Final Data Analysis Project

Tomoki Okuno

March 19, 2024

Predicting the Final Matchday of Japan Football League 2023

Introduction

Soccer (i.e., European football) is one of the most popular sports in the world in terms of population and market size. Soccer is highly strategic and has become more complex in recent years, with offensive and defensive styles and tactical trends changing with the times. This project focuses on the Japan Football League 1, abbreviated as J1, the top professional tier within Japan's soccer hierarchy. Comprising 18 clubs, J1 sees each club competing both at their home venue and at their opponent's (away avenue). Consequently, every club participates in 34 (= 17 clubs $\times$ 2 games) matches out of a total of 306 (= 18×17). Matches are typically held weekly, amounting to 9 games per week. We call the day 'matchday'.

This project aims to predict the results of the 9 matches scheduled for the final matchday of the 2023 J1 season without considering time effects. To achieve this, data from all preceding matchdays within the same season are utilized through a hierarchical Bayesian model. The study compares two distinct Bayesian models, calculating several performance metrics for evaluation. As in other professional sports, there are betting markets that aim to predict the outcome of matches. By accurately forecasting these outcomes, we may enhance our prospects of success in such betting endeavors.

Methods

Dataset

We designate the 18 clubs as T_1 through T_{18} based on their ranking in the 2023 season, with T_1 being the top-ranked and T_{18} the lowest. The data includes the number of goals scored by each team as follows:

$$y_{ij1} : \text{Goals scored by } T_i \text{ when playing at home against } T_j$$

$$y_{ij0} : \text{Goals scored by } T_i \text{ when playing away against } T_j$$

for $i \neq j = 1, \ldots, 18$. Note that in soccer, each goal corresponds to one point. The final matchday comprised the following 9 matches, with the first team listed playing at home: 1) T_5 vs T_{18}, 2) T_6 vs T_{17}, 3) T_7 vs T_3, 4) T_{10} vs T_9, 5) T_{12} vs T_4, 6) T_{13} vs T_2, 7) T_{14} vs T_8, 8) T_{15} vs T_{11}, and 9) T_{16} vs T_1. Consequently, we assume that 18 y_{ijk}'s, which are the target of our predictions, are missing. Examples include $y_{5,18,1}$ and $y_{18,5,0}$ in the match between T_5 (home) and T_{18} (away).

Models

Model 1. We explore two different models. The first model is Poisson-based, with y_{ijk} representing the outcome, as it is a positive integer with virtually no upper limit. The specific model with priors is as follows:

$$
\begin{aligned}
y_{ijk} \mid \lambda_{ijk} &\sim \text{Poisson}(\lambda_{ijk}), \quad i \neq j = 1, \ldots, 18, \\
\log(\lambda_{ijk}) &= \alpha \times k + \beta_i - \gamma_j, \quad \alpha \sim N(m, c), \\
\beta_i \mid \sigma_\beta^2 &\sim N(0, \sigma_\beta^2), \quad \sigma_\beta \sim \text{InvGamma}(a_1, b_1), \\
\gamma_j \mid \sigma_\gamma^2 &\sim N(0, \sigma_\gamma^2), \quad \sigma_\gamma \sim \text{InvGamma}(a_2, b_2),
\end{aligned}
$$

where the fixed effect α can be interpreted as the home advantage, which is expected to be positive. In addition to the physical advantage of playing games in their own stadium, with shorter travel time and on familiar turf on the field, there is also the psychological advantage of playing games with the support of a larger number of club fans. For the random effects, β_i represents the attacking strength of team i, while γ_j reflects the defensive strength of team j. In other words, this model assumes that a team with strong offensive capabilities is more likely to score against a team with weaker defensive skills.

