Scuola universitaria professionale
della Svizzera italiana

SUPSI

University of Applied Sciences and Arts of Southern Switzerland
Department of Innovative Technologies

Bayesian Data Analysis

HOCKEY GAMES ANALYSIS

Andrea Wey
andrea.wey@student.supsi.ch

Christian Berchtold
christian.berchtold@student.supsi.ch

hp test gamma-poisson ok

chiedere come hanno deciso la rope e cosa rappresenta

descrtittiva e conclusioni del test ok

post predictive check nel notebook, mancano nel report (mostrano un lack of fit in realtà)

prior sensitivity: ok averla fatta, di fatto il modello è robusto (chiedere se se ne sono accorti)

implementazione di weekend vs weekday un po' convoluta: ogni giorno è stimato a sè stante, e poi fanno la media dei giorni wend e wday. Sarebbe stato meglio lavorare subito con variabile binaria.

Post predictive check: nel secondo box (destra) il modello non fitta bene, sembra che però non se ne accorgano.

Direi 5.5.

Professor: Giorgio Corani
SUPSI, Lugano Switzerland

Assistant: Marco Forgione
SUPSI, Lugano Switzerland

22/08/2023

Table of Contents

List of Figures

List of Tables

1. Data

The dataset we decided on using is about hockey games of the Swiss National League. Since every year there is a new season, we decided to take multiple seasons and concatenate them together from 2015-2016 to 2021-2022 to have more data at our disposal. The dataset is composed of 17 columns and 2450 rows. We began modifying the data by calculating the days of rest for each team before each game, as we used this information for modelling. The data was split on the `Resultat` column to get the end result for the game, and separate home and away goals. This was the final data frame (`hockey_df.head(5)` shown only):

Home	Away	Home_goals	Home_rest_days	Tag	Away_goals	Away_rest_days
EHC Biel-Bienne	EV Zug	1.0	0.0	Sa	0.0	0.0
EHC Biel-Bienne	Fribourg-Gottéron	0.0	0.0	Fr	3.0	2.0
EHC Biel-Bienne	HC Ambri-Piotta	3.0	0.0	Di	0.0	0.0
EHC Biel-Bienne	HC Ambri-Piotta	0.0	0.0	Fr	2.0	5.0
EHC Biel-Bienne	Genève-Servette HC	0.0	0.0	Sa	4.0	0.0

Table 1: Transformed data

Furthermore, when dealing with the modelling, it's necessary to provide some *observed* data. To facilitate this, we transformed the data frame in Table 1 into a flattened version (Table 2), consolidating it into a single column for all the teams. We also added a `team_code` column, which will be used in both the unpooled and hierarchical models.

Team	Goals	days_of_rest	day_of_the_week	team_code
EHC Biel-Bienne	1.0	0.0	Sa	0
EHC Biel-Bienne	0.0	0.0	Fr	0
EHC Biel-Bienne	3.0	0.0	Di	0
EHC Biel-Bienne	0.0	0.0	Fr	0
EHC Biel-Bienne	0.0	0.0	Sa	0

Table 2: Flattened data

As shown in Figure 1, the histogram illustrates the distribution of goals scored during home and away games. It is apparent from the figure that the mean of the distribution for home-field goals is higher than that of away goals. This visualization serves a descriptive purpose and does not directly influence our hypothesis testing or prior selection; rather, it provides an overview of how the data is distributed, setting the stage for subsequent analyses.

