6回終了時点での失点数を元に
プロ野球の順位決定要素を考察する

芝浦工業大学　数理科学研究会
BV17057　西脇 友哉

平成 30 年 2 月 7 日

目 次

1　導入

1.1　前提

- スペースの都合により, 球団名の略称の表記を以下の通りに使用する.

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

- 試合比率, 累積試合比率, 年間勝率（全て後述）などの 1 以下の値をとる指標は有効数字 3 桁で表記しているため, 足し合わせた際に本来和が 1 になるはずの箇所が（丁度）1 にならないなどの誤差が生じることがある.
- 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が（例：打率 2 割 8 分 6 厘→ .286）, 本研究では指標を統計的に処理することから数値としての側面を意識し, 省略せずに表記している.
- 文章中の n はある一試合における 6 回終了時点での失点数を表す.
- 文章中の (p-q-r) は成績が「p 勝 q 敗 r 分」であることを表す.
- 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.

1.2　研究背景

MLB（Major League Baseball）と NPB（Nippon Professional Baseball Organization）での **QS**（クオリティ・スタート）が持つ価値の違いについて興味があり, 「多くの試合で 6 回までを無失点に抑えられるものの勝率が悪い球団」「6 回までを無失点に抑えたときの勝率は高いがそういった試合が少ない球団」などが存在することを発見した. そこで, **6 回 n 失点** ($n = 0, 1, 2, \cdots$) という局面について, 各 n に対し試合数と勝率という 2 項目での評価を基にして考察を図った.

1.3　問題提起 ～6 回という位置づけについて～

延長戦に入らなければ野球の攻防は 9 回で終了するため, 6 回終了時点であれば 1 試合の 2/3 が終わったことになる. この段階で試合の大勢が見えてきていることも多いが, 場合によっては 7 回以降の逆転劇も起こりうる. それでも, 初回や 3 回でのビハインドよりも 6 回でのビハインドを覆す方が難しいことは明らかである. また, 9 回を無失点に抑えて負けるというケースはほとんど無い[1]のに対し, 6 回までを無失点に抑えたからといって必ずしも勝てるとは限らない[2]. そこに私は野球の面白さを感じているのだが, この状況は野球以外のスポーツや他の物事にも応用可能で, ある長期の計画に対して全行程の 2/3 が終了した時点での進行状況から全体の成功の可否を考察することは有意義である. また野球そのものにおいても, 首脳陣が 6 回を軸に試合の組み立てを行うことで采配に根拠が生まれやすくなる[3]ので, この研究の意義や独創性は確かに存在する.

[1] 9 回まで両チーム無得点のまま延長戦に入り, 10 回以降に失点して負けるという場合のみ.
[2] 負ける要因としては打線が得点出来なかった場合, 好投した先発投手に替わって登板した救援投手が失点して逆転負けを喫する場合など.
[3] 例えば, ある条件を満たしていたことで 6 回終了時点でその試合が敗色濃厚であると分かれば, チーム内で力量の劣る投手（いわゆる敗戦処理）を躊躇無く起用出来る.

1.4 目標

過去の成績 ('14〜'17) を元に, 2018 年シーズン以降各球団がリーグ優勝を目指すにあたって達成すべき基準を定め, 表の形式でまとめることを最終目標とする.

2 下準備

2.1 リーグ優勝に求められる年間成績

まず, 6 回 n 失点についての考察を行う前に, リーグ優勝にはどれ程の成績が必要になるか考えなければならない. 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 0.596, 中央値は 0.589 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ではない.[4] 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.[5] 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**判別分析**の結果から優勝ラインを求めることにした.

2.2 マハラノビス距離を用いた一次元の判別分析

2.2.1 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, ここで用いる関数の定義である.

表 1: '07〜'17 の 1 位, 2 位チーム

		'17	'16	'15	'14	'13	'12	'11	'10	'09	'08	'07
	パ	ソ	日	ソ	ソ	楽	日	ソ	ソ	日	西	日
1 位群	セ	広	広	ヤ	巨	巨	巨	中	中	巨	巨	巨
	パ	西	ソ	日	オ	西	西	日	西	楽	オ	ロ
2 位群	セ	神	巨	巨	神	神	中	ヤ	神	中	神	中

表 2: '07〜'17 の 1 位, 2 位チームと年間勝率

		'17	'16	'15	'14	'13	'12	'11	'10	'09	'08	'07
	パ	0.657	0.621	0.647	0.565	0.582	0.556	0.657	0.547	0.577	0.543	0.568
1 位群	セ	0.633	0.631	0.539	0.573	0.613	0.667	0.560	0.560	0.659	0.596	0.559
	パ	0.564	0.606	0.560	0.563	0.529	0.533	0.526	0.545	0.538	0.524	0.555
2 位群	セ	0.561	0.507	0.528	0.524	0.521	0.586	0.543	0.553	0.566	0.582	0.549

x :年間勝率（変数）

μ_k :k 位群の平均

σ_k :k 位群の標準偏差

D_k :k 位群からのマハラノビス距離

$D_k^{'}$:k 位群からのユークリッド距離

[4]「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.

