Seven Outliers Produce the False Impression of Skin Tone Bias in Soccer Referees: A Bayesian Logistic Regression Analysis

Authors:

Richard D. Morey*[1] , Eric-Jan Wagenmakers[†2]

Affiliations

[1] University of Groningen.

[2] University of Amsterdam.

*Correspondence to: richarddmorey@gmail.com.

†This work was supported by an ERC grant from the European Research Council.

Abstract

Are soccer referees more likely to give red cards to dark skin toned players than to light skin toned players? We addressed this question using Bayesian logistic regression, modeling a player's red-card record as a function of seven predictors. Referee strictness and player aggression were important nuisance/control predictors. The key inference concerns the player skin tone predictor. Initial analysis showed that for this predictor, the 95% credible interval does not overlap with zero; furthermore, the inclusion of the skin tone predictor is supported by the Bayes factor. These results seem to reinforce the hypothesis that soccer referees are biased against players with dark skin tone. However, closer inspection revealed that the data are contaminated by seven players who collect red cards at a disproportionally high rate. Removing these seven outliers –0.3% of the complete data set– eliminates the bias effect entirely.

The abstract should be about 100-150 words.

One Sentence Summary

After removing seven outliers –0.3% of the complete data set– a Bayesian logistic regression model no longer revealed any evidence for the assertion that soccer referees are more likely to give red cards to players with darker skin tone.

Results

Our analysis aims to address the two crowdstorm research questions:

1. "Are soccer referees more likely to give red cards to dark skin toned players than light skin toned players?"
2. "Are soccer referees from countries high in skin-tone prejudice more likely to award red cards to dark skin toned players?"

Data Preprocessing

The mean IAT score and the mean explicit bias score were z-transformed before they were used as a predictor. The skin tone scores from the two raters were first z-transformed and then averaged, so that these ratings are represented by only a single predictor.

Before proceeding, we excluded cases for which the IAT score was unknown. Importantly, we did not exclude cases for which skin color was unknown. These data are still useful as they provide information on referee strictness.

Model Construction: General

Our initial analysis was a probit regression on the probability of getting a red card; in our final analysis we changed this to a logistic regression, in order to be able to report the odds ratio of the effect. Below we discuss the result of the logistic regression.

The logistic regression modeled the probability of getting a red card as a function of the following predictors:

1. Player aggression (playerEffect);
2. Referee strictness (refEffect);
3. IAT score for the referee's country (IAT);
4. Explicit bias score for the referee's country (Ebias);
5. Player skin tone (ratingVals). This was quantified by the average of the z-scores for the two raters.
6. Interaction between IAT score and player skin tone.
7. Interaction between explicit bias score and player skin tone.

We decided to leave out league country as a predictor, because any differences due to this predictor will be accounted for by player aggression and referee strictness. Other possible predictors were deemed superfluous or uninformative and were not added to the regression equation. For instance, player position is not all that informative after player aggression has already been included in the model. Predictors 3 and 4 are included to keep the interaction terms (i.e., predictors 6 and 7) interpretable. As we will detail below, the research questions can be addressed by assessing the importance of predictors 5, 6, and 7.

Note that predictor 1 (player aggression) and predictor 2 (referee strictness) were not available directly but were estimated from the data set; that is, strict referees hand out many red cards, and aggressive players tend to collect many red cards. These are important nuisance/control predictors, and an appropriate assessment of bias may require that these predictors are present in the regression model.

Model Construction: Specific

For each player-referee dyad, we modeled the logit of the probability of receiving a red card. In terms of the BUGS language, for each dyad *i* we have **cards[i] ~ dbin(p[i],games[i])**, and we model **logit(p[i])** as a function of seven predictors.

For each of *k* referees, we assume that they have a level of strictness governed by a group-level normal distribution centered at zero: **refEffect[k] ~ dnorm(0,precRef)**. Similarly, for each of *j* players, we assume that they have a level of aggression which is also governed by a group-level normal distribution: **playerEffect[j] ~ dnorm(mu0,precPlayer)**. Note that the **mu0** parameter sets the overall baseline for a player receiving a red card. The second term in the dnorm specification indicates the precision (i.e., 1/variance) and it is assigned a **dgamma(1,1)** prior. Parameter **mu0** is assigned a more informative prior: **mu0 ~ dnorm(-2,1)** prior. Although somewhat informative, these priors do not affect the key comparisons of interest.

The impact of the remaining five predictors is quantified by their beta regression coefficients or slopes. These five predictors are **slpIAT**, **slpEbias**, **slpColor**, **slpColIATint**, and **slpColEbiasint**. For simplicity, each predictor is assigned an independent standard normal prior (e.g., **slpIAT ~ dnorm(0,sd=x)**). In our first analysis, we used a relatively wide, uninformative prior with sd = 1; in a second analysis, we studied the robustness of our conclusions and used a more peaked prior with sd = 0.2. As outlined below, the critical test concerns the effects of **slpColor**, **slpColIATint**, and **slpColEbiasint**.

