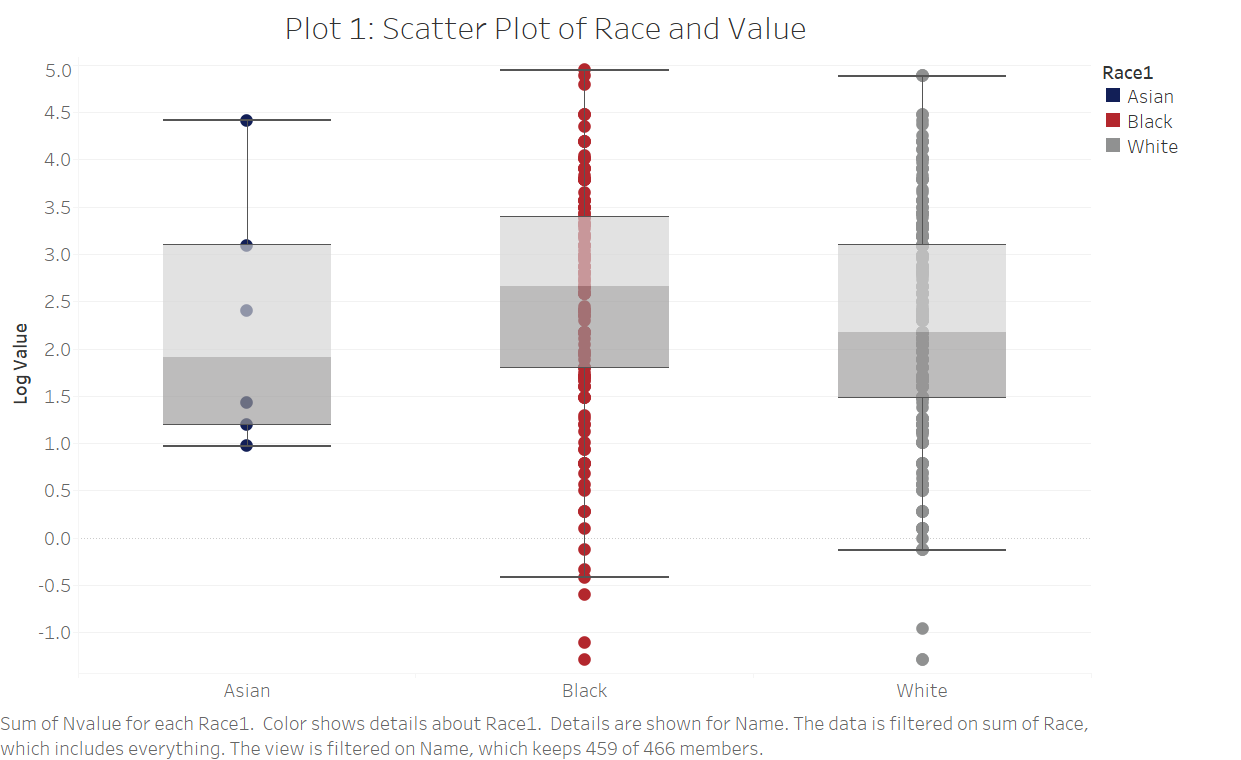
Impact of Race on Player Value in the English Premier League

After an outbreak of racist insults hurled at players during matches in the 2019-2020 European football season, many leagues implemented anti-racism campaigns. Among these leagues was the English Premier League, considered by many to be the best domestic league in the world. The league put forward it’s “No Room for Racism” campaign which included lifetime bans for fans who used racial slurs at games and fines for teams. Many saw this as a huge step forward for football, and a model for other leagues to follow suit. While this campