Racism in Soccer

Authors: G. Christensen[1]*

Affiliations

[1]Swarthmore College.

*Correspondence to: gchrist1@swarthmore.edu

Abstract

Data from major European soccer leagues gives inconclusive evidence about racism in the form of redcards being given to players with skin of a darker tone. Players with darker skin do receive more redcards on average, even after controlling for a handful of player characteristics such as position. However, the difference is quite small, and the inability to adequately control for potentially omitted variables gives me little confidence that the observed differences are real, i.e., that the relationship is causal and darker skinned players receive more redcards because they have darker skin, instead of for some other unrelated reason.

One Sentence Summary

Players with darker skin receive slightly more redcards than players of lighter skin, but this correlation should be viewed with skepticism and likely not given a causal interpretation.

Results

A simple linear regression of redcards on skintone rating shows that players with darker skin tone are more likely to have received a redcard. (Very few players received more than one redcard from the same referee, so I simplified redcards to a binary variable.) Going from a completely light-skinned player (0 on a scale from 0 to 1) to completely dark skinned (1 on a scale of 0 to 1) increases the probability of a player receiving a redcard from a given referee by 0.3 percentage points. (The coefficient from a bivariate linear probability model with standard errors clustered by player is 0.003, standard error 0.001). A player with the lightest skin has a 1.17% chance of getting a redcard from a given referee, so in percentage terms the increase may seem fairly large.

However, obvious caution should be taken when interpreting these results. Players of certain skin tone may be statistically more (or less) likely to play a certain position, play in a certain league, or play in more games, and players with those characteristics may be more likely to receive a redcard, and if we do not control for these factors our estimates will be biased. The dataset allows us to control for a handful of these potential omitted variables, such as player position, height, weight, games played under each referee, goals scored, and victories. Linear probability models that control for these available variables indicate there is a 0.3 percentage point (coefficient .0034, standard error .0014, p-value 0.013) increase in the likelihood of receiving a redcard for a dark skinned (1) player compared to a light skinned (0) player. A logit model gives similar results, indicating an odds ratio of 1.30, standard error 0.14, p-value 0.015.

Unfortunately, the data does not include the name or nationality of the league in which each of the player-referee interactions took place. The leagues included in the data are the top leagues in four European countries, and a league variable is indeed in the dataset, but this only refers to the league of the first league in which a player appeared, and players frequently play in different leagues. The lead researchers on the paper have suggested that controlling for league with fixed effects is thus inadequate to address potential omitted variable bias arising from different leagues. The suggestion was made to instead control for referee country of origin. While this eliminates the potential for omitted variable bias from referees of certain national origins giving more (or fewer) redcards while also refereeing players with darker (or lighter) skin, I believe this still does not solve the problem of omitted variable bias by league. It also introduces its own small additional problem—99 of the 161 countries of referee origin (4% of the player-referee interaction observations) gave no redcards, so nonlinear models such as logit and probit regressions that attempt to condition on referee country of origin drop these observations by necessity. The dataset contains 146,028 player-referee pair observations, but in all regressions, I drop observations with missing values, which typically results in 115,603 observations being used.

Initial Approach

My initial approach was to control for player position, height, weight, games played under each referee, goals scored, and victories, as well as league of play, and to cluster standard errors by player. I investigated numerous different measurement of skin tone, as there were two independent measurements, but I noticed no significant differences in outcomes using minimum, maximum, or only the first or second rating, so I ultimately report only the average of the two ratings. I used both linear probability models and a logistic regression model. In this economist's experience, these models have rarely given significantly different results, so I prefer the linear model, despite its obvious shortcomings, for its ease of interpretation, but always also run a nonlinear logistic or probit regression as a robustness check. As requested, I have reported the logistic results as odds ratios, though I personally believe these to be often misunderstood. The linear model gives a coefficient of 0.0026 (standard error 0.0015, p-value 0.073). The logit model gives an odds ratio of 1.25 (standard error 0.147, p-value 0.057).

For both the question of implicit and explicit racism by referee country of origin, I simply interacted the racism variable with the skin tone score and repeated the same linear and logistic regression. Neither of these regressions shows anything remotely significant. Referees from more explicitly or implicitly racist countries do not appear to give more redcards to soccer players.

Final Approach

My final approach is to simply replace the league fixed effects with fixed effects for referee country of origin, and again cluster standard errors by player. I believe that standard errors should almost certainly be clustered in two non-nested dimensions: by player and by referee. User-written Stata code is available to implement simple versions of this procedure (see http://www.kellogg.northwestern.edu/faculty/petersen/htm/papers/se/se_programming.htm, I learned of this code through the anonymous feedback procedure that was part of the crowdstorming aspect of this paper) but the functionality is limited, so I do not implement this in my final approach—initial tests indicate that standard errors change only slightly when the multiple dimensions of clustering are properly accounted for. I again run both a linear probability model and a logistic regression, and report the logistic regression in terms of odds ratios, as per request. The linear model gives a coefficient of 0.0032 (standard error 0.0014, p-value 0.024). The logit model gives an odds ratio of 1.341 (standard error 0.135, p-value 0.004).

For the questions of implicit and explicit racism by referee country of origin, fixed effects for referee country of origin cannot be used, as they do not vary with the country-level racism scores. Thus I simply control for neither league of play nor country of referee origin. Of the four estimates for implicit and for explicit racism, using linear and logistic regression, only one of the interaction coefficients is statistically significant at standard levels, with a p-value of 0.081. I do not have much confidence in these estimates.

Conclusion

I do not draw strong conclusions from my analysis because I believe the dataset has serious limitations. Put less strongly, I myself haven't been able to think of a method to overcome what I believe are limitations in the data. To the extent I am able to draw a conclusion from the data, it is that there may be some small amount of prejudice against players with darker skin in major European soccer league refereeing (an odds ratio of receiving a redcard of 1.3 for a player of very dark skin compared to a very light skinned player, or an increase of 0.3 percentage points off a basis of 1.2%). This should be interpreted as an observed correlation and not as a causal relationship. I find a relationship between neither implicit nor explicit prejudice in referee country of origin and referee redcard behavior.