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

Professor: Giorgio Corani
SUPSI, Lugano Switzerland

Assistant: Marco Forgione
SUPSI, Lugano Switzerland

22/01/2023

Table of Contents

List of Figures

List of Tables

1. Problem definition

The problem consisted in searching for a dataset on which perform bayesian analysis. These were the precise tasks to perform:

- Description of the data.
- At least one hypothesis test for each student.
  - discuss the choice of the prior and likelihood
  - evaluate prior sensitivity (i.e., re-run with a different and reasonable prior)
  - draw some conclusions based on the analysis of the posterior
- At least one hierarchical and one one non-hierarchical regression (or classification) model presenting:
  - the choice of the prior and likelihood;
  - the posterior distribution of the mean for a group (i.e., a specific hospital, rider etc) of choice
  - the predictive distribution for the same group, i.e., the probability distribution for a future measure for that group. The predictive distribution has to be computed by a cycle implemented by the students, without using pymc3 functions such as sample_posterior_predictive
  - the posterior distribution of a novel group, for which there are currently not yet observations (for instance, an hospital not yet included in the trial).
- Moreover, the project should include:
  - Sensitivity analysis with respect to a couple of different prior choices
  - An example of model comparison (WAIC) to compare two candidate models (for instance hierarchical and non-hierarchical).
  - For the chosen model, discuss the convergence diagnostic.
  - PyMC3 code

2. 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. The dataset is composed by 17 columns and 2458 rows. For the purpose of this project we transformed our data so that it resulted in this way:

- dataset containing only games relative to HC Lugano
- result is represented in difference of goals
- a “Place” column, which represents if game was played at home (0) or away (1)
- a “Win” column, which represents if game was won at home (1) or lost (0)
- a “Difference” column, which represents the number of resting days before that match
- the opponents’ team name

Figure 1 shows the distribution of the difference of goals for the team HC Lugano, against the 12 other teams. Value -1 is the highest bar, showing that HC Lugano is most probable to lose a game by 1 point, on the other hand, we see that the bars on the right side are higher on average that the ones on the left side, suggesting that the team is winning more games than losin them.

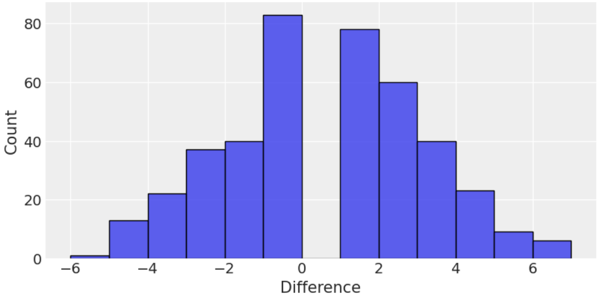


Figure 1: Distribution of difference of goals

3. Models

The models used were the ones seen in class with professors G. Corani and M. Forgione, namely pooled, unpooled, hierarchical and varying-slope models.

4. Hypothesis

4.1 1st Hypothesis

The first hypothesis we came up with was whether the team does perform better in their own stadium in respect of playing away.

In this case for the prior we expect a Normal distribution, since the results should be distributed this way.

4.2 Non-hierarchical models

4.2.1 Pooled

We chose to predict the results with a linear regression which ignores the team, therefore we defined 2 priors: one on the slope and the second one on the intercept.

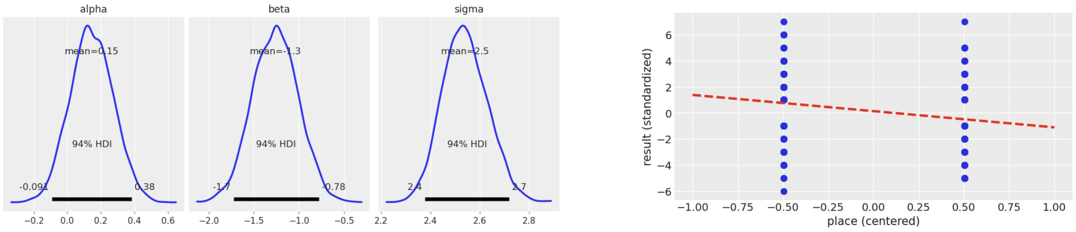
$$Y \sim N(\alpha + \beta X, \sigma)$$

Y: difference of goals (412 measures)
X: home or away (0 or 1)
α : intercept. (on centered data, it represents the mean result for HC Lugano).

