求められる. ちなみに'18 西武の総失点数 653 は, $\mu + 1.34\sigma$ に相当する.

表 4.3　'18 西武

年間勝率	総得点	総失点	チーム OPS
.624	792	653	.806

4.4　OPS と得点の線形回帰モデル

年間総得点数の変動に関する考察を OPS に帰着させるために, OPS と得点の間にある関係を線形関数による回帰式で表す必要がある. そこで, **最小 2 乗法**を用いて**回帰直線**を導出し, 単回帰分析を行う.

4.4.1　最小 2 乗法の手順

標本サイズ n の 2 変数の観測値 $(x_i, y_i), i = 1, 2, \ldots, n$ に対して説明変数 x と目的変数 y の間に成り立つ直線関係を表す**線形回帰モデル**

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \cdots, n.$$

を考える. ここで, β_0, β_1 を**回帰係数**, ε_i を**誤差**といい, y_i の**予測値** $\hat{y}_i$ は

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{4.7}$$

で与えられる. 従って, 実測値 y_i と予測値 $\hat{y}_i$ の差は $e_i = y_i - \hat{y}_i$ で与えられ, このとき e_i は**残差**と呼ばれる.

最小 2 乗法とは, 以下に示す**残差平方和** $S(\hat{\beta}_0, \hat{\beta}_1)$ が最小となるように β_0, β_1 の推定値 $\hat{\beta}_0, \hat{\beta}_1$ を求める方法である.

$$S(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \tag{4.8}$$

式 (4.8) は $\hat{\beta}_0, \hat{\beta}_1$ を変数とする 2 次関数であることから, S を $\hat{\beta}_0, \hat{\beta}_1$ で偏微分し 0 とおく.

$$\frac{\partial S}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$
$$\frac{\partial S}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

これらを移項して, **正規方程式**と呼ばれる $\hat{\beta}_0, \hat{\beta}_1$ を未知数とする線形方程式

$$\begin{cases} n\hat{\beta}_0 + \left(\sum_{i=1}^{n} x_i \right) \hat{\beta}_1 & = \sum_{i=1}^{n} y_i \\ \left(\sum_{i=1}^{n} x_i \right) \hat{\beta}_0 + \left(\sum_{i=1}^{n} x_i^2 \right) \hat{\beta}_1 & = \sum_{i=1}^{n} x_i y_i \end{cases} \tag{4.9}$$

を解くと, $\overline{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i, \overline{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i$ として, 以下の β_0, β_1 の推定値が得られる.

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \tag{4.10}$$

$$\hat{\beta}_1 = \frac{n\left(\sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \tag{4.11}$$

4.4.2　数値の代入と単回帰式の計算

OPS を説明変数 x, 年間総得点を目的変数 y へそれぞれ対応させると, 以下のような計算結果を得る.

$$n = 96, \quad \overline{x} = 0.692, \quad \overline{y} = 550.990 \tag{4.12}$$

$$\sum_{i=1}^{n} x_i = 66.406, \quad \sum_{i=1}^{n} y_i = 52895, \quad \sum_{i=1}^{n} x_i^2 = 46.086, \quad \sum_{i=1}^{n} x_i y_i = 36874.46 \tag{4.13}$$

式 (4.12), 式 (4.13) をそれぞれ式 (4.10), 式 (4.11) へ代入すると,

$$\hat{\beta}_1 = \frac{96 \times 36874.46 - 66.406 \times 52895}{96 \times 46.086 - 66.406^2} = 1890.903 \tag{4.14}$$

$$\hat{\beta}_0 = \frac{1}{96}(52895 - 1890.903 \times 66.406) = -757.003 \tag{4.15}$$

となり, 単回帰直線

$$\hat{y} = -757.003 + 1890.903x \tag{4.16}$$

を得る. OPS と総得点の散布図と回帰直線を平面上に表すと図 4.3 のようになる.

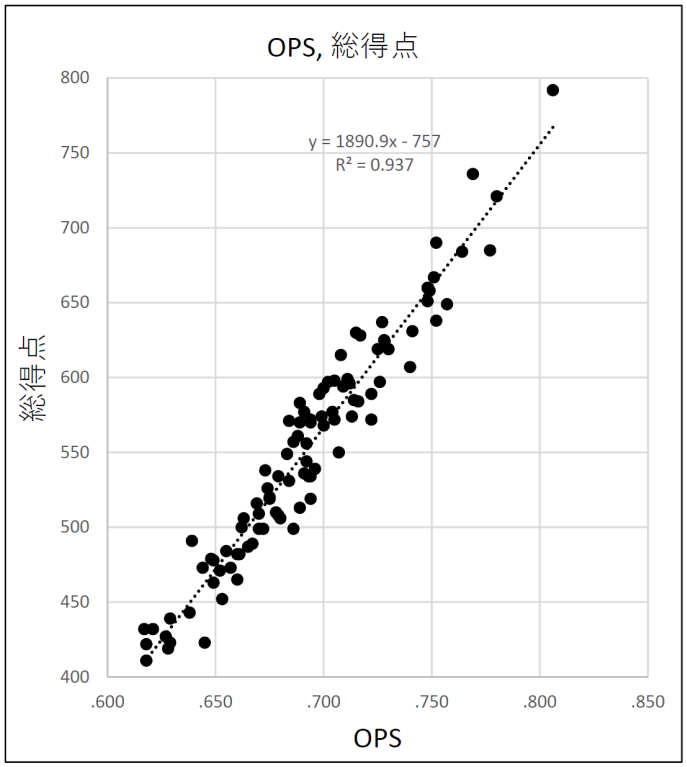

図 4.3　OPS, 総得点の散布図と回帰直線 ('11～'18)

4.4.3　寄与率

寄与率を用いて, 推定された回帰直線の適合度を評価する.

