Extrapolation Using Bayesian Linear Regression: Weekly Swim Times in High School

Dhyey Dharmendrakumar Mavani (ddm2149)
Sophomore (Bachelors') student, Columbia Visiting Students Program Spring 2023
CS, Math and Statistics Triple Major, Amherst College Class of 2025
STAT GU 4224: Bayesian Statistics

1. Data & Motivation

For this project, I used the "swim.dat", which is part of "A First Course in Bayesian Statistics" book by Peter J. Hoff. The file "swim.dat" contains data on the amount of time, in seconds, it takes each of the four high school swimmers to swim 50 yards. Each swimmer has six times recorded. The data is recorded on a biweekly basis. The data can be accessed at http://www2.stat.duke.edu/~pdh10/FCBS/Exercises/swim.dat for open-source usage. The swim times are recorded for weeks 1, 3, 5, 7, 9, and 11 for four different swimmers.

My goal for this project is to perform data analysis on each swimmer separately by fitting a Bayesian linear regression model with swimming time as a response variable and week as an explanatory variable using a variety of flat and informed priors. Furthermore, I plan to obtain a posterior predictive distribution for each swimmer's time to swim if we were to record the next observation in the experiment, or more concretely, if they were to swim 2 weeks after the last observation noted in the dataset. I plan to rank players based on which ones I recommend the coach to be chosen to represent the school in the tournament 2 weeks after the last observation (or week 13) by computing the conditional probabilities of having the best time for each swimmer from the posterior predictive distribution generated earlier. Finally, I plan to perform Bayesian linear regression on the swim times data with the swimming time as a response variable and the week as an explanatory variable for each swimmer and for the group as a whole.

2. Priors Descriptions

Let's first consider how to select priors for our model deliberately. With the goal in mind, it seems feasible to me to use a variety of priors including one informed by our general knowledge of the situation at hand and one bearing very little information (flat). For the determination of informed prior, I assumed that high school students' general swimming times in the given situation range from 22 to 24 seconds. This means that the prior expectation of our y-intercept would be 23 in this case. In this prior, I also assumed that the training from week to week will not make much difference since all high school swimmers are already at the top of their form, so my prior expectation on week-by-week changes would be 0. This means that I have $\boldsymbol{\beta} = (23,\ 0)^T$.

Assuming we have no covariance with the coefficients of $\boldsymbol{\beta}$, but considering the uncertainty of our prior values, we say that we are 95% confident that the y-intercept falls in the range [22, 24]. We let $\boldsymbol{\Sigma}_0(1,1) = 1/4$ because this satisfies our condition of a 95% confidence interval that 2 standard deviations from 23 lie in 1 unit change. We also expect that training has a relatively mild effect on time (even though centered around 0), so we let, $\boldsymbol{\Sigma}_0(2,2) = 0.5$, which is just an arbitrarily chosen small variance for the training effect shifts. The rest of the entries in the matrix $\boldsymbol{\Sigma}_0$ are initialized to 0 because of our assumption of no covariance among those aspects. For our expectation of the variability of measurements, let's similarly set $\sigma_0^2 = 1/4$ and only lightly center this prior with $\nu_0 = 1$.

Moreover, I will also consider flat prior in the context of this situation so that we can see how that affects our interpretations and conclusions for the model at hand. I used a flat prior for the regression coefficients, which is equivalent to assuming that all possible values of the coefficients are equally likely before observing the data. This is implemented in the code by setting the prior mean for the regression coefficients to a reasonable value and the prior variance to a very large value, which effectively places almost no constraint on the values of the coefficients. The prior for the error variance is still an inverse-gamma distribution with respective parameters, which is conjugate to the likelihood and allows for efficient sampling using the Gibbs sampler.

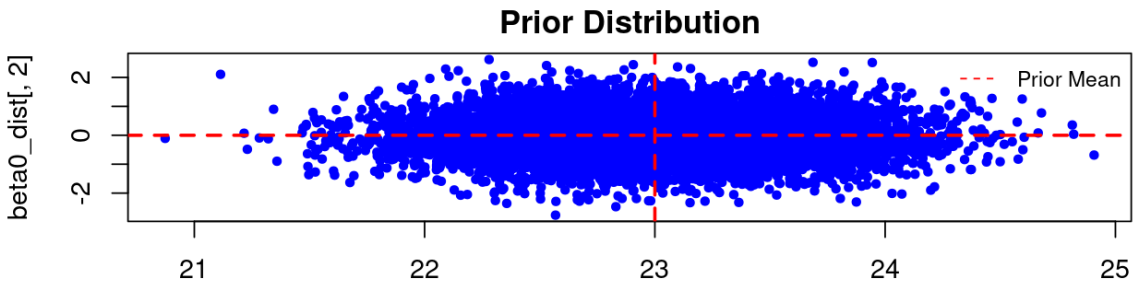
In the next part, we generate the posterior predictive distribution with different priors in mind.