Model 2. The second model defines the outcome as $d_{ij} = y_{ij1} - y_{ji0}$, representing the score difference for T_i (home team) versus T_j (away team). This outcome is expected to follow a normal distribution with a positive mean, reflecting the home advantage. Therefore, a normal Bayesian model is considered:

$$
\begin{aligned}
d_{ij} &\sim N(\mu_{ij}, \sigma^2), \quad i \neq j = 1, \ldots, 18, \\
\mu_{ij} &= \beta_i - \gamma_j, \\
\beta_i \mid \alpha, \sigma_\beta^2 &\sim N(\alpha, \sigma_\beta^2), \quad \alpha \sim N(m, c), \quad \sigma_\beta \sim \text{InvGamma}(a_1, b_1), \\
\gamma_j \mid \sigma_\gamma^2 &\sim N(0, \sigma_\gamma^2), \quad \sigma_\gamma \sim \text{InvGamma}(a_2, b_2),
\end{aligned}
$$

where α in Model 2 represents the home advantage. Although the same notations are used as in Model 1, these parameters are not directly comparable with those of Model 1. We intend to compare $\widehat{y}_{ij1} - \widehat{y}_{ij0}$

obtained from Model 1 and $\widehat{d}_{ij}$ from Model 2 against the true values in the final matchday.

Hyperparameters are determined based on my 25-year experience in soccer. Both models are fitted using Markov Chain Monte Carlo (MCMC) methods with 30,000 samples distributed across 3 chains. Each chain comprises 11,000 iterations, with a burn-in period of 1,000 iterations and a thinning rate of 1. If convergence is not deemed satisfactory, diagnostics such as autocorrelation and time-series plots will be examined, and additional iterations will be pursued accordingly.

Performance Metrics

We assess performance using three metrics: mean absolute error (MAE), mean squared error (MSE), and concordance (Conc). Let d_{ij} represent the true score difference between the home team i and the away team j. Furthermore, let $\widehat{d}_{ij}^{(1)} = \widehat{y}_{ij1} - \widehat{y}_{ij0}$ denote the predictions from Model 1, and $\widehat{d}_{ij}^{(2)}$ denote those predicted from Model 2. The absolute error (AE) and the square error (SE) are defined as $|\widehat{d}_{ij}^{(m)} - d_{ij}|$ and $(\widehat{d}_{ij}^{(m)} - d_{ij})^2$, respectively. Therefore, MAE and MSE are calculated as follows:

$$\text{MAE} = \frac{1}{9} \sum_{(i,j) \in D_{34}} \left| \widehat{d}_{ij}^{(m)} - d_{ij} \right|, \quad \text{MSE} = \frac{1}{9} \sum_{(i,j) \in D_{34}} \left(\widehat{d}_{ij}^{(m)} - d_{ij} \right)^2,$$

for $m = 1, 2$ (Model number), where D_{34} is the set of 9 games on the last matchday (34th day). Regarding concordance, we categorize outcomes into three levels: home win, draw, and home lose (away win). We classify the predicted score difference as a home win if $\widehat{d}^{(m)}ij \geq 0.5$, a home lose if $\widehat{d}^{(m)}ij \leq -0.5$, and a draw otherwise (i.e., $|\widehat{d}_{ij}^{(m)}| < 0.5$). Then, concordance is the proportion for matched results out of nine. For sensitivity analysis, we explore how these metrics change with different hyperparameters.

Results

Data Description

The distribution of the number of goals, separated by home and away games, is shown in **Figure 1**, utilized in Model 1. Additionally, **Figure 2** shows the difference in goals between home and away for Model 2. Both figures indicate that teams typically score more points at home stadiums, suggesting the home advantage.

Prior Information

Based on my soccer knowledge, the home advantage is estimated to be worth one point on average, with a range from 0.5 to two points having a 95% chance. This determination guides the choice of parameters (m, c) in both models. The other parameters (a_1, b_1) and (a_2, b_2) for offensive and defensive strength in both

models were established by considering scenarios where professional teams are likely to have outcomes such as 8-0 or 0-8. Refer to **Table 1** for detailed specifications.