[5]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, この 10 年間で年間勝率 0.539 を上回っていながら 2 位以下に終わったチームは延べ 18 チーム存在する. なお, 私個人の意見としては年間勝率が 0.539 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.

新たなサンプル（2018 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, $D_1^{'} = \mu_1 - x, D_2^{'} = x - \mu_2$ 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.

2.2.2 優勝ラインの計算

NPB の公式 HP から収集したデータを基に,

$$\mu_1 = 0.596, \quad \sigma_1 = 0.043,$$
$$\mu_2 = 0.548, \quad \sigma_2 = 0.024$$

を得た. 優勝ラインは $D_1 = D_2$ となる勝率より求められるので,

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

を得る. さらに x について解くと,

$$\sigma_2(\mu_1 - x) = \sigma_1(x - \mu_2)$$
$$x(\sigma_1 + \sigma_2) = \sigma_2\mu_1 + \sigma_1\mu_2$$

両辺を $\sigma_1 + \sigma_2 \neq 0$ で割り,

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

そこに $\mu_1, \mu_2, \sigma_1, \sigma_2$ の数値を代入すると,

$$x = \mathbf{0.565}$$

を得る. これを優勝ラインとして定める. 年間勝率 0.565 に十分近い数値であり, かつ現実的な引き分け数の中で[6]勝敗成績を考えると, (78-60-5)[7]を目安とするのが妥当である.

3 年間勝率 0.565 の達成に必要な要素

優勝ラインが定まったので, 次はそこに到達するための要素が必要となる. 今回は冒頭に述べた通り, 6 回までの攻防の結果から得られた情報を基に考察する. 具体的には, 「6 回 n 失点の試合を何試合生み出すことが出来たのか（量）」「6 回を n 失点に抑えたときにどれ程の割合で勝利したか（質）」がその要素であり, 便宜上それらを以下のように名付ける.

用語 1 (試合比率)**.** 年間の公式戦試合数[8]に対して 6 回 n 失点の試合数が占める割合を, **試合比率**と呼ぶことにする.

用語 2 (状況別勝率)**.** 「6 回 n 失点のとき」という特定条件下での勝率を, 年間勝率と区別して**状況別勝率**と呼ぶことにする.

[6]勝率は（勝利数）÷（勝利数+敗戦数）, もしくは（勝利数）÷（試合数-引き分け数）によって求められるため, 0.565 に近づける過程で引き分け数を調整すれば (78-60-5) よりも近い数値を得られる可能性はある. しかし, プロ野球の公式戦で引き分けの試合は各チーム年に数試合程度であるという実情を踏まえれば, 引き分け数が不自然に多い勝敗成績はモデルとして相応しくない.

[7]ちなみに'14 ソフトバンク（1 位）の成績は (78-60-6).

[8]～'14:144 試合, '15～:143 試合

使用例：'17 楽天 ($n = 1$) について考える．このケースでの対戦成績は, (21-7-0), 計 28 試合であった．この場合, 試合比率は $28 \div 143 = 0.196$, 状況別勝率は $21 \div (21 + 7) = 0.750$ となる．

試合比率と状況別勝率を用いると, 6 回までの攻防の結果に対して 2 つの評価基準を使用することができ, それらを年間勝率に影響を及ぼす要因として捉えることもできる．また, 試合比率を用いることで, 状況別勝率と同様に数値が 0 以上 1 未満の範囲に収まり, 両者を比較しやすくなる．以下にその例として 2015 年の千葉ロッテマリーンズ（パ・リーグ 3 位）の公式戦成績を示した．

表 3: 6 回までの失点と試合の勝敗（'15 ロッテ）

n (6 回までの失点)	勝	敗	分	試合数	試合比率	累積試合比率	状況別勝率
0	17	2	0	19	0.133	0.133	0.895
1	18	9	0	27	0.189	0.322	0.667
2	20	10	0	30	0.210	0.531	0.667
3	13	14	0	27	0.189	0.720	0.481
4 以上	5	34	1	40	0.280	1.000	0.128
計 (年間勝率)	73	69	1	143	1.000	1.000	0.514