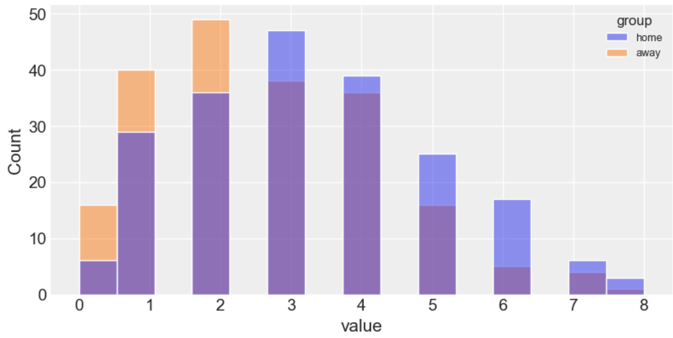


Figure 1: Distribution of goals

2. Hypothesis testing

2.1 1st Hypothesis

In sports, the concept of *home-field advantage* is often debated: do teams really perform better and score more when playing at their home stadium compared to playing away? We aim to investigate this question. To check for this supposed difference, we used the methods seen in class [1] to accept or reject our null hypothesis.

We define our null and alternative hypotheses as follows:

- **H0 (Null Hypothesis)**: Teams score the same number of goals on average, whether at home or away.
- **H1 (Alternative Hypothesis)**: Teams score more goals on average when playing at home.

ok, conjugate Gamma Poisson model

The goals scored cannot be negative, therefore it is appropriate, in our opinion, to use a Gamma distribution for setting the prior. Given that we're working with data that only takes integer values, we logically use a Poisson distribution for our likelihood. Online data indicates [2] that the mean goals scored in a hockey game are 5.51 for both teams combined. Dividing this by two results in an approximate mean for each team. Since we assume that the goals scored by each team follow a Poisson distribution, the variance is equal to the mean. Hence, the standard deviation (SD) is computed as:

$$SD = \sqrt{\text{mean}} = \sqrt{\frac{5.51}{2}} \approx 1.66 \tag{1}$$

From this, we can derive the $alpha$ and $beta$ parameters to set our priors.
Using the relations:

$$\frac{\alpha}{\beta} = \mu, \quad \frac{\alpha}{\beta^2} = \sigma^2$$

We deduce:

$$\alpha = \frac{\mu^2}{\sigma^2}, \quad \beta = \frac{\mu}{\sigma^2}$$

Our derived values for $alpha$ and $beta$ are $\alpha = 4.03$ and $\beta = 1.47$.
Finally, our probabilistic model will have this structure:

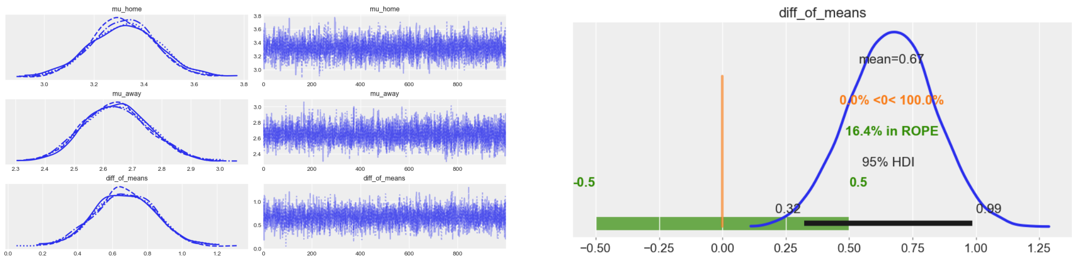
$$\begin{aligned} \mu_{\text{home}} &\sim \text{Gamma}(4.03, 1.47), & \vec{y}_{\text{home}} &\sim \text{Poisson}(\mu_{\text{home}}) \\ \mu_{\text{away}} &\sim \text{Gamma}(4.03, 1.47), & \vec{y}_{\text{away}} &\sim \text{Poisson}(\mu_{\text{away}}) \end{aligned}$$

2.1.1 Outcomes

Looking at the PyMC3 model outputs, we see that it converges well, but most importantly that evidence points towards the Alternative hypothesis: with 83.6% of the posterior distribution outside the ROPE, there's reasonably strong evidence suggesting there exists home-field advantage.

However, the 16.4% in the ROPE means there's still some uncertainty.

ok. Question: how did you set the rope?



