Football Analytics

Alexandre Andorra
PyMC Labs

Maximilian Göbel
Bocconi University

First Draft: April 16, 2024
This Draft: April 24, 2024

Abstract

.

Acknowledgments: For helpful comments we thank, without implicating, ...

Key words: football analytics, Bayesian statistics

JEL codes:

1 Introduction

2 Extracting a Player's "Skill"

What is a player's innate skill, respectively ability (PIA)? This is a crucial question any recruiter, team manager or head coach has to find an answer to. The big problem is: PIA is latent, i.e. it is unobserved. One can only observe a player's *performance*, which is a convoluted variable, contaminated by many other variables, mostly originating from strength of his team mates and the strength differential to the opponent's team. Hence, assessing PIA adequately, is a skill of its own, which requires experience and knowledge.

We propose a data-driven supplement to the above described discretionary approach. Our model is based on the above described assumption, that in order to assess PIA, we need to filter out team effort from a player's observed performance. In particular, our model reads as follows:

$$Y_{i,t} = f\left(\alpha_i + g\left(\mathbf{X}_{i,t}\right)\right) \ , \tag{1}$$

where $Y_{i,t}$ is a player's observed performance, $\mathbf{X}_{i,t}$ is a vector of team performance indicators, which may include indicators measuring the performance delta between the player's team and the opponent to be faced in match t. $g\left(\cdot\right)$ is a function that combines the individual features of $\mathbf{X}_{i,t}$ in potentially any kind of linear or non-linear manner.

Our key variable of interest however, is α, which represents that structural part of $Y_{i,t}$ that cannot be explained by the covariates, $g\left(\mathbf{X}_{i,t}\right)$.

$f\left(\cdot\right)$ is again some kind of transformation function that depends on the nature of $Y_{i,t}$. If $Y_{i,t}$ is for example a binary variable, $f\left(\cdot\right)$ is a sigmoid function, whereas if $Y_{i,t}$ is categorically distributed with multiple discrete classes, $f\left(\cdot\right)$ is a softmax function.

Now, what is $Y_{i,t}$ in particular? It can be any kind of observed performance metric. Yet, different positions within the team probably require different performance metrics. For a forward, $Y_{i,t}$ might be a binary variable $Y_{i,t} \in \{0, 1\}$ indicating whether a forward has scored (at least) one goal in match t or not. It could however also be a player's total number of goals scored in a match t. For a defender, $Y_{i,t}$ might be the number of shots (on target) admitted or the number of duels won. A midfielder might be measured against the number of assists given and the goalkeeper might be evaluated against the number of attempts denied.

Different $Y_{i,t}$ may then also require different covariates $\mathbf{X}_{i,t}$. While $Y_{i,t}$ being the probability of scoring a goal, it may suffice to include only aggregate team statistics, it may be necessary to include individual team-mate statistics when it comes to $Y_{i,t}$ being the number of shots (on target) admitted by a defender.

Hence, data availability is crucial. Yet, our model is flexible enough to work with widely available aggregate team statistics and proprietary measurements.

We showcase this with data sourced from two different data sets: first, we use data on every

Premier League match[1] since its inception in 1992/93 to 2020/21. This data, though being very comprehensive, is scarcely populated with player-specific information. Still, it serves to construct aggregate team statistics and allows to observe the scoring board and corresponding strikers. Our second data set is less comprehensive in the time dimension, but comes with much more detailed player specific data.[2] This data set allows us to for example observe duels won, the outcome of dribblings, interceptions, passes to assist shots, the reason for substitution (e.g. injury), or the result of a goal attempt.

3 Predicting a Player's Susceptibility to Injury

The La Liga data also allows us to model a player's probability of injury during a match.[3] Again, we can resort to Equation (1). Our dependent variable is a binary indicator, leveraging the data field $\texttt{substitution}_{\texttt{o}}\texttt{utcome}['\texttt{injury}']$. The covariates $\mathbf{X}_{i,t}$ in this analysis are potentially of much higher interest, and may include many more player-specific variables than the exercise in Section 2. Some variables of interest may include number of duels, differentiated by duel type ($\texttt{duel}_{\texttt{t}}\texttt{yp}$), number of dribblings ($\texttt{dribble}_{\texttt{o}}\texttt{utcome}$), or number of injuries in the past.

Now, there are other metrics that might be important, but that we do not observe, such as proxies for fatigue (e.g. distance covered during match), nutrition, or lifestyle. One can argue that these unobserved factors are actually *confounders*, i.e. variables that affect both $Y_{i,t}$ and $\mathbf{X}_{i,t}$. In such a case, if someone were interested in the causal effect of any of the covariates ($\mathbf{X}_{i,t}$), caution is warranted, as inference on the structural parameters of $\mathbf{X}_{i,t}$ are invalid in the presence of *confounders.*

α_i is now not to be interpreted as PIA, but rather as the general predisposition of a player to injury.