Primary and Sensitivity Results

As shown in **Figure 3**, both models demonstrated satisfactory convergence when employing the MCMC algorithm with 30,000 samples. However, when a smaller number of samples, such as 3,000, were utilized, the fixed effect α did not converge well, especially in Model 1.

Posterior summaries for the fixed effect α and the random effects β_i and γ_j are provided in **Table 2**. As expected, the home advantage exhibited a significantly positive effect on both the number of goals (Model 1) and goal differences (Model 2). Interestingly, the teams with the highest final rankings did not necessarily excel in both offense and defense; some teams demonstrated strengths in either attack or defense, while others are more balanced. The characteristics of each team showed slight variations between the two models.

Table 3 displays the actual outcomes of the 9 games on the final matchday and the predictions, based on posterior means, for Models 1 and 2. Model 2 yielded a lower MSE (1.44) compared to Model 1 (1.53), while maintaining the same MAE (1.00) and achieving a concordance of 2 games matched out of 9 games (22%). Sensitivity analysis, where hyperparameters are modified independently and simultaneously, revealed the robustness of both models in terms of performance metrics, except for the case where m (prior mean for the home advantage α) approached zero, resulting in a concordance of $5/9 = 56\%$.

Discussion

In this study, we attempted two distinct models to analyze soccer outcomes in the Japanese professional league, incorporating the home advantage and each team's offensive and defensive capabilities as random effects. However, both models exhibited less than optimal fit. As mentioned in the introduction, soccer dynamics are evolving, necessitating adjustments for various factors such as weather conditions, seasonal effects, mutual tactical compatibility, individual player performances, and club budgets.

The findings underscore the complexity of accurately predicting soccer outcomes. Future research should refine models by considering additional factors to improve predictive accuracy and gain deeper insights into the sport's dynamics. Moreover, exploring alternative modeling techniques or more sophisticated statistical methods may yield further advancements in understanding soccer outcomes.

Reference

MEIJI YASUDA J1 LEAGUE, 2023 Stats, https://www.jleague.co/stats/dashboard/j1/2023/

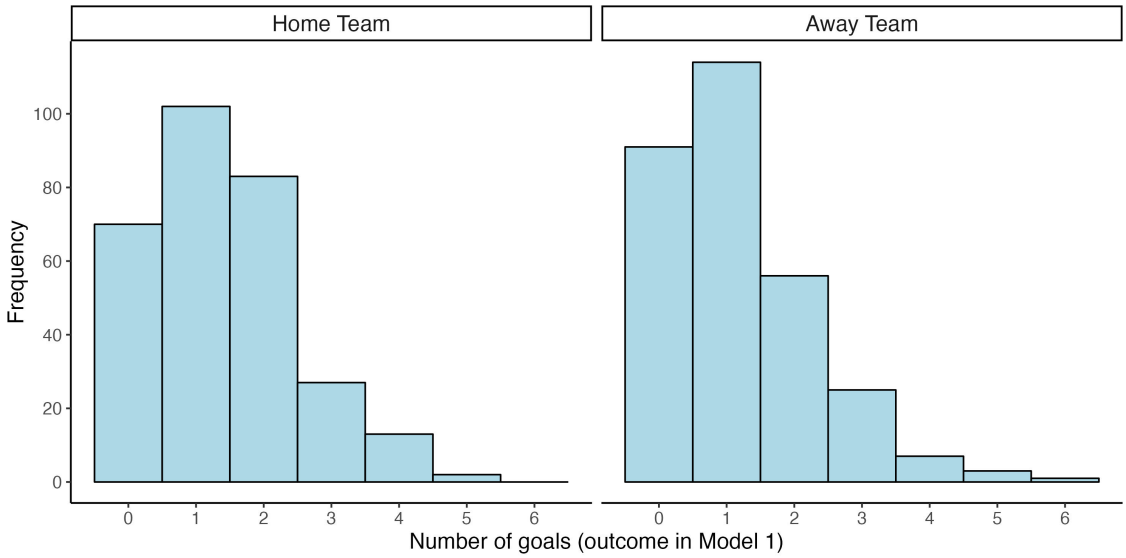

Figure 1. Observed number of goals for each match ($n = 297$).