We use the centered covariate $X_t = X - \bar{x}$. We have no prior knowledge and thus we use fall back to weakly informative, data-dependent, priors:

$$\begin{aligned}
Y &\sim N(\alpha_t + \beta X_t, \sigma) && \text{likelihood} && (1)\\
\alpha &\sim N(\bar{y}, 2s_y) && \text{prior on the intercept} && (2)\\
\beta &\sim N(0, 2.5\frac{s_y}{s_x}) && \text{prior on the slope} && (3)\\
\sigma &\sim \text{HalfNormal}(0, 1.5s_y) && \text{prior on the dev std} && (4)
\end{aligned}$$

Results: From the posterior in figure 2a we understand that the value for the slope is negative, meaning that our hypothesis is indeed correct: apparently the selected team (HC Lugano, but this is probably for every team) is performing better in its own stadium compared to games in foreign ones. We clearly see from figure 2b that the games played away have the results more towards the lower side, suggesting the first hypothesis is true.



(a) Posterior distributions

(b) Pooled model fitted line

Figure 2: Pooled model

4.2.2 Unpooled

For each team we have a different independent intercept $\alpha_1, \alpha_2, ..\alpha_j ... \alpha_{12}$

$$Y \sim N(\alpha_{\text{team}[i]} + \beta X, \sigma)$$

$\alpha_{\text{team}[i]}$: the intercept for the team where the i-th observation has been made. The slope β is equal for all teams; we assume that the difference between home and away does not depend on the teams.

$$\begin{align} Y &\sim N(\alpha_{\text{team}} + \beta X_t, \sigma) && \text{likelihood} \tag{5}\\ \alpha_j &\sim N(\bar{y}, 2s_y)\ j = 1, 2..., 12 && \text{prior on the intercept} \tag{6}\\ \beta &\sim N(0, 2.5\frac{s_y}{s_x}) && \text{prior on the slope} \tag{7}\\ \sigma &\sim \text{HalfNormal}(0, 1.5s_y) && \text{prior on the dev std} \tag{8} \end{align}$$

Result: with the unpooled model we see once again our first hypothesis is further confirmed. In image 3, we compare the pooled and unpooled model, and we see that the pooled model is best at fitting the intercept for each team. Quite interestingly, it seems that HC Lugano and Genève-Servette HC are on the same level of strength.

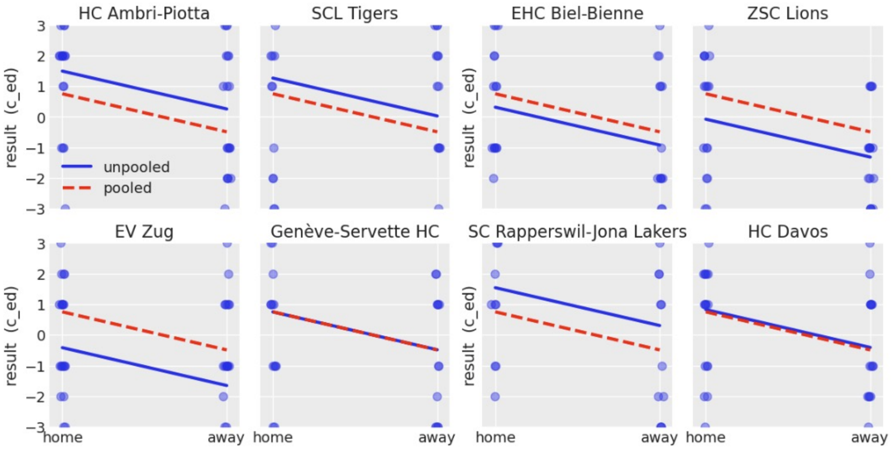


Figure 3: Comparing unpooled with pooled model

4.3 Hierarchical models

4.3.1 Hierarchical/varying-intercept

A different intercept for each team, modeling also the population of intercepts, its estimates are a compromise between the unpooled and the pooled model.

$$
\begin{aligned}
Y &\sim N(\alpha_{j[i]} + \beta X_t, \sigma) && \text{likelihood} && (9)\\
\beta &\sim N(0, 2\frac{s_y}{s_x}) && \text{prior on the slope} && (10)\\
\sigma &\sim \text{HalfNormal}(0, 1.5 s_y) && \text{prior on the dev std} && (11)\\
\alpha_j &\sim N(\mu_\alpha, \sigma_\alpha)\ j = 1, 2..., 12 && \text{population of intercepts} && (12)\\
\mu_\alpha &\sim N(\bar{y}, 10 s_y) && \text{prior on the mean intercept} && (13)\\
\sigma_\alpha &\sim \text{HalfNormal}(5 s_y) && \text{prior on the std of intercepts} && (14)
\end{aligned}
$$

Broad priors on the parameters of the population of intercepts: this is a **varying-intercept** model, since the intercept is different for each team, while the slope is unique for all teams.

Result: In figure 4 we perform the same comparison, with the added hierarchical model. In fact, we see that this model is a compromise between the pooled and unpooled models.