Inference: General

The first research question, “Are soccer referees more likely to give red cards to dark skin toned players than light skin toned players?”, can be assessed by quantifying the importance of predictor 5, player skin tone.

The second research question, “Are soccer referees from countries high in skin-tone prejudice more likely to award red cards to dark skin toned players?”, can be assessed by quantifying by the importance of predictors 6 and 7 (interaction between implicit and explicit bias and player skin tone).

To quantify the importance of predictors 5, 6, and 7, we compare the full regression model (that includes all of the predictors) to three simplified regression models: (1) model A has all predictors except 5; (2) model B has all predictors except 6; (3) model C has all predictors except 7. Whenever the full model outperforms its simplified versions, this constitutes evidence for the inclusion of the associated predictor.

We fit the full regression model to the data using JAGS code provided on the OSF website. Parameter estimation results can be gauged by plotting the posterior distributions for the relevant regression coefficients. Each posterior distribution can be summarized by its mean and a 95% credible interval. For hypothesis testing, we compute three Bayes factors based on the comparison between the full model against each of three simplified models that omit one of the three key predictors.

Because the simple models are nested in the full model, we can compute the Bayes factor using the Savage-Dickey density ratio; that is, we compute the ratio between the prior and the posterior ordinate at the value under test (i.e., beta = 0). This also allows a visual representation of the test outcomes.

Inference: Specific

We fit the full regression model to the data using JAGS routines for Markov chain Monte Carlo (MCMC). We used 10,000 iterations in one chain. Visual inspection revealed fast convergence, so no burn-in was used. In our first analysis, we used relatively wide, uninformative prior distributions on the key predictors (i.e., N(0,sd=1)). A step-by-step analysis is provided in the accompanying html file “AllPlayers.html”.

Figure 1 shows the prior (red) and posterior (black) distribution for the effect of skin tone (i.e., predictor 5). As can be seen, the posterior is substantially away from 0 (95% credible interval: [0.037, 0.159]). Nevertheless, the Bayes factor for its inclusion equals a relatively modest 4.6. If the prior distribution would have assigned mass to only positive values (implementing a one-sided test), this Bayes factor would increase to about 9. The posterior distribution and the Bayes factor provide a summary of the evidence for including the skin tone predictor. To satisfy the requirements of this crowdstorm project we also derived an odds ratio and its 95% credible interval: 1.394 [1.135, 1.714].

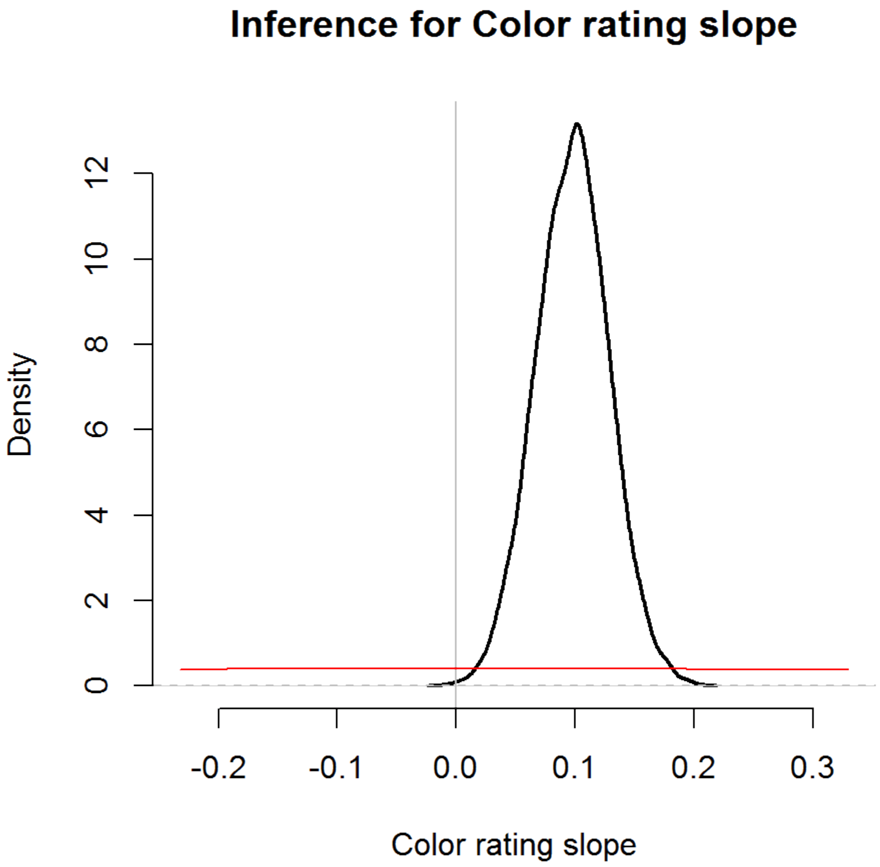

Figure 1. Posterior distribution for the regression coefficient on the skin tone predictor.

