Relative age effect in elite soccer

Abigail Titzer, Yushan Zhao, Molly Thorbahn, Net Zhang

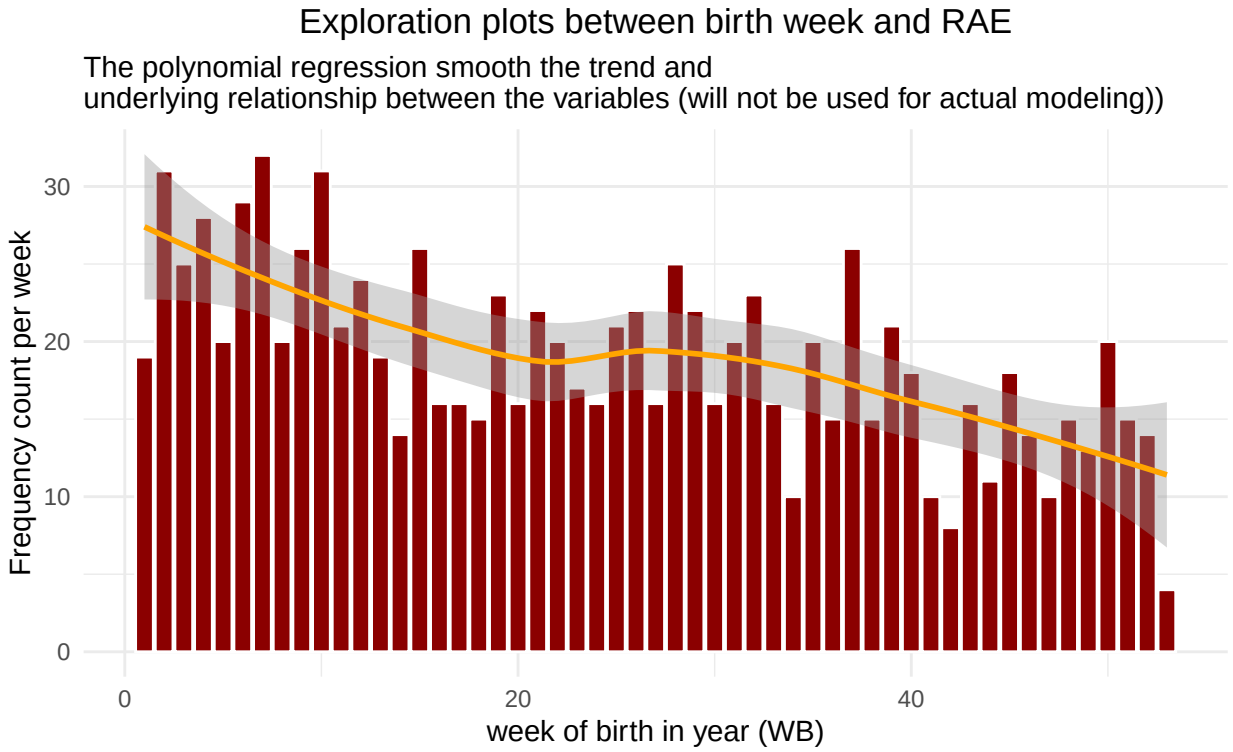
4/3/2021

I: Introduction

In the article, The relative age effect in European elite soccer: *A practical guide to Poisson regression modeling*, the authors explore the Relative Age Effect (RAE) and address frequent misuse in research involving it. To do so, the authors introduce the Poisson regression modeling method and apply it to Elite European soccer data. Through this, they can explore common sources for bias in RAE research and show how more data and knowledge can be extracted with the Poisson model.

The Relative Age Effect is a phenomenon that occurs widely within schools and sports teams. Children are grouped by age into cohorts, where those older in said groups are seen to have an advantage over the younger group members. Researchers contend that due to this, those that are older are given more positive regard and thus have higher levels of confidence. Combined with this and timed selectional bias, those older within the cohorts are more likely to succeed academically and athletically.

The original authors of this paper contend that those using RAE do not always get as much information from their data as possible, either by ending analyses too early or by comparing past results from the past that were analyzed differently. The solution to this issue would be to use the Poisson model more frequently in RAE research.



To demonstrate their solution, they use the Poisson model to analyze data from Big Five European soccer leagues. They use two indices to showcase RAE bias, Wastage, and Discrimination. The primary data is domestic players from the Big Five European leagues in the 2016/2017 season.

