プロ野球投手の統計的分析
セイバーメトリクスの役割

2006MI110 村田一樹
指導教員：木村美善

1　はじめに

現在，セイバーメトリクスという野球の概念が注目されている．メジャーリーグでは野球の WEB サイトに大きく取り上げられるほどセイバーメトリクスの統計的手法を用いた指標は認識されている. それに対し日本プロ野球ではあまり認識されておらずその名前も知らない野球ファンは多い．本研究では日本のプロ野球投手に着目した. その理由としては, 投手の能力を測る「防御率」,「奪三振率」等の指標があるが, これらの値が本当にその投手の能力を測るものであるかと疑問を持ったからである. 従来の指標とは異なり, 運の要素をあまり含まず選手の本当の価値を表す指標がセイバーメトリクスの指標であるならば, これからの野球の戦略や経営面で大きく影響するものではないかと考えた.

2　セイバーメトリクス

野球には打率や打点、防御率など多くの価値基準・指標基準が存在するが, 統計的視点から選手の評価方法や戦略を考え, 分析してできた様々な指標の総称をセイバーメトリクスという.([1] 参照)

3　データについて

プロ野球データリーグ [4] と KazmixWarld–BaseballData[5] から 2007 年度と 2008 年度で規定投球回に到達している先発投手データを用いた. 従来の指標からは「防御率」,「被本塁打」,「奪三振」,「与四球」,「与死球」,「失点」,「自責点」,「投球回」,「勝率」,「完投」,「完封」,「無四死球」,「旧年棒」,「年棒率」の 14 項目を用いた. セイバーメトリクスの指標からは「WHIP」,「RSAA」,「被 BABIP」,「DIPSera」,「K.BB」,「QS%」6 項目を用いた.([2],[3] 参照)

4　指標の比較

従来の指標とセイバーメトリクスの指標をそれぞれ重回帰分析を用いて比較を行った. 目的変数を「勝率」として分析を行い, 変数選択にはステップワイズ法を用いた. 回帰診断を行うことにより, データのあてはまりや外れ値の検定も行った.

4.1　従来の指標「目的変数：勝率」

「勝率」を目的変数とし, 説明変数は「防御率」,「被本塁打」,「奪三振」,「与四球」,「与死球」,「失点」,「自責点」,「投球回」,「完投」,「完封」,「無四球」,「旧年棒」,「年棒率」を用いて分析を行った.

4.1.1　分析結果

変数選択して得られた回帰式の残差を用いて, 回帰診断を行い外れ値を発見したので,2007 年の和田 (H) のデータを削除して分析を行った．重回帰分析の結果を表 1 に示した. このときの決定係数は 0.5641 であり, 調整済み決定係数は 0.5341 であった.
「年棒率」が有意水準 1 %で効いており．「自責点」,「失点」が有意水準 5 %で効いており, そして「防御率」が有意水準 10 %で効いているという結果になった．この時の回帰診断では, QQ プロットに外れ値がないことから正規性が満たされている．等分散であり, クックの距離にも問題は見られない．推定式は $x1$=防御率，　$x2$=被本塁打，　$x3$=失点，　$x4$=自責点，　$x5$=完封，　$x6$=年棒率とおくと,

$$\hat{y} = 0.845495 - 0.089546x_1 - 0.004982x_2 - 0.008445x_3 + 0.009402x_4 + 0.021883x_5 + 0.042482x_6.$$

表 1　従来の指標 (目的変数：勝率) の重回帰分析の結果

項目	回帰係数	標準誤差	t-統計量	p-値	
防御率	-0.089546	0.044651	-2.005	0.0517	.
被本塁打	-0.004982	0.002986	-1.668	0.1031	
失点	-0.008445	0.003509	-2.407	0.0208	*
自責点	0.009402	0.004413	2.13	0.0393	*
完封	0.021883	0.013068	1.674	0.1018	
年棒率	0.042482	0.013991	3.036	0.0042	**
定数項	0.845495	0.08391	10.076	1.55E-12	***

4.2　セイバーメトリクスの指標「目的変数：勝率」

「勝率」を目的変数とし, 説明変数は「WHIP」,「RSAA」,「DIPSera」,「被 BABIP」,「K.BB」,「QS%」を用いて分析を行った.

4.2.1　分析結果

重回帰分析の結果を表 2 に示した. このときの決定係数は 0.6016 であり, 調整済み決定係数は 0.5646 である.
「RSAA」,「DIPSera」が有意水準 5 %で効いており．「被 BABIP」が有意水準 10 %で効いているという結果となった．回帰診断結果は QQ プロットに外れ値がないことから正規性が満たされており, 等分散性であり, クックの距離にも問題は見られない．推定式は
x_1=RSAA,x_2=DIPSera,x_3=被 BABIP,x_4=QS %とおくと,

$$\hat{y} = 1.5534 + 0.0041x_1 - 0.1111x_2 - 1.3924x_3 - 0.0034x_4.$$

4.3　重回帰分析の考察

結果として勝率に関して大きな決定係数が得られたのはセイバーメトリクスの指標であった．決定係数に関し

表 2　セイバーメトリクスの指標「目的変数：勝率」の重回帰分析の結果

項目	回帰係数	標準誤差	t-統計量	p-値	
RSAA	0.004165	0.001910	2.180	0.01474	*
DIPSera	-0.111113	0.044527	-2.495	0.01649	*
被 BABIP	-1.392433	0.813353	-1.712	0.09410	.
QS%	-0.003414	0.002035	-1.677	0.10072	
定数項	1.553443	0.439144	3.537	0.000983	***

