Bayesian hierarchical binomial regression model suggests soccer referees are more likely to give red cards to dark skin toned players

Author: Erikson Kaszubowski[1]*

Affiliations

[1]Universidade Federal da Fronteira Sul – *Campus* Cerro Largo – Brazil; Universidade Federal de Santa Catarina – Brazil

*E-mail: erikson84@yahoo.com.br

Abstract

In order to verify if soccer referees have skin tone bias, this article analyzed dyadic data from 1,433 soccer players who played in the first male divisions of England, France, Germany and Spain in 2012 and 2013, and 2,906 referees they encountered during their professional career. With the counts of red cards as the outcome variable and controlling for player position and referee country of origin, the effect of skin tone in the probability of a player receiving a red card was estimated using a bayesian hierarchical binomial regression model. The model revealed a linear trend in the effect, with darker toned players being more likely to receive a red card.

One Sentence Summary

Soccer referees are more likely to give red cards to dark skin toned players.

Results

To examine the question if soccer referees are more likely to give red cards to dark skin toned players, we model the probability of a player receiving a red card from a given referee during the course of a single game using a binomial logistic regression corrected for overdispersion (Gelman & Hill, 2007). Different versions of the model were fit to analyze the effect of various combinations of input variables in the estimation of the main effect of interest, the player skin tone.

Preliminary versions of the model were fit using the *lme4* (version 1.1-6) package for R (Bates, Maechler, Bolker & Walker, 2014). The results from these fits suggested that many input variables – like player height, weight, age, games won or lost – don't have a significant influence in predicting the probability of a player receiving a red card; nor they changed significantly the estimation of player skin tone effect.

After choosing a final set of predictors, the definitive models were implemented and fitted using *Stan* (version 2.3.0) (Stan Development Team, 2014). Unmodeled parameters were given weakly informative priors. Each model was fit using four chains, each with 200 warm-up and 500 effective iterations, and no thinning. All models exhibited good convergence (Rhat <=1.05 and n_eff > 100 for the great majority of the parameters). The resulting parameters of interest are presented as points estimates (the mean of the chain samples) together with 95% credible intervals.

Model Description

Although soccer rules describe the kind of fouls that should receive a yellow or a red card, the referee has some degree of freedom in deciding if a specific action constitutes a mild or serious foul. This freedom of choice, the fact that a direct red card is independent of any previous

card given to a player and the rule that a player may receive at most one red card in a single game make it reasonable to model the occurrence of a red card as the outcome of a Bernoulli trial.

$$y_{ij} \sim Bernoulli(\theta_{ij})$$

The double subscript is justified by the way the data is organized: we have a data set in which each player is paired with a referee. The probability that a player *i* receives a red card (*y*) from a referee *j* in the course of a game follows a Bernoulli distribution with parameter θ_{ij}. In our data set, we also have information about how many games happened between the player-referee dyad. We don't have more information about how much time the player stayed in-game, so we will assume that every game in the dyads has the same weight. In order to take advantage of the format of the data set, we generalize the model as a binomial distribution.

$$y_{ij} \sim Binomial(N_{ij}^{games}, \theta_{ij})$$

In other words, we expect the total count of red cards in a player-referee dyad to follow a binomial distribution with the number of trial *N* equal to the number of games and probability of "success" (a red card) θ.

We want to know if the player skin tone has any influence in the probability that a referee gives him a red card. To assess this, we model the parameter θ using the inverse logit link function applied to the linear combination of a baseline probability of a red card and the effect of player skin color, controlling the effects of different predictors in different versions of the model (described below). We add an error term to account for possible overdispersion in the binomial model.

