"Tobit regression suggests a somewhat greater likelihood of receiving red cards when skin tone is dark"

Authors: Lammertjan Dam[1], Eric Molleman[2], Laetitia B. Mulder[2*], Bernard A. Nijstad[2], Floor Rink[2], Susanne Täuber[2].

Affiliations

[1]University of Groningen, Faculty of Economics and Business, Department of Economics, Econometrics, and Finance.

[2]University of Groningen, Faculty of Economics and Business, Department of Human Resource Management and Organizational Behavior.

*Correspondence to: l.b.mulder@rug.nl

Abstract

We transformed the number of red cards to a frequency score of "red cards per game per player" (bounded between 0 and 1). As negative binomial or Poisson regression erroneously assumes the same number of games played for each player, we chose to perform a Tobit regression which assumes a normal distribution that is truncated at 0 and 1. We preferred Tobit over logistic regression as there was a huge peak in the distribution at zero red cards, whereas for a logistic distribution all frequencies preferably are larger than zero. In our initial analysis we used total number of red cards and in our final analysis we used the number of straight red cards. In both cases, we found that skin color was weakly and (marginal) significantly related to the number of red cards received. Implicit or explicit prejudice against dark people did not moderate this effect.

One Sentence Summary

A Tobit regression method showed that skin color was weakly related to the number of red cards received, but this was not moderated by skin-tone prejudice as determined by referee country.

Results

Initial Approach

Data transformations

We transformed the main dependent measure, the number of red cards, to a frequency score of “red cards per game per player”. This was done by standardizing the number of games to one, by calculating the number of red cards per game. By construction the new variable ranges from 0 to 1, and can be interpreted as the probability of receiving a red card. In our initial approach, we made no distinction in the “type” of red card (i.e. a direct red or twice yellow where treated equally).

We also standardized (z-values) all continuous predictors before analysis, except for skin color. For skin color, we took the average of the two skin color ratings, given high agreement. This variable ranges from 0 to 1. We did not exclude any cases from the data analysis. However, some cases were lost due to missing values (for research question 1 the N was 351,339 and for research question 2 the N was 351,160, where N is the total number of games played).

Choice for analysis

At first sight, methods using count data may seem appropriate to use (i.e. a negative binomial or Poisson regression). However, these methods assume that the number of trials (in this case “games played”) is the same for every observation, whereas each observation is based on a different number of games. As such, these distributions erroneously assume the same number of games played for each player. Controlling for the number of games does not solve this problem as then this model would still make it possible to produce, for example, a “predicted” or “fitted” value of 5 red cards for a player who has only played 2 games. To solve this issue, we transformed the dependent variable and bounded it between 0 and 1. As such, the frequencies were between 0 and 1 which made the sample bounded, and thus OLS inappropriate. We therefore chose to perform a Tobit regression (using STATA). Tobit regression assumes that the "underlying" latent variable can be any continuous value, but if it is lower than 0 or higher than 1, the actual observed value is 0 or 1 respectively (Greene, 2008). In other words, Tobit assumes a normal distribution that is truncated at 0 and 1. With regard to the current data, the idea is that the latent variable is the tendency with which a player makes fouls. This cannot be observed, but we observe red card frequencies, which can only be between 0 and 1. It implies that not all zeros are the same. After all, two players with very low but different tendencies (e.g., one player with zero cards in one game and another player with zero cards in ten games) will both exhibit a red card frequency of 0 (the same goes for the upper bound of 1). So, we only observe values for these tendencies between 0 and 1 –even if the underlying tendency is very high, the observed frequency can never be higher than 1 or lower than 0. So it is as if the underlying tendency variable is truncated.

Note that we did not choose for a logistic regression. Although this would seem a logical choice, we deemed it hard to follow this approach because there is a huge peak in the distribution at zero red cards. For a logistic distribution it is more pragmatic when all frequencies are strictly larger than zero, otherwise the log odds will be minus infinity for a huge part of the sample.