4　年間勝率を重回帰式で表現する

表 3 のような形式で各球団の成績を並べても大まかな傾向をつかむことは可能だが, 数式として表現することでより精密な分析を行うことが出来る．今回はその手段として重回帰分析を用いた．

4.1　手順

$n = 0, 1, 2, 3$ のとき, また $n \geq 4$ のときに分類し, n を固定する．その後, **試合比率** (x_1) と**状況別勝率** (x_2) を説明変数, **年間勝率** (Y) を目的変数として重回帰分析を行うと, 以下のような形式で重回帰式が得られる．

ただし, Y_n は 6 回 n 失点時の年間勝率, $Y_n^{'}$ は 6 回 n 失点以上の時の年間勝率を表す．

$$
\begin{aligned}
Y_0 &= a_1 x_1 + a_2 x_2 + a_0 \\
Y_1 &= b_1 x_1 + b_2 x_2 + b_0 \\
Y_2 &= c_1 x_1 + c_2 x_2 + c_0 \\
Y_3 &= d_1 x_1 + d_2 x_2 + d_0 \\
Y_4^{'} &= e_1 x_1 + e_2 x_2 + e_0
\end{aligned}
$$

偏回帰係数及び定数項 a, b, c, d, e の具体的な数値を求めることにより重回帰式は完成される．

4.2　多重共線性

重回帰分析において, 非常に強い相関関係のある説明変数が含まれているときに起こる状態を, **多重共線性**という．多重共線性が認められると, データの値のわずかな変化によって偏回帰係数の推定値が大きく変わってしまうことがあり, 結果的に予測がうまくいかなくなることがある．

試合比率 x_1 と状況別勝率 x_2 の相関係数 r は

$$
r = \begin{cases}
0.12 & (n = 0) \\
-0.02 & (n = 1) \\
0.08 & (n = 2) \\
0.13 & (n = 3) \\
-0.13 & (n \geq 4)
\end{cases}
$$

であり, どの n についてもほぼ無相関であったため, 今回は問題はないものとする.

5　重回帰式の導出

今回は**標本分散共分散行列**と**重回帰式による分散共分散行列**との比較による方法を採用する.

定義 5.1 (分散共分散行列)**.**

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

記号で表すと

$$\begin{pmatrix} Var(y) & Cov(y,x_1) & Cov(y,x_2) \\ Cov(y,x_1) & Var(x_1) & Cov(x_1,x_2) \\ Cov(y,x_2) & Cov(x_1,x_2) & Var(x_2) \end{pmatrix}$$

5.1　$n=0$ の場合（計算過程含む）

5.1.1　偏回帰係数 a_1, a_2 の導出

$n=0$ のとき, 標本分散共分散行列を計算すると以下のようになる.

$$\begin{pmatrix} Var(y) & Cov(y,x_1) & Cov(y,x_2) \\ Cov(y,x_1) & Var(x_1) & Cov(x_1,x_2) \\ Cov(y,x_2) & Cov(x_1,x_2) & Var(x_2) \end{pmatrix} = \begin{pmatrix} 0.005546 & 0.001589 & 0.003664 \\ 0.001589 & 0.001984 & 0.000487 \\ 0.003664 & 0.000487 & 0.007896 \end{pmatrix}$$

次に, 重回帰式による分散共分散行列

$$\begin{pmatrix} Var(Y_0) & Cov(Y_0,x_1) & Cov(Y_0,x_2) \\ Cov(Y_0,x_1) & Var(x_1) & Cov(x_1,x_2) \\ Cov(Y_0,x_2) & Cov(x_1,x_2) & Var(x_2) \end{pmatrix}$$

を求めたい. Y_n は完成した（偏回帰係数の出揃った）重回帰式に x_1, x_2 の数値を代入して得られる値であるので, 現時点では $Var(Y_0)$ を求めることは出来ない. また, $Var(x_1), Var(x_2), Cov(x_1,x_2)$ は既に求めている. 故に, $Cov(Y_0,x_1)$ と $Cov(Y_0,x_2)$ の導出についてのみを考えればよい.

