不均一分散を仮定した Massey のレーティング手法と その統計的性質について

2020SS069 高橋李功

指導教員：塩濱敬之

1　はじめに

Massey のレーティング手法はスポーツなどのレーティングを得点差から推定する手法として知られている [1]. Massey のレーティングは, プレイヤー i と j の試合の得点差 y_k が, 両プレイヤーのレート差 $r_i - r_j$ と一致し, $y_k = r_i - r_j$ であることを仮定している.

本研究では, 通常の Massey のレーティングの誤差項に仮定される等分散性や無相関性を取り除き, 不均一分散を仮定した手法を考案し, その統計的性質を明らかにする. さらに, 不均一分散を仮定した手法を実際の統計データに適用し, その実用性を検証することを目的としている.

本研究は Massey のレーティングに関する先行研究 [2] で課題として挙げられていた問題についての研究である.

2　Massey のレーティング

m プレイヤーが n 試合行った場合, 次式を得る.

$$y = Xr + \varepsilon \tag{1}$$

X は $n \times m$ の行列で, 対戦の組み合わせを表しており, この式は対戦したプレイヤーのレート差 $r_i - r_j$ が得点差 y になるということを意味している. ε は誤差である.

また, レーティングベクトル r の推定値 $\hat{r}$ は次式で求められる.

$$\hat{r} = (X^TX)^{-1}X^Ty \tag{2}$$

ここで, $X^TX = M, X^Ty = p$ とすると, $\hat{r} = M^{-1}p$ となる. ベクトル p は得失点差を表しており, 行列 M の対角成分は試合数, 非対角成分はそれぞれのチームとの対戦回数を表している.

しかし, M は正則行列ではないため M^{-1} は一意に定まらない. そこで, M の任意の 1 行を第 i 行として, M の第 i 行成分をすべて 1, p の第 i 成分を 0 とする操作を行うことで, Massey のレーティングの合計値と平均値は 0 となる. 以上の手順によって得られる Massey のレーティングを $\hat{r} = \tilde{M}^{-1}\tilde{p}$ と表す.

3　一般化最小二乗法

等分散性, 無相関性の仮定を満たさない線形モデルに対して, 最良線形不偏推定量を得ることができる一般化最小二乗法 (Generalized least squares; GLS) について説明する [3]. (1) 式の r を通常の最小二乗法 (Ordinary least squares; OLS) で推定する場合, 誤差項 $\boldsymbol{\varepsilon}$ には無相関性, 等分散性が仮定され, $\boldsymbol{\varepsilon}$ の分散共分散行列 $\Omega = E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T]$ は分散 σ^2 を用いて $\Omega = \sigma^2 I_n$ と表される. ここで, I_n は $n \times n$ の単位行列である. 一方で, 無相関性, 等分散性を仮定しない一般化最小二乗法では次式のように表される.

$$\Omega = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{1n} & \sigma_{2n} & \cdots & \sigma_n^2 \end{bmatrix}$$

Ω は半正定値対称行列であるが以下では Ω が正定値対称行列である場合を仮定する. 正定値行列 Ω は $\Omega^{-1} = \Psi\Psi^T$ となる $n \times n$ 行列 Ψ が存在し, Ψ は逆行列 Ψ^{-1} を持つ. このような行列 Ψ^T を (1) 式の左からかけて次式を得る.

$$\Psi^T y = \Psi^T Xr + \Psi^T \boldsymbol{\varepsilon} \tag{3}$$

$\tilde{X} = \Psi^T X$, $\tilde{y} = \Psi^T y$, $\tilde{\boldsymbol{\varepsilon}} = \Psi^T \boldsymbol{\varepsilon}$ とおくと, $\tilde{y} = \tilde{X}r + \tilde{\boldsymbol{\varepsilon}}$ と表せる. この式の誤差項 $\tilde{\boldsymbol{\varepsilon}}$ の分散共分散行列は $E[\tilde{\boldsymbol{\varepsilon}}\tilde{\boldsymbol{\varepsilon}}^T] = I_n$ となる. よって (3) 式は無相関性, 等分散性を満たす. ($\boldsymbol{\varepsilon} \sim N(\mathbf{0},\ \Omega)$) を仮定し, (3) 式に最小二乗法を適用すれば一般化最小二乗法による推定値 $\hat{\boldsymbol{r}}_{GLS} = (X^T\Omega^{-1}X)^{-1}X^T\Omega^{-1}y$ を得る.