Initial Approach

For the first version of the model, we used a transformed combination of the skin tone rating. The values from the two raters were averaged and rounded to the most central value. This

way we could work with only one variable for player skin tone, assuming a conservative rating when the raters didn't agree. The new, averaged variable agreed with 83% of the ratings from the first rater and with 93% from the second rater. The skin tone rating was added to the model as a categorical predictor to account for possible non-linearity of skin tone effect. We also added the player position and the country league as non-nested group predictors. Adding player position to the model follows an obvious reason: some field positions are more likely to receive a red card because their role make them commit more fouls. The league country also has a clear reason to be in the model: some leagues are more strict in applying the rules, so the referees are more likely to give red cards to different players. We didn't add any interaction term to the regression because the model suffers from sparse data: red cards are, after all, rare events, and the estimated effects for each predictor is very small, as we shall see in the results. The mode can be formalized as:

$$y_{ij} \sim Binomial(N_{ij}^{games}, \theta_{ij})$$
$$\theta_{ij} = logit^{-1}(\beta_0 + \beta_i^{color} + \beta_i^{position} + \beta_{ij}^{league} + \epsilon_{ij})$$
$$\beta_{position} \sim N(0, \sigma^2_{position})$$
$$\beta_{league} \sim N(0, \sigma^2_{league})$$
$$\beta_{color} \sim N(0, \sigma^2_{color})$$
$$\epsilon_{ij} \sim N(0, \sigma^2_{\epsilon})$$

All cases with missing data were excluded from the analysis. To make the model simpler, we assumed that the missingness mechanism was completely at random and performed a complete-case analysis, retaining 115457 cases (79%). This assumption is strong and might be far from the truth, since a quick analysis shows that most of the missing data comes from the English and French leagues.

The skin tone coefficients estimated from this first model suggest an almost linear trend, which can be verified in the graph below.

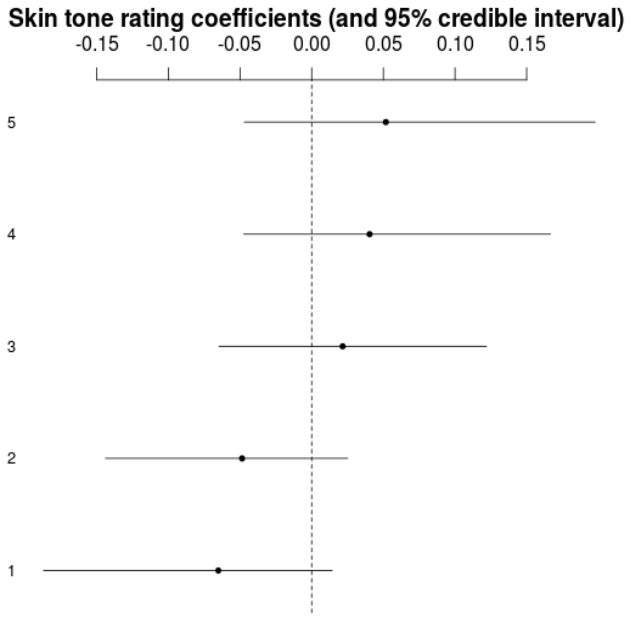

Figure 1 – Skin tone rating coefficients with 95% credible interval for the first model.

The seemingly linear trend in the coefficients suggested a change to the model. We included the skin tone rating as a numerical predictor in the second level of the model, but kept it as a categorical predictor in the first level. The formal definition remained almost the same as before, with this change:

$$\beta_{color} \sim N(\gamma_0 + \gamma_1 \times meanRating, \sigma^2_{color})$$

This change allows the categorical estimates to be pulled to linearity if it is supported by the data (Gelman & Hill, 2007). The resulting coefficients didn't change much from the first version, but the credible intervals now excludes 0 for the two extremes of the 5-point scale.