Their analysis found that the Poisson pooled model shows that those born at the end of the cohort year are less likely to become an elite big five European soccer player than those born at the beginning of the cohort year. The Discrimination index indicates that those born at the very beginning of the year are around twice more likely to become elite European players than those born at the very end of the year. The two indices of Wastage show that there are many members within cohorts that potential is not unveiled. The authors then add extensions to their regression by adding independent variables. Through this, they still found that RAE was the best explanation for their results and that there were no other significant factors. The results of this study widely support the results of earlier RAE research; however, the Poisson model more easily models multiple explanatory variables at once. They suggest that the Poisson model could be beneficial to further RAE research because of its ability to do so.

In this replication, we will model the Poisson regression with the same data from the Big Five European leagues' soccer clubs. We will compute the same indices, Wastage, and Discrimination, used in the study and calculate the RAE bias frequency-wise and value-wise. Through these, we will replicate the results that support previous RAE research and that RAE research can be more thoroughly investigated with Poisson regression.

II: Modeling

2.1 Methods

Our goal is to model the Relative Age Effecct (RAE) for frequency using the birth-week number (W_B). Here, the birth-week number (W_B) denote the week in which the player was born. For the sake of compatibility, we will transform W_B into time of birth, illustrating how far through the competation year a player's birthday is :

$$t_B = (W_B - 0.5)/52$$

In other words, we scaled the player's birth-week number into the interval of $(0, 1)$. The data collect the information on the 1000 top professional soccer players in the major leagues. We think it is reasonable to apply the poisson regression model here since the probability of having a successful professional athlete is quite low. The overall model is denoted as below:

$$Frequency = e^{\beta_0 + \beta_1 * t_B}$$

This Poisson regression model contains 204 football clubs with 6644 players of the "Big 5 European Leagues" from 2016 to 2017. Domestic players mean their nationality is the same as the league they play in. There are 882, 734, 609, 882, and 868 domestic players in England, France, Germany, Italy, and Spain, 3975 of 6644.

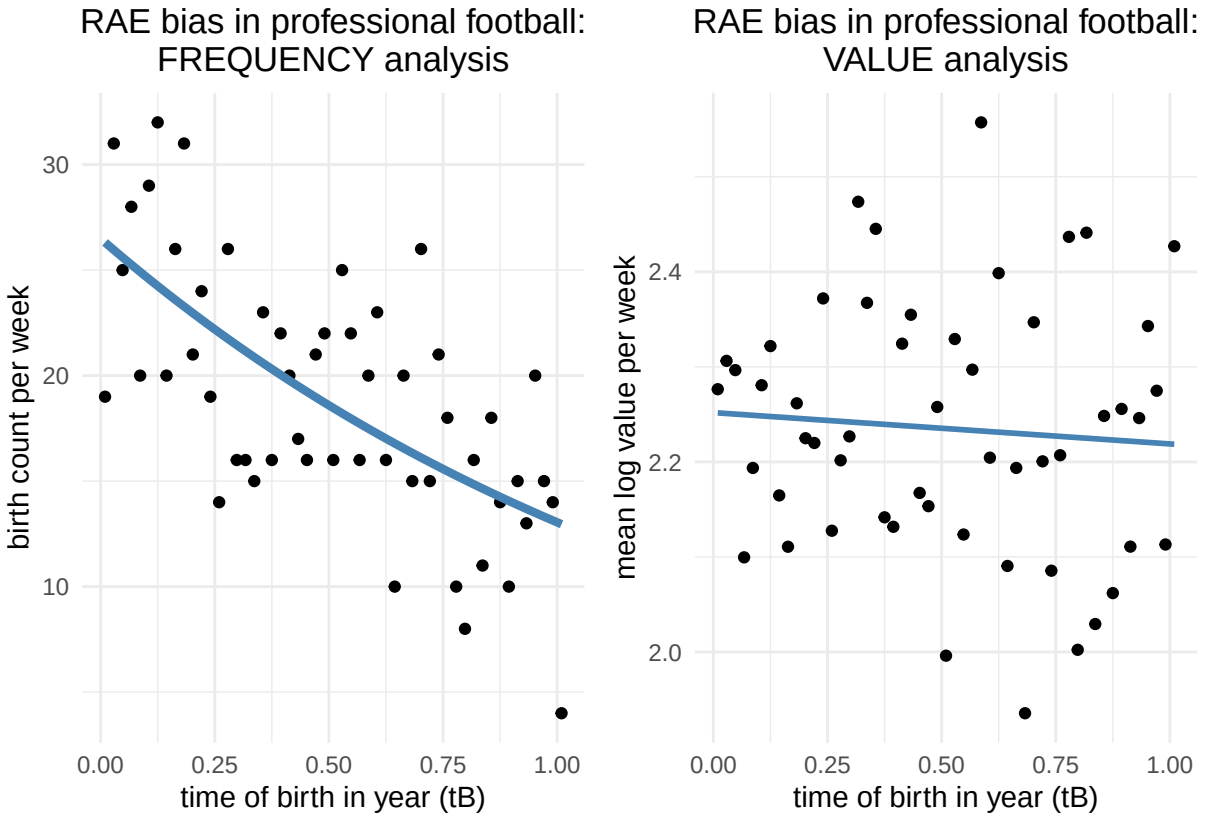
2.2 Findings

From our model output, we get the estimate values of $\hat{\beta_0} = 3.2768$ and $\hat{\beta_1} = -0.7083$. Therefore, our overall fitted model is denoted as below:

$$\hat{Frequency} = e^{3.2768 - 0.7083 * t_B}$$

Table 1: Table 1. Model results: Overall and country-by-country analysis

Country	beta_0	beta_1	AIC
Argentina	0.9480	-1.0332	100.1744
Netherlands	0.8256	-0.9554	72.6025
Turkey	0.2097	0.0817	45.6419
Belgium	0.6231	-0.7516	59.6048
Spain	1.3626	-0.9953	143.1613
Italy	0.9125	-0.7415	118.4851
Germany	0.6832	-0.1381	107.1367
Portugal	0.5031	-0.4505	57.9986
Brazil	0.9775	-0.2442	134.9844
Russia	0.5982	-0.7704	47.5641
France	0.8445	-0.3907	130.8024
all	3.2768	-0.7083	315.8524



From the plot above, we can conclude that the birth frequency per week gets smaller with time as the year grows since the slope regression line we fitted goes down and the slope coefficient for t_B is -15.8 (df=50, $p<0.01$, significantly smaller than zero). This means for per RAE, there are fewer later-born players than earlier-born players.

Also, we can find that players born early of the year have a more significant probability of becoming the top level of European soccer than players born at the end of the year from the blue line we fitted in the Poisson model. Returning to our model, consider two players born at the beginning of year and end of the year, where $t_{B1} = 0$ and $t_{B2} = 1$. From the model formula, we get the expect player frequencies are 16.4 and 6.8, take their ratio, these illustrate players who born at the start of the year has almost 2.4 times more likely to become top-level soccer than those born at the end of the year.

Morever, from Table 1 we can see that country Turkey and Russia has the best model fit with the lowest AIC around 45 and 47 respectively. The overall model has the AIC of 315.8524 which is not super ideal. From the table output, we conclude that players from Spain, Belgium, Argentina Brazil, France contribute the major variance of our model. And we are planning on addressing this in the last commentary section.

III: Commentary & Future work

We think that the methods used in the paper "The relative age effect in European elite soccer: A practical guide to Poisson regression modeling" were the correct choices because the Poisson Regression Model is ideal for predicting a response using one or more explanatory variables. This article is dedicated to exploring the Relative Age Effect (RAE), which looks very in depth at age, birth date, and other age-related factors to determine how a child will succeed academically and athletically. The nature of the 'problem' being researched aligns very well with what the model was designed to do.

One of the (often violated) assumptions in Poisson Models in general are that the mean and variance are equal. While this study did violate this assumption, it was dealt with by using a dispersion parameter/coefficient. This parameter/coefficient is an aggregated ratio of the variance divided by the mean. In general, if this number is greater than 1 by a sufficient amount, the likelihood is either that variables are missing or the assumptions of the Poisson model are not being met. Thankfully, the dispersion parameter/coefficient of this study was 0.985, very close to the ideal of 1. The study did not check the independence assumption between the observations, but we think it is safe to assume that the players' birth is independent.

Overall, we think the article did a decent job applying the Poisson regression model fitting the players' Relative Age Effect (RAE). They didn't check the independence assumption between the observations, but we think it is safe to assume that the players' birth is independent. Also, we agree with choosing the Poisson distribution to map the frequency of the player's birth. Since these are elite professional athletes with great talent and dedicate work ethic, it is scarce to happen in the city.

We didn't reproduce the paper's exact results, but our conclusions towards the problem are the same. There are minor differences between the estimated parameters of our models and the article's results, but the sign of the parameters are all the same. We had a slighter difference with the author because we handle the t_B in the model more carefully. Since the package `lubridate` from R provides excellent functionality for dealing with date objects, we believe our calculations are more precise than the authors because we take into account the difference in total days within a year is fluctuant.

Lastly, we think it is good to add the player's nationality into the Poisson regression model as a categorical predictor. We could certainly reduce the categories down to 4 or 5 and make the rest of the countries as others. By doing this, we could provide an estimation of selected countries more easily, and the AIC of the overall model could be reduced significantly.