The left and right panels correspond to y_{ij1} and y_{ij0}, respectively. The goals in the 9 matches held on the final matchday are not included.

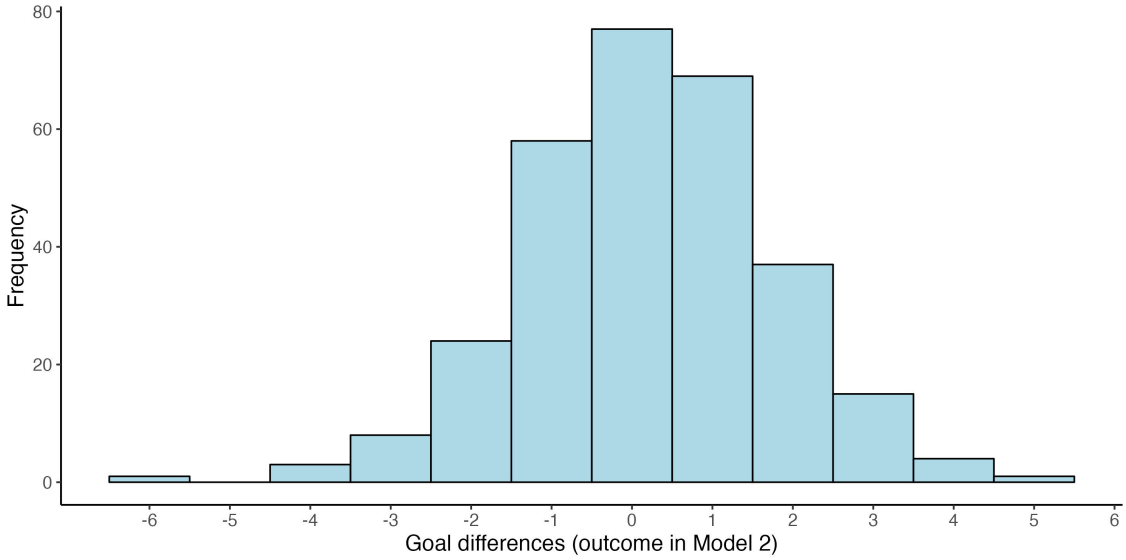

Figure 2. Observed goal differences between home and away teams for each match ($n = 297$).

This corresponds to $d_{ij} = y_{ij1} - y_{ji0}$. The differences in the 9 games on the final matchday are not included.

Table 1. Prior information based on my knowledge in soccer.

Parameter		Model 1 (Poisson)		Model 2 (Normal)	
		Value	Process	Value	Process
Home advantage† α	m	0	$\ln 1$	1	-
	c	0.12	$[(\ln 2 - \ln 0.5)/4]^2$	0.14	$[(2-0.5)/4]^2$
Offensive ability* σ_β	a_1	2.84	$\left(\frac{\ln 8 - \ln 0.01}{4}\right)^2 \frac{1}{a_1 - 2} = 2\left(\frac{\ln 8 - \ln 0.01}{4}\right)$	3	$\left(\frac{8-0}{4}\right)^2 \frac{1}{a_1 - 2} = 2\left(\frac{8-0}{4}\right)$
	b_1	5.13	$\frac{b_1}{a_1 - 1} = \frac{\ln 8 - \ln 0.01}{4}$	4	$\frac{b_1}{a_1 - 1} = \frac{8-0}{4}$
Defensive ability* σ_γ	a_2	2.84	$a_2 = a_1$	3	$a_2 = a_1$
	b_2	5.13	$b_2 = b_1$	4	$b_2 = b_1$

† The home advantage is estimated to have a mean value of 1 and a range from 0.5 to 2 with a 95% probability. The range method is utilized to compute c.