公式 1 (分散と共分散の公式)**.**

$$Cov(x, ay+bz+c) = aCov(x,y) + bCov(x,z) \quad (a,b,c : const.)$$

証明.

$$\begin{aligned} u &:= ay+bz+c \\ Cov(x,u) &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(u_i-\overline{u}) \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})\{(ay_i+bz_i+c)-(a\overline{y}+b\overline{z}+c)\} \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})\{a(y_i-\overline{y})+b(z_i-\overline{z})\} \\ &= a\cdot\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}) + b\cdot\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(z_i-\overline{z}) \\ &= aCov(x,y)+bCov(x,z) \end{aligned}$$

□

この公式を利用して, 重回帰式 $Y_0 = a_1x_1 + a_2x_2 + a_0$ と x_1 の共分散 $Cov(Y_0, x_1)$ は

$$\begin{aligned} Cov(Y_0, x_1) &= Cov(a_1x_1 + a_2x_2 + a_0, x_1) \\ &= a_1Cov(x_1, x_1) + a_2Cov(x_2, x_1) \\ &= a_1Var(x_1) + a_2Cov(x_1, x_2) \\ &= 0.001984a_1 + 0.000487a_2 \end{aligned}$$

と計算される. さらに, 重回帰式 Y_0 と x_2 の共分散 $Cov(Y_0, x_2)$ は

$$\begin{aligned} Cov(Y_0, x_2) &= Cov(a_1x_1 + a_2x_2 + a_0, x_2) \\ &= a_1Cov(x_1, x_2) + a_2Var(x_2) \\ &= 0.000487a_1 + 0.007896a_2 \end{aligned}$$

上記より, 重回帰式 $(n = 0)$ による分散共分散行列は

$$\begin{pmatrix} Var(Y_0) & 0.001984a_1 + 0.000487a_2 & 0.000487a_1 + 0.007896a_2 \\ 0.001984a_1 + 0.000487a_2 & 0.001984 & 0.000487 \\ 0.000487a_1 + 0.007896a_2 & 0.000487 & 0.007896 \end{pmatrix}$$

となることがわかる.

また, 先に求めた標本分散共分散行列は

$$\begin{pmatrix} 0.005546 & 0.001589 & 0.003664 \\ 0.001589 & 0.001984 & 0.000487 \\ 0.003664 & 0.000487 & 0.007896 \end{pmatrix}$$

であった. ここで, 重回帰式がデータと一致しているならば, 2 つの分散共分散行列も一致するはずである.

$$\begin{aligned} Cov(y, x_1) &\leftrightarrow Cov(Y_0, x_1) \\ Cov(y, x_2) &\leftrightarrow Cov(Y_0, x_2) \end{aligned}$$

の対応関係に注目すると

$$\begin{aligned} 0.001589 &= 0.001984a_1 + 0.000487a_2 \\ 0.003664 &= 0.000487a_1 + 0.007896a_2 \end{aligned}$$

となる. これを行列で表現すると[9]

$$\begin{pmatrix} 0.001589 \\ 0.003664 \end{pmatrix} = \begin{pmatrix} 0.001984 & 0.000487 \\ 0.000487 & 0.007896 \end{pmatrix} \cdot \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$$

両辺の先頭に逆行列をかけると

$$\begin{pmatrix} 0.001984 & 0.000487 \\ 0.000487 & 0.007896 \end{pmatrix}^{-1} \cdot \begin{pmatrix} 0.001589 \\ 0.003664 \end{pmatrix} = \begin{pmatrix} 0.001984 & 0.000487 \\ 0.000487 & 0.007896 \end{pmatrix}^{-1} \cdot \begin{pmatrix} 0.001984 & 0.000487 \\ 0.000487 & 0.007896 \end{pmatrix} \cdot \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$$

$$\begin{pmatrix} 0.001984 & 0.000487 \\ 0.000487 & 0.007896 \end{pmatrix}^{-1} \cdot \begin{pmatrix} 0.001589 \\ 0.003664 \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$$

$$\begin{pmatrix} 0.001984 & 0.000487 \\ 0.000487 & 0.007896 \end{pmatrix}^{-1} = \begin{pmatrix} 511.692801 & -31.574884 \\ -31.574884 & 128.598993 \end{pmatrix}$$

より,

$$\begin{cases} a_1 &= 0.697598 \\ a_2 &= 0.421006 \end{cases}$$

を得る.

[9]そのまま連立一次方程式として解くことも可能だが非常に計算の手間がかかる.

5.1.2 定数項 a_0 の導出

定数項 a_0 は平均値 $\overline{y}, \overline{x}_1, \overline{x}_2$ を用いて

$$a_0 = \overline{y} - a_1 \times \overline{x}_1 - a_2 \times \overline{x}_2$$

で求められる.

$$\begin{cases} \overline{y} & = 0.500112 \\ \overline{x}_1 & = 0.183991 \\ \overline{x}_2 & = 0.821384 \end{cases}$$

より,

$$a_0 = 0.025953$$

を得る.

