Bayesian clustering of referee bias in soccer reveals skin color prejudice

Authors: Silvia Liverani[1]

Affiliations

[1]Department of Mathematics, Brunel University, Uxbridge UB8 3PH

Abstract

We applied a Dirichlet process Bayesian clustering method to the data. This method aims at identifying subgroups of the player-referee dyads which had a higher chance of incurring into red cards. Our results show that there is a higher relative risk of incurring in a red card for players with darker skin. Moreover, our analysis suggested that there might be a higher chance of incurring in red cards also for other subgroups of the players, in particular those who have been rated as “neither dark nor light skin”. Our concerns about implicit and explicit bias scores led us to use only the subset of the data for which a large sample size was available. We found no particular association between a higher relative risk of red cards and these bias scores, but we noted that referees from specific countries appear to have a bias which reflected in the relative risk of giving red cards.

One Sentence Summary

Darker skin players appear to have a higher relative risk of incurring in red cards, but we also found this for other subgroups of the players, in particular those who have been rated as “neither dark nor light skin”.

Results

We have carried out a Dirichlet process Bayesian cluster analysis which aims at identifying subgroups of the player-referee dyads which had a higher chance of incurring into red cards.

The statistical method that we employed is also referred to as profile regression (Molitor et al, 2010; Liverani et al, 2014) and it is implemented in the R package PReMiuM. Standard regression methods encounter issues when they are used to make inferences with covariates which are highly correlated. As the covariates available for this study are highly correlated (such as the skin scores provided by the two raters), we chose to use profile regression. This method addresses this issue by using a profile formed from a sequence of covariate values and then clusters them into groups associated via a regression model to the relevant outcome.

We chose to use this method for the reason above, but also to provide an alternative point of view for this problem in the context of crowdstorming. We accept and agree that clustering is not the first method that one would apply to answer the questions asked in this exercise, but it can provide what we believe are interesting answers. These answers do not necessarily blend well in a comparative table with regression models, but they provide additional insight on the underlying process. Partecipating in this project we have learned how even regression can be applied in many different ways, with different model choices and assumptions, giving such different results. Within this project, we position ourselves in a unique way, carrying out the analysis with a method which is completely different from those used by most other partecipants, but also provides a unique insight into the issue under study.

We chose to cluster the minimum number of covariates to shade light on the underlying process. By using other covariate the result can be subdivided in additional clusters. We felt this was unnecessary to answer the questions asked in this project. To answer the first question we included the response variable red cards, and the skin score ratings as covariates. We used the number of games played as the denominator for the Binomial distribution. For the second question we also included the mean implicit bias score and the mean explicit bias score for part (a) and (b) respectively. We did not transform any of the variables.

We removed all dyads which had missing values for the skin scores. For question 2(a) we also excluded all dyads for which the sample size for race IAT was less than 5,000 and similarly for question 2(b) we excluded all dyads for which the sample size for implicit bias was less than 5,000. This choice was due to noticing that the implicit and explicit scores had a high variability for countries with small sample sizes and small variability for countries with high sample size. This is particulary noticeable when the log of the sample size is plotted vs. the implicit or explicit bias score. This raised doubts on the validity of such scores for countries with small sample sizes and ultimately resulted in the choice of excluding the dyads with players from countries with sample sizes smaller than 5,000. The threshold was chosen visually by identifying when, in the plot of the log of the sample size vs. the implicit and explicit bias scores, the variability of such scores is no longer dependent on the sample size. Unfortunately, this resulted on including the bias scores only for few countries.

Initial Approach

In our initial approach, due to the size of the dataset, we decided to include in our clustering approach only those dyads where red cards, our response variable, had been given.

It was noted by all three reviewers of our approach that controls (ie. player-referee dyads for which no red cards were incurred) were necessary for the analysis.

Final Approach

Following the feedback round, we decided to include all player-referee dyads in our analysis.

The results of our final approach show that there are clusters which include mostly players with darker skin and which have a higher relative risk of incurring in red cards. The full results are provided in the tables below.

Regarding the first question, our method identified three clusters with a higher relative risk of incurring in red cards. One of them, notably, is formed exclusively by darker skin players and in the context of this crowdstorming exercise, we chose to provide the details of this cluster for comparison with other methods. However, it is important to note again how our results cannot be directly compared with the results obtained by carrying out regression analysis as our model is specified differently, and the underlying assumptions are also different. However, it will be interesting to see what has been learnt using the different methods. Interestingly we also identified two more clusters with a relative risk greater than one, and they both correspond to clusters which include players which have "neither dark nor light skin" according to the raters. There is therefore a suggestion that not only dark skin players might face prejudice on soccer fields, but also other subgroups of players.

Regarding the second question, we have concerns about the use of the implicit and explicit bias scores as we have discussed earlier. This led us to analyse only a subset of the data for this question. Our results show that players with dark skin scores were more likely to incurr in a red card. However, there is no clear relation between this and implicit nor explicit bias scores. It transpired that regardless of their bias scores referees from certain countries of origin were more likely to give red cards to darker skin players.