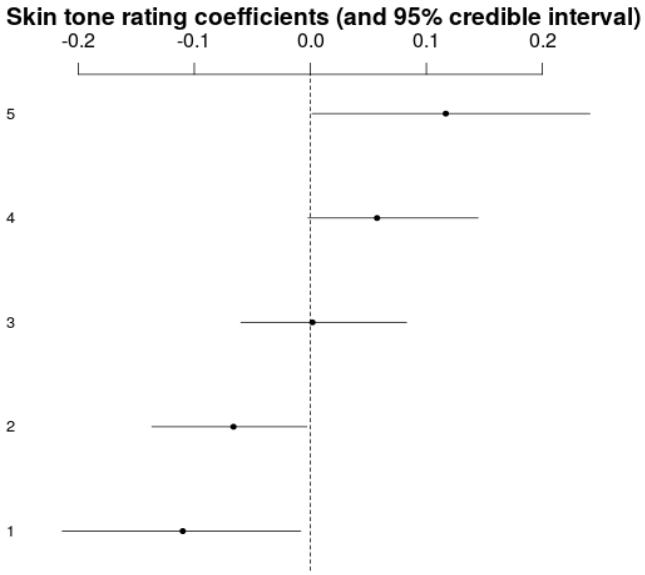

Figure 2 – Skin tone rating coefficients with 95% credible interval for the second model

Transforming the coefficients to odds ratio, we have: 0.89 (0.80-0.90) for level 1; 0.93 (0.87, 0.99) for level 2; 1.00 (0.94, 1.08) for level 3; 1.05 (0.99, 1.15) for level 4 and 1.12 (1.00, 1.27) for level 5. These ratios compare to the mean odds baseline, which is 0.0039 (0.0037, 0.0042). In terms of probabilities, the results do not differ much because the odds are very small. The model predicts that the mean probability of any player receiving a red card in a game is 0.0039. A player with light skin tone rating (level 1) has 11% less chance to receive a red card than average (0.0034); while a dark skin toned player (level 5) has 12% more chance to receive a red card than average (0.0043).

To answer the second research question, we selected the cases with mean skin tone rating higher than 3, resulting in a total of 15,953 cases. The skin tone variable was excluded from the model predictors, since the data are reduced to dark skin toned players. Referee country of origin was included to estimate its effect in the model. We included the countries implicit and explicit

racial bias data as second level predictors, multiplying the original values by 10. So, for the second research question, the model was defined as follows:

$$y_{ij}^{DarkSkin} \sim Binomial(N_{ij}^{games}, \theta_{ij})$$

$$\theta_{ij} = logit^{-1}(\beta_0 + \beta_j^{refCountry} + \beta_i^{position} + \beta_{ij}^{league} + \epsilon_{ij})$$

$$\beta_{refCountry} \sim N(\gamma_0 + \gamma_1 \times meanIAT_i + \gamma_2 \times meanExp_i, \sigma^2_{country})$$

$$\beta_{position} \sim N(0, \sigma^2_{position})$$

$$\beta_{league} \sim N(0, \sigma^2_{league})$$

$$\epsilon_{ij} \sim N(0, \sigma^2_{\epsilon})$$

The estimated country effects didn't vary much ($\sigma_{country}$ = 0.07(0.00, 0.27)), but all 95% credible intervals included 0. The estimates for the second level of the model suggests that the implicit racial bias has bigger weight on country effect (γ_1 = 0.08 (-0.43, 0.67), γ_2 = 0 (-0.07, 0.08)), but the credible interval is too big to provide convincing evidence.

Final Approach

For the final approach, the model was redefined to account for suggestions from peer review and changes in the variables definitions. The main difference was the exclusion of league country variable, because the data for the dyads do not come only from the player current league. The skin tone rating changed to a 0-1 scale that kept the original five levels. The values from the two raters were combined in a single mean value, without rounding. The mean rating was included directly as a numerical predictor, because the initial approach suggested that skin tone effect followed a linear trend. Finally, we included the referee country of origin in the model. The cases with missing data in the variables of interest were excluded as before. The model used data from 116,014 cases (79,4% of the total).