Note that for an "average" player/referee combination, where other covariates are 0, every standard deviation increase in standardized skin tone rating results in an increase in probability of receiving a red card of .00035 [.00013, .00059]. This is a rather modest increase. Figure 2 shows the relation between skin tone rating and the probability of getting a red card, confirming the impression that the effect, if present, is small.

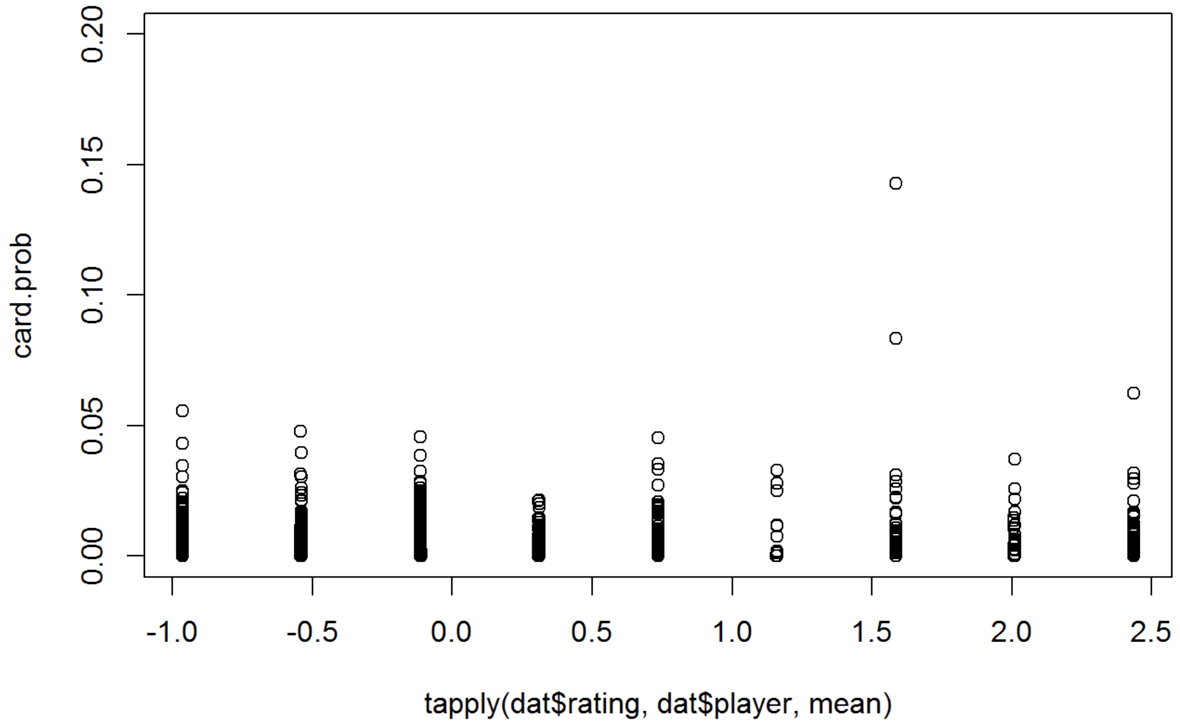

Figure 2. The observed proportion of matches in which a player received a red card, separately for each skin-tone rating.

Figure 3 shows the prior and posterior distribution for the interaction between IAT score and skin tone (i.e., predictor 6). As can be seen, the posterior is relatively close to zero, with most mass on negative values (95% credible interval: [-0.131, 0.049]). The Bayes factor against inclusion of the predictor is 15.0. To satisfy the demands of this crowdstorm project we also derived an odds ratio and its 95% credible interval: 0.806 [0.493, 1.302].