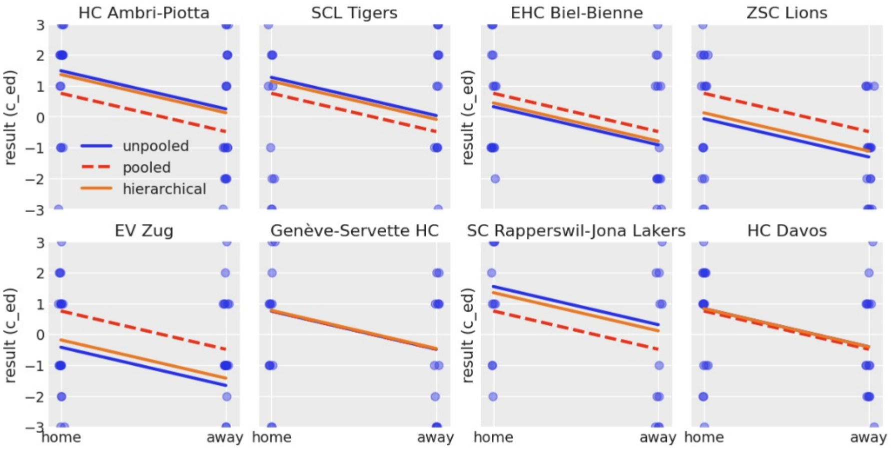


Figure 4: Comparing hierarchical with pooled and unpooled models

4.3.2 Varying-intercept & varying-slope

Much like varying-intercept model, this one also takes into variable the slope of the fitted line, in order to find the best fit for each opponent team.

$$
\begin{aligned}
Y &\sim N(\alpha_{j[i]} + \beta X_t, \sigma) && \text{likelihood} && (15)\\
\beta_j &\sim N(\mu_\beta, \sigma_\beta)\ j = 1, 2..., 12 && \text{population of slopes} && (16)\\
\mu_\beta &\sim N(0, 10 s_y) && \text{prior on the mean slope} && (17)\\
\sigma_\beta &\sim \text{HalfNormal}(5 s_y) && \text{prior on the std of slopes} && (18)\\
\alpha_j &\sim N(\mu_\alpha, \sigma_\alpha)\ j = 1, 2..., 12 && \text{population of intercepts} && (19)\\
\mu_\alpha &\sim N(\bar{y}, 10 s_y) && \text{prior on the mean intercept} && (20)\\
\sigma_\alpha &\sim \text{HalfNormal}(5 s_y) && \text{prior on the std of intercepts} && (21)\\
\sigma &\sim \text{HalfNormal}(0, 1.5 s_y) && \text{prior on the dev std} && (22)
\end{aligned}
$$

Result: In figure 5 we perform the same comparison, with the varying-slope model instead of the hierarchical one. We expected to see a major difference in the slope, instead, is is very similar to the hierarchical one.

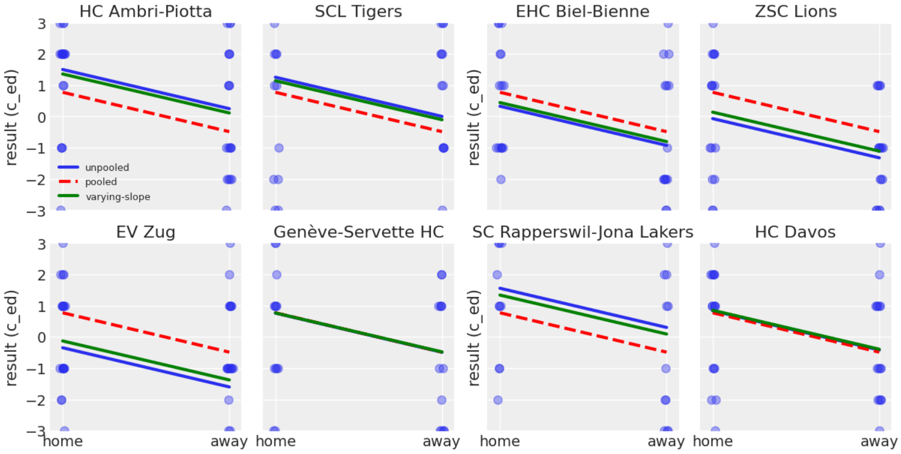


Figure 5: Comparing varying-slope with pooled and unpooled models

4.4 2nd Hypothesis

The second hypothesis was about the resting days: does a team perform worse if the resting days are too low, or even too much?

