STAT 775: Bayesian Statistics

Problem Set 4

DUE: 12:00 pm (noon) on Friday 2 May *Tyler Teichmiller*

4.1. The UEFA Women's Championship is an quadrennial European football championship. The 2025 tournament will take place later this year in July and features teams from 16 European countries. The 16 teams are divided into four groups of four teams each. The tournament consists of two stages, a group stage and a knockout stage. During the group stage, teams play a single match against every team in their group and earn three points for each win, zero points for each loss, and one point for each draw. The two teams with the highest number of points advance to the knockout stage. See the Wikipedia page for the exact tournament structure. Note, in particular, the rules about breaking ties.

Please build a Bayesian predictive model to estimate the probability that each team advances past the group stage and into the knockout stage. You should specifically report a point estimate and uncertainty interval for each team's changes of moving on to the knockout stage. Make sure that you justify your modeling choices and evaluate the sensitivity of your model output to those choices.

To build your model, you may use the following information:

- The country
- Current FIFA Rankings for each team
- Results for matches played in past UEFA Women's Championships (ging back to 2005) and the FIFA rankings for the participating teams prior to the tournament.

You may find FIFA rankings by going back many years at this website and you may find results from previous tournaments on Wikipedia.

You should submit a short writeup that includes

- An introduction summarizing your overall approach and top-line conclusions.
- A section in which you clearly specify the full probabilistic model and describe how you used it to compute the requested probabilities. Be sure to state all modeling assumptions and justify them!
- A section containing summarizing your overall results. Be sure to include enough evidence that any MCMC simulations have converged and that your final model is reasonable. You should also include a short description of any sensitivity analyses, model diagnostics, and model assessments you performed

- A discussion where you summarize your findings, comment on limitations, and suggest avenues for future enhancement.

You should also submit a compressed archive containing your code and data.

Solution:

1 Introduction

This problem tasked me with building a bayesian predictive model to estimate the probability that each team advances past the group stage and into the knockout stage of the Women's UEFA Euro 2025 tournament.

The Women's UEFA Euro 2025 soccer tournament will consist of 16 European teams. To start, teams are placed into 4 groups based on some criteria and some randomization. The 4 teams in a group first play against each other in a round-robin tournament, where each team plays against every other team in the group. Then, points are assigned for wins, draws, and losses, and the top two teams move on to the knockout stage of the tournament. There are some criteria that will break ties if multiple teams finish with the same number of points that will establish who is the second place team in the group.

My approach to this problem was to model the outcome of individual games and use that information to form probabilities that a team will advance. More formally, I model

$$G_H \sim \text{Poisson}\left(\lambda_H\right)$$

$$G_A \sim \text{Poisson}\left(\lambda_A\right)$$

where G_H is the number of goals the Home Team scores and G_A is the number of goals the away team scores.

The parameters λ_H and λ_G are defined in relation to some underlying parameters and the standardized ranking difference of the teams that are playing. More details can be found in the Model section.

We can then do MCMC simulations to get approximations of the posterior-predictive distributions of the number goals the home and away team score in each match in 2025. In each of these simulations, we can deduce the final ordering of teams within groups based on the number of wins, draws, losses, and other tie-breaking criteria defined by the tournament. If we save these final positions, we now have samples from the posterior-predictive distribution of ordered finishes in each group, which gives us the posterior-predictive distribution of advancing.

From this, we can obtain point estimates and intervals for the probability of advancing for each team in the UEFA Women's Euro 2025 tournament.

Overall, the team with the highest average posterior predictive probability of advancing out of the group stage in 2025 is Spain, with a probability of 0.857. The team with the lowest

average posterior predictive probability of advancing out of the group stage in 2025 is Poland, with a probability of 0.0675.