以降 $n = 1, 2, 3$ と $n \geq 4$ の場合についても重回帰式を導出するが, 全て $n = 0$ の場合と同様に計算できる.

5.2 $n = 1, 2, 3$ 及び $n \geq 4$ の場合（概略）

5.2.1 $n = 1$ の場合

$n = 1$ のとき, 標本分散共分散行列は

$$\begin{pmatrix} Var(y) & Cov(y, x_1) & Cov(y, x_2) \\ Cov(y, x_1) & Var(x_1) & Cov(x_1, x_2) \\ Cov(y, x_2) & Cov(x_1, x_2) & Var(x_2) \end{pmatrix} = \begin{pmatrix} 0.005546 & 0.000110 & 0.005380 \\ 0.000110 & 0.000817 & -0.000058 \\ 0.005380 & -0.000058 & 0.009688 \end{pmatrix}$$

であり, ここから

$$\begin{cases} b_0 & = 0.086575 \\ b_1 & = 0.174188 \\ b_2 & = 0.556363 \end{cases}$$

を得た.

5.2.2 $n = 2$ の場合

$n = 2$ のとき, 標本分散共分散行列は

$$\begin{pmatrix} Var(y) & Cov(y, x_1) & Cov(y, x_2) \\ Cov(y, x_1) & Var(x_1) & Cov(x_1, x_2) \\ Cov(y, x_2) & Cov(x_1, x_2) & Var(x_2) \end{pmatrix} = \begin{pmatrix} 0.005546 & 0.000465 & 0.004814 \\ 0.000465 & 0.001032 & 0.000245 \\ 0.004814 & 0.000245 & 0.009358 \end{pmatrix}$$

であり, ここから

$$\begin{cases} c_0 & = 0.168943 \\ c_1 & = 0.330155 \\ c_2 & = 0.505778 \end{cases}$$

を得た.

5.2.3 $n = 3$ の場合

$n = 3$ のとき, 標本分散共分散行列は

$$\begin{pmatrix} Var(y) & Cov(y, x_1) & Cov(y, x_2) \\ Cov(y, x_1) & Var(x_1) & Cov(x_1, x_2) \\ Cov(y, x_2) & Cov(x_1, x_2) & Var(x_2) \end{pmatrix} = \begin{pmatrix} 0.005546 & -0.000040 & 0.004279 \\ -0.000040 & 0.001179 & 0.000567 \\ 0.004279 & 0.000567 & 0.015951 \end{pmatrix}$$

であり, ここから

$$\begin{cases} d_0 &= 0.403880 \\ d_1 &= -0.166196 \\ d_2 &= 0.274144 \end{cases}$$

を得た.

5.2.4 $n \geq 4$ の場合

$n \geq 4$ のとき, 標本分散共分散行列は

$$\begin{pmatrix} Var(y) & Cov(y, x_1) & Cov(y, x_2) \\ Cov(y, x_1) & Var(x_1) & Cov(x_1, x_2) \\ Cov(y, x_2) & Cov(x_1, x_2) & Var(x_2) \end{pmatrix} = \begin{pmatrix} 0.005546 & -0.002124 & 0.003603 \\ -0.002124 & 0.002267 & -0.000473 \\ 0.003603 & -0.000473 & 0.006022 \end{pmatrix}$$

であり, ここから

$$\begin{cases} e_0 &= 0.646553 \\ e_1 &= -0.825338 \\ e_2 &= 0.533394 \end{cases}$$

を得た.

5.3 得られた重回帰式

以上の計算結果より, $n = 0, 1, 2, 3$ のときの重回帰式

$$\begin{aligned} Y_0 &= 0.697598x_1 + 0.421006x_2 + 0.025953 \\ Y_1 &= 0.174188x_1 + 0.556363x_2 + 0.086575 \\ Y_2 &= 0.330155x_1 + 0.505778x_2 + 0.168943 \\ Y_3 &= -0.166196x_1 + 0.274144x_2 + 0.403880 \end{aligned}$$

がそれぞれ得られた. 6 回までに 4 失点以上した試合については

$$Y_4^{'} = -0.825338x_1 + 0.533394x_2 + 0.646553$$

となった.

6 重回帰式の有用性

求めた重回帰式がデータによくあてはまっているかどうかはこの先の分析の精度に根本的に関わってくる. 重回帰式の有用性を評価するのにあたり**残差**や**重相関係数**を用いることが出来る.

