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

西脇 友哉

芝浦工業大学 数理科学研究会

May 19, 2019

発表の目次

研究背景

- プロ野球のデータ分析 (Part 4)
- 昨年 11 月の芝浦祭及び今年 3 月の第 8 回サイエンス・インカレで発表した OPS の研究[1]に付随する内容
- データ分析にはデータの収集が付きものだが, プロ野球のデータは無限に使えるわけではない[2]. その上収集には手間もかかる
- シミュレーション（仮想実験）を行えば大量のデータを獲得できるのではないかと考えた

[1]西脇友哉,『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〜』, 2018

[2]楽天球団が参入する以前などの古いデータは, さすがに使いづらい

本研究のスタンス

- プログラムは MATLAB で記述する
- 方向性：「現在の環境が維持されると仮定してプロ野球のペナントレースを 6000 年繰り返した場合, どのような結果が得られるか？」[3]
- 公式戦全試合を再現したいのであれば野球ゲームの“ペナントモード”をオート進行すれば良い
- 発表者はプログラミングが大の苦手なので,
なるべく簡素な設定でシミュレーションを行いたい
- チームの年間成績のみを再現し, 1 つ 1 つの試合結果と個人成績は完全に無視する

[3] もちろん現実に観測できる訳は無いので, コンピューター上で擬似的に結果を得る.

研究の進行手順

1. データの収集（※先行研究で実施済）
2. モデリング（シミュレーションの計画を立てる）
3. 実装（プログラムを組む）
4. 分析, 考察

準備

データの構成要素

1. 基本情報（サンプル番号, 年度, リーグ）
2. 得点数
    - 出塁率 (OBP) On-base percentage
    - 長打率 (SLG) Slugging percentage
    - OPS (On-base Plus Slugging)
3. 失点数
4. 勝率（ピタゴラス勝率）
5. 順位

用語の紹介

用語 (実データ)

実際のプロ野球の試合結果に基づくデータ

用語 (仮想データ)

実データの数値を元にシミュレーションを行い, 擬似的に生成したデータ

野球の指標

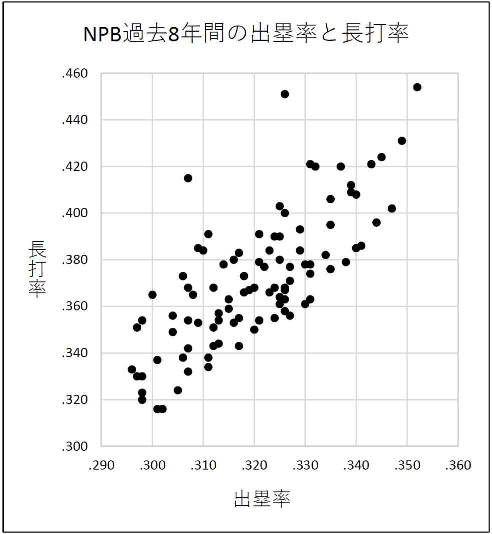
$$\text{出塁率 (OBP)} = (\text{安打} + \text{四球} + \text{死球}) \div (\text{打数} + \text{四球} + \text{死球} + \text{犠飛})$$

$$\text{長打率 (SLG)} = \text{塁打} \div \text{打数}$$

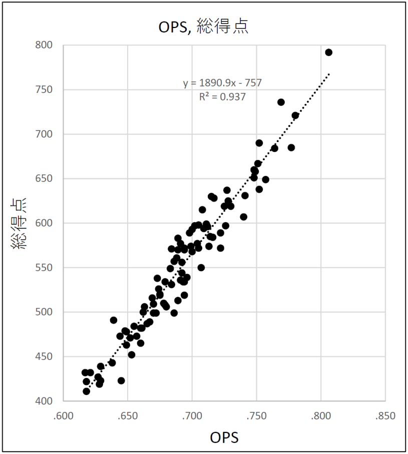
$$\text{OPS} = \text{出塁率} + \text{長打率}$$

出塁率と長打率の散布図 ('11～'18)

NPB の公式 HP より収集した実データ



OPS, 総得点の散布図と回帰直線 ('11～'18)