寄与率（回帰分析）

標本サイズ n の 2 変数の観測値 $(x_i, y_i), i = 1, 2, \ldots, n$ が与えられたとき, 推定された回帰直線から計算された y_i の予測値を $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ とする. そして, 実測値 y_i, 予測値 $\hat{y}_i$ および残差 $e_i = y_i - \hat{y}_i$ のそれぞれの平方和を

$$SS_T = \sum_{i=1}^{n} (y_i - \overline{y})^2, \quad SS_R = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2, \quad SS_E = \sum_{i=1}^{n} e_i^2 \tag{4.17}$$

とするとき $\left(\text{ただし } \overline{y} \text{ は } y_i \text{ の平均値 } \overline{y} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} y_i \right)$, SS_T を**総変動**, SS_R を**回帰変動**, SS_E を**残差変動**と呼ぶ. このとき, それぞれの変動には次の関係

$$SS_T = SS_R + SS_E$$

がある. このような関係式のことを**回帰分析の変動分解**という. このとき, 回帰変動 SS_R は推定された回帰直線が当てはまっている度合いを表しており, 残差変動 SS_E は推定された回帰直線が当てはまっていない度合いを表す.

回帰変動が総変動に占める割合を計算することで, 推定された回帰直線の適合度を要約した指標が**寄与率（決定係数）**である. したがって, 寄与率は

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T} \tag{4.18}$$

である. 寄与率 R^2 の範囲は $0 \leq R^2 \leq 1$ であり, 1 に近づくほど良く当てはまっていると解釈される.

式 (4.17) について, データより $SS_T = 576029.0, SS_R = 539755.8, SS_E = 36237.2$ を式 (4.18) に代入し, 寄与率 $R^2 = 0.937$ を得た. ここから, 式 (4.16) の回帰直線はデータに良く当てはまっていると言える.

4.4.4　「要求得点数」から必要な OPS を求める

式 (4.16) の左辺（目的変数 y）に 4.3 で求めた**要求得点**を代入し, OPS の目標値を設定する.

$$629.23 = -757.003 + 1890.903x \tag{4.19}$$
$$585.79 = -757.003 + 1890.903x \tag{4.20}$$
$$542.35 = -757.003 + 1890.903x \tag{4.21}$$

をそれぞれ計算すると, 表 4.4 の様な形式で OPS の達成目標を定めることが出来る.

表 4.4　要求得点から計算した OPS の達成目標

要求得点	条件の厳しさ	OPS
629.23	比較的厳しい	.733
585.79	中間	.710
542.35	比較的緩い	.687

4.5　出塁率・長打率向上目標（分析の例）

この節では千葉ロッテマリーンズを例として考察を行う. '18 ロッテの OPS が .679 であったのに対し, NPB 平均程度の総失点数 550.99[*7]を想定した場合の要求得点は 629.23 で, その要求得点を満たすために必要な OPS は .733 であった. この差を埋めることを考えたいが, 必要な OPS を求めただけでは出塁率と長打率が一意に定まらない. そこで, 極端な例ではあるが出塁率 (OBP) と長打率 (SLG) の上げ方を次の 3 通りのモデルで考える.

1. 出塁率と長打率の比率を維持しつつ OPS を上昇させる
2. 出塁率のみを上昇させ, 長打率は現状維持
3. 長打率のみを上昇させ, 出塁率は現状維持

1. の場合, 出塁率と長打率をそれぞれ 0.733/0.679 倍する. データより, '18 ロッテは

$$(\mathrm{OBP}, \mathrm{SLG}, \mathrm{OPS}) = (0.324, 0.355, 0.679) \tag{4.22}$$

であったから, $(\mathrm{OBP}, \mathrm{SLG}, \mathrm{OPS}) = (x_m, y_m, 0.733)$ に対し

$$x_m + y_m = 0.733, \quad x_m : y_m = 0.324 : 0.355 \tag{4.23}$$

という条件で x_m, y_m を求めれば良い. 式 (4.23) を解くと, $(x_m, y_m) = (0.350, 0.383)$ を得る.

2. 及び 3. の場合, 目標とする OPS と'18 ロッテの OPS の差 $(0.733 - 0.679 = 0.054)$ を, 2. では出塁率のみに, 3. では長打率のみに足し合わせる. これにより 2. の計算結果は $(x_m, y_m) = (0.378, 0.355)$, 3. の計算結果は $(x_m, y_m) = (0.324, 0.409)$ となる.

以上のモデルを, $x = \mathrm{OBP}, y = \mathrm{SLG}$ として xy 平面上に図示すると, 図 4.4 の様になる. この図において, 点 M は'18 ロッテ, 点 Q, P, R はそれぞれモデル (1.)～(3.) から求めた点を表すものとすると, 点 Q は半直線 OM と $x + y = 0.733$ の交点であり, $\mathrm{MP} \parallel y$ 軸, $\mathrm{MR} \parallel x$ 軸 である.