(a) Convergence check (b) Posterior distribution

Figure 2: Hypothesis 1 model

2.1.2 Prior sensitivity

The choice of prior can significantly influence the posterior distribution, especially when the data is not highly informative. To check the robustness of our results, we conduct a prior sensitivity analysis. For this, we doubled the standard deviation of our prior distribution and examined how this change impacts our conclusions.

sensato raddoppiare la std.
Ma come si spiega che parti con una prior più incerta e finisci con conclusioni più forti??
anyway there is 0 probability of the home factor having a negative effect

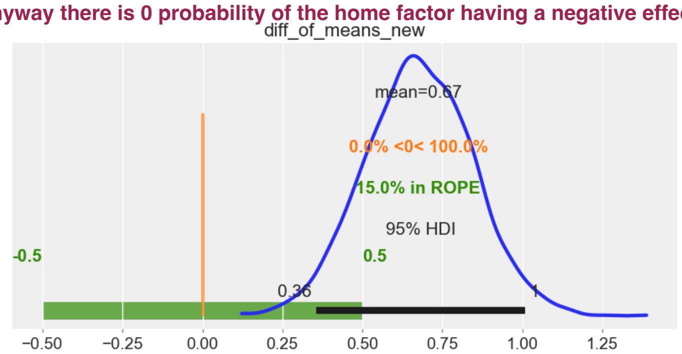


Figure 3: Prior sensitivity posterior distribution

The change in the prior has shown some impact on the ROPE test. The results now point even more strongly towards the Alternative hypothesis. In this new model, only 15% of the distribution falls within the ROPE — compared to the 16.4% of the previous model — as shown in Figure 5. This suggests that our conclusions are somewhat sensitive to the choice of prior, indicating the importance of carefully selecting the latter.

2.1.3 Conclusion

According to our findings, the 95% HDI for the mean difference between home and away games does not overlap largely with the ROPE, and the effect size is considerable. Therefore, we reject our null hypothesis that stated the home-field advantage has no significant impact on the number of goals scored. Our data suggest that playing on home soil is both a statistically and practically significant factor in the number of goals scored.

2.2 2nd Hypothesis

Another factor that can be analyzed is how the day of the week on which a match takes place affects a team's performance.

Our hypothesis wants to check if teams perform better or worse depending on the day of the week. To investigate our hypothesis, we use a similar approach as before, examining the differences in means and utilizing posterior plots to evaluate the hypothesis.

The null and alternative hypotheses are:

- **H0 (Null Hypothesis)**: The day of the week does not have any significant impact on the average goals scored by teams.
- **H1 (Alternative Hypothesis)**: There is a variation in the average goals scored by teams depending on the day of the week.

We proceed with the statistical modelling approach:
Given that the data of goals scored cannot be negative, it is appropriate to use a Gamma distribution for setting the prior, as was the case of the first hypothesis.
We follow the same prior setting procedure of Hypothesis 1. Hence, the standard deviation (SD) is computed as:

$$SD = \sqrt{\text{mean}} = \sqrt{\frac{5.51}{2}} \approx 1.66 \tag{2}$$

From this, we can derive the $alpha$ and $beta$ parameters for our distribution.
Finally, our probabilistic model will have this structure:

$$\begin{aligned} \mu_{\text{Monday}} &\sim \text{Gamma}(4.03, 1.47), & \vec{y}_{\text{Monday}} &\sim \text{Poisson}(\mu_{\text{Monday}}) \\ &\vdots & &\vdots \\ \mu_{\text{Sunday}} &\sim \text{Gamma}(4.03, 1.47), & \vec{y}_{\text{Sunday}} &\sim \text{Poisson}(\mu_{\text{Sunday}}) \end{aligned}$$

We then calculate the difference between the weekday mean and weekend (we consider Friday to be part of the weekend).

