Title:
Counting across borders: a unified model for the effect of league and age on player output

Opta experience:
I have worked with Opta's XML data before

Area of study:
We know that all leagues are not equal. Football varies from competition to competition in both style and substance; however, in public analytics, little work has been done to effectively quantify these differences. Understanding and quantifying these differences can help lighten the load on decision makers, and allow human judgement and intuition to be deployed more effectively.

Understanding the effect of different factors on player output is critical to accurately evaluating player performance. Whether that is team performance in different competitions and different leagues, or in assessing prospective transfer targets, accounting for footballing context is key. Intuitively, we know that scoring goals in the Norwegian Eliteserien is less impressive than scoring in the Premier League; however, this is rarely quantified. Likewise, the same patterns of thought are less frequently applied to other domains. Statistics like tackles made or dribbles won are less clearly linked to league quality and are therefore harder to reason about without data.

Clearly it is not wise to go off numbers alone, although being able to translate player performance from one league to another helps us prevent the misuse of statistics, and deliver more accurate information to those who need it.

Method

I believe this question lends itself well to Bayesian analysis. Using historical statistical performance of players, we can estimate the effect of factors like age, league and position on player output, as well as the underlying ability of the players themselves. A model such as this is well suited to the modelling framework Stan, which allows us to estimate even very complex models via Markov Chain Monte Carlo sampling.

The model I propose is relatively simple, but potentially very powerful in the questions it allows us to answer. Let us suggest that a player's rate of a given action on the football pitch per 90 minutes played (for instance, goals per 90) is the product of the player's ability, the effect of the league, the effect of their age, and the effect of their position. With appropriate priors and constraints on these parameters, we can estimate the effect of each factor in turn. And advantage of a flexible modelling framework like Stan is that we can add extra sophistication, like allowing the factors to change over time. In other words, we can see how the age curve and league effects vary over the past few seasons, as well as how players have performed vs their age cohort.

This model is also relatively straightforward to validate. We can test variants of the model with certain factors excluded against each other in a cross validation, as well as testing against simpler models such as the assumption that output is constant across leagues.