2 Model

2.1 Full Model

Note that there are 24 games played in total in the group stage in the tournament. Let i be the game identification number from 1 to 24.

$$G_{Hi}|\lambda_{Hi} \sim \text{Poisson}\left(\lambda_{Hi}\right) \text{ independently for } i = 1, \ldots, 24$$

$$G_{Ai}|\lambda_{Ai} \sim \text{Poisson}\left(\lambda_{Ai}\right) \text{ independently for } i = 1, \ldots, 24$$

G_{Hi} = Goals Home Team scores in game i

G_{Ai} = Goals Away Team scores in game i

A poisson distribution makes sense for the number of goals, as the poisson distribution is discrete and nonnegative. Then, let

$$\beta_0|\mu_0, \sigma_0^2 \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right)$$

$$\beta_1|\mu_1, \sigma_1^2 \sim \mathcal{N}\left(\mu_1, \sigma_1^2\right)$$

$$\lambda_{Hi} = \exp(\beta_0 + \beta_1 * \tilde{x}_i)$$

$$\lambda_{Ai} = \exp(\beta_0 + \beta_1 * -\tilde{x}_i)$$

Where

- x_i = (Home Team Ranking - Away Team Ranking) in game i
- $\tilde{x}_i = \frac{x_i - mean(\mathbf{x})}{sd(\mathbf{x})}$, where $\mathbf{x}$ is the vector of x_i's

Note $\tilde{x}_i$ is just the standardized difference in Pre-Tournament World Ranking.

This model states that the number of goals a team scores follows a Poisson distribution with an average value of λ_{Hi} or λ_{Ai}, depending on if it is the home or away team. This model is assuming that there is a direct log-linear relationship between the average number of goals a team scores, and the standardized difference in pre-tournament world ranking of the team against their opponent.

Then, we can use Stan to get samples the posterior distribution of $\beta_0, \beta_1|G_{Ai}, G_{Hi}, \mathbf{x}$. We can also use Stan to get posterior predictive draws of $G^{\star}_{Ai}|\beta_0, \beta_1, \mathbf{x}$ and $G^{\star}_{Hi}|\beta_0, \beta_1, \mathbf{x}$. From these posterior predictive draws of goals, we can compute the number of wins, losses, and draws each team has, which leads to computing directly the number of points each team has. After handling some tiebreaking criteria where teams tie for the second place, the top two teams advance to the knockout stage. We can save whether or not each team advances to the

knockout stage based on the individual draw of goals for every draw from the posterior. We can then use a Monte Carlo approximation to obtain the approximate posterior predictive probability that a team advances. In order to get interval estimates, a Bayesian bootstrap was needed.

To see how I collected data on previous tournaments, check `pset4_data_collection.R`. To see the implementation of this model in Stan, check `pset4t2.stan`. To see the work to transform the predicted goals scored to group points, and then to advancement probabilities, check `pset4_model.R`.

2.2 Hyperparameter Choice

Our hyperparameters that need fixed values to be chosen are $\mu_0, \sigma_0^2, \mu_1, \sigma_1^2$.

Let's start with μ_1. μ_1 is the average value of the slope coefficient β_1. Note that the parameters λ_{Ai} and λ_{Hi} are the average number of goals the away and home team will score in game i, respectively. Since $\lambda_{Hi}, \lambda_{Ai}$ are defined using β_1, we need to think about how μ_1 value would effect the values of lambda. A reasonable guess is $\mu_1 = 0$. This means that we are guessing that the standardized rank difference has on average no effect on the number of goals the home team and away team will score, meaning there is a universal average number of goals defined only by the value of β_0.

Then, let's define $\sigma_1^2 = 0.3^2$. This will give sufficient probability to values of β_1 that are reasonable. That is, a value of β_1 two standard deviations below the mean is -0.6. Since $\tilde{x}_i$ is defined by taking home minus away ranking, an increase in one unit of $\tilde{x}$ is an increase in the ranking difference by one standard deviation, which means the away team is becoming more highly ranked while the home team stays constant. Then, every 1 unit increase in $\tilde{x}$ would result in a $(\exp(-0.6) - 1) \times 100\% = -45.11\%$ change in the average number of goals. This seems reasonably dispersed, but not so dispersed that it is uninformative.