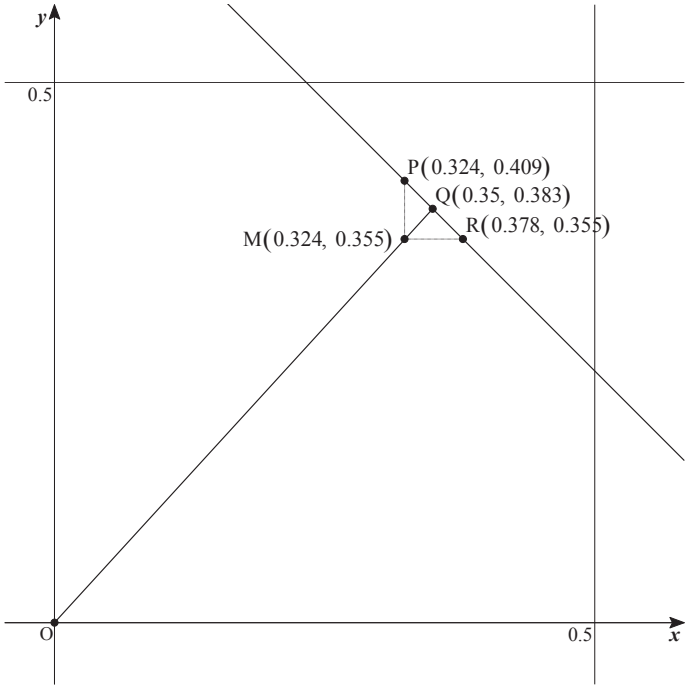

図 4.4　出塁率と長打率の向上モデル（'18 ロッテの場合）

[*7] 表 4.2 を参照

第 5 章

手法 B ～判別分析～

手法 B では, 出塁率と長打率の 2 次元データに対して直接**判別分析**を行い, 「優勝出来るか否か」を線引きする境界線を求める.

5.1 マハラノビス距離による 2 次元の判別分析

4.1.2 で扱った 1 次元のマハラノビス距離を, 以下のように 2 次元へ拡張する.

マハラノビス距離による 2 次元の判別分析

1 変量のマハラノビス距離 D

$$D = \frac{|x - \overline{x}|}{\sqrt{s^2}} \tag{5.1}$$

について, D の 2 乗 D^2 をとる.

$$D^2 = \frac{(x - \overline{x})^2}{s^2} \tag{5.2}$$

$$= (x - \overline{x})(s^2)^{-1}(x - \overline{x}) \tag{5.3}$$

式 (5.3) をベクトルと行列の積であると見なせば

$$(x - \overline{x}) \longleftrightarrow (x_1 - \overline{x}_1, x_2 - \overline{x}_2) \tag{5.4}$$

$$(s^2)^{-1} \longleftrightarrow \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix}^{-1} \tag{5.5}$$

という対応が考えられる. そこで, 2 変量のマハラノビス距離の 2 乗を以下のように定義する.

$$D^2 := \begin{pmatrix} x_1 - \overline{x}_1 & x_2 - \overline{x}_2 \end{pmatrix} \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix}^{-1} \begin{pmatrix} x_1 - \overline{x}_1 \\ x_2 - \overline{x}_2 \end{pmatrix} \tag{5.6}$$

なお,

$$\begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix} \tag{5.7}$$

は分散共分散行列[a]である. グループ G_1 のマハラノビス距離の 2 乗を D_1^2, グループ G_2 のマハラノビス距離の 2 乗を D_2^2 とするとき, 境界線は

$$D_1^2 - D_2^2 = 0 \tag{5.8}$$

である.

[a] 定義 2.1 を参照

以下, x_1 は出塁率, x_2 は長打率を表し, 1 位の 2(リーグ) × 8(年) = (延べ)16(球団) は G_1, 2 位～6 位の 5(球団) ×

2(リーグ) × 8(年) = (延べ)80(球団) は G_2 に属するものとする.（つまり表 4.1.2 の「1 位群」が G_1 に相当する.）

5.1.1　D_1^2 及び D_2^2 の計算

マハラノビス距離を求めるにあたり, G_1, G_2 について以下のデータを使用する.

$$\begin{cases} \overline{x}_1^{(1)} &= 0.332 \\ \overline{x}_2^{(1)} &= 0.396 \\ s_{11}^{(1)} &= 0.000207 \\ s_{12}^{(1)} &= s_{21}^{(1)} = 0.000349 \\ s_{22}^{(1)} &= 0.000934 \end{cases} \qquad \begin{cases} \overline{x}_1^{(2)} &= 0.318 \\ \overline{x}_2^{(2)} &= 0.367 \\ s_{11}^{(2)} &= 0.000149 \\ s_{12}^{(2)} &= s_{21}^{(2)} = 0.000208 \\ s_{22}^{(2)} &= 0.000669 \end{cases} \tag{5.9}$$

まず G_1 とのマハラノビス距離の 2 乗 D_1^2 について, 式 (5.9) を式 (5.6) に代入すると

$$D_1^2 = \begin{pmatrix} x_1 - 0.332 & x_2 - 0.396 \end{pmatrix} \begin{pmatrix} 0.000207 & 0.000349 \\ 0.000349 & 0.000934 \end{pmatrix}^{-1} \begin{pmatrix} x_1 - 0.332 \\ x_2 - 0.396 \end{pmatrix} \tag{5.10}$$

