Final Project, Task 1: Description of paper and model

Demorest, Lemire and Wilson

1. A brief description of the background motivation of your chosen article and a *single* research question you will choose to investigate that was presented in the paper. Note that your article may address more than 1 question but I want you to focus on just 1 for this project.

Chosen Paper: Our chosen paper is *Relative age and birthplace effect in Japanese professional sports: a quantitative evaluation using a Bayesian hierarchical Poisson model* by Hideaki Ishigami (published in the Journal of Sports Sciences in 2016).

Background Motivation: 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 the relative age effect in athletes in North America and Europe. However this paper and our analysis extends this research to include Asian countries. This is an important step in geographically extending the scope of inference regarding the relative age effect. 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 effect is statistically significant, for example using a χ^2 test. Although it is beyond the scope of this paper, these results impact parents' decisions surrounding red shirting and informs public understanding of factors influencing success in athletics (particularly professional athletics).

Single Research Question: What is the relative age effect on soccer players in Japan between the ages of 23 and 25?

2. A brief description of the methods used by the authors of your paper to address the question of interest you are focusing on.

Data

The participants in the original study were professional baseball and soccer players, but we will focus only on professional soccer. The Japan Professional Football League (J. League) consists of 40 teams, representing a total of 1013 players registered in the 2012 season and we will focus on players between the ages of 23 and 25, for a total of 227.

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. Birth date was

treated as monthly data and because school year and competitive season both begin in April in Japan, relative age was coded as 0 (April) to 11 (March).

An athlete's birthplace is defined as the prefecture the player was born in. A prefecture is a first-order administrative district in Japan, and there are 47 prefectures in total.

The author's model includes the total number of male children born in each month as an offset term, which is a variable whose coefficient is fixed at one. The variable in the estimated equation is the total number of male births over the years when the sampled players were born.

Methods

Becoming a soccer player can be seen as an "event", and a number of them can be regarded as a "count of events". Thus, the author applied a Bayesian hierarchical Poisson regression model.

3. The full probability model used in the paper written out in mathematical notation. That is, all likelihood and prior components.

Likelihoood

Our data are denoted by y_i which represents the number of professional sports players. The subscript i indicates birth month. Our likelihood, or data generating mechanism, is described by the following distribution:

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

Priors

need to add stuff about priors are for alpha and beta Here we have 'priors' and below we have 'prior'

Our priors on λ_i are defined by the exponential relationship below where θ_i is the total number of men born over the years during the sample in month i, an offset term to make the months comparable since some months have different numbers of births. 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. The coefficient β measures the relative age effect (RAE). For example, April would be month 1, with $RA_1 = 0$. The exponential term gives the probability of becoming a professional male soccer player.

$$\lambda_i = \theta_i exp\left\{ \alpha + \beta RA_i \right\}$$

Prior

The authors found the posterior estimate for $\beta|y$ to be $\beta|y \sim \text{Normal}(-0.0934, 0.214^2)$ which we will use as our prior.

We did not see an updated posterior for α so we assume the prior in the paper, $\alpha \sim \text{Normal}(\mu, \sigma^2)$ where $\mu \sim \text{Normal}(0, 100^2)$ and $\sigma^2 \sim \text{Uniform}(0, 100)$.

Question for Katie: We are confused by the posteriors being on (what seems like) hyper-parameters. Are these two "layers" of hyperparameters because our initial model was a two layer hierarchical model? (i.e. random effects varying by month, i, and random effects varying by location, j, which we excluded). The "two layers" we're seeing are: hyperparameters α and β and (hyper)-hyperparameters μ and σ^2. Can you help clarify?