仮想的に再現した打撃成績に基づくプロ野球の勝率予測

芝浦工業大学 数理科学研究会 BV17057 西脇 友哉

令和元年 5 月 19 日

研究背景

プロ野球の分析を行う際に収集できるデータの量は限られているため, MATLAB を用いたシミュレーションによって大量のデータを擬似的に生成したいと考えた.[*1] 精密な結果を得るためには公式戦全試合を再現すればよいが, これは“野球ゲームのペナントモードをオート進行する”という形で既に実装されていると言っても過言では無い. そこで, 簡易的な方法として年間成績のみの再現を図り, 1 つ 1 つの試合結果は無視した.

1 方針

1.1 準備

シミュレーションの準備として, 正規乱数を使用したいデータに対し QQ プロットの作成とコルモゴロフ・スミルノフ検定を行い, **正規性**（データが正規分布に従うこと）を仮定する.

1.2 得点数の再現

2 次元正規乱数で OBP（出塁率）と SLG（長打率）の組を出力する. 式 (1) が表す OPS[*2]及び式 (2) の単回帰式を用いて得点数を推定する.

$$\mathrm{OPS} = \mathrm{OBP} + \mathrm{SLG} \tag{1}$$

$$(\text{得点数}) = -757.003 + 1890.903 \times \mathrm{OPS} \tag{2}$$

1.3 ピタゴラス勝率の最適化

OBP 及び SLG とは異なる正規乱数で失点数を出力し, 勝率を**ピタゴラス勝率**で近似する.

$$(\text{ピタゴラス勝率}) = \frac{1}{1+\rho^x} \quad \left(\rho = \frac{\text{失点数}}{\text{得点数}}\right) \tag{3}$$

一般的に ρ^x の指数部は $x = 2$ として用いられるが, 厳密には競技の種類によって異なる指数を要する. 本研究では, 式 (4) の**平均絶対偏差** (MAD[*3]) を最小化する x をピタゴラス勝率の最適な指数 x^* とする. (m はサンプル数, ω はシーズン中の勝率)

$$\mathrm{MAD}(x) = \frac{1}{m}\sum_{i=1}^{m}\left|\omega_i - \frac{1}{1+\rho_i^x}\right| \tag{4}$$

NPB では $x^* = 1.69$, $\mathrm{MAD}(x^*) = 0.0230$ と計算される.

2 主結果

2.1 生成した仮想データ

6 チームを 1 リーグ, 2 リーグ 12 チームを 1 年度分として扱い, データを区切る. また, 各リーグについてピタゴラス勝率の降順をシーズンの順位とする. 今回は延べ 6000 チームの成績を出力したため, 500 年分のデータに相当する. ここでは紙面の都合上 1 リーグの 1 年度分のみを抜粋して表 1 に示した.

表 1　シミュレーション結果の抜粋（仮想データ 1 年分）

順位	OBP	SLG	OPS	得点	失点	勝率
1	.345	.423	.768	695	488	.645
2	.324	.411	.736	634	520	.583
3	.327	.373	.700	567	474	.575
4	.302	.368	.670	509	569	.453
5	.332	.360	.692	551	684	.410
6	.289	.350	.639	452	630	.363

2.2 考察（再現性の評価）

仮想データを実データと比較し, 結果の妥当性を考察する. その判断基準として, 順位別に勝率の平均をとった. 表 1 の様に上位と下位の差が開いたサンプルもあるが, データ数を増やすと表 2 の様に平均値が実データに近づく.

表 2　順位別平均勝率

順位	実データ	仮想データ	絶対誤差
1	.6068	.6059	.0009
2	.5476	.5539	.0062
3	.5136	.5149	.0012
4	.4801	.4830	.0029
5	.4474	.4472	.0002
6	.4053	.3967	.0086

今後の課題

今回はモデリングとプログラムの実装に終始してしまった部分があるので, 本研究の結果を利用した分析や, 異なる条件でシミュレーションを行った際の精度の比較などを行いたい.

参考文献

[1] NPB. jp 日本野球機構, `http://npb.jp/`, 最終アクセス: 2019.4.28

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

[3] Amy N. Langville, Carl D. Meyer (訳) 岩野和生, 中村英史, 清水咲里, レイティング・ランキングの数理, 共立出版, 2015

[*1] 以下, シミュレーションによって擬似的に生成したデータを「仮想データ」, 過去の試合結果に基づくデータを「実データ」と呼ぶことにする.

[*2] On-base Plus Slugging の略語で, 打撃の成績を評価する指標の一つ.

[*3] MAD : mean absolute deviation