2.2.1 Outcomes

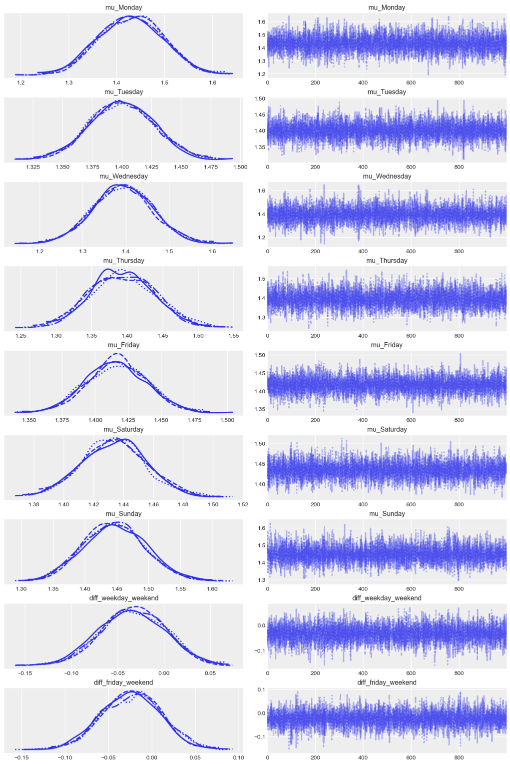
The PyMC3 model converges well, as we see from Figure 4a.

Looking at Figure 4b, we see that the average difference, labelled as "diff_weekday_weekend", is -0.029. This negative value hints that weekends might have slightly higher scores than weekdays. However, this effect is really small. Importantly, the 95% HDI for this difference is between -0.094 and 0.039, which includes zero. This means that statistically speaking, we can't say the difference is important.

filosoficamente errato

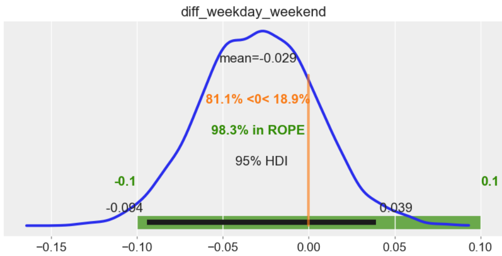
For the ROPE test, we picked a narrow range; the reason being we think that scores should not really differ between weekdays and weekends. Most of the 95% HDI falls inside this ROPE range, prompting us to think that the difference in scores is not significant in real-world terms.

conclusion is a bit stronger: you can actually claim that the two are practically equivalent



(a) Convergence check

come mai gestite ogni giorno separamente e poi
nel test valutate wend vs wday?
allora meglio fare wend vs wday già nel modello



(b) Posterior distribution

Figure 4: Hypothesis 2 model

2.2.2 Prior sensitivity

In our sensitivity analysis, similarly to the original model, we observe that the 95% HDI mostly falls within the ROPE interval. This strengthens the argument that the day of the week doesn't significantly affect the number of goals scored. The distribution is slightly shifted to the left, hinting that weekends (from Friday to Sunday) have a higher goal count, but this is neither statistically nor practically significant given the narrow ROPE interval.

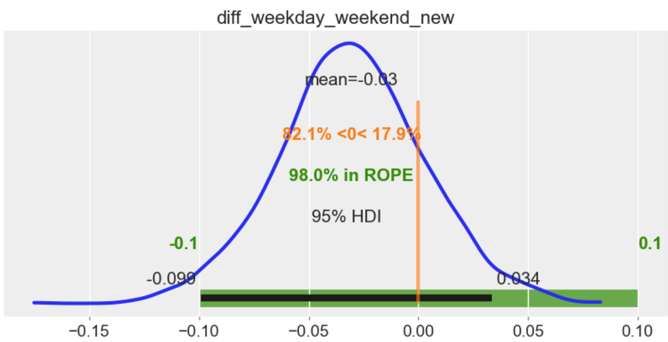


Figure 5: Prior sensitivity posterior distribution

2.2.3 Conclusion

Based on the results of our analysis, we can make a decision regarding our initial hypothesis. The 95% HDI largely overlaps with the ROPE, and the observed effect size is minimal. Therefore, we fail to reject the null hypothesis that states "the day of the week has no significant effect on the number of goals scored." In other words, our analysis suggests that the day of the week is neither a statistically nor practically significant factor in the number of goals scored in the matches under study.