となる. ここで,

$$\begin{pmatrix} 0.000207 & 0.000349 \\ 0.000349 & 0.000934 \end{pmatrix}^{-1} = \begin{pmatrix} 13056 & -4879 \\ -4879 & 2894 \end{pmatrix} \tag{5.11}$$

より,

$$D_1^2 = 13056x_1^2 + 2894x_2^2 - 9758x_1x_2 - 4805x_1 + 948x_2 + 610 \tag{5.12}$$

を得る. 次に, G_2 とのマハラノビス距離の 2 乗 D_2^2 について,

$$D_2^2 = \begin{pmatrix} x_1 - 0.318 & x_2 - 0.367 \end{pmatrix} \begin{pmatrix} 0.000149 & 0.000208 \\ 0.000208 & 0.000669 \end{pmatrix}^{-1} \begin{pmatrix} x_1 - 0.318 \\ x_2 - 0.367 \end{pmatrix} \tag{5.13}$$

となる.

$$\begin{pmatrix} 0.000149 & 0.000208 \\ 0.000208 & 0.000669 \end{pmatrix}^{-1} = \begin{pmatrix} 11858 & -3687 \\ -3687 & 2641 \end{pmatrix} \tag{5.14}$$

を用いると,

$$D_2^2 = 11858x_1^2 + 2641x_2^2 - 7374x_1x_2 - 4835x_1 + 406x_2 + 694 \tag{5.15}$$

を得る.

5.1.2　境界線

式 (5.8) の左辺に, 式 (5.12), (5.15) を代入することで

$$(D_1^2 - D_2^2 =)1198x_1^2 + 253x_2^2 - 2384x_1x_2 + 30x_1 + 542x_2 - 84 = 0 \tag{5.16}$$

が境界線として得られる. ただし, この関数は 2 次曲線の形式であるため[*1], 出塁率のデータの範囲に準拠し, 定義域を $x \geq 0.320$ とする. [*2]

[*1] 描画すると双曲線であることが分かる.

[*2] 本来は漸近線から定義域を設定するのが適切であると考えられるが, ここでは簡単のため y 軸と平行な直線を基に定義域を判断した.

5.2　境界線に到達する最小の OPS ～非線形計画法～

式 (5.16) で得られた境界線が持つ意味を分かりやすくするため, 出塁率と長打率がそれぞれどのような値を取れば最小の OPS で境界線に到達できるかを考察する. そのためには, 以下の問題を**非線形計画法**で解けば良い.

$$\text{目的関数}：x_1 + x_2 \to \text{最小化} \tag{5.17}$$

$$\text{制約条件}：\begin{cases} 1198x_1^2 + 253x_2^2 - 2384x_1x_2 + 30x_1 + 542x_2 - 84 = 0 \\ x_1 \geq 0.320 \end{cases} \tag{5.18}$$

これを Mathematica で解くと,

$$(x_1, x_2) = (0.333, 0.374), \quad \min(x + y) = 0.707 \tag{5.19}$$

という結果になる. つまり, OPS ≥ 0.707 は境界線に到達するための必要条件であると言える.

境界線及び線形方程式 $x_1 + x_2 = 0.707$ を $x = \mathrm{OBP}(x_1), y = \mathrm{SLG}(x_2)$ として xy 平面上に図示すると, 図 5.1 の様になる. なお, 点 P の座標は $P(0.333, 0.374)$ であり, 白抜きの円は出塁率と長打率に関する散布図を表す.

