Relative Age Effect in Japanese Professional Soccer

Demorest, Lemire and Wilson

12/5/2021

Introduction

The Relative Age Effect (RAE) is a term used to describe how those born early in the academic year tend to have an advantage both athletically and academically. An earlier birth is typically associated with increased physical ability and this advantage may occur because those who are older are typically more physically, emotionally or cognitively developed than those who are younger. Much research has been done to look at RAE in North American and European athletes. However very little research has attempted to extend findings surrounding RAE to other parts of the world, specifically Asia. This is why we are interested in Hideaki Ishigami's work in the paper *Relative age and birthplace effect in Japanese professional sports: a quantitative evaluation using a Bayesian hierarchical Poisson model* (published in the Journal of Sports Sciences in 2016). In our paper we will replicate Ishigami's modelling process using simulated data. In order to answer the question: what is the RAE on soccer players in Japan between the ages of 23 and 25 in the 2012 season?

Our simulation and modeling process will mirror the author's study design in which he uses data sourced from the Japan Professional Football League (J. League) from the 2012 soccer season. The J. League data consist of 40 teams, representing a total of 1013 players registered in the 2012 season and since we will focus on players between the ages of 23 and 25, we have 227 observations. The author characterized "becoming a professional sports player" as an event and thus we are dealing with discrete count data. Using a poisson regression model it is possible to estimate the magnitude of the RAE on the likelihood of becoming a professional athlete. More details on the poisson model can be found below. The authors then ran an MCMC algorithm [1] using JAGS (Plummer, 2012) and confirming with Stan (Stan Development Team, 2013).

The work replicated in our paper is an important step in extending the RAE's scope of inference. Additionally this paper and our analyses allow the magnitude of relative age to be quantified. In contrast many other analyses have simply stated whether an association is statistically significant using, for example, a χ^2 test. Finally this paper also includes an offset term in our model to capture the differences in birth rates by month which has also often been left out in other analyses. All of these factors make this topic and methodology valuable in the literature on RAE.

Model

Our Bayesian model consists of our likelihood which describes the data generating process, an offset term which converts our "counts" into "rates", and a prior. We will go into further detail below on all the components of this model.

Data

Each of our observations is the total number of professional players born in a given month. The school year in Japan begins April and ends on March of the following calendar year, which corresponds with the competitive season of most professional sports. Since school year and competitive season both begin in April in Japan, relative age was coded as 0 (April) to 11 (March). Therefore in total we have 12 observations which sum to

[1] The MCMC ran for 25,000 iterations with five chains where the first 5,000 samples were discarded as burn-in. Only every 100 iteration was saved for a total of 1,000 MCMC samples.

our total number of professional players ($\sum_{i=0}^{11} y_i = 227$). *Can say more about this dependending on whether we find data and use it or whether we generate - clarify with Katie first.*

Likelihoood

As previously mentioned becoming a soccer player can be seen as an "event", and thus a number of them can be regarded as a "count of events". To capture this data generating process the author used a poisson distribution for the likelihood with an offset term.

Poisson Distribution Our data are denoted by y_i which represents the number of professional sports players between ages 23 and 25 in the year 2012 who are born in a given month (where $\sum_{i=0}^{11} y_i = 227$). We have:

$$y_i \sim \text{Poisson}(\lambda_i) \text{ for months } i = 1, \ldots, 12;$$

Where our likelihood function is:

$$L(\lambda; y_0, ...y_{11}) = \prod_{i=0}^{11} e^{-\lambda} \frac{\lambda^{y_i}}{y_i!}$$

And our log-likelihood is:

$$l(\lambda; y_0, ...y_{11}) = -11\lambda - \sum_{i=0}^{11} ln(y_i!) + ln(\lambda) \sum_{i=0}^{11} y_i$$

Offset Term The author points out that many analyses of RAE assume that births are uniformly distributed across the year (i.e. that the same number of children are born in every month of the year). However we know that this is not true. Thus we want our model to take into account the number of total births in a given month when assessing whether the number of professional players born in that same month are over-represented. One way to do this is to compare the monthly rates for becoming a professional athlete rather than comparing the number of professional athletes. We can accomplish this by using an offset term in our model. Note for a poisson regression model for counts we have $log(\lambda_i) = \alpha + \beta y_i$. However if we want a poisson regression model for rates we can use $log(\frac{\lambda_i}{\theta_i}) = \alpha + \beta y_i$, where θ_i is the number of Japanese males born in that month in the years of 1987-1989 (which corresponds to professional male athletes who are 23-25 in the year 2012). This is equivalent to $log(\lambda_i) - log(\theta_i) = \alpha + \beta y_i$ or $\lambda_i = \theta_i \times e^{\alpha + \beta y_i}$ (which is how we see the offset term represented in Ishigami's paper). To put it even more plainly, our rate $\frac{\lambda_i}{\theta_i}$ is exponential. Note that interpreting α and β is the same as for a poisson regression model for counts except we multiply the expected counts by θ_i.