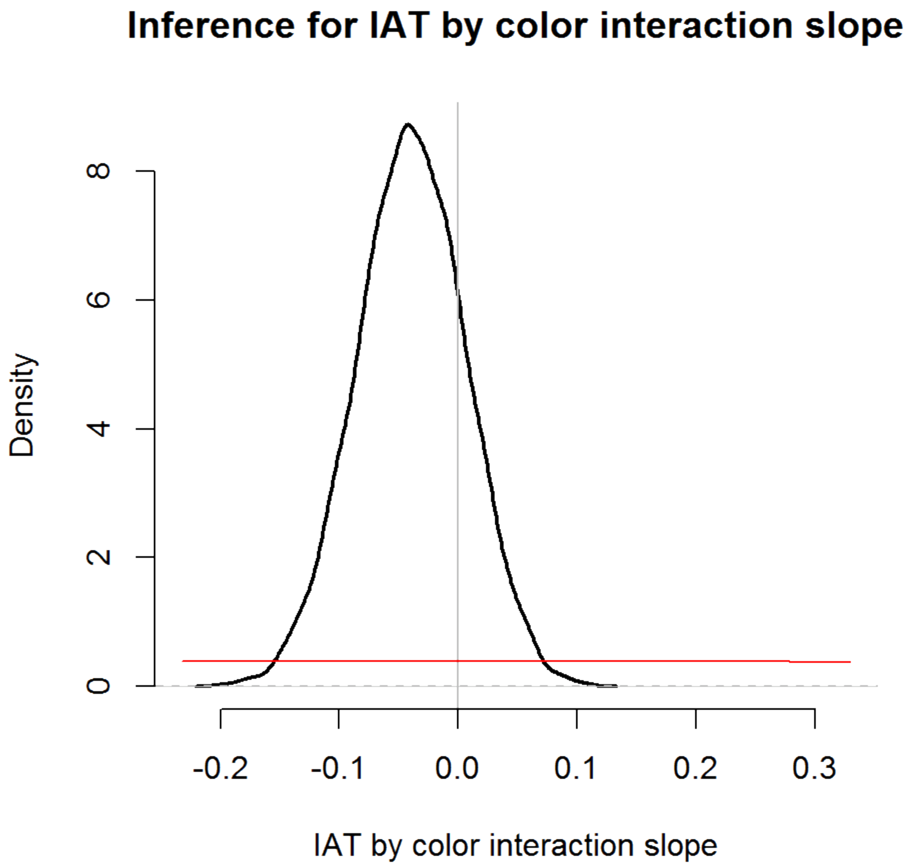

Figure 3. Posterior distribution for the regression coefficient on the interaction between IAT score and skin tone.

Figure 4 shows the prior and posterior distribution for the interaction between explicit bias score and skin tone (i.e., predictor 7). As can be seen, the posterior is relatively close to zero, with most mass on positive values (95% credible interval: [-0.051, 0.125]). The Bayes factor against inclusion of the predictor is 15.2. To satisfy the demands of this crowdstorm project we also derived an odds ratio and its 95% credible interval: 1.316 [0.701, 2.395].

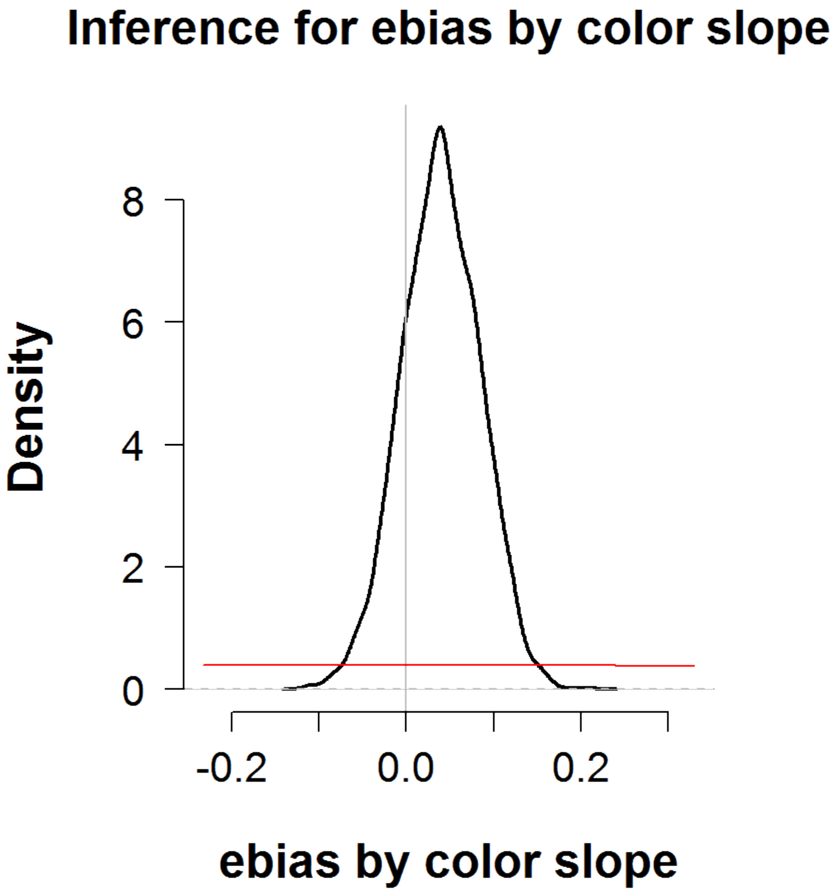

Figure 4. Posterior distribution for the regression coefficient on the interaction between explicit bias score and skin tone.

In sum, our first analysis provided some evidence in favor of a skin tone bias, and evidence against the importance of implicit and explicit bias measures. However, the specification of our prior distribution was done without the help of much substantive knowledge, and after looking at the data (!) it became clear to us that our prior may have been too wide. Consequently, below we explore the robustness of our conclusions against a more narrow prior.

Sensitivity Analysis: Exploring the Effect of A More Peaked Prior

In our second analysis we changed the prior distribution on the three crucial slopes to be N(0, sd=.2) instead of N(0, sd=1).

The parameter estimation results are similar to those obtained with the wide prior. For the effect of skin tone, the 95% credible interval equals [0.035, 0.152] ; for the interaction between skin tone and IAT, the 95% credible interval equals [-0.122, 0.051]; finally, for the interaction between skin tone and explicit bias, the 95% credible interval equals [-0.051, 0.118].