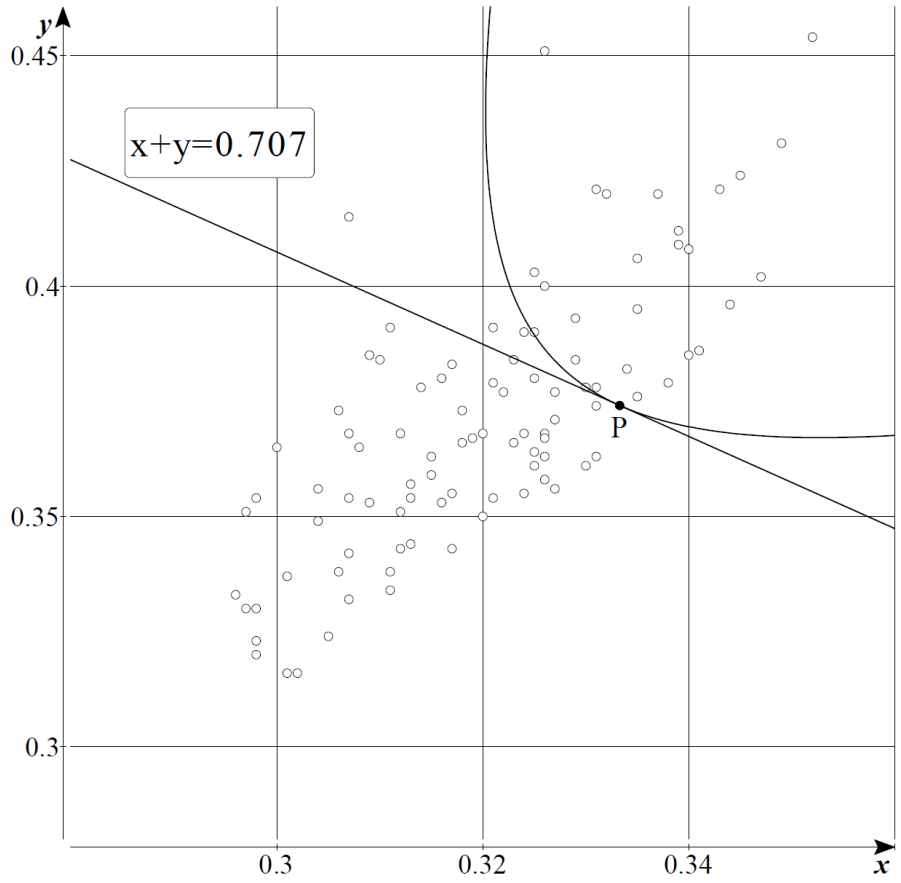

図 5.1　マハラノビス距離による境界線

5.3　誤判別率の導出と手法 B の有用性

2 次元の判別分析の場合は与えられたデータを 2 次曲線によって分離するため, **誤判別**の発生は免れないものであり, それは図 5.1 においても例外ではない. ここでの誤判別とは, グループ G_1 に属しているにも関わらず誤ってグ

ループ G_2 と判別されることと, 逆にグループ G_2 に属しているにも関わらず誤ってグループ G_1 と判別されることを指す. そこで**誤判別率**を表 5.1 のように定義すると, 表 5.2 のように算出される.

表 5.1　誤判別率の定義

	グループ G_1	グループ G_2
サンプル数	n_1	n_2
誤判別数	m_1	m_2
誤判別率	m_1/n_1	m_2/n_2

表 5.2　誤判別率

	グループ G_1	グループ G_2
サンプル数	16	80
誤判別数	6	16
誤判別率	0.375	0.200

表からはグループ G_1, G_2 ともに誤判別率がやや高く, 境界線 (5.16) による判別の精度はあまり良くないことが読み取れる. そこで手法に問題点があったと考え, 「出塁率と長打率の情報のみから優勝の可否を判別するのは流石に無理があるのではないか」という仮説を立てた. この仮説は本研究の意義そのものを否定しているものではなく, （第 4 章で扱った“手法 A”のような考察の順序立てをせずに）出塁率と長打率から直接「優勝の可否」を判断しようとした点について, 手順や発想に飛躍があったのではないかという意味である.

誤判別されたサンプルの内訳を表 5.3 に, それに該当する各球団の OPS を表 5.4 に示した. 5.2 節で, G_1 と判別されるためには OPS が.707 以上でなければならないことを示したが, この表から実際に G_2 で誤判別された全ての球団がそれを満たしていることを確認できる.

表 5.3　誤判別されたサンプル

本来属している群	順位	誤判別数	内訳
G_1	1	6	’15 ヤ, ’14 巨, ’12 日, ’12 巨, ’11 ソ, ’11 中
G_2	2	6	’18 ソ, ’18 ヤ, ’17 西, ’16 ソ, ’14 オ, ’14 神
	3	3	’18 日, ’18 巨, ’14 広
	4	3	’16 西, ’15 西, ’13 ソ
	5	3	’16 ヤ, ’14 西, ’13De
	6	1	’14 ヤ

表 5.4　誤判別されたサンプルの OPS

G_1	順位	1								
	球団	’15 ヤ	’14 巨	’12 日	’12 巨	’11 ソ	’11 中			
	OPS	.699	.712	.678	.693	.707	.628			
G_2	順位	2						3		
	球団	’18 ソ	’18 ヤ	’17 西	’16 ソ	’14 オ	’14 神	’18 日	’18 巨	’14 広
	OPS	.777	.749	.752	.727	.716	.711	.722	.728	.757
	順位	4			5			6		
	球団	’16 西	’15 西	’13 ソ	’16 ヤ	’14 西	’13De	’14 ヤ		
	OPS	.730	.741	.748	.709	.713	.715	.751		

5.4　手法 B の反省点と手法 A との比較

手法 B で行った判別分析には守備面の要素が全く考慮されていなかったため, G_2 のうち得点力は高いが失点も多い球団は G_1 と誤判別されやすく, 逆に G_1 のうち得点力が平凡であっても投手力や守備力が優れていたために優勝出来た球団は G_2 と誤判別されやすいという傾向が見られた. また'11, '12 の優勝球団（4 球団）が全て誤判別された背景には, その時期に低反発球が使用されていたために打球が飛びづらく, 結果的に OPS の平均も下がっていたことが考えられる. このように誤判別される原因が複数考えられる状況から, 出塁率と長打率から優勝の可否を判別しても安定して良い予測精度をもたらすことは難しいと言える.

今回の“手法 B”の代わりに得点数に対する判別分析から考察を展開していれば良い結果を得られていた可能性はあるが, 得点数を出塁率と長打率で説明したければ回帰分析を用いれば良く, 「得点数が p 以上であるか否か」などとあえて量的な基準を設けて判別分析を行う意味は薄い.

以上を加味して手法 A と手法 B を比較すると, 分析手順は手法 B の方が簡潔ではあるが, 手法 A の方が論理の飛躍は起こりにくいと考えられる.

研究結果のまとめ

主成分分析の結果

出塁率と長打率を 0.36 : 0.93 の割合で評価するとデータの分散を最大化できるが, 出塁率と長打率を 1 : 1 の割合で評価する OPS の方が指標としての利便性は高い.

手法 A の分析結果

手法 A から得られた結論を簡潔に表現すると, 以下の様になる.