3. Modelling

regression!

After the two hypothesis tests, we focused on carrying out the modelling part. Here we introduced a co-variate that represents the days of rest each team had before the game. We want to know if the days of rest could affect the expected goals; we linearly combine this variable together with the Gamma prior.

chiedere dove hanno preso ispirazione per Poisson
exp_goals ~ Poisson (lambda = mu_teams +....)

$$expected\_goals = \mu_{\text{teams}} + \mu_{\text{rest}} * true\_days\_rest$$

We use the model to explore the relationship between the number of "rest days" a team has and the resulting number of goals they score. Specifically, the model consists of:

- We assume that the average number of goals, represented by μ, follows a Gamma distribution. In simpler terms, we use the Gamma distribution to describe how the average number of goals can vary.

$$\mu \sim \text{Gamma}(4.03, 1.47)$$

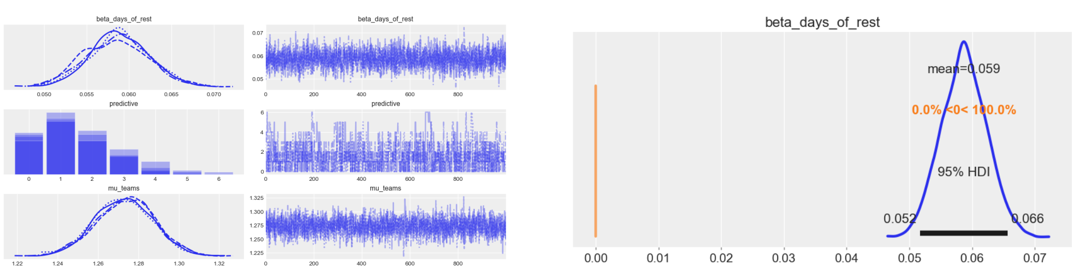
  The numbers 4.03 and 1.47 are parameters that we derived from online research [2] and then calculated, as shown in section 2..

- The observed number of goals, denoted as $\vec{y}$, is assumed to follow a Poisson distribution, which is commonly used for count-based data. The average rate, λ, for this distribution, is influenced by the "expected_goals" variable, which could be affected by various factors like team performance or morale. Formally, this is given as:

$$\vec{y} \sim \text{Poisson}(\lambda = \text{expected\_goals})$$

3.1 Results and discussion

The pooled, unpooled, and hierarchical models all converged effectively, underlining the robustness of the results. The parameter $\beta_{\text{days\_of\_rest}}$ came out as statistically significant, with a 95% HDI that does not include zero. This suggests that the variable "days of rest" does influence the number of goals scored, at least within the framework of these models.