We chose to predict the results with a *Bernoulloi* distribution which ignores the team, therefore we defined 2 priors: one on the slope and the second one on the intercept.

$$Y \sim f(k;p) = \begin{cases} p & \text{if } k = 1, \\ 1-p & \text{if } k = 0. \end{cases}$$

with

$$p = \alpha + \beta X$$

Y: difference of rest days (412 measures)
X: win or not (1 or 0)
α : intercept. (on centered data, it represents the mean rest days for HC Lugano).

We use the centered covariate $X_t = X - \bar{x}$. We have no prior knowledge and thus we use fall back to weakly informative, data-dependent, priors:

We developed a pooled model in order to immediately answer the question, and to our surprise, it did not seam to be true.

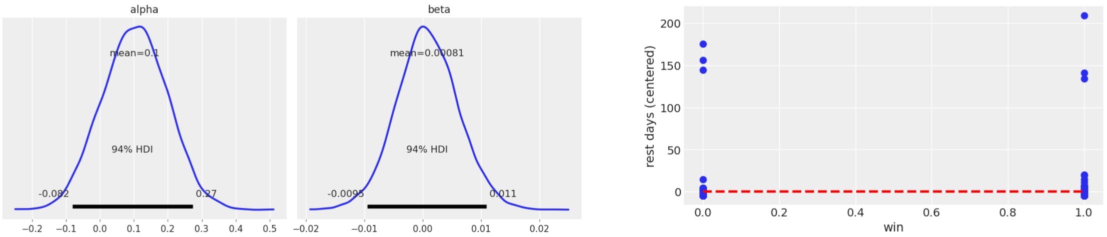
$$Y \sim f(k;p) = \begin{cases} p & \text{if } k = 1, \\ 1-p & \text{if } k = 0. \end{cases} \qquad \text{likelihood} \qquad (23)$$

$$p = \alpha + \beta X \qquad \text{probability of winning} \qquad (24)$$

$$\alpha \sim N(\bar{y}, 2s_y) \qquad \text{prior on the intercept} \qquad (25)$$

$$\beta \sim N(0, 2.5\frac{s_y}{s_x}) \qquad \text{prior on the slope} \qquad (26)$$

Figure 6a we already see that the slope is near 0, meaning we will have a nearly flat line. From figure 6b we understand that no matter how many rest days the team HC Lugano got, it did not influence the game results, as we see the red dashed line being parallel to the 0 value.



(a) Posterior distributions

(b) Pooled model fitted line

Figure 6: Pooled model

From our perspective, it did not made sense to perform this analysis also using unpooled or hierarchical models, as we believed we would have introduced the different teams' power as variables, and not only the rest days. To our advice, this model is actually valid, as the rest days actually should not influence no the games results, since the teams still practice between games.

5. Predictions

The predictions were calculated only for the first hypothesis. In table 1 we show how the hierarchical and varying-slope models predict the mean goal difference. In this case, we used HC Ambri-Piotta as it is the most anticipated and famous match for the fans of these 2 teams. The predictions shows that HC Lugano is more prone to win against HC Ambri-Piotta, with a mean over 1 for all models.

	Hierarchical	**Unpooled**	**Varying-slope**
count	8000	8000	8000
mean	1.33	1.51	1.67
std	2.52	2.46	2.49
min	-8.32	-7.92	-7.60
25%	-0.38	-0.12	-0.01
50%	1.31	1.53	1.67
75%	3.02	3.13	3.35
max	11.69	9.70	11.14

Table 1: Predictions for an existing team

In the case of the prediction for a new team, still HC Lugano is predicted to win with a mean difference of goals of +0.89. See table 2.

	Hierarchical
count	8000
mean	0.89
std	2.64
min	-9.75
25%	-0.87
50%	0.86
75%	2.68
max	9.51

Table 2: Predictions for a new team

6. Conclusions

From the comparison plot we understand that the unpooled model is the best one, even if by very little.

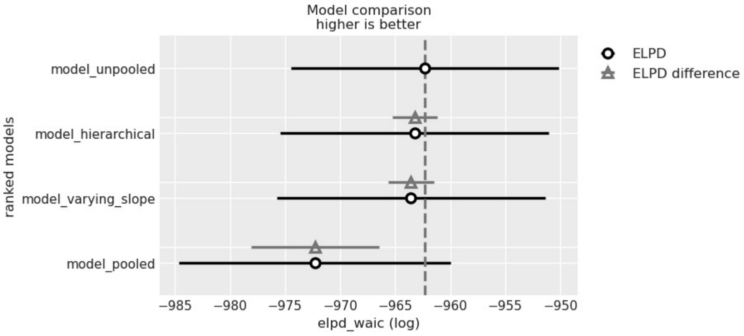


Figure 7: Model comparison

All in all, this project helped us in understanding how the bayesian models work. Moreover, it was interesting discovering how the prediction we got were actually reasonable, confronting them with the actual performances of the teams in real life.