1. NPB 平均程度の総失点数 (=551) のチームが優勝するには目安として 629 以上の総得点が必要で, これを達成するためには**チーム全体の OPS が.733 以上**であることが望ましい.
2. {(NPB 平均) − 0.5 × (標準偏差)} 程度の総失点数 (=513) のチームが優勝するには目安として 586 以上の総得点が必要で, これを達成するためには**チーム全体の OPS が.710 以上**であることが望ましい.
3. {(NPB 平均) − (標準偏差)} 程度の総失点数 (=475) のチームが優勝するには目安として 542 以上の総得点が必要で, これを達成するためには**チーム全体の OPS が.687 以上**であることが望ましい.

手法 B の分析結果

出塁率と長打率を基に優勝球団と非優勝球団を判別する境界線は, 図 5.1 の様になる. このとき, 出塁率 : .333, 長打率 : .374 であれば, 最小の OPS(.707) で境界線に到達できる. ただし, この境界線は誤判別率が高いため, 用いる変数の設定などを改善する必要がある.

今後の課題

勝率予測の考察方法の改善

第 III 部において, 手法 A の考察は特定の球団を例としたケースバイケースによるものであったため現状では一般化出来ておらず, 手法 B については問題点を定量的に分析できていなかった. 今後はそれぞれの分析方法の妥当性を精査し, 手法を提示するだけでなく, 定量的に手法の評価を行いたい.

年度による OPS のばらつき

5.4 節でも述べたように, 試合で使われる公式球の違いや「投高打低」及び「打高投低」の傾向の違いによって OPS の分散が大きくなる. これを解消するにはデータを正規化するか, OPS+ という指標を用いて OPS のリーグ平均からの傑出度をもとに打撃成績を評価することを検討する必要がある.

規定打席

当初はセ・パ両リーグの**規定打席到達者**の成績を対象とすることを考えたが, その場合規定打席未到達者による影響を完全に排除してしまうことになり, かといって全選手を平等に扱おうとすると打席数の違いによる誤差が生じかねないため, '17 芝浦祭の QS の研究と同様にチームの成績を対象とした. 前回は失点と自責点の違いの処理を課題に挙げ, 今回は規定打席未到達者の成績の処理方法が課題となったが, 「野球の研究をする際にはチーム成績と個人成績の扱いを考えなければならない」ということについては前回と今回で共通していた. 今後この課題の解決を図るにあたっては, 必要に応じて多変量解析以外の分野を取り入れることを検討する必要があるように思う.

補足資料

次頁以降の表 5.6〜5.8 に以下のプロ野球データを掲載した.

表 5.5　データ集に掲載した項目（表記を本文から変更していないものは割愛）

記号	意味
年	年度 (西暦下 2 桁)
#	順位
T	球団名 (Team)
BA	打率 (Batting Average)
OBP	出塁率
SLG	長打率
p_1	第 1 主成分得点
p_2	第 2 主成分得点
RS/G	1 試合平均得点
RS	年間総得点 (Runs Scored)
RA	年間総失点 (Runs Allowed)
Py	ピタゴラス勝率
W%	年間勝率

ただし紙面の都合上, 各年度について上段をパシフィック・リーグ, 下段をセントラル・リーグとした.

表 5.6　公式戦データ（’18〜’16）

年	#	T	BA	OBP	SLG	OPS	GPA	NOI	p_2	p_1	RS/G	RS	RA	Py	W%
’18	1	L	.273	.352	.454	.806	.272	503	-0.164	0.551	5.50	792	653	.595	.624
	2	H	.266	.326	.451	.777	.259	476	-0.141	0.538	4.76	685	579	.583	.577
	3	F	.251	.329	.393	.722	.246	460	-0.165	0.485	4.09	589	586	.503	.529
	4	Bs	.244	.308	.365	.673	.230	430	-0.155	0.452	3.74	538	565	.476	.471
	5	M	.247	.324	.355	.679	.235	442	-0.174	0.448	3.71	534	628	.420	.421
	6	E	.241	.307	.368	.675	.230	430	-0.153	0.454	3.61	520	583	.443	.414
	1	C	.262	.349	.431	.780	.265	493	-0.169	0.528	5.01	721	651	.551	.582
	2	S	.266	.347	.402	.749	.257	481	-0.178	0.500	4.57	658	665	.495	.532
	3	G	.257	.325	.403	.728	.247	459	-0.157	0.493	4.34	625	575	.542	.486
	4	DB	.250	.307	.415	.722	.242	445	-0.136	0.498	3.97	572	642	.443	.475
	5	D	.265	.325	.380	.705	.241	452	-0.166	0.472	4.15	598	654	.455	.447
	6	T	.253	.330	.361	.691	.239	450	-0.177	0.456	4.01	577	628	.458	.440
’17	1	H	.259	.331	.421	.752	.254	471	-0.156	0.512	4.43	638	483	.636	.657
	2	L	.264	.332	.420	.752	.254	472	-0.158	0.512	4.79	690	560	.603	.564
	3	E	.254	.324	.390	.714	.243	454	-0.161	0.481	4.06	585	528	.551	.550
	4	Bs	.251	.316	.380	.696	.237	443	-0.157	0.469	3.74	539	598	.448	.444
	5	F	.242	.313	.357	.670	.230	432	-0.163	0.446	3.53	509	596	.422	.420
	6	M	.233	.297	.351	.648	.221	414	-0.150	0.435	3.33	479	647	.354	.383
	1	C	.273	.345	.424	.769	.261	486	-0.168	0.520	5.11	736	540	.650	.633
	2	T	.249	.327	.371	.698	.240	451	-0.171	0.464	4.09	589	528	.554	.561
	3	DB	.252	.311	.391	.702	.238	441	-0.148	0.477	4.15	597	598	.499	.529
	4	G	.249	.318	.373	.691	.236	442	-0.162	0.463	3.72	536	504	.531	.514
	5	D	.247	.300	.365	.665	.226	422	-0.148	0.449	3.38	487	623	.379	.428
	6	S	.234	.306	.338	.644	.222	419	-0.163	0.426	3.28	473	653	.344	.319
’16	1	F	.266	.340	.385	.725	.249	468	-0.178	0.482	4.30	619	467	.637	.621
	2	H	.261	.341	.386	.727	.250	470	-0.178	0.483	4.42	637	479	.639	.606
	3	M	.256	.326	.363	.689	.237	447	-0.173	0.456	4.05	583	582	.501	.514
	4	L	.264	.335	.395	.730	.250	467	-0.169	0.489	4.30	619	618	.501	.457
	5	E	.257	.324	.368	.692	.238	447	-0.169	0.460	3.78	544	654	.409	.443
	6	Bs	.253	.317	.355	.672	.231	435	-0.167	0.446	3.47	499	635	.382	.407
	1	C	.272	.343	.421	.764	.260	483	-0.167	0.517	4.75	684	497	.654	.631
	2	G	.251	.310	.384	.694	.236	438	-0.150	0.470	3.60	519	543	.477	.507
	3	DB	.249	.309	.385	.694	.235	437	-0.149	0.471	3.97	572	588	.486	.493
	4	T	.245	.312	.351	.663	.228	429	-0.164	0.440	3.51	506	546	.462	.457
	5	S	.256	.331	.378	.709	.243	457	-0.172	0.472	4.13	594	694	.423	.451
	6	D	.245	.309	.353	.662	.227	427	-0.160	0.441	3.47	500	573	.432	.414