The first model fitted in the final approach estimated the coefficient for player skin tone rating as a fixed effect. It can be formally defined as:

$$y_{ij} \sim Binomial(N_{ij}^{games}, \theta_{ij})$$

$$\theta_{ij} = logit^{-1}(\beta_0 + \beta_1 \times raterMean_i + \beta_i^{position} + \beta_j^{refCountry} + \epsilon_{ij})$$

$$\beta_{position} \sim N(0, \sigma^2_{position})$$

$$\beta_{refCountry} \sim N(0, \sigma^2_{country})$$

$$\epsilon_{ij} \sim N(0, \sigma^2_{\epsilon})$$

As the mean skin tone rating now has values from 0 to 1, the baseline now represents the first level of skin tone (lighter). The odds a lighter skin toned player receives a red card in a game is 0.0031 (0.0025, 0.0038). As the odds are very small, again, transforming them to probabilities do not change much their value.

The coefficient for β_1 estimated by the model was 0.27 (0.09, 0.45). This means that players with dark skin tone (fifth level) have an odds ratio of 1.31 (1.09, 1.57) of receiving a red card when comparing to players with light skin tone. In terms of probabilities, a player in the lighter end of the scale has 0.31% chance of receiving a red card during a game, while another player in the darker end of the scale, holding all else constant, has 0.41% chance.

Comparing to the the initial approach, the predicted difference between the probabilities of players in the extreme of the scale did not change much. The models also agree that the effect of skin tone is positive with a high degree of certainty – a darker skin toned player has a higher probability of receiving a red card.

For the second research question, the model was expanded to allow the β_1 coefficient to vary by referee country of origin. In the second level of the model, the countries implicit and explicit racial bias varibles were included as predictors for the skin tone rating coefficients. The final model definition is as follows:

$$y_{ij} \sim Binomial(N_{ij}^{games}, \theta_{ij})$$

$$\theta_{ij} = logit^{-1}(\beta_0 + \beta_{refCountry[j]} \times raterMean_i + \beta_i^{position} + \beta_j^{refCountry} + \epsilon_{ij})$$

$$\beta_{refCountry[j]} \sim N(\gamma_0 + \gamma_1 \times meanIAT_j + \gamma_2 \times meanExp_j, \sigma^2_{rater})$$

$$\beta_{position} \sim N(0, \sigma^2_{position})$$

$$\beta_{refCountry} \sim N(0, \sigma^2_{country})$$

$$\epsilon_{ij} \sim N(0, \sigma^2_{\epsilon})$$

This model makes more conservative prediction than the previous: a player with 0 skin tone rating has 0.0029 (0.0022, 0.0036) mean odds of receiving a red card, while a player with 1 skin tone rating has 0.0035 (0.0025, 0.0049) mean odds of receiving a red card. The point estimate of the coefficient for most countries (76%) were above 0, but in only two cases the 95% credible interval did not include 0.

The uncertainty in country level data is also evident in the second level coefficients. The expected coefficient of a country with no racial bias (IAT =0 and Exp=0), which correspond to the γ_0 coefficient, is 0.17 (-2.34, 2.51). The estimate of implicit racial bias coefficient had a different sign than the estimate of the initial approach: -0.05 (-0.83, 0.77). The explicit racial bias coefficient now showed a higher estimate, with less uncertainty: 0.06 (-0.06, 0.19). The variation in the estimates between models, together with their high uncertainty, implies that the model and the data do not provide enough evidence to answer the second research question.

Predictive Model Checking

To assess the fitted model, we conducted some predictive model checks to verify discrepancies from the predicted values to the original data. The tests were conducted using the estimates from the final approach extended model (with varying skin tone rating coefficients).

As expected, the model performs poorly in predicting the true counts above zero. Red cards are rare events and the model predicts very small probabilities even in the most extreme

condition. So, instead of checking if the model can predict correctly the number of red cards in a player-referee dyad, we analyzed if simulated datasets present distinct features from the original data.

First, we verified if the simulated datasets had an equivalent number of red card counts. The original data had a total of 1480 red cards. The mean count from 2000 simulated datasets was 1478, with a 95% range of 1372-1590. The original data count is well within the predicted datasets.