In contrast to the parameter estimation results, the Bayes factor results are affected by narrowing the prior. Specifically, the Bayes factor for the inclusion of the skin tone predictor now equals 19.9. If the prior was one-sided, this would equal approximately 39. Thus, although the effect remains small, the Bayes factor does support its presence. The narrow priors also decrease the Bayes factor support against the inclusion of the two bias measures, from about 15 to 3.4.

Inference After Eliminating Outliers: Skin Tone Bias Disappears

Inspection of Figure 2 suggests that some of the effect of skin tone may be due to the presence of outliers, that is, players who are disproportionally likely to receive red cards. Figure 5 shows the histogram of the proportion of matches in which players received a red card:

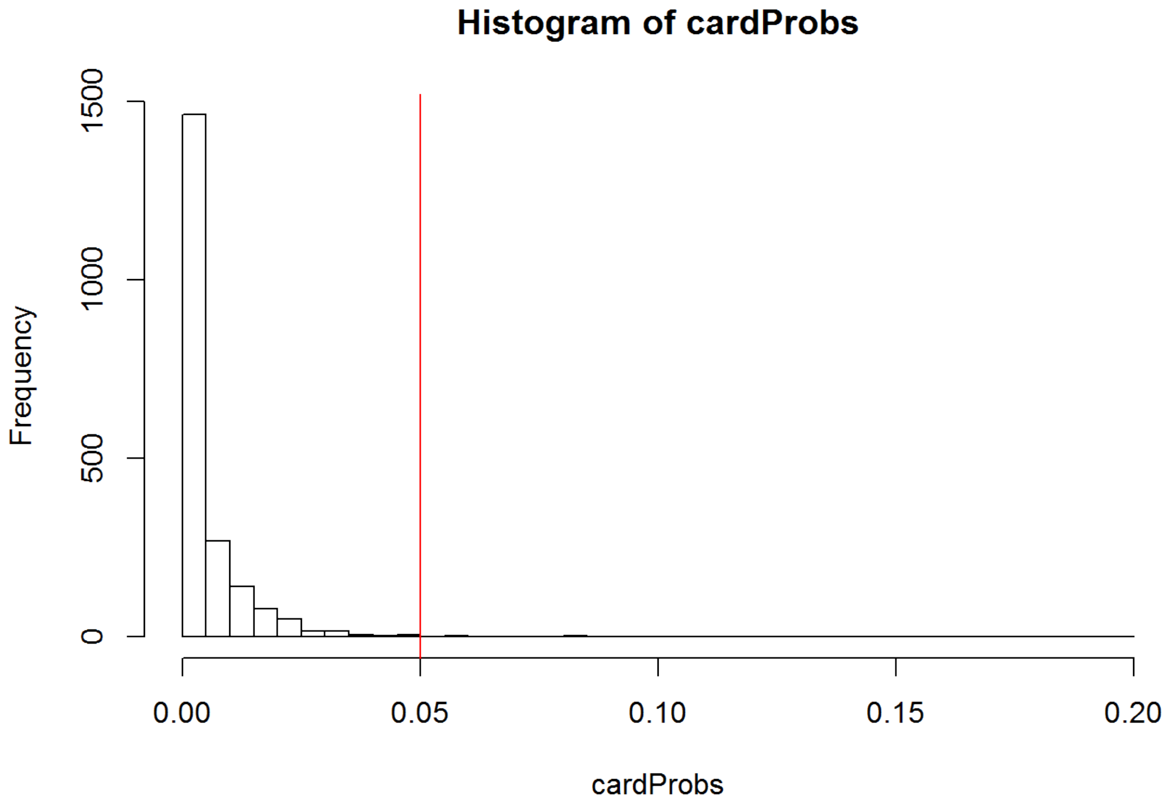

Figure 5. Histogram of the proportion of matches in which players received a red card. Only seven out of 2053 players had a proportion larger than 5%. We deemed these seven players to be outliers and excluded them from consideration in the later analysis.

Figure 5 shows that the players who have a red-card rate higher than 5% are not representative for the population of players under consideration. For infrequent events and highly skewed distributions, the presence of outliers can greatly affect the results. In order to obtain a more robust and more reliable impression of referee bias we eliminated from subsequent analysis all seven players who have a red-card rate higher than 5%, hence removing 0.3% of the complete data set.