* In matches between professional teams, occurrences of results like 8 vs. 0 or 0 vs. 8 are observed at times, resulting in a mean value of $(\ln 8 - \ln 0.01)/4 = 1.67$ and $(8-0)/4 = 2$ for σ_β and σ_γ using the range method in Models 1 and 2, respectively. Their standard deviations are assumed to be twice their means. In Model 1, a score of 0 is replaced by 0.01 to avoid taking the natural logarithm of 0.

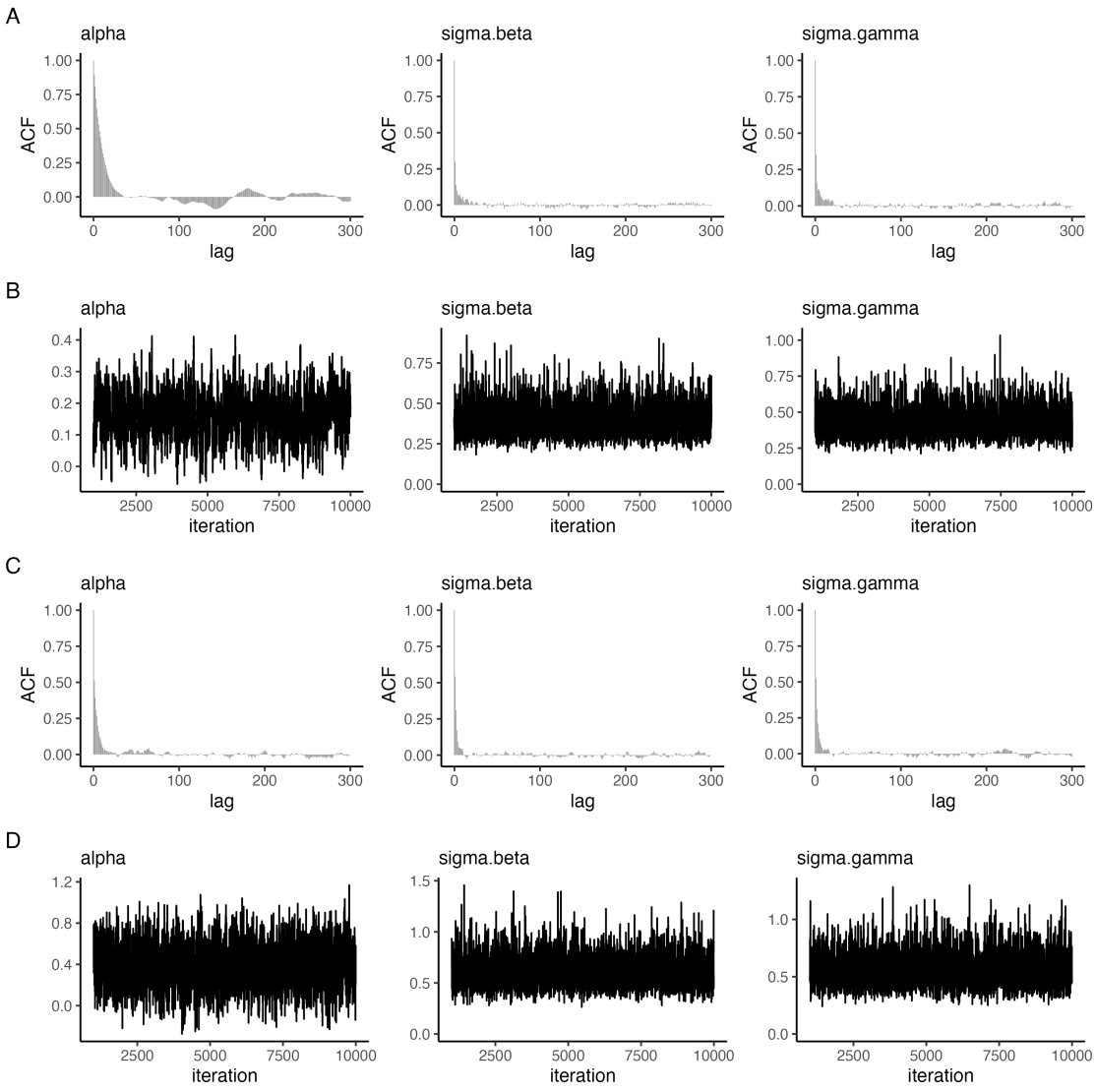

Figure 3. Convergence diagnosis in Models 1 and 2.