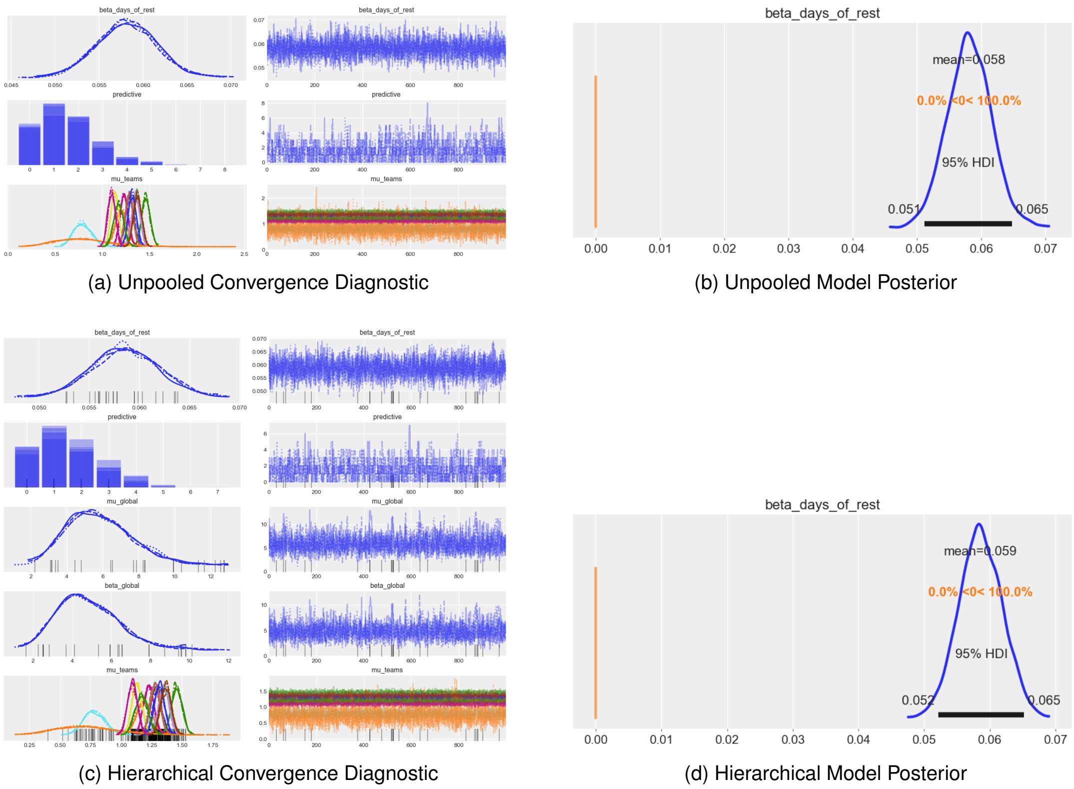


(a) Convergence Diagnostic

(b) Pooled Model Posterior

Figure 6: Diagnostic and Posterior Plots for the Pooled Model

In Figure 6 we see that the mean value for $\beta_{\text{days\_of\_rest}}$ is 0.059. This suggests that each additional day of rest is associated with an increase in the expected number of goals by approximately 0.059. While this effect is statistically significant, its practical relevance appears to be limited.



(a) Unpooled Convergence Diagnostic

(b) Unpooled Model Posterior

(c) Hierarchical Convergence Diagnostic

(d) Hierarchical Model Posterior

Figure 7: Diagnostic and Posterior Plots for Unpooled and Hierarchical Models

Posterior plot checks in Figure 7 confirm the results found with the pooled model. In these models, the "days of rest" factor also appears to influence goal-scoring, although the practical significance remains negligible due to the minor scale of the impact.

3.2 Novel Prediction

Our focus then turned to using the hierarchical model to predict the performance of a novel team. This choice was motivated by the hierarchical model's ability to incorporate both global and team-specific effects, making it well-suited for generalization to new teams.

We conducted two versions of this prediction:

basic: modello addestrato con la covariata, ma che usa solo l'intercetta per fare predizioni?/

1. **Basic Version**: Here, we used only the global μ from the hierarchical model's posterior to forecast the novel team's performance. While this approach provides a point estimate, it may not capture the effects of variables like "rest days".

2. **Advanced Version**: In this method, we integrated the $\beta_{\text{days\_of\_rest}}$ parameter into our prediction to account for the effect of rest days on the expected number of goals. This is a more comprehensive prediction that considers the specific characteristics of the new team, leading to potentially more accurate forecasts.

In our results, we observed some major disparities in the precision of the two prediction versions, prompting us to present both for a fuller understanding of the model's capabilities and limitations.