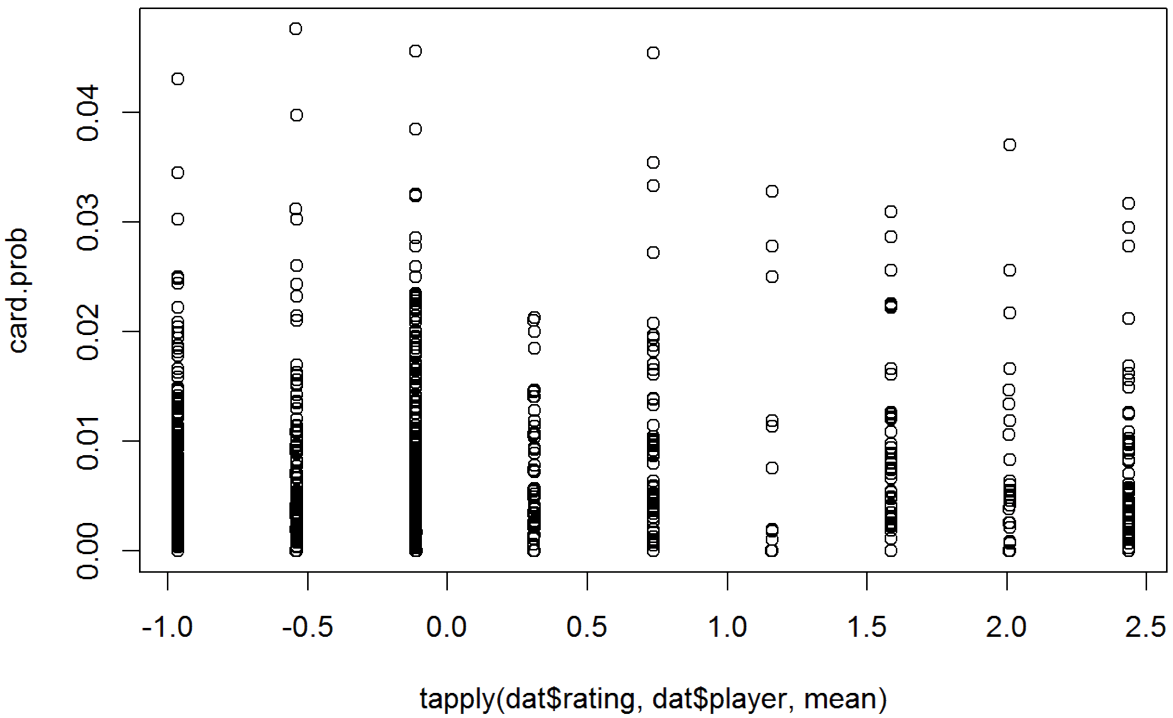

Figure 6. The observed proportion of matches in which a player received a red card, separately for each skin-tone rating and after excluding seven outlying players.

Figure 6 shows the relation between skin tone rating and the probability of getting a red card after the seven outlying players have been removed. If anything, Figure 6 suggests that the red card proportion is higher for players with light skin tone than with dark skin tone. The analyses below confirm this visual impression. A step-by-step analysis is provided in the accompanying html file “Without7Outliers.html”.

After excluding the outliers, we re-ran our analyses. For the broad N(0,1) prior, the results are as follows. Figure 7 shows the prior (red) and posterior (black) distribution for the effect of skin tone (i.e., predictor 5). As can be seen, the posterior is now almost perfectly centered around 0 (95% credible interval: [-0.077, 0.050]). The Bayes factor for its exclusion equals 28.3. To satisfy the demands of this crowdstorm project we also derived an odds ratio and its 95% credible interval: 0.956 [0.771, 1.184]. In sum, after including seven outlying players the effect of skin tone bias is completely eliminated.