表 5.7　公式戦データ（'15〜'13）

年	#	T	BA	OBP	SLG	OPS	GPA	NOI	p_2	p_1	RS/G	RS	RA	Py	W%
'15	1	H	.267	.340	.408	.748	.255	476	-0.169	0.503	4.52	651	491	.637	.647
	2	F	.258	.330	.378	.708	.243	456	-0.171	0.472	4.27	615	581	.528	.560
	3	M	.257	.320	.368	.688	.236	443	-0.165	0.459	3.90	561	563	.498	.514
	4	L	.263	.335	.406	.741	.252	470	-0.165	0.500	4.38	631	573	.548	.500
	5	Bs	.249	.321	.354	.675	.233	439	-0.171	0.446	3.60	519	548	.473	.433
	6	E	.241	.311	.338	.649	.224	424	-0.168	0.428	3.22	463	612	.364	.407
	1	S	.257	.322	.377	.699	.239	448	-0.164	0.468	3.99	574	518	.551	.539
	2	G	.243	.313	.354	.667	.229	431	-0.164	0.443	3.40	489	443	.549	.528
	3	T	.247	.317	.343	.660	.228	431	-0.171	0.434	3.23	465	550	.417	.496
	4	C	.246	.312	.368	.680	.232	435	-0.158	0.456	3.51	506	474	.533	.493
	5	D	.253	.313	.344	.657	.227	428	-0.167	0.434	3.28	473	504	.468	.446
	6	DB	.249	.306	.373	.679	.231	430	-0.150	0.458	3.53	508	598	.419	.437
'14	1	H	.280	.344	.396	.740	.254	476	-0.177	0.494	4.22	607	522	.575	.565
	2	Bs	.258	.334	.382	.716	.246	461	-0.173	0.477	4.06	584	468	.609	.563
	3	F	.251	.321	.379	.700	.239	447	-0.162	0.469	4.12	593	569	.521	.518
	4	M	.251	.314	.378	.692	.236	440	-0.156	0.466	3.86	556	642	.429	.465
	5	L	.248	.329	.384	.713	.244	457	-0.168	0.477	3.99	574	600	.478	.450
	6	E	.255	.327	.356	.683	.236	446	-0.176	0.450	3.81	549	604	.452	.444
	1	G	.257	.321	.391	.712	.242	451	-0.158	0.481	4.14	596	552	.538	.573
	2	T	.264	.335	.376	.711	.245	460	-0.176	0.472	4.16	599	614	.488	.524
	3	C	.272	.337	.420	.757	.257	477	-0.162	0.513	4.51	649	610	.531	.521
	4	D	.258	.325	.364	.689	.237	446	-0.171	0.457	3.96	570	590	.483	.479
	5	DB	.253	.317	.383	.700	.238	445	-0.157	0.472	3.94	568	624	.453	.472
	6	S	.279	.339	.412	.751	.256	476	-0.167	0.507	4.63	667	717	.464	.426
'13	1	E	.267	.338	.379	.717	.247	464	-0.178	0.476	4.36	628	537	.578	.582
	2	L	.257	.331	.363	.694	.240	452	-0.177	0.458	3.96	570	562	.507	.529
	3	M	.262	.331	.374	.705	.242	456	-0.173	0.468	3.97	572	584	.490	.521
	4	H	.274	.339	.409	.748	.255	475	-0.168	0.504	4.58	660	562	.580	.514
	5	Bs	.256	.323	.366	.689	.237	445	-0.169	0.458	3.56	513	529	.485	.475
	6	F	.256	.326	.368	.694	.239	449	-0.171	0.461	3.71	534	604	.439	.451
	1	G	.262	.326	.400	.726	.247	459	-0.159	0.491	4.15	597	508	.580	.613
	2	T	.255	.326	.358	.684	.236	445	-0.174	0.452	3.69	531	488	.542	.521
	3	C	.248	.319	.367	.686	.235	441	-0.165	0.458	3.87	557	554	.503	.489
	4	D	.245	.315	.359	.674	.232	435	-0.164	0.449	3.65	526	599	.435	.454
	5	DB	.262	.325	.390	.715	.244	455	-0.162	0.481	4.38	630	686	.458	.448
	6	S	.253	.327	.377	.704	.241	453	-0.168	0.470	4.01	577	682	.417	.407