	Count	Mean	Std Dev	Min	2.5%	Median	97.5%	Max
Basic	4000	1.2505	1.2578	0	0	1	4	9
Advanced	4000	7.2073	35.6933	0	0	2	25.03	559

Table 3: Summary statistics for two predictions

farsi spiegare le equazioni dei modelli basic e advanced

Table 3 gives us important insights:
Basic prediction:

- This prediction indicates a more modest expected number of goals, with a mean of 1.2505.
- The standard deviation is fairly low at 1.2578, suggesting that outcomes are generally close to the mean.
- The median of 1 suggests that more often than not, one can expect a single goal.
- The 97.5 percentile value is 4, meaning that there's a 2.5% chance of scoring 4 or more goals.

Advanced prediction:

- The mean of 7.2073 is drastically higher, implying a more optimistic outlook.
- A much larger standard deviation (35.6933) indicates more uncertainty in the prediction.
- The 97.5 percentile value of 25.03 suggests a very wide range of possible outcomes.
- The maximum value of 559 is an extreme outlier, suggesting the model might overestimate in some scenarios.

The basic and advanced models give very different predictions. This makes us wonder if the advanced model is maybe too complicated or sensitive to certain things in the data. The basic model is more steady and consistent, but it might be missing some important details. The advanced model has a lot more variation, making it less reliable. The really high maximum value in the advanced model could mean it's picking up random noise instead of real patterns. Because of all this, we might need to adjust the advanced model to get better predictions.

3.3 WAIC comparison

In order to compare the predictive accuracy of our three models — unpooled, hierarchical, and pooled — we employed the Watanabe-Akaike Information Criterion (WAIC). A lower WAIC value suggests a better balance and thus a more desirable model for out-of-sample prediction.

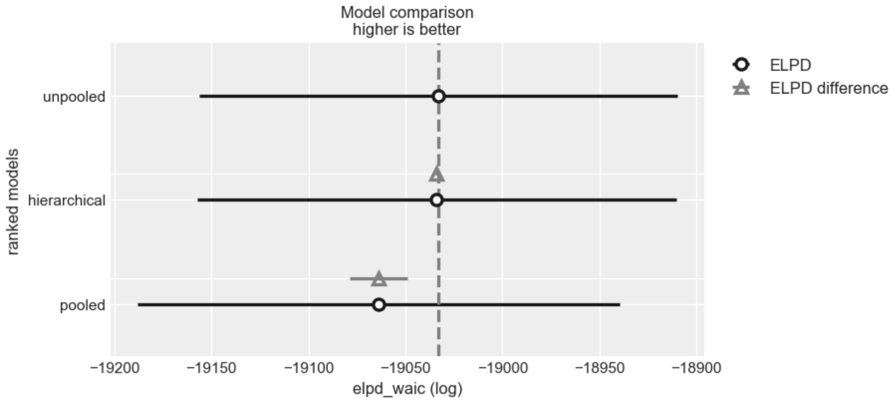


Figure 8: Comparison of WAIC values across models

Upon examination of the WAIC values, all three models demonstrated similar predictive accuracy. However, some distinctions can be noted. The unpooled model ranked highest in terms of predictive accuracy, as indicated by its lowest WAIC value. It was closely followed by the hierarchical model, which had a nearly identical WAIC value. The pooled model lagged slightly behind the other two, indicating that it may be less preferable for making out-of-sample predictions.

4. Conclusions

Based on our hypothesis testing and statistical models, it appears that the home-field advantage is indeed a real phenomenon. However, as for our second hypothesis concerning the impact of the days of the week on the game's score, the evidence suggests that they do not have a statistically or practically significant influence, given the minute scale of the observed effects.

In our study, most of our models worked well, especially in terms of fitting the data properly and making good guesses for future outcomes. However, regarding novel prediction, the model that was supposed to show how "rest days" affect the game didn't work as we hoped, leaving us with the most basic model taking first place.

References

[1] G. Corani, “Bayesian data analysis and probabilistic programming,” SUPSI, 2023.

[2] [Online]. Available: https://www.sport12x.com/en/mathematics/ice-hockey.