$X^T\Omega^{-1}X = M, X^T\Omega^{-1}y = p$ とすると, $\hat{\boldsymbol{r}}_{GLS} = M^{-1}p$ となる. M の任意の第 i 行を全て 1, p の第 i 成分を 0 とする操作をした Massey のレーティングを $\hat{\boldsymbol{r}}_{GLS} = \tilde{M}^{-1}\tilde{p}$ と表す.

X の第 i 列を全て 0 にする操作をした行列を $\tilde{X}$ とすると, $\hat{\boldsymbol{r}}_{GLS} = \tilde{M}^{-1}\tilde{X}^T\Omega^{-1}y = \tilde{M}^{-1}\tilde{X}^T\Omega^{-1}Xr + \tilde{M}^{-1}\tilde{X}^T\Omega^{-1}\boldsymbol{\varepsilon}$ と表せる. 真のレーティング $r_i (i = 1, \cdots, m)$ の合計が 0 の仮定の下で, 不均一分散を仮定した Massey のレーティング $\hat{\boldsymbol{r}}_{GLS}$ の標本分布は次のように与えられる.

$$(\hat{\boldsymbol{r}}_{GLS} - \boldsymbol{r}) \sim N(\mathbf{0}, \tilde{M}^{-1}\tilde{X}^T\Omega^{-1}\tilde{X}\tilde{M}^{-1T})$$

4　シミュレーション

実際のデータでは GLS を用いた推定に必要な分散共分散行列 Ω は不明である. そのため, データから推定した Ω を使って実行する. これを実行可能な一般化最小二乗法 (Feasible generalized least squares; FGLS) という. 本研究では Ω の推定にノンパラメトリック回帰を用いる. また簡単のために Ω は対角行列として扱う.

以下のシミュレーションで OLS を用いた通常の Massey と GLS を用いた不均一分散を仮定した Massey と FGLS を用いた実行可能な不均一分散を仮定した Massey を比較する.

4.1 シミュレーション 1

このシミュレーションではチーム数を 11 チーム ($m=11$) に設定し, 総当たり戦 55 試合 ($n=\binom{11}{2}=55$) と, すべてのチームがそれぞれ一回以上試合を行うようにランダムに 25 試合を選択した $n=25$ の場合をシミュレーションする. レーティングベクトルを $r=(r_1,r_2,r_3,\ldots,r_{11})^T$, スコアを $r=(5,4,3,2,1,0,-1,-2,-3,-4,-5)^T$ とした. 得点差の誤差項は正規分布に従い, 得点差の誤差分散はレート差が大きいと分散も大きい場合 $\sigma^2=3+\frac{(r_i-r_j)^2}{10}$ $(3\le\sigma^2\le 13)$ とした.

各シミュレーションは 10000 回繰り返し, OLS, GLS, FGLS を $\mathrm{RMSE}_j=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{r}_{ji}-r_j)^2}$ で比較する,

その結果を図 1 上に示した. どのプレイヤーに対しても RMSE が小さいことから, GLS, FGLS の Massey は誤差に不均一分散を仮定したデータに対して OLS の Massey よりも良い推定値を与える事がわかる.

次に, すべてのチームがそれぞれ一回以上試合を行うようにランダムに 25 試合を選択したときをシミュレーションする. その結果を図 1 下に示した.