表 5.8　公式戦データ（'12〜'11）

年	#	T	BA	OBP	SLG	OPS	GPA	NOI	p_2	p_1	RS/G	RS	RA	Py	W%
'12	1	F	.256	.315	.363	.678	.233	436	-0.162	0.452	3.54	510	450	.562	.556
	2	L	.251	.316	.353	.669	.230	434	-0.167	0.443	3.58	516	518	.498	.533
	3	H	.252	.304	.349	.653	.224	420	-0.157	0.435	3.14	452	429	.526	.508
	4	E	.252	.307	.332	.639	.221	418	-0.166	0.421	3.41	491	467	.525	.500
	5	M	.257	.320	.350	.670	.232	437	-0.172	0.442	3.47	499	502	.497	.481
	6	Bs	.241	.301	.337	.638	.220	413	-0.159	0.423	3.08	443	525	.416	.425
	1	G	.256	.326	.367	.693	.238	448	-0.171	0.460	3.71	534	354	.695	.667
	2	D	.245	.311	.334	.645	.223	422	-0.169	0.424	2.94	423	405	.522	.586
	3	S	.260	.325	.361	.686	.237	445	-0.172	0.454	3.47	499	514	.485	.511
	4	C	.233	.297	.330	.627	.216	407	-0.158	0.415	2.97	427	454	.469	.462
	5	T	.236	.302	.316	.618	.215	407	-0.167	0.404	2.85	411	438	.468	.423
	6	DB	.233	.298	.320	.618	.214	405	-0.162	0.406	2.93	422	571	.353	.351
'11	1	H	.267	.323	.384	.707	.241	451	-0.162	0.475	3.82	550	351	.711	.657
	2	F	.251	.304	.356	.660	.226	423	-0.155	0.442	3.35	482	418	.571	.526
	3	L	.253	.318	.366	.684	.235	440	-0.164	0.456	3.97	571	522	.545	.504
	4	Bs	.248	.307	.342	.649	.224	421	-0.162	0.430	3.32	478	518	.460	.504
	5	E	.245	.298	.323	.621	.215	406	-0.161	0.409	3.00	432	464	.464	.482
	6	M	.241	.301	.316	.617	.214	406	-0.166	0.403	3.00	432	533	.396	.406
	1	D	.228	.298	.330	.628	.217	408	-0.158	0.415	2.91	419	410	.511	.560
	2	S	.244	.312	.343	.655	.226	426	-0.167	0.433	3.36	484	504	.480	.543
	3	G	.243	.298	.354	.652	.223	416	-0.150	0.438	3.27	471	417	.561	.534
	4	T	.255	.307	.354	.661	.227	425	-0.158	0.441	3.35	482	443	.542	.493
	5	C	.245	.305	.324	.629	.218	413	-0.167	0.412	3.05	439	496	.439	.441
	6	B	.239	.296	.333	.629	.216	407	-0.155	0.418	2.94	423	587	.342	.353

参考文献

[1] http://npb.jp/：NPB. jp 日本野球機構, 最終アクセス日：2018.10.15

[2] http://baseballdata.jp/：データで楽しむプロ野球, 最終アクセス日：2018.10.15

[3] https://baseball-data.com/：プロ野球データ Freak, 最終アクセス日：2018.9.13

[4] https://ja.wikipedia.org/wiki/：Wikipedia—野球の各種記録, 最終アクセス日：2018.5.15

[5] https://bellcurve.jp/statistics/glossary/660.html：統計 WEB—統計用語集—因子負荷量, 最終アクセス日：2018.5.19

[6] 小西貞則, 多変量解析入門—線形から非線形へ, 岩波書店, 2010

[7] 森田浩, 図解入門ビジネス 多変量解析の基本と実践がよ〜くわかる本, 秀和システム, 2014

[8] 室淳子, 石村貞夫, Excel でやさしく学ぶ多変量解析 [第 2 版], 東京図書株式会社, 2007

[9] 下川敏雄, 実践のための基礎統計学, 講談社, 2016

[10] 廣瀬英雄, 実例で学ぶ確率・統計, 日本評論社, 2014

[11] データスタジアム株式会社, 野球 × 統計は最強のバッテリーである セイバーメトリクスとトラッキングの世界, 中央公論新社, 2015

[12] 有馬哲, 石村貞夫, 多変量解析のはなし, 東京図書株式会社, 2002

[13] 石村貞夫, 石村光資郎, 入門はじめての多変量解析, 東京図書株式会社, 2007

[14] 井上雅裕, 陳新開, 長谷川浩志, システム工学—定量的な意思決定法—, 株式会社オーム社, 2018