Conclusion

We applied a Dirichlet process Bayesian clustering method to the dataset provided. Our results show that there is a higher relative risk of incurring in a red card for players with darker skin. Moreover, our analysis suggested that there might be a higher chance of incurring in red cards also for other subgroups of the players, in particular those who have been rated as "neither dark nor light skin".

Our concerns about implicit and explicit bias scores led us to use only the subset of the data for which a large sample size was available. Therefore, only a few countries were included. We found no particular association between a higher relative risk of red cards and these bias scores, but we noted that referees from specific countries appear to have a bias which reflected in the relative risk of giving red cards.

Tables

Question 1

Nine clusters were identified by our method when we included the two skin scores as covariates. Each row of the following tables represents the characteristics of the clusters. Two clusters of outliers, constituted by one and seven dyads respectively, have been removed by these table for simplicity.

The following table includes the median relative risk of each cluster, and the lower and upper bound for the 95% credible interval of the relative risk. The last column shows the number of dyads allocated in each cluster. The relative risk is computed with reference to the cluster with the highest proportion of light skin scores, as can be checked in the following table.

	Lower bound	**Median**	**Upper bound**	**Cluster Size**
1	1.00	1.00	1.00	32522.00
2	0.38	0.39	0.39	57578.00
3	1.42	1.42	1.43	1384.00
4	0.03	0.04	0.04	22.00
5	0.30	0.30	0.30	15080.00
6	1.42	1.43	1.43	413.00
7	0.50	0.50	0.50	10408.00
8	0.16	0.16	0.16	26.00
9	1.70	1.71	1.72	7180.00

The following table provides the median proportion of skin scores for each cluster.

	Rater 1					**Rater 2**				
	1	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
1	0.99	0.01	0.00	0.00	0.00	0.99	0.01	0.00	0.00	0.00
2	0.30	0.69	0.01	0.00	0.00	0.06	0.91	0.03	0.00	0.00
3	0.00	0.88	0.12	0.00	0.00	0.00	0.23	0.75	0.02	0.00
4	0.05	0.34	0.52	0.05	0.04	0.04	0.69	0.13	0.09	0.04
5	0.02	0.17	0.81	0.00	0.00	0.00	0.22	0.65	0.13	0.00
6	0.00	0.13	0.86	0.01	0.00	0.00	0.05	0.63	0.31	0.00
7	0.00	0.00	0.00	0.91	0.09	0.00	0.00	0.01	0.73	0.26
8	0.04	0.03	0.03	0.03	0.86	0.03	0.04	0.03	0.18	0.71
9	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	1.00

Question 2(a)

Six clusters were identified by our method when we included the two skin scores and the mean implicit score. Each row of the following tables represents the characteristics of the clusters. Two clusters of outliers, constituted by one and two dyads respectively, have been removed by these table for simplicity.

The following table includes the median relative risk of each cluster, and the lower and upper bound for the 95% credible interval of the relative risk. The last column shows the number of dyads allocated in each cluster. The relative risk is computed with reference to the cluster with the highest proportion of light skin scores, as can be checked in the following table.

	Lower bound	**Median**	**Upper bound**	**Cluster Size**
1	1.00	1.00	1.00	17,002
2	0.33	0.33	0.34	24,888
3	0.92	0.94	0.96	608
4	0.56	0.56	0.56	5,165
5	1.19	1.19	1.19	2,427
6	1.14	1.15	1.15	3,837

The following table provides the median proportion of skin scores for each cluster.

	Rater 1					**Rater 2**				
	1	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
1	1.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
2	0.33	0.66	0.01	0.00	0.00	0.07	0.90	0.03	0.00	0.00
3	0.02	0.93	0.05	0.00	0.00	0.01	0.05	0.94	0.00	0.00
4	0.01	0.02	0.95	0.02	0.00	0.00	0.13	0.67	0.20	0.00
5	0.00	0.00	0.01	0.93	0.06	0.00	0.00	0.01	0.93	0.06
6	0.00	0.00	0.00	0.24	0.76	0.00	0.00	0.00	0.13	0.87

The following table provides the median proportion of implicit bias scores for each cluster. Only 8 countries were included in the analysis as a result of the restriction of a sample size greater than 5,000. The 8 scores are ordered according to their corresponding implicit bias score.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	0.00	0.59	0.30	0.05	0.00	0.01	0.03	0.02
2	0.00	0.50	0.35	0.07	0.00	0.01	0.03	0.03
3	0.00	0.63	0.21	0.07	0.01	0.03	0.03	0.01
4	0.00	0.33	0.52	0.07	0.00	0.01	0.03	0.03

5	0.00	0.31	0.52	0.07	0.01	0.02	0.04	0.04
6	0.01	0.33	0.49	0.08	0.00	0.02	0.03	0.05

Question 2(b)

Six clusters were identified by our method when we included the two skin scores and the mean explicit score. Each row of the following tables represents the characteristics of the clusters.