てはセイバーメトリクスの指標は従来の指標に勝っているが, 各説明変数の p-値が従来の指標に比べ低いという結果となった. その理由はセイバーメトリクスの指標は従来の指標を合成して出来ているために重回帰分析を行うと各指標の個性が上手く現れてこなかったことが考えられる.

5　多重共線性の問題

重回帰分析を行ってみると, セイバーメトリクスの指標は単体でも大きな内容を保有していることがわかる. そのためにセイバーメトリクスの各指標の間に大きな相関が存在するのではないかと考えられる. 変数が厳密に線形従属でなくても, 変数間に強い相関関係が存在する場合には, 係数の推定値の分散が大きくなり, 推定結果の信頼性が低下してしまう. データは重回帰分析で用いた目的変数を「勝率」としたものを用いる. まず多重共線性の有無を調べた, その後にリッジ回帰による重回帰分析の結果と比較する. 目的変数を「勝利」とした従来の指標のデータについて同じ分析と比較を行う.

5.1　リッジ回帰結果:セイバーメトリクスの指標

重回帰分析の結果に基づき, 目的変数を「勝利」とし, 説明変数に「RSAA」,「DIPSera」,「被 BABIP」,「QS %」を用いて分析を行った. 説明変数間の相関行列の最小の固有値は 0.1112 であり,VIF は 6.7 であった. したがって多重共線性の心配はそれほどなく, リッジ回帰を行う必要はないといえる.

5.2　リッジ回帰結果:(従来の指標)

目的変数に「勝利」を用い, 説明変数に「防御率」,「被本塁打」,「失点」,「自責点」,「完投」,「年棒率」を用いて分析を行った. 説明変数間の相関行列の最小の固有値は 0.0296 という 0 に近い値となり VIF は 8.8 であったので多重共線性の疑いがある.

リッジ回帰のプロットを図 7.2 に示した.

図 7.2 から多重共線性の存在が確認出来る. 図 7.2 から回帰係数を縮小する k の値は 0.4 付近と考えられる.

5.3　考察

予想とは異なりセイバーメトリクスの指標には多重共線性の心配は見られなかった. これはセイバーメトリクスの指標の一つ一つが多くの内容を保有しつつも個々が別々の意味を持っていることを示していると考えられる. 従来の指標で多重共線性が発見された大きな理由としては「失点」と「自責点」という相手に点を取られるという点で似ている指標が存在していることが理由に挙げられる. さらに点を取られた場合に変動する指標である「防御率」も含まれているため, 多重共線性への影響が考えられる.

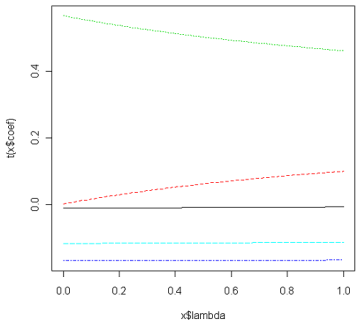

図 1　リッジ回帰プロット：従来の指標

6　セイバーメトリクスの役割

いくつかセイバーメトリクスの役割については次のことが.

- 選手の総合的な評価に適している.
  「QS %」従来の指標では分かりにくい安定感を表す指標であり,「RSAA」はリーグ平均と比べた失点率「RSAA」であり一つの指標でおおまかな選手の能力を掴める. 特に RSAA については勝率に対する決定係数も高く信頼性もある.
- 長期の評価で有効である
  セイバーメトリクスは統計学から出来た指標であるので短期間でのデータでは分析を行おうとしても値が高すぎることや, 低すぎることがあり外れ値となってしまう.
- 選手の隠れた能力を発見する.
  主成分分析を分析してみてみると選手の以外な能力を見つけることができる.2008 年度のライオンズの岸は従来の指標ではとりわけ目立つ指標はないが, セイバーメトリクスの指標を見ると K.BB が高く QS %も高い値なので一人で試合を作る力を持った選手であるといえる.

7　おわりに

本研究を終えて, セイバーメトリクスは, ファンが野球を楽しむためのアイテムであったり, マネージャーがチームの方針を決定するアイテムのひとつであったりするものであり, これから日本でももっと普及する可能性があると考える.

参考文献

[1] 高橋知也：「セイバーメトリクスについての分析」, 南山大学数理情報学部数理科学科卒業論文,2008

[2] プロ野球データリーグ, http://npbdl.web.fc2.com/

[3] Kazmix Warld–Baseball Data, http://www.kazmix.com/