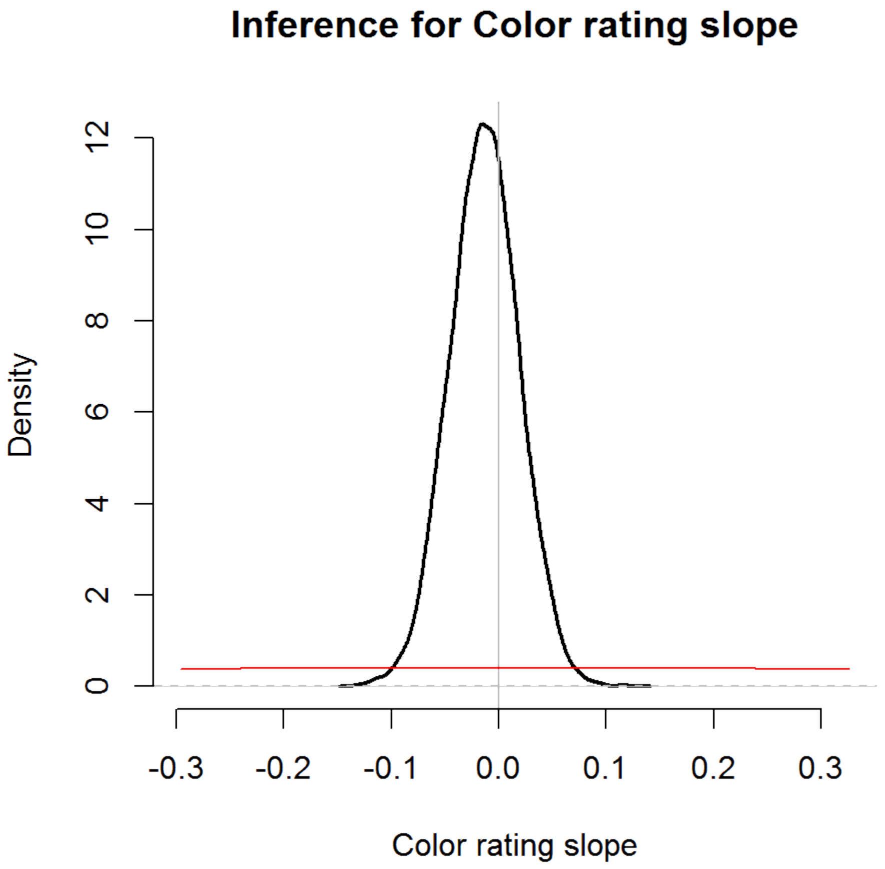

Figure 7. Posterior distribution for the regression coefficient on the skin tone predictor after excluding seven outlying players.

Figure 8 shows the prior and posterior distribution for the interaction between IAT score and skin tone (i.e., predictor 6). As can be seen, the posterior is relatively close to zero, with most mass on positive values (95% credible interval: [-0.047, 0.174]). The Bayes factor against inclusion of the predictor is 8.6. To satisfy the demands of this crowdstorm project we also derived an odds ratio and its 95% credible interval: 1.445 [0.777, 2.564].

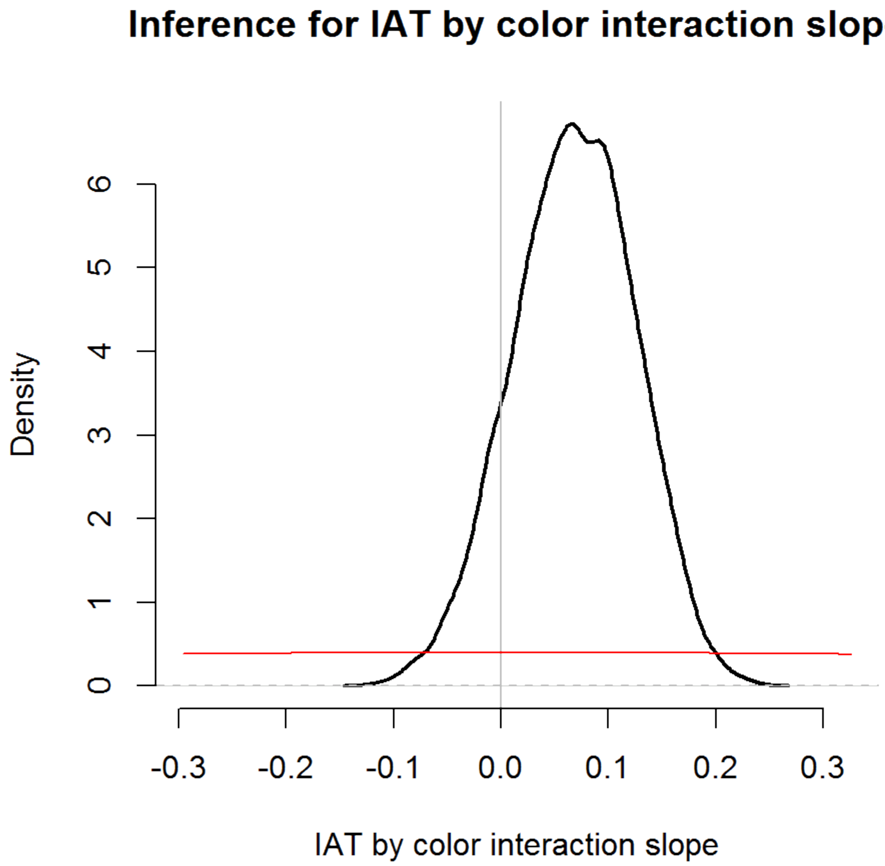

Figure 8. Posterior distribution for the regression coefficient on the interaction between IAT score and skin tone after excluding seven outlying players.

Figure 9 shows the prior and posterior distribution for the interaction between explicit bias score and skin tone (i.e., predictor 7). As can be seen, the posterior is relatively close to zero, with most mass on negative values (95% credible interval: [-0.191, 0.043]). The Bayes factor against inclusion of the predictor is 7.4. To satisfy the demands of this crowdstorm project we also derived an odds ratio and its 95% credible interval: 0.587 [0.263, 1.351].