Data aggregation

Because the level of analysis for this question is player, we initially aggregated the data to the player level, however, for the subsequent analysis we did not do so as it hardly affected the results. However, since the number of times that a player encounters the same referee can affect the average frequency of receiving a red card, we ran standard model specifications *and* model specifications with frequency weights to each player-referee dyad. Frequency weighing entailed that, for example, a player-referee dyad that has 10 encounters is weighted 10 times more than a player-referee dyad that only had one encounter. Because this weighted method blows up the number of observations, we adjusted the standard errors for clusters of repeated observations at different levels. This adjustment generally does not affect the point estimate, but may affect the significance of the effect. We therefore ran different model specifications depending on the clusters (either club or country and either player or referee) and frequency weights (yes vs. no). Across all possible model specifications, the results did not differ substantially. The results we report for both research questions are based on the models in which we used frequency weights and models in which we clustered for country (research question 1) or for player (research question 2).

Covariates

We aimed to rule out possible alternative explanations for the effect of skin color on red cards. Therefore, we tested which variables correlated significantly with both skin color and red cards. These were: the number of games played ($r_{\text{reds-games}} = .40$ and $r_{\text{skin-games}} = -.06$), number of victories ($r_{\text{reds-victories}} = .36$ and $r_{\text{skin-victories}} = -.06$), number of defeats ($r_{\text{reds-defeats}} = .40$ and $r_{\text{skin-defeats}} = -.03$), height ($r_{\text{reds-height}} = .07$ and $r_{\text{skin-height}} = -.06$). Further, skin color and/or red cards also depended on league and position. English, French and German leagues correlated significantly with skin color (respectively $r = .06$, $r = .25$ and $r = -.19$). The German league also correlated significantly with red cards ($r = -.08$), being a keeper, midfielder or attacker all correlated significantly with skin color (respectively $r = -.11$, $r = -.09$ and $r = .13$). Being a defender or midfielder also correlated significantly with red cards (respectively $r = .18$ and $r = -.06$).

As number of games played, number of victories and number of defeats are naturally related (playing more games results in more victories and defeats), we only needed one of them to be a covariate, and we chose for number of games. Further, we included age, height, league (three dummies: France, Germany and Spain), and position (three dummies: defender, midfielder and attacker) as covariates. For league one may argue that this is a misleading variable as it only represents the *current* league (while the dataset includes referee encounters throughout players' professional careers, which also means previous leagues). However, we reasoned that, although league may be a variable hard to interpret, the fact that it correlates with skin color and red cards or both (for whatever reason) one will still rule out a possible alternative explanation by controlling for it. We used the same covariates for research question 1 and research question 2.

Results

For research question 1, across all four model specifications, we found that the effect of Skin color was statistically significant at a 10% significant level, regression weight $b = 0.0005$, 95% conf: [-0.0000468, 0.0011317], $t = 1.80$, $p < .10$ (standardized regression weight beta = 0.20, $t = 1.87$, $p =$ is 0.061, 95% Ci = [-0.009713,0.4135025]). Since the skin color ranks from 1 to 5, it means that the difference in frequency of red cards between a player with skin color 1 and a player with skin color 5 is 4*0.0005 = 0.002, which is 0.2%. So a player with a very dark skin color has 0.2% higher likelihood of receiving a red card compared to a player with a very light skin. To test research question 2, we tried to reproduce the results for the aggregated data with the disaggregated data, either clustered on player or referee observations. This latter made little difference and we report the results clustered on player. Now, the main effect of skin color on frequency of cards is even larger ($b = 0.0278$, $t = 2.59$, $p < .01$). We subsequently included the two prejudice variables in specifications 5 and 6 and the first interaction term (IAT × skin color) in specification 7. This interaction-term did not yield a significant effect, regression weight of the interaction $b = .0005$, $t = 0.13$, estimate = + 0.0005, 95% conf: [-0.0074375, 0.0085091]. This result suggests that the relationship between skin color and the average frequency of red cards does not depend on this implicit prejudice measure. The second interaction term (Explicit × Skin color), which was entered in specification 8, was also insignificant, regression weight of interaction *b* = .0016, *t* = 0.40, estimate = + 0.0016, 95%, conf: [-,.0061072, 0.0092735]. This result suggests that the relationship between skin color and the average frequency of red cards does not depend on this explicit prejudice measure.

Final Approach

In our final approach, we decided to use straight red cards as a dependent variable rather than total red cards. The reason for this is that the most usual red cards were a simple result of two yellow cards in answer to incidents that are not deemed severe. In contrast, a straight red card is given in answer to an incident that is regarded as severe. It says more about the "drasticness" of a referee decision when he gives a straight red card instead of the alternative of a yellow card. Further, at the request of the project organisers, we re-scaled the skin tone variable to values that ranges from 0 to 1.