4 Performance Above Replacement

In baseball, a frequently used measure of player performance is the concept of *wins above replacement* (WAR) (see e.g. Baumer et al. (2015)). We adapt this framework for evaluating the performance of football players, and propose a new concepts: *performance above replacement* (PAR).

The concept of PAR rests on a simple regression model, in which the observed performance of a player of interest (Y_i) is compared to his *expected* performance ($\hat{Y}_i$). The expected performance, in turn, is inferred from the performance of a reference group (R_i), where $i \notin R_i$.

We thus regress the observed performance ($Y_{j,t}$) of each player ($j \neq i$) in the reference group on some set of covariates ($\mathbf{X}_{j,t};$). The model thus reads as follows:

$$Y_{j,t} = \alpha_j + g\left(\mathbf{X}_{j,t}; \theta\right) + \varepsilon_j \ , \quad \text{for} j \in R_i \tag{2}$$

[1] See: kaggle.com

[2] See: github.com/statsbomb/

[3] Of course, injuries also often happen during training, which is why we stress the focus on *injury during a match here.*

where $Y_{j,t}$ is a player's observed performance, $\mathbf{X}_{j,t}$ is a vector of covariates. These can for example include team performance indicators, measuring the performance delta between the player's team and the opponent to be faced in match t, or player-specific features. $g(\cdot)$ is a function that combines the features of $\mathbf{X}_{j,t}$ in potentially any kind of linear or non-linear manner. In case of a linear function, the vector θ collects the coefficients on each feature in $\mathbf{X}_{j,t}$. But $g(\cdot)$ may also be a tree-based algorithm or even a deep-learning model in which case θ collects the corresponding tree splitting-rules, respectively network weights.

α_j represents that part of $Y_{j,t}$ that can neither be explained by the covariates $(g(\mathbf{X}_{j,t}))$ nor by random fluctuations, e.g. luck, such as a goal being scored because of a deflection $(\varepsilon_{j,t} \sim N\left(0, \sigma^2\right))$.

Here, we are interested in the vector of coefficients/weights θ, which we use to determine the difference between player i's observed performance $(Y_{i,t})$ and his expected performance $(\hat{Y}_{i,t})$:

$$\delta_i = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(Y_{i,t} - \hat{Y}_{i,t}\right)^2} \ , \tag{3}$$

where $\hat{Y}_{i,t} = \mathbf{X}_{i,t}\,\theta'$. Accordingly, if $\delta_{i,t} > 0$, player i's performance is better than expected, while the opposite holds for $\delta_{i,t} < 0$.

Player i's performance above replacement (PAR_i) is then measured as:

$$PAR_i = \Pr\left(\boldsymbol{\Delta}_i \leq \delta_i\right) \ , \tag{4}$$

where $\boldsymbol{\Delta}_i$ is the collection of all $\delta_{j \neq i}$. In essence, PAR_i is player i's position within the empirical distribution of how much the performance of each player in the reference group (R_i) deviates from his expected performance.

Comparing Apples to Apples. Now, the observed performance Y could in principle be anything from shots on target via duels won to completed passes into the opponent's third of pitch. Yet, for the estimates to be representative of the "true" value, one should be very considerate when constructing player i's reference group, R_i. Simply shuffling players regardless of their position into reference group R_i, if i is a striker and the performance metric Y is shots on target, might lead to estimates with large uncertainties.

Nonetheless, there are two ways out of this misery: either one constructs for example a position-specific reference group, or one opts for a multilevel modeling approach. The latter is implemented by making certain model parameters, such as α_j in Equation (2) vary by position, as in Yurko et al. (2019). If the dependent variable $(Y_{j,t})$ is for example the number of shots on target, one could make α_j follow a binomial distribution with position-specific parameters, such that:

$$\alpha_{j,z} \sim B\left(n_z, p_z\right) \qquad \text{for } z = \{\text{goalkeeper, defender, ..., left-winger, striker}\} \ .$$

The reason why a normal distribution would be a poor choice, comes from the fact that that the dependent variable lives in $Y_{j,t} \in \mathbb{Z}^{0+}$. Hence, we need a prior distribution that is discrete and lives within the positive domain including zero, i.e. with support $[0, +\infty]$.

However, and in line with Yurko et al. (2019), when opting for such a multilevel modeling approach, estimating PAR_i in Equation (4), requires the reference group $\boldsymbol{\Delta}_i$ to only include *reasonable alternatives* to player i, e.g. players with a similar field position.

References

Baumer, B. S., Jensen, S. T., and Matthews, G. J. (2015). openWAR: An Open Source System for Evaluating Overall Player Performance in Major League Baseball. *Journal of Quantitative Analysis in Sports*, 11(2):69–84.

Yurko, R., Ventura, S., and Horowitz, M. (2019). nflWAR: A Reproducible Method for Offensive Player Evaluation in Football. *Journal of Quantitative Analysis in Sports*, 15(3):163–183.