Panel A: Autocorrelation plots for α, σ_β, and σ_γ in Model 1. Panel B: Corresponding time-series plots. Panel C: Autocorrelation plots for α, σ_β, and σ_γ in Model 2. Panel D: Corresponding time-series plots.

Table 2. Posterior Summary for fixed and random effects in Models 1 and 2.

	Model 1 (Poisson)					Model 2 (Normal)				
	Mean	SD	2.5%	97.5%	P(>0|Y)	Mean	SD	2.5%	97.5%	P(>0|Y)
α	0.16	0.07	0.03	0.29	0.99	0.39	0.19	0.03	0.78	0.98
β_1	0.35	0.15	0.06	0.63	0.99	0.85	0.33	0.22	1.51	1.00
β_2	0.40	0.14	0.12	0.67	1.00	1.04	0.34	0.39	1.72	1.00
β_3	0.01	0.15	-0.29	0.31	0.54	0.71	0.33	0.08	1.35	0.99
β_4	0.00	0.16	-0.32	0.31	0.51	0.57	0.32	-0.06	1.22	0.96
β_5	0.03	0.15	-0.28	0.33	0.58	0.49	0.33	-0.14	1.15	0.94
β_6	0.01	0.16	-0.31	0.32	0.52	0.62	0.33	-0.02	1.28	0.97
β_7	-0.06	0.16	-0.38	0.24	0.35	0.06	0.33	-0.58	0.71	0.58
β_8	0.21	0.15	-0.08	0.51	0.92	0.50	0.32	-0.13	1.14	0.94
β_9	-0.02	0.16	-0.33	0.28	0.45	0.40	0.32	-0.22	1.04	0.90
β_{10}	-0.11	0.16	-0.44	0.20	0.26	0.06	0.33	-0.59	0.71	0.58
β_{11}	0.04	0.16	-0.28	0.34	0.60	0.47	0.32	-0.17	1.11	0.93
β_{12}	0.32	0.14	0.03	0.60	0.98	0.19	0.33	-0.46	0.84	0.72
β_{13}	-0.06	0.16	-0.39	0.26	0.36	0.06	0.33	-0.58	0.70	0.58
β_{14}	0.08	0.16	-0.23	0.38	0.70	-0.17	0.34	-0.84	0.47	0.31
β_{15}	0.03	0.16	-0.28	0.33	0.58	-0.36	0.34	-1.04	0.30	0.14
β_{16}	-0.02	0.16	-0.35	0.30	0.46	0.21	0.33	-0.44	0.86	0.74
β_{17}	-0.17	0.17	-0.51	0.15	0.15	-0.14	0.33	-0.79	0.50	0.34
β_{18}	-0.21	0.18	-0.57	0.12	0.11	-0.16	0.33	-0.80	0.48	0.31
γ_1	0.23	0.17	-0.10	0.58	0.91	0.75	0.33	0.12	1.42	0.99
γ_2	0.02	0.16	-0.29	0.34	0.55	0.36	0.32	-0.25	1.00	0.87
γ_3	0.28	0.18	-0.06	0.65	0.94	0.20	0.32	-0.41	0.83	0.74
γ_4	0.29	0.18	-0.05	0.66	0.95	0.33	0.32	-0.28	0.97	0.85
γ_5	0.15	0.17	-0.17	0.49	0.81	0.25	0.31	-0.36	0.87	0.79
γ_6	0.10	0.17	-0.22	0.44	0.72	-0.01	0.31	-0.61	0.61	0.49
γ_7	-0.06	0.16	-0.36	0.26	0.35	0.10	0.31	-0.51	0.72	0.63
γ_8	-0.14	0.16	-0.44	0.17	0.18	0.09	0.32	-0.52	0.72	0.61
γ_9	0.14	0.17	-0.18	0.48	0.79	0.22	0.32	-0.39	0.86	0.75
γ_{10}	-0.01	0.16	-0.32	0.31	0.47	0.10	0.31	-0.50	0.72	0.62
γ_{11}	-0.15	0.16	-0.45	0.17	0.18	-0.29	0.31	-0.91	0.32	0.18
γ_{12}	-0.39	0.15	-0.67	-0.10	0.00	0.05	0.31	-0.55	0.66	0.57
γ_{13}	-0.11	0.15	-0.41	0.19	0.24	0.02	0.31	-0.59	0.62	0.53
γ_{14}	-0.14	0.15	-0.44	0.16	0.17	0.40	0.32	-0.20	1.04	0.90
γ_{15}	-0.31	0.15	-0.59	-0.02	0.02	0.12	0.31	-0.48	0.73	0.64
γ_{16}	-0.39	0.14	-0.67	-0.11	0.00	-0.72	0.32	-1.36	-0.12	0.01
γ_{17}	-0.14	0.15	-0.44	0.16	0.17	-0.09	0.31	-0.70	0.54	0.39
γ_{18}	-0.32	0.15	-0.60	-0.03	0.02	-0.55	0.32	-1.19	0.06	0.04
σ_β	0.39	0.09	0.26	0.59		0.59	0.14	0.37	0.92	
σ_γ	0.42	0.09	0.28	0.63		0.55	0.13	0.34	0.85	

