Statistics in the football analytics: a composite model for the probability of scoring a goal

Danny Paganin [1] and Bruno Scarpa [2]

Statistical models are becoming more and more popular in the football analytics thanks to the availability of massive data that captures all the events generated during a match. In particular, through computer vision it is possible to retrieve "tracking data" (historical series of movements of each player) and "on the ball data" (every touch of the ball is recorded). In this work we analyze all the shots, which are part the "on the ball data", in the five major European championships of the one season 2020-21. The objective of this analysis is the study of the probability of scoring a goal. We use two random random variables to calculate the probability: Y_i which is equal to 1 if the shot i hits the goal, 0 otherwise; Z_i, is equal to 1 if the shot i turns into a score. The purpose of the analysis is to study $Pr(Z_i)$, which is defined as the product between the probability that a shot hits the goal, $Pr(Y_i)$ and the probability that the shot turns into a score since it is on goal, $Pr(Z_i|Y_i = 1)$. We compare inferential results and predictive performance of a joint model versus separate model for the two probability. Note that we considered only the shots that hit the goal for the study of the conditional probability, because the event $(Z_i = 0|Y_i = 0)$ is impossible to occur, as it is not possible to score a goal without hitting the goal. We study then two probabilities simultaneously through the use of composite likelihood (Lindsay, 1988), a pseudo-likelihood and compare results with separate models. Thanks to the composite likelihood, we can assume the same effect for the variables shared by the two probabilities, obtaining an increase in precision on the

[1]Department of Statistical Sciences, University of Padua, Via C. Battisti 241, Padova, Italy, paganindanny12@gmail.com

[2]Department of Statistical Sciences, University of Padua, Via C. Battisti 241, Padova, Italy, scarpa@stat.unipd.it

parameter estimates of the shared variables.

References

B. G. Lindsay. Composite likelihood methods. *Contemporary mathematics*, 80(1):221–239, 1988.