6.1　残差の考察

データへのあてはまりが良いということは,

$$(\text{残差}) = (\text{実測値}) - (\text{予測値})$$

が小さいということであると言える.

ここで, **予測値**とは, 試合比率や状況別勝率のデータから見込まれる年間勝率を重回帰式から計算した値の事である. また, **実測値**とは, NPB の公式 HP に掲載された実際の年間勝率の事である.[10] つまり, 実測値は年度と球団が同じであれば異なる n に対しても同じ値が用いられる.

6.1.1　予測値と残差の導出方法

6 回無失点 ($n = 0$) の場合を考えると,

$$(\text{予測値 } Y_0) = 0.697598x_1 + 0.421006x_2 + 0.025953$$

である. （ここまではどの球団でも同じ）

例えば'17 オリックスについては

$$x_1 = 0.133, \quad x_2 = 0.684$$

であるから,

$$(\text{予測値}) = 0.407$$

となり, 実測値（年間勝率の実際のデータ y）$= 0.444$ より, 残差は 0.037 である.

(予測値) < (実測値) より, $n = 0$ の重回帰式は'17 オリックスの年間勝率を実際よりも低く見積もったといえる.

6.2　重相関係数

「予測値と実測値がどの程度一致しているのか」という視点で考える場合, 重相関係数を用いるのが有効である. 重相関係数を求めるには, 予測値と実測値の相関係数を計算すればよい.

各 n について重相関係数 R を計算した結果,

$$R = \begin{cases} 0.691 & (n = 0) \\ 0.737 & (n = 1) \\ 0.683 & (n = 2) \\ 0.461 & (n = 3) \\ 0.814 & (n \geq 4) \end{cases}$$

となり, $n = 0, 1, 2$ については R はおおむね 0.7 前後の値であったが, $n = 3$ のみ著しく低い値であった.

7　仕上げ

7.1　変数の消去

「1.3 目標」で述べた通り優勝ノルマを表形式でまとめるためには, 各 n について試合比率と状況別勝率を共に具体的な数値として示さなければならない. 得られた重回帰式だけでは変数が 2 つ残っている状態なので, 2 つの説明変数のうちのどちらかを固定（消去）する必要がある.

では試合比率と状況別勝率のうちどちらを固定するべきか. 2 つの説明変数の特徴について, **状況別勝率**は打線の援護や救援投手の出来など複数の要因が絡み, **先発投手の活躍だけで数値が上昇するとは限らない**のに対し, **試合比率**は決定要素のほとんどが先発投手の出来によると言える. このような性質から, 複雑な変数を先に処理するために状況別勝率を固定することにした.

[10] '17 西武：0.564, '14 中日：0.479 など

7.2　状況別勝率の数値設定

優勝ライン到達基準に対応した状況別勝率を各 n について求めたい. 先述の通り優勝ラインは「ぎりぎりで優勝出来る年間勝率のボーダーライン」である. 「2 下準備」で優勝ライン（年間勝率）を求めた際にはマハラノビス距離を用いたが, 今回は状況別勝率が判明しているデータの年数が少なすぎる[11]ため, マハラノビス距離を用いてサンプルの標準偏差を考慮に加える意味合いは薄いと考えた. そこで, （「マハラノビス距離による判別分析」と同じ）1 位, 2 位チームの'14〜'17 のデータの平均をとった.

（対象チーム）･･･'17：ソ, 西, 広, 神　'16：日, ソ, 広, 巨　'15：ソ, 日, ヤ, 巨　'14：ソ, オ, 巨, 神

求める状況別勝率を $\mu^{'}$ とすると,

$$\mu^{'} = \begin{cases} 0.858512 & (n=0) \\ 0.766482 & (n=1) \\ 0.602555 & (n=2) \\ 0.502667 & (n=3) \\ 0.248065 & (n \geq 4) \end{cases}$$

7.3　試合比率の計算

上記の内容によって 2 つの説明変数のうちの片方を消去したので, 最後に試合比率の具体的な数値を単純な四則演算によって求めたい. それには「5.3」で求めた各重回帰式に Y_n と x_1 の値を代入すればよい. ここで, 重回帰式の左辺（目的変数）に該当する年間勝率の予測値は優勝ラインに等しいため, n の値に関わらず Y に 0.565 を代入する. また, x_2 には前述の $\mu^{'}$ を代入する.