Priors

Our priors on λ_i are defined by the exponential relationship: $\lambda_i = \theta_i e^{(\alpha + \beta RA_i)}$. The intercept term α is a baseline probability of becoming a male professional soccer player after controlling for the relative age effect. RA_i is the relative age of those born in month i (where $RA_i \epsilon (0, 1, ...11)$. The coefficient β measures the relative age effect (RAE). For example, April would be mo[illegible] 0, with $RA_1 = 0$. The exponential term gives the rate of becoming a professional male soccer player for that month.

Hyperparameter α

Hyperparameter β Katie's comment about this: *alpha and beta are partial regression coefficients in your glm that are of interest. I'm wondering if alpha has a subscript on it (i.e., is it a "random intercept" term?) because then that heterogeneity in the different intercepts gets modeled by assuming htey all come from one population distr that's a normal with mean mu and var sigma2... the mean and sig2 need to also be estimated from the data (similar to the MH wihtin Gibbs logit hierarcical model example). so you put a prior on those hyper parmas. We'll talk more about this and see chapters 5 and 6 in gelman.*

Full Model

Given our likelihood and prior our full model can be expressed as follows:

$$P(\lambda_i|\vec{y}) \propto \prod_{i=0}^{11} e^{-\lambda} \frac{\lambda^{y_i}}{y_i!} \times \theta_i e^{(\alpha+\beta RA_i)}$$

Look into whether our posterior is Gamma, and if this is even our posterior of interest: https://stats.stackexchange.com/questions/26199/how-do-i-calculate-a-posterior-distribution-for-a-poisson-model-with-exponential

Show visual of prior and likelihood with our chosen alpha and beta hyperparams

Data Simulation Using Model

Data for y_i

The author did not provide the data for the birth months of the 227 soccer players aged 23-25 in the 2012 season. However he did refer to the Japan Professional Football League (J. League) website which contains these data. We sought out these data on the given website and **insert result of search here and what our data are possibly in a code snippet where we store the values in a vector - Connor**

Data for θ_i

The author did not provide the data used in the offset term and therefore we sought out data sources which would provide values for θ_i, the monthly number of Japanese males born in 1987-1989. These are the birth years for 23-25 year old athletes in 2012. We used birth data by country and year provided by United Nations Statistics Division and the sex ratio for the year 2015 provided by Statista. **insert result of search here and what our data are possibly in a code snippet where we store the values in a vector - James**

Simulated α_i and β_i

to simulate we want:

- distribution for monthly birth data in Japan for males born 87-89

- distribution of birth months for our 227 observations

or are we simulating alpha & beta

- she mentioned this when looking at paper

- and in her comment on our "assign 1"

I think it's the latter, but if it's the former what dist do we generate from (not uniform because that's something that separates our model from others).

Right now our JAGS code is generating a posterior for Beta but since we are not doing the hierarchical model shouldn't alpha and beta simply be hyperparams that we input and our posterior is for lambda?

but isn't beta our param of interest because it's the RA_i

RA_i is the "relative age of those born in month i" but based on how it is in the model it seems like it should be a categorical variable (or `as.factor()`) i.e. if April is 0 and March is 11... these numeric values are arbitrary.

Show a visual of data that we generated

Results

Implementing Bayesian Model using JAGS *can include code snippets and describe our implementation*

Bayesian Model *include visual for Density of Beta* `plot(jags_out)[2]`

Assessing Convergence `plot(jags_out)[1]`

with $\hat{R}$ The authors used the Gelman-Rubin statistic (Gelman & Rubin, 1992), $\hat{R}$, which assesses convergence by comparing the estimated between-chains and within-chain variances for each model parameter. Our results for

Assessing Convergence with Geweke's Diagnostic Additionally, the authors used the Geweke's convergence diagnostic (Geweke, 1992) to check the convergence of the MCMC algorithms. The Geweke convergence diagnostic is a test for equality of the means of the first and last part of a Markov chain, typically the first 10% and the last 50%. If the samples are drawn from a stationary distribution of the chain, then the two means are equal and Geweke's statistic has a standard normal distribution. The test statistic is a standard Z-score: the difference between the two sample means divided by its estimated standard error. The standard error is estimated from the spectral density at zero, and so takes into account any autocorrelation and the Z-score is calculated under the assumption that the two parts of the chain are asymptotically independent. Both of these tests are available in the `coda` package in R.

Posterior Predictive Checks

Assessing Posterior Fit to Our Data generate plots with posterior predictive dist and have our data on it too?

Other ideas for PPCs?

Conclusion

Can keep notes of our thoughts/critiques as we go along so it's easier to write this section at the end:

- it was good they included the offset term (he notes in the paper "the assumption that the distribution of the number of births is uniform across the months of the year is clearly invalid" p146)
- didn't like that he didn't include data when he said where he got it

References

Gelman, Andrew, and Cosma Rohilla Shalizi. 2012. "Philosophy and the Practice of Bayesian Statistics." *British Journal of Mathematical and Statistical Psychology* 66 (1): 8–38. https://doi.org/10.1111/j.2044-8317.2011.02037.x.

Ishigami, Hideaki. 2016. "Relative Age and Birthplace Effect in Japanese Professional Sports: A Quantitative Evaluation Using a Bayesian Hierarchical Poisson Model." *Journal of Sports Sciences* 34 (2): 143–54. https://doi.org/10.1080/02640414.2015.1039462.