We did similar Tobit Regressions in which we used frequency weights for games and clustered at referee ID level. As mentioned earlier, we did not aggregate the data to the player level this time. For research question 1, we again found that the effect of Skin color was statistically significant at a 10% significant level, estimate *b* = 0.037, 95% conf: [-0.001774, 0.075522], *t* = 1.87, *p* = .061. Since the skin color ranks from 0 to 1, it means that the difference in frequency of straight red cards between a player with skin color 0 and a player with skin color 1 is 1*0.037 = 0.037, which is 3.7%. So a player with a very dark skin color has 3.7% higher likelihood of receiving a straight red card compared to a player with a very light skin. Upon request of the project organizers we provided the standardized regression weight for calculating the effect size, β = 0.20, 95% conf: [-0.009713, 0.4135025].

For research question 2 we found, similar to our initial approach, that the main effect of skin color on frequency of cards was more significant, *b* = 0.038, 95% conf: [0.000549, 0.075636], *t* = 1.99, *p* < .05. The IAT x skin color interaction-term did not yield a significant effect, regression weight of the interaction *b* = -0.0031, 95% conf: [-0.03792, 0.031633], *t* = -0,18, *p* = .86. This result suggests that the relationship between skin color and the average frequency of straight red cards does not depend on this implicit prejudice measure. The Explicit × Skin color interaction term was also insignificant, regression weight of interaction b = 0.010086, 95%, conf: [-0.02347, 0.04364], *t* = 0.59, *p* = .56. This result suggests that the relationship between skin color and the average frequency of straight red cards does not depend on this explicit prejudice measure.

Additional analyses

As an additional analysis we were asked to do our final analysis again but now without controlling for league. Also, due to the extensive online discussion with the other teams, we had become more convinced that a logit regression would also be an appropriate method (maybe even better) so we considered to switch to a logit regression instead. However, for the current dataset, we judged all analyses, also the logit, to have pros and cons. More specific, compared to the Logit regression, the Tobit regression is more conservative and gives larger confidence intervals and thus a lower probability of a Type I error, but a higher probability of a Type II error. We considered that the dataset had many pitfalls due to the doubtful validity of some variables. For example, not only league but also field position might be subject to change over a career and the prejudice measure was merely a proxy measured at country level instead of referee level. This, together with the "fraughtness" of the research question, led us to conclude that a conservative test is appropriate. Also considering that we were the only group to have chosen the Tobit regression, we decided it to be a good thing to represent the Tobit in the paper and stick to it.

So, we performed the same Tobit regression as in our final analysis only now without the league dummies as covariates. For research question 1, we again found that the effect of Skin color was statistically significant, but now at a 5% significant level, standardized estimate β = 0.28, 95% conf: [0.0071586,   0.5581066], t = 2.01, p = .044. For analyzing research question 2 (using the IAT scores) we also found a main effect of skin color, β = 0.38, 95% conf: [0.0065263,   0.618087], t = 2.42, p < .05. The IAT x skin color interaction-term did not yield a significant effect, standardized regression weight of the interaction β = -0.16, 95% conf: [-0.36191, 0.043263], t = -1,54, p = .12. For analyzing research question 2 (using the explicit scores) we again found a main effect of skin color, β = 0.32, 95% conf: [0.04435,   0.591146], t = 2.28, p < .05. The Explicit × Skin color interaction term was insignificant, regression weight of interaction β = -0.08962, 95%, conf: [-0.27389, 0.094646], t = -0.95, p = .34.

Conclusion

Based on Tobit regressions clustered for country or for referee and in which frequency weights were used, it was found that skin tone color was weakly and related (significantly when league was not included as covariate and marginally significant when league was included as covariate) to the number of red cards received (their total of red cards or direct red cards). Although this may indicate that referees are biased against darker skin-toned players, causal inferences cannot be made. Also, based on Tobit regressions clustered for player or referee in which frequency weight were used, it was not found that the implicit or explicit prejudice against dark people based on the country of the referee moderated this effect. Further research in which the referee’s implicit and explicit bias (at the personal level, not at the country level) are tested, is needed, to further interpret the skin color-red cards link.

References and Notes

William H. Greene, Econometric Analysis, 6th ed., 2008 New Jersey: Prentice.