よって, 求めた重回帰式は

$$\begin{aligned}
(Y_0 =)0.565 &= 0.697598x_1 + 0.421006 \times 0.858512 + 0.025953 \\
(Y_1 =)0.565 &= 0.174188x_1 + 0.556363 \times 0.766482 + 0.086575 \\
(Y_2 =)0.565 &= 0.330155x_1 + 0.505778 \times 0.602555 + 0.168943 \\
(Y_3 =)0.565 &= -0.166196x_1 + 0.274148 \times 0.502667 + 0.403880 \\
(Y_4^{'} =)0.565 &= -0.825338x_1 + 0.533394 \times 0.248065 + 0.646553
\end{aligned}$$

となり, x_1 について解くと,

$$x_1 = \begin{cases} 0.255 & (n=0) \quad \cdots \text{約 36 試合} \\ 0.301 & (n=1) \quad \cdots \text{約 43 試合} \\ 0.278 & (n=2) \quad \cdots \text{約 40 試合} \\ -0.143 & (n=3) \quad \cdots \text{約-20 試合} \\ 0.259 & (n \geq 4) \quad \cdots \text{約 37 試合} \end{cases}$$

を得る. なお, 上記の試合数の整数表記は, 年間公式戦試合数である 143 に試合比率をかけた概数である.

8　優勝ノルマの提示 (1)

以上の結果を基に, 2018 年度以降に各球団がリーグ優勝を目指すにあたって達成すべき基準をまとめた.

なお, ここでの勝利数及び敗戦数は, それらが整数の値をとる範囲で**状況別勝率**が $\mu^{'}$ に最も近づくように設定されている. また, この表における状況別勝率は $\mu^{'}$ とは区別され, (勝) ÷ (勝 + 敗) の値を表す.

[11] 年間勝率は 11 年分のデータ, 試合比率や状況別勝率は 4 年分のデータを利用

表 4: リーグ優勝達成のためのノルマ（ver. 1）

n	勝	敗	分	試合数	試合比率	状況別勝率	$\mu^{'}$
0	31	5	0	36	0.255	0.861	0.859
1	33	10	0	43	0.301	0.767	0.766
2	24	16	0	40	0.278	0.600	0.603
3	-10	-10	0	-20	-0.143	0.500	0.503
4 以上	9	27	1	37	0.259	0.250	0.248
計 (結果)	87	48	1	136	0.951	0.644	
計 (理想)	78	60	5	143	1	0.565	

8.1 問題点, 改善点

リーグ優勝達成のための基準をまとめたものの, 現実に適用出来ない（使い物にならない）結果が出てしまった. 失敗した点として最も致命的であるのは「$x_1(n=3)$ が負の値をとってしまった」ことである. その根本的な原因として考えられるのは重相関係数の低い重回帰式をそのまま分析に利用したことであり, 現実的な基準を作るためには重回帰式による予測の精度を高める必要がある. 今回は $n \geq 4$ の重回帰式の重相関係数が 0.8 を超えていた[12]ことからヒントを得て, 重回帰分析の対象に「6 回 3 失点以上」「6 回 1 失点以内」などを加え, n の範囲を広げるという発想に至った.

9 （再）重回帰分析

「5 重回帰式の導出」と同様の手法で, 「$n \leq 1$」「$n \leq 2$」「$n \leq 3$」「$n \geq 1$」「$n \geq 2$」「$n \geq 3$」について重回帰分析を行ったところ, 以下の式が得られた. （$Y_n^{'}$ は 6 回 n 失点以上の時, $Y_n^{''}$ は 6 回 n 失点以内の時の年間勝率をそれぞれ表す.）

$$
\begin{aligned}
Y_1^{''} &= 0.341276x_1 + 0.699308x_2 - 0.152921 \\
Y_2^{''} &= 0.392982x_1 + 0.764796x_2 - 0.239816 \\
Y_3^{''} &= 0.453333x_1 + 0.803326x_2 - 0.324042 \\
Y_1^{'} &= -0.416564x_1 + 0.906432x_2 + 0.452775 \\
Y_2^{'} &= -0.516019x_1 + 0.841936x_2 + 0.529790 \\
Y_3^{'} &= -0.584973x_1 + 0.709203x_2 + 0.567921
\end{aligned}
$$

なお, 各重回帰式について x_1（試合比率）, x_2（状況別勝率）間の相関係数 r は

$$
r = \begin{cases}
0.17 & (n \leq 1) \\
-0.24 & (n \leq 2) \\
0.40 & (n \leq 3) \\
-0.25 & (n \geq 1) \\
0.02 & (n \geq 2) \\
-0.05 & (n \geq 3)
\end{cases}
$$

であり, 今回も**多重共線性**は認められなかった.