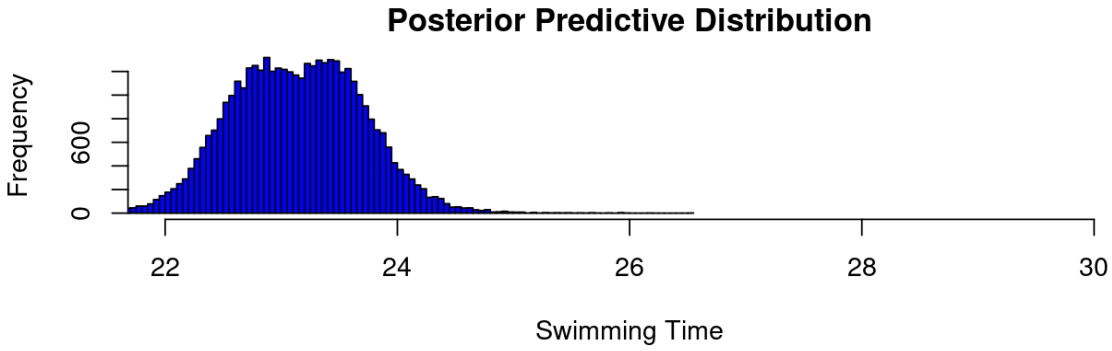
3. Posterior Predictive Distribution

Now, let's Y_j^* represent the posterior predictive swim times on the next time step for player j, where j is either 1, 2, 3, or 4. We try to create the posterior predictive distribution for the next time step using the Gibbs Sampler method and compute $Prob(Y_j^* = min\{Y_1^*, \dots, Y_4^*\} \mid Y_1, Y_2, Y_3, Y_4)$ for j = 1, 2, 3, and 4, which represents the probabilities of the swimmer j having the best time among the four in the next tournament for each of the four swimmers.

(a) "Informed" prior:

Firstly, let's consider the case of "informative" prior developed in the last section. We have the following prior and posterior predictive distribution graphs in this case:





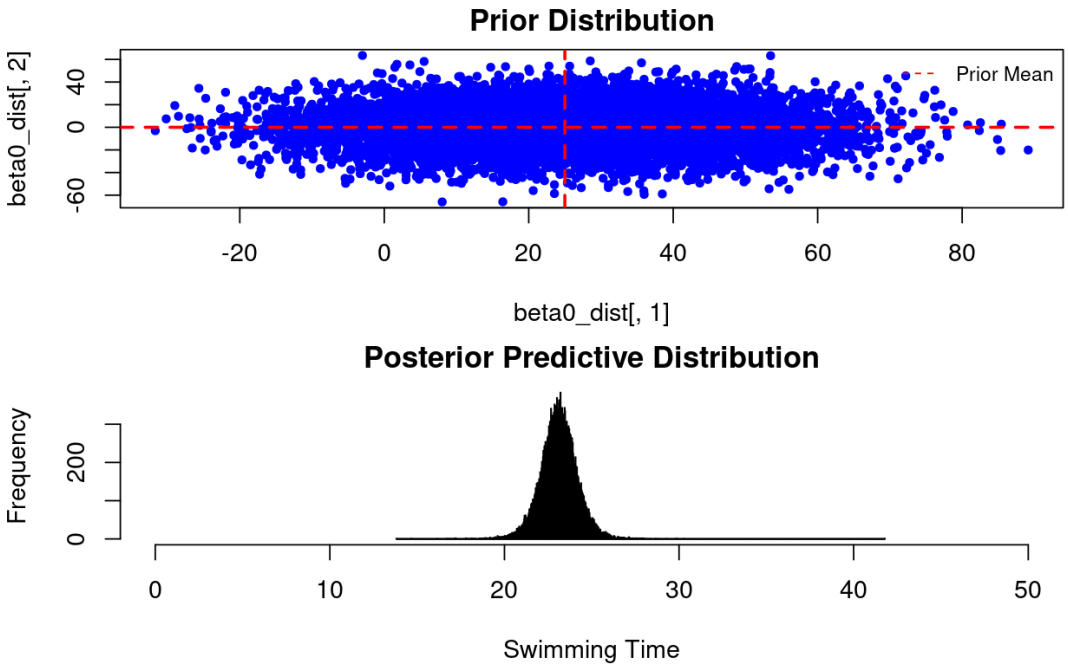
Furthermore, in this case of “informed” prior, we found the probabilities of each player outperforming all others at week 13 or the following observation. These probabilities are outlined in the table below:

```
best_times
     1      2      3      4
0.6571 0.0195 0.2977 0.0257
```

Based on my statistical analysis, I noticed that with our posterior predictive dataset in the case when we have an “informative” prior, swimmer 1 turned out to be the fastest about 66% of the times, so we recommend that the school coach selects swimmer 1 to represent the school in the upcoming race.