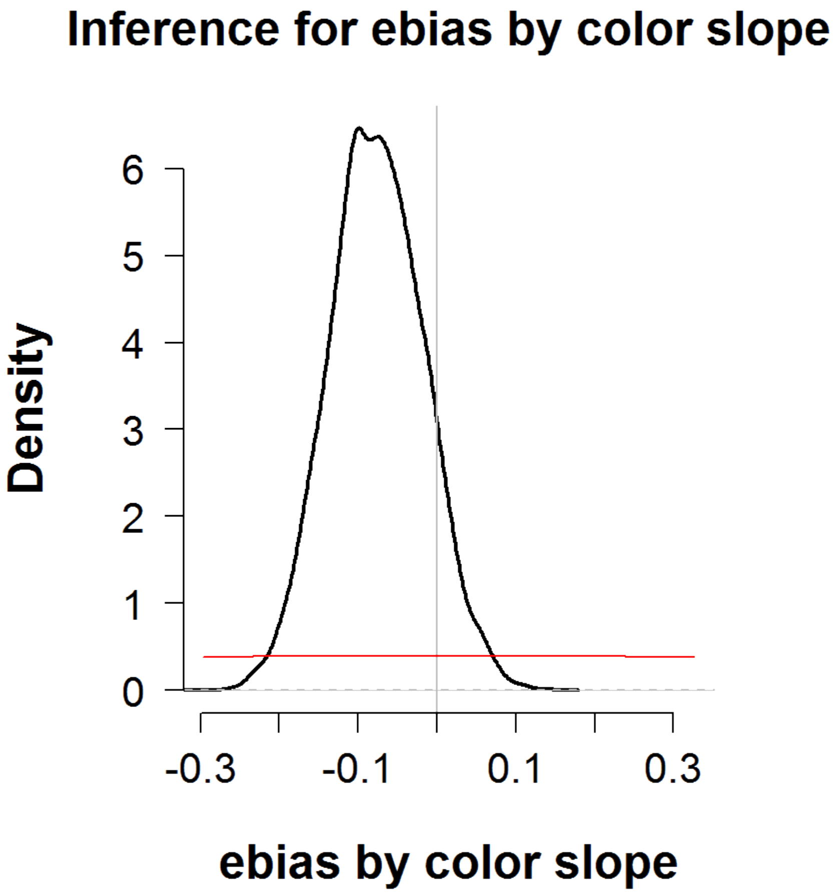

Figure 9. Posterior distribution for the regression coefficient on the interaction between explicit bias score and skin tone after excluding seven outlying players.

Additional Analyses

The analyses conducted above suggest several others. For instance, one can analyze the data set without outliers using narrow priors. This will change the Bayes factors but it will hardly affect the parameter estimates. The crowdstorm project does not center on Bayes factors, and (also in the interest of time) we omit the narrow-prior analysis here, at least for the moment.

Another more specific analysis could focus on the outlying players: one could assess whether they have unequal impact on the results using influence functions. Outlying players who have managed a > 5% red-card rate across many matches should affect the results more than those who have managed to retain such a high rate across fewer matches.

Another interesting analysis is to focus on yellow cards instead of on red cards. For the expression of bias, yellow cards are potentially more informative than red cards: in match play, red cards are often precipitated by a particularly violent or unorthodox event (e.g., hitting, biting, head-butting, etc.) that leaves little room for alternative interpretation. In contrast, yellow cards are usually given in more ambiguous situations that allow bias to exert a stronger influence on the decision. Another advantage of analyzing yellow cards is that these are more common than red cards and hence afford greater power.

Although all of these analyses are interesting and important, time constraints force us to omit them here.

Difference between Initial and Final Approach

Our final approach was identical to our initial approach except for three differences. First, we changed the regression link function. That is, our initial approach used the probit link, whereas our final approach used the logit link. We made this change in order to be able to summarize our effect as an odds ratio. The second change was that we explored the effects of using a more peaked prior in addition to our initial choice. The final change was that we eliminated seven outliers.

Caveats

In future work, more time and expertise should be invested in the specification of appropriate priors for the crucial predictors. The priors do not matter much for parameter estimation, but they do affect the results for the Bayes factors.

We are relatively certain that, in order to reach reliable conclusions, the seven outliers need to be excluded from consideration. However, we are still somewhat reluctant to conclude that there is no evidence for bias in this data set. Our reluctance comes from the fact that there is uncertainty about the likelihood. This uncertainty is worrisome in any modeling attempt, and the present crowdstorming project can be seen as an attempt to check the robustness of the conclusions against different likelihoods. In this respect the choice of predictors is particularly relevant; we made one choice but others are possible. Even if we had included all available predictors, this does not exclude the presence of other important control predictors that could change the results.

Conclusion

Using Bayesian logistic regression, our initial analyses supported the conjecture that soccer referees are slightly biased to give red cards to players with a dark skin tone. After excluding seven outlying players, however, this effect was eliminated in its entirety. This serves as an illustration and a warning that for low-probability events and skewed distributions, a few extreme observations may seriously distort statistical inference.

The current analysis process also demonstrates a potential drawback of the preregistration of analyses, in that it was difficult perhaps to foresee the presence and impact of outliers. Nevertheless it is evident (to us) that these outliers needed to be removed, as otherwise the model would be seriously misspecified. This document presents the analysis in the order in which they were conceived and conducted. We discovered the impact of outliers at the very last moment. It is interesting to examine the extent to which the conclusions from alternative crowdstorm approaches are affected by the removal of outliers.

Data, code, and output are available on the Open Science Framework at https://osf.io/wfvpc/.