OPS($\hat{y}$) と総得点数[4](x) の単回帰式は

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

[4]各チームにおける 1 年間の得点数

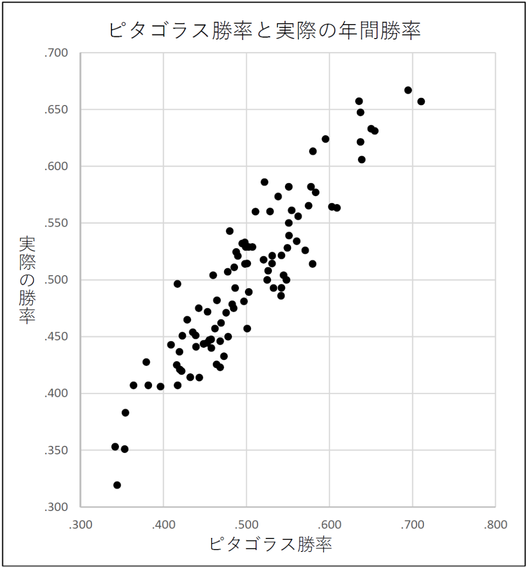
ピタゴラス勝率とは

- ビル・ジェームズにより提唱
- 得点と失点から勝率を推定するために用いられる
- 名前の由来は式の形がピタゴラスの定理と似ている事
- '11〜'18 の 8 年間について, 実際の勝率との相関係数は $r_{xy} = 0.920$（母相関係数に対する 95% 信頼区間は $[0.883, 0.946]$）

ピタゴラス勝率

$$\text{ピタゴラス勝率} = \frac{(\text{総得点})^2}{(\text{総得点})^2 + (\text{総失点})^2} = \frac{1}{1+\rho^2} \quad \left(\rho = \frac{\text{総失点}}{\text{総得点}}\right)$$

ピタゴラス勝率と実際の年間勝率の散布図 ('11〜'18)



ピタゴラス勝率の最適化 (1/3)

厳密には競技の種類によって異なる指数を要するため,
ρ^2 を ρ^x で置き換えて最適化する

平均絶対偏差 MAD(mean absolute deviation)

$$\begin{cases} \omega_i & = \text{チーム } i \text{ のシーズン中の勝率} \\ \rho_i & = \dfrac{\text{チーム } i \text{ の失点数}}{\text{チーム } i \text{ の得点数}} \end{cases} \tag{1}$$

とする. リーグに m チームいるとき, x の与えられた値に対する平均絶対偏差は

$$\mathrm{MAD}(x) = \frac{1}{m} \sum_{i=1}^{m} \left| \omega_i - \frac{1}{1 + \rho_i^x} \right| \tag{2}$$

となる.

1. 一般化したピタゴラス勝率

$$\frac{1}{1+\rho^x} \quad \left(\rho = \frac{失点数}{得点数}\right)$$

2. 平均絶対偏差（MAD）

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

- 指数 x の最適な値 x^* を求めたい
- 方法：x の十分多くの異なる値（1 から 4 まで 0.002 刻みの x）に対して MAD(x) を計算し, MAD(x) を最小にする x を特定する

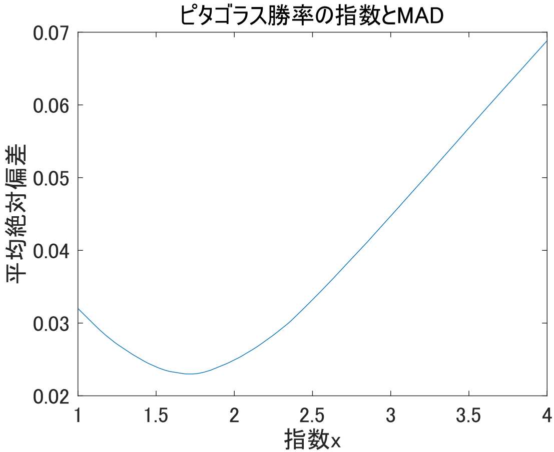


Figure : x に対する MAD(x) の関係

計算結果：$x^* = \mathbf{1.69}$, MAD$(x^*) = 0.0230$

勝率に影響する要素（イメージ）