Next, let's think about μ_0. μ_0 is the average value of β_0. Since the average value of β_1 is 0, we need to select μ_0 so that on average, the value of the average number of goals the teams will score is reasonable. $\mu_0 = 0.5$ is reasonable, becuase it means that on average, the number of goals scored by a team will be close to $e^{0.5} = 1.64$. We can set $\sigma_0^2 = 0.3$ for reasons similar to before for σ_1^2.

So, our hierarchical model will start with

$$\beta_0 | \mu_0, \sigma_0^2 \sim \mathcal{N}\left(0.5, 0.3^2\right)$$

$$\beta_1 | \mu_1, \sigma_1^2 \sim \mathcal{N}\left(0, 0.3^2\right)$$

Team	Group	Point Estimate	95% Interval
Iceland	A	0.74025	(0.72675, 0.75375)
Switzerland	A	0.332	(0.31725, 0.3465)
Norway	A	0.638	(0.623, 0.65275)
Finland	A	0.28975	(0.27525, 0.30425)
Belgium	B	0.3455	(0.331, 0.36025)
Spain	B	0.857	(0.846, 0.86775)
Portugal	B	0.256	(0.2424938, 0.27)
Italy	B	0.5415	(0.526, 0.55675)
Denmark	C	0.48375	(0.46825, 0.499)
Germany	C	0.74755	(0.7345, 0.761)
Poland	C	0.0675	(0.06, 0.0755)
Sweden	C	0.701	(0.6875, 0.715)
Wales	D	0.0795	(0.07125, 0.088)
France	D	0.57525	(0.56025, 0.59075)
England	D	0.7585	(0.74525, 0.77175)
Netherlands	D	0.58675	(0.5715, 0.602)

Table 1: Posterior Predictive Probabilities of Advancing Past Group Stage

3 Results

3.1 Numerical Summaries

Our main quantities of interest are the predictive probabilities that each individual team advances past the group stage, as well as intervals for the predicted probability. We can see in Table 1 that the team with the highest posterior predictive probability of advancing is Spain with a probability of 0.857. The team with the lowest posterior predictive probability of advancing is Poland with a probability of 0.0675.

We also are concerned with the estimate of the posterior mean of β_0 and β_1. The posterior mean of β_0 is 0.2141, with a 95% interval of (0.0849, 0.3371). The posterior mean of β_1 is -0.4075, with a 95% interval of (-0.5164, -0.3006).

So, the posterior mean for the average number of goals a team will score when playing a team of the same rank is $\exp(0.2141 - 0.4075 * (0)) = 1.2387$.

3.2 Model Diagnostics

First, let's do some sensitivity analysis for the hyperparameters.

If we instead use

$$\beta_0 | \mu_0, \sigma_0^2 \sim \mathcal{N}\left(1, 0.3^2\right)$$
$$\beta_1 | \mu_1, \sigma_1^2 \sim \mathcal{N}\left(0, 0.3^2\right)$$

Then the posterior mean for the average number of goals a team will score when playing a team of the same rank is $\exp(0.2375 - 0.3993 * (0)) = 1.2605$. This is overall very similar to the real hyperparameters chosen. It leads to similar point estimates and confidence intervals.

If we instead use

$$\beta_0|\mu_0, \sigma_0^2 \sim \mathcal{N}\left(0.5, 0.3^2\right)$$
$$\beta_1|\mu_1, \sigma_1^2 \sim \mathcal{N}\left(0, 0.01^2\right)$$

Then the posterior mean for the average number of goals a team will score when playing a team of the same rank is $\exp(0.2954 - 0.0109 * (0)) = 1.3436$. The estimated β_1 is now very shrunk towards zero, which is a sign of a prior that is too restrictive.