[12]「6.2」を参照

また, 年間勝率に対する予測値と実測値の重相関係数 R は

$$R=\begin{cases}0.856 & (n\leq 1)\\ 0.920 & (n\leq 2)\\ 0.951 & (n\leq 3)\\ 0.983 & (n\geq 1)\\ 0.945 & (n\geq 2)\\ 0.875 & (n\geq 3)\end{cases}$$

となり, $n=1,2,3$ 及び $n\geq 4$ のときに比べて軒並み高い数値となった. 特に $n\leq 3$, $n\geq 1$, $n\geq 2$ での R の値は極めて高く, 精度の高い予測が期待できる.

「7.2 状況別勝率の数値設定」や「7.3 試合比率の計算」と同様に $\mu^{'}$ を求め,

$$\mu^{'}=\begin{cases}0.815686 & (n\leq 1)\\ 0.747285 & (n\leq 2)\\ 0.701257 & (n\leq 3)\\ 0.507883 & (n\geq 1)\\ 0.421942 & (n\geq 2)\\ 0.337001 & (n\geq 3)\end{cases}$$

を得た. また, 重回帰式の変数を消去した後にそれぞれを x_1 について解き,

$$x_1=\begin{cases}0.433 & (n\leq 1) & \cdots\text{約 62 試合}\\ 0.595 & (n\leq 2) & \cdots\text{約 85 試合}\\ 0.719 & (n\leq 3) & \cdots\text{約 103 試合}\\ 0.835 & (n\geq 1) & \cdots\text{約 119 試合}\\ 0.619 & (n\geq 2) & \cdots\text{約 89 試合}\\ 0.413 & (n\geq 3) & \cdots\text{約 59 試合}\end{cases}$$

を得た.

10　優勝ノルマの提示 (2)

以上を基に改めて優勝ノルマを設定したい. その際に, 各 n の具体的な勝敗数をどのように定めるかが問題となるが, 各 n の試合数の合計がなるべく 143 に近い値を取るようにしたい. そこで, 「9 （再）重回帰分析」より試合数の和が約 144 であった $n\leq 2$ と $n\geq 3$ を軸に決めていくことにする. n の範囲を組み合わせて考えた結果,

- $n=0\cdots$「$n\leq 2$」−「$n=2$」−「$n=1$」
- $n=1\cdots$「$n\geq 1$」−「$n\geq 2$」
- $n=2\cdots$「$n\leq 2$」−「$n\leq 1$」
- $n=3\cdots$「$n\leq 3$」−「$n\leq 2$」
- $n\geq 4\cdots$「全体 (78-60-5)」−「$n\leq 3$」

の計算結果から勝敗数を導くことにした. （「」内は勝敗数の計算に用いる重回帰式の範囲）
以下が表の改訂版である.

表 5: リーグ優勝達成のためのノルマ（ver. 2）

n	勝	敗	分	試合数	試合比率	累積試合比率	状況別勝率
0	29	4	0	33	0.231	0.231	0.879
1	22	7	1	30	0.210	0.441	0.759
2	12	10	1	23	0.161	0.602	0.545
3	9	10	0	19	0.133	0.735	0.474
4 以上	6	29	3	38	0.266	1	0.171
計	78	60	5	143	1	1	0.565

11 今後の計画

11.1 展望

利用したデータの年数をさらに増やし,「$n = m$」よりも「$n \leq m$」や「$n \geq m$」($m = 0, 1, 2 \cdots$) の重回帰式の方が精度が良くなった理由を考察し, リーグ優勝の他に CS[13]出場（3 位以内）や最下位回避（5 位以内）のための基準も設定していく. 優勝ラインと同様に, CS 出場や最下位回避に必要な年間勝率をそれぞれ「CS 出場ライン」「最下位回避ライン」と名付けるならば, 判別分析によってそれらは「0.505」「0.433」という値になる.

11.2 課題

元々QS は投手個人の記録という側面があるが, 今回はチームの失点を対象に資料を収集したため, 先発投手一人一人の成績まで踏み込んだ考察を行うことが出来なかった. 今後選手個人に着目していく場合は, QS の達成条件が「6 回以上かつ**自責点** 3 以内」であることに注意し, 失点と自責点の違いを正しく処理する方法を考えなければならない.

参考文献

[1] `http://npb.jp/` : NPB. jp 日本野球機構

[2] 森田 浩 著, 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

[3] 室 淳子, 石村貞夫 著, 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

[13]CS : クライマックスシリーズ. 日本シリーズへの出場権を懸けたポストシーズンゲーム. 各リーグのペナントレース上位 3 球団が出場可能. 制度の詳細については Web サイトなどを参照されたい.