(b) “Flat” prior:

Secondly, let's consider the case of “flat” prior developed in the last section. We have the following prior and posterior predictive distribution graphs in this case:



Furthermore, in this case of “flat” prior, we found the probabilities of each player outperforming all others at week 13 or the next observation. These probabilities are outlined in the table below:

```
best_times
     1      2      3      4
0.4550 0.1052 0.2983 0.1415
```

On the other hand, I noticed that with our posterior predictive dataset in the case when we have used a flat prior, swimmer 1 still turned out to be the most probable to be the fastest about 46% times in the swim time by week 13, but the differences among the conditional probabilities of the players 1 and 3 became smaller by around 20%, so although we still recommend that the school coach selects swimmer 1 to represent the school in the upcoming race, we would highly suggest running some further statistical or qualitative analysis to make an informed decision of selection of player 1 or player 3.

4. Bayesian Linear Regression

For the Linear Regression part, I first decided to regress swim time vs week it was measured with all the swimmers together while using “informative” prior. After doing so, I learned that the mean value of the intercept was around 22.9 with a 95% confidence interval of (22.7, 23.2). Furthermore, the slope coefficient (or beta) turned out to have a mean of around 0.03 with a 95% confidence interval of (-0.002, 0.07), and since this interval contained 0 in it, we cannot be sure that there is an improvement in swim times as the week progresses for the swimmers in

general. This is in line with our prior belief in general. Now, let's try to see how is the situation with the individual swimmers with the same prior.

(a) Swimmer 1

I learned that the mean value of the intercept was around 23.2 with a 95% confidence interval of approximately (22.9, 23.4). Furthermore, the slope coefficient (or beta) turned out to have a mean of around -0.04 with a 95% confidence interval of (-0.06, -0.008), and since this interval did not contain 0 in it, but lies completely on the negative side, we can say with 95% confidence that there is a deterioration of performance in terms of the swim times as the week progresses for the swimmer 1.

(b) Swimmer 2

I learned that the mean value of the intercept was around 23.1 with a 95% confidence interval of approximately (22.8, 23.3). Furthermore, the slope coefficient (or beta) turned out to have a mean of around 0.04 with a 95% confidence interval of (0.0006, 0.08), and since this interval did not contain 0 in it, but lies completely on the positive side, we can say with 95% confidence that there is an improvement of performance in terms of the swim times as the week progresses for the swimmer 2.

(c) Swimmer 3

I learned that the mean value of the intercept was around 22.7 with a 95% confidence interval of approximately (22.5, 23.0). Furthermore, the slope coefficient (or beta) turned out to have a mean of around 0.01 with a 95% confidence interval of (-0.03, 0.03), and since this interval contained 0 in it, we cannot be sure that there is an improvement in swim times as the week progresses for the swimmer 3.

(d) Swimmer 4

I learned that the mean value of the intercept was around 23.6 with a 95% confidence interval of (23.1, 23.8). Furthermore, the slope coefficient (or beta) turned out to have a mean of around -0.01 with a 95% confidence interval of (-0.04, 0.06), and since this interval contained 0 in it, we cannot be sure that there is an improvement in swim times as the week progresses for the swimmer 4.

This analysis tells us that overall our prior beliefs hold, and gives us more insight into the relationship between swim times and week for swimmers in general and for each of the swimmers individually. Even after doing this analysis, I realized that since swimmer 1 has a much higher probability of having the best time among the four swimmers at the future time step of interest, I would recommend the coach to select swimmer 1 for the upcoming tournament.

But, I would definitely say there is much more room for discovery in further analysis which can be done using hierarchical models and even more data regarding the other recorded characteristics of swimmers in the same time period.

5. Acknowledgment

I would like to acknowledge Professor Dobrin Marchev's time and effort invested in my learning process throughout the Spring 2023 semester as part of the STAT GU 4224 Bayesian Statistics course. This project would not have been possible without continuous feedback from the professor on my understanding of the concepts during office hours and after class.

Also, I would like to acknowledge the efforts made by Mr. Jitong Qi (TA for the course) by answering my burning questions during the weekly office hours. This really helped me build a strong understanding of the course content both in terms of the theory and implementation in R.

6. References

[1] Hoff, P. D. (2009). A First Course in Bayesian Statistical Methods. Springer Science & Business Media. Along with the code from: https://pdhoff.github.io/book/