The following table includes the median relative risk of each cluster, and the lower and upper bound for the 95% credible interval of the relative risk. The last column shows the number of dyads allocated in each cluster. The relative risk is computed with reference to the cluster with the highest proportion of light skin scores, as can be checked in the following table.

	Lower bound	**Median**	**Upper bound**	**Cluster Size**
1	1.00	1.00	1.00	15,783
2	0.18	0.23	0.23	24,551
3	0.30	0.31	0.32	860
4	0.19	0.21	0.21	5,165
5	1.14	1.18	1.19	2,310
6	0.94	0.99	1.00	3,693

The following table provides the median proportion of skin scores for each cluster.

	Rater 1					**Rater 2**				
	1	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
1	0.96	0.04	0.00	0.00	0.00	0.94	0.06	0.00	0.00	0.00
2	0.38	0.62	0.00	0.00	0.00	0.13	0.85	0.02	0.00	0.00
3	0.08	0.89	0.03	0.00	0.00	0.02	0.22	0.75	0.00	0.00
4	0.00	0.01	0.97	0.02	0.00	0.00	0.14	0.67	0.20	0.00
5	0.00	0.00	0.01	0.94	0.05	0.00	0.00	0.01	0.94	0.05
6	0.00	0.00	0.00	0.24	0.76	0.00	0.00	0.00	0.13	0.87

The following table provides the median proportion of explicit bias scores for each cluster. Only 7 countries were included in the analysis as a result of the restriction of a sample size greater than 5,000. The 7 scores are ordered according to their corresponding explicit bias score.

	1	**2**	**3**	**4**	**5**	**6**	**7**
1	0.31	0.01	0.00	0.60	0.03	0.00	0.05
2	0.36	0.01	0.00	0.53	0.03	0.00	0.07

3	0.27	0.02	0.01	0.58	0.03	0.00	0.08
4	0.54	0.01	0.00	0.34	0.04	0.00	0.07
5	0.53	0.02	0.00	0.32	0.04	0.01	0.08
6	0.51	0.02	0.01	0.34	0.03	0.00	0.08

Data and Output

Please upload to the Open Science Framework (OSF): https://osf.io/ and include both your initial analyses before the feedback round and your final analyses. Links to these files will appear in the published manuscript.

Instruction for Uploading to the Open Science Framework:
In addition to these instructions here is a brief video on the OSF:
https://www.youtube.com/watch?feature=player_embedded&v=c6lCJFSnMcg

1. **Create an account**. Visit the site (www.opencienceframework.org). Each contributor to the crowdstorm should create a personal account, by clicking the 'create an account or sign in' button in the top right corner.

2. **Create the project**. One contributor should go to the Dashboard by clicking the link on the top of the page. Create a new project by clicking the 'New Project' button. For the title, write: "Crowdstorming a dataset: Do soccer referees give more red cards to dark skin toned players? Analyses by [Team Member Names]", and then click the 'Create New Project' button.

3. **Add collaborators (Optional)**. The project creator can add collaborators by clicking the 'add' link just below the project title. Type in the name of any other team members of yours (just last name may be enough) and add them. If they have not registered, they will not appear in the search. Do not add them until they are registered. Now all your team members have editing privileges for the project.

4. **Using project space**. The project space includes tags, a wiki, files, and components/nodes. You can use any of these features as they are useful for documenting your research. Nodes operate like folder. These may be most useful to define discrete components of the research process, particularly if they have independent contributor lists, or if you'd like to be able to cite those components independently, or if you'd like to keep some parts of the project private while other parts are public (e.g., a data node that stays private until the article is accepted for publication). Each node has the same features as the project - unique contributor list, tags, wiki, files (a new "project" node creates a project within a project). Project nodes might be useful, for example, if your project consists of multiple studies. Until you click the "make public" button on the top right of any project or node, the project page is private. Only the collaborators can access the materials.

5. **Upload files**. When in the main project, go to the Files tab. Click the Upload button, or simple drag and drop files onto the webpage. The file appears in the upload list. Click the blue Start button to upload the file. If you revise a document and then upload it again with the identical filename, the OSF will retain a version history of the file. You will be able to access any prior version, and when you download it. The date the file was uploaded will be appended automatically to the filename (probably in the same way you manage file edits in your local directory).

6. **Make project public**. Click the 'Make public' button on the top right of the project space. If you have multiple components to make public, each one must be made public manually.

References and Notes

J. T. Molitor, M. Papathomas, M. Jerrett and S. Richardson (2010) Bayesian Profile Regression with an Application to the National Survey of Childrens Health, Biostatistics, 11, 484-498.
Liverani, S., Hastie, D. I., Azizi, L.,

Papathomas, M. and Richardson, S. (2014) PReMiuM: An R package for Profile Regression Mixture Models using Dirichlet Processes. Forthcoming in the Journal for Statistical Software. Available at http://uk.arxiv.org/abs/1303.2836