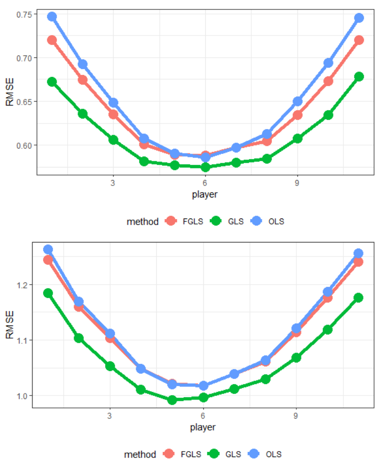

図 1 OLS, GLS, FGLS の RMSE(上:n=55)(下:n=25)

試合数が少ない場合でも GLS, FGLS のほうが良い推定値が得られた.

4.2 シミュレーション 2

より実際のデータに近い状況で推定を行うために, 2021-22 V. LEAGUE DIVISION1 MEN V レギュラーラウンドの結果 [4] を分析し, それを模倣した設定でシミュレーションを行った. 実際の V リーグを OLS の Massey のレーティング手法で分析した後, レーティングと得点差の誤差の 2 乗を 2 次式で回帰した. それを図 2 に示した. この情報を元に, レーティングを $r=(-1.78,-1.75,-1.02,-0.15,0.00,0.88,0.88,0.92,0.95,1.07)^T$, 誤差分散を $\sigma^2=4.02-0.38(r_i-r_j)^2$ と設定し, 10000 回のシミュレーションを行った. 真のレート差で 6 クラスに分け, クラスごとの $\mathrm{RMSE}_{class}=\sqrt{\frac{1}{n_{class}}\sum_{i,j\in class}^{\#class}((\hat{r}_i-\hat{r}_j)-(r_i-r_j))^2}$ を図 3 に示した. どのクラスに対しても RMSE が小さいことから,

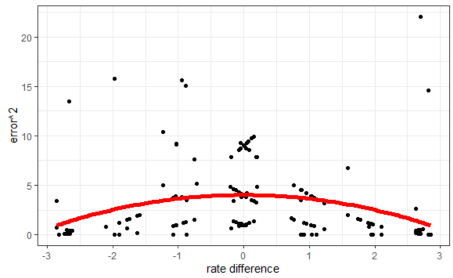

図 2 V リーグのレート差と誤差の関係

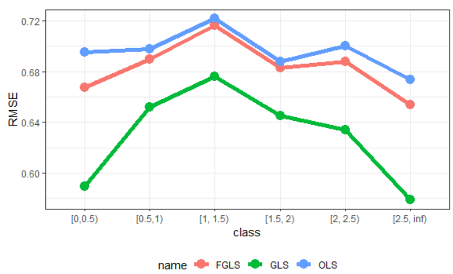

図 3 クラスごとのレート差の RMSE

GLS, FGLS の Massey は OLS の Massey よりも良い推定値を与える事がわかる.

また, サッカーの J リーグやバスケットボールの B リーグの分析も行ったが, 誤差をレーティング差から推定することはできなかった. 得点差に制限のないスポーツは誤差の分散が大きくなってしまい, 規則性を見出すのが難しいと考えられる.

5 おわりに

特定の条件下では FGLS を用いた不均一分散を仮定した Massey が OLS を用いた通常の Massey を上回る性能を発揮できることが分かった. 今後の研究では誤差項に従属性を仮定した (Ω の非対角成分も考慮した) Massey の性質を明らかにしていきたい.

参考文献

[1] Massey, K.: Statistical models applied to the rating of sports teams. Bluefield College, 1997.

[2] 黒木裕鷹・塩濱敬之：「Massey のレーティング指標の統計的性質とそのネットワーク分析への応用」, 日本行動計量学会, 行動計量学, 49 巻 (2022) 2 号, p.237-251

[3] 森 裕一・黒田 正博・足立 浩平：『最小二乗法・交互最小二乗法』, 共立出版株式会社, 東京, 2017

[4] V・レギュラーラウンド ラウンド一覧 バレーボール, https://www.vleague.jp/round/list/318, (アクセス日：2023 年 10 月 7 日)