If we instead use

$$\beta_0|\mu_0, \sigma_0^2 \sim \mathcal{N}\left(0.5, 0.01^2\right)$$
$$\beta_1|\mu_1, \sigma_1^2 \sim \mathcal{N}\left(0, 0.3^2\right)$$

Then the posterior mean for the average number of goals a team will score when playing a team of the same rank is $\exp(0.4925 - 0.3226 * (0)) = 1.634$. The estimated β_0 is now very shrunk towards 0.5, which is a sign of a prior that is too restrictive.

We also need to make sure that our final model is reasonable. The final estimated β_0 and β_1 using the original hyperparameter values gave us

$$\hat{\lambda}_{Hi} = \exp(0.2141 - 0.4075 * \tilde{x}_i)$$

This means that for an increase in ranking difference by one standard deviation, the average number of goals the home team scores changes by $\exp(-0.4075) - 1) \times 100\% = -33.46\%$. This makes sense. Let some home team's rank be 5. An increase in standardized ranking difference means that the rank of the away team is getting closer to the rank of the home team from above (e.g. rank 10 to rank 8), or is getting smaller than the home team (e.g. rank 4 to rank 2). This means the away team is getting better. We should expect the average number of goals a team plays to be less when they play a comparably better team. Thus, our model makes logical sense.

Finally, we need to make sure our MCMC iterations actually converged to the posterior distribution. Based on the `R` output, each parameter's r-Hat values were less than 1.01, with effective sample sizes of at least 2000. These diagnostics do not guarantee that the posterior distribution is correct, but we are pretty confident that the MCMC iterations converged.

4 Discussion

4.1 Summary

In summary, in section 2.1 we fit a hierarchical model to predict the number of goals a team will score based on their own world rank and the world rank of their opponent. We did

this assuming a log-linear relationship between the average number of goals scored and the standardized world ranking difference. We allowed the slope and intercept of that relationship be random variables, putting normal priors on both values. We argued in section 2.2 for specific hyperparameter values for the priors on the slope and intercept. We then fit a Stan model that gives us draws from the posterior distribution of $\beta_0, \beta_1 | G_{Ai}, G_{Hi}, \mathbf{x}$. Additionally, it gave us draws from the posterior predictive distributions $G^{\star}_{Hi} | \beta_0, \beta_1, G_{H_i}, \mathbf{x}$ and $G^{\star}_{Ai} | \beta_0, \beta_1, G_{A_i}, \mathbf{x}$. From these draws, we were able to compute Monte Carlo approximations for the probability of a team advancing through the tournament. The team with the highest predicted probability of advancing is Spain, with a point estimate probability of 0.857. The team with the lowest predicted probability of advancing is Poland, with a point estimate probability of 0.0675.

4.2 Limitations

One major limitation of my framework is that it does not directly estimate the probability of advancing at all. It does, but through other variables that directly impact the probability of advancing, like the number of goals scored in a match by each team. It would have been a lot less code to do do a logistic regression on a binary variable which records whether or not a team advances.

Another limitation is that I had to define the variable λ_{Hi} and λ_{Ai} to ensure that the rate parameter of the Poisson distribution was positive. This was the workaround that I came up with, but it came at a large cost: having to assume the log-linear relationship described before.

Another major limitation is that it was very time consuming to manually tiebreak in my R code, so I simplified the criteria to only rely on the head to head game result. If the head to head game was a tie, I randomly selected one of the two teams. This was not a realistic simplification because it would not be what the tournament would really do.

The final major limitation of my work was that it required some code that I forgot how to do, like how to scrape data off of websites.

4.3 Future Work

Future work could address the limitations of my work by devising a way to sample directly from the posterior distribution of the probability of advancing using something like a Dirichlet Process, like we talked about in class. Other work could be done to investigate the relationship between λ_{Hi}, λ_{Ai}, and $\tilde{x}_i$ that I assumed to be log-linear.