Second, we verified if the number of zeros were correctly predicted by the model. The original data has 114,557 zeros; the simulated datasets have a mean number of zeros equal to 114,600. Again this feature of the dataset doesn't seem to be a problem for the model.

Third, we verified the number of cases with one and two red cards. The data has 1434 cases of only one red card and 23 cases with two red cards. The mean number of cases with one red card in the replicated datasets is 1430, and the mean number of cases with two red cards is 23. The numbers are very close; the model seems to predict well the sparsity of red card counts.

This seemingly good results hide a problem: the same predictive checks from a “null” model (with only the baseline) show similar results. To solve this issue, we analyzed the mean proportion of red card per game grouped by skin tone rating level. The resulting graph below shows that the model captures the difference of the mean proportion between the different levels, even though it displays greater variation in the higher end of the scale.

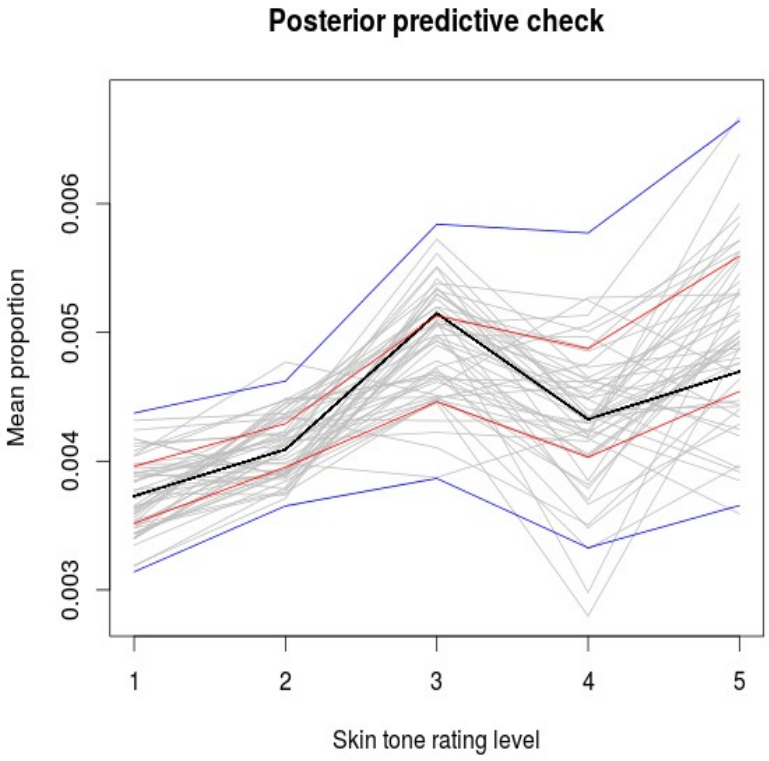

Image 3 – Posterior predictive check of mean proportion of red cards per game grouped by skin rating level, with 50% interval (in red) and 95% interval (in blue). The dark line represents the values from the original dataset, the lighter lines are 50 random lines from the simulated datasets.

Conclusion

The skin tone rating coefficient estimated by the model in the final approach suggests that players with darker skin tone have 31% higher odds of receiving a red card than players with lighter skin tone. The 95% credible interval for the estimate (1.09, 1.57) excludes zero, which implies that there is a great certainty the sign of the coefficient is positive. These results suggests that soccer referees are more likely to give red cards to players with darker skin tones. The effect is small, though, specially if compared to the effects of player position or referee country of origin.

References

Bates, Dounglas; Maechler, Martin; Bolker, Ben & Walker, Steven. (2014). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.1-6. http://CRAN.R-project.org/package=lme4

Gelman, Andrew & Hill, Jennifer. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York, USA: Cambridge University Press.

Stan Development Team. (2014). *RStan: the R interface to Stan, Version 2.3*. http://mc-stan.org/rstan.html.