Note: The interpretation of each parameter differs between the two models.

Table 3. Absolute and square errors, concordance and their means for each last match in Models 1 and 2.

	True Results						Model 1 (Poisson)						Model 2 (Normal)			
#	Home (i)	Away (j)	y_{ij1}	y_{ij0}	d_{ij}	WDL	$\widehat{y}_{ij1}$	$\widehat{y}_{ji0}$	$\widehat{d}_{ij}^{(1)}$	AE_1	SE_1	$Conc_1$†	$\widehat{d}_{ij}^{(2)}$	AE_2	SE_2	$Conc_2$†
1	12	4	0	2	-2	Lose	1.24	1.52	-0.28	1.72	2.96	Draw (×)	-0.12	1.88	3.53	Draw (×)
2	5	18	2	1	1	Win	1.69	0.71	0.98	0.02	0.00	Win (✓)	1.04	0.04	0.00	Win (✓)
3	15	11	0	1	-1	Lose	1.44	1.43	0.01	1.01	1.02	Draw (×)	-0.07	0.93	0.86	Draw (×)
4	10	9	1	0	1	Win	0.95	1.01	-0.06	1.06	1.13	Draw (×)	-0.15	1.15	1.32	Draw (×)
5	13	2	3	1	2	Win	1.11	1.69	-0.58	2.58	6.66	Lose (×)	-0.30	2.30	5.29	Draw (×)
6	16	1	0	1	-1	Lose	0.94	2.12	-1.19	0.19	0.03	Lose (✓)	-0.54	0.46	0.22	Lose (✓)
7	7	3	0	1	-1	Lose	0.85	1.10	-0.25	0.75	0.56	Draw (×)	-0.14	0.86	0.74	Draw (×)
8	14	8	0	1	-1	Lose	1.49	1.46	0.03	1.03	1.07	Draw (×)	-0.28	0.72	0.52	Draw (×)
9	6	17	1	1	0	Draw	1.39	0.78	0.61	0.61	0.37	Win (×)	0.69	0.69	0.48	Win (×)
									Mean	1.00	1.53	22.2%		1.00	1.44	22.2%

† ✓, Matched with the true result (WDL); ×, Unmatched.

AE, absolute error; SE, square error; Conc, concordance.

Table 4. Sensitivity analysis results for different priors.

	Hyperparameter						Performance metric		
	m	c	a_1	b_1	a_2	b_2	MAE	MSE	Conc
Model 1	$0(m_0)$	$0.12(c_0)$	$2.84(a_{10})$	$5.13(b_{10})$	$2.84(a_{20})$	$5.13(b_{20})$	1.00	1.53	22%
(Poisson)	$\ln 2$	c_0	a_{10}	b_{10}	a_{20}	b_{20}	1.01	1.54	22%
	m_0	$10c_0$	a_{10}	b_{10}	a_{20}	b_{20}	1.00	1.53	22%
	$\ln 2$	$10c_0$	$10a_{10}$	$10b_{10}$	$10a_{20}$	$10b_{20}$	1.04	1.60	22%
	m_0	c_0	$0.001a_{10}$	$0.001b_{10}$	$0.001a_{20}$	$0.001b_{20}$	1.04	1.42	22%
	m_0	c_0	$100a_{10}$	$100b_{10}$	a_{20}	b_{20}	1.00	1.50	22%
	$\ln 0.001$	c_0	a_{10}	b_{10}	a_{20}	b_{20}	0.97	1.59	56%
Model 2	$1(m_0)$	$0.14(c_0)$	$3(a_{10})$	$4(b_{10})$	$3(a_{20})$	$4(b_{20})$	1.00	1.44	22%
(Normal)	$2m_0$	c_0	a_{10}	b_{10}	a_{20}	b_{20}	1.02	1.45	22%
	m_0	$10c_0$	a_{10}	b_{10}	a_{20}	b_{20}	0.99	1.41	22%
	$2m_0$	$10c_0$	$10a_{10}$	$10b_{10}$	$10a_{20}$	$10b_{20}$	0.99	1.47	22%
	m_0	c_0	$0.001a_{10}$	$0.001b_{10}$	$0.001a_{20}$	$0.001b_{20}$	1.07	1.46	22%
	m_0	c_0	$100a_{10}$	$100b_{10}$	a_{20}	b_{20}	0.98	1.43	22%

MAE, mean absolute error; MSE, mean square error; Conc, concordance.

JAGS Code

```
# Model 1.
model {
  for (i in 1:18) {
    for (j in 1:18) {
      for (k in 1:2) {
        y[i, j, k] ~ dpois(lambda[i,j,k])
        log(lambda[i, j, k]) <- alpha * (2 - k) + beta[i] - gamma[j]
      }
    }
    beta[i] ~ dnorm(0, tau.beta^2)
  }

  alpha ~ dnorm(m, prec)
  c <- 1 / prec

  for (j in 1:18) {
    gamma[j] ~ dnorm(0, tau.gamma^2)
  }

  tau.beta ~ dgamma(a1, b1)
  sigma.beta <- 1 / tau.beta

  tau.gamma ~ dgamma(a2, b2)
  sigma.gamma <- 1 / tau.gamma
}

# Model 2.
model {
  for (i in 1:18) {
    for (j in 1:18) {
        d[i, j] ~ dnorm(mu[i, j], tau.e)
        mu[i, j] <- beta[i] - gamma[j]
    }
    beta[i] ~ dnorm(alpha, tau.beta^2)
  }

  alpha ~ dnorm(m, prec)
  c <- 1 / prec

  tau.e ~ dgamma(ea, eb)
  sigma <- 1 / sqrt(tau.e)

  for (j in 1:18) {
    gamma[j] ~ dnorm(0, tau.gamma^2)
  }

  tau.beta ~ dgamma(a1, b1)
  sigma.beta <- 1 / tau.beta

  tau.gamma ~ dgamma(a2, b2)
  sigma.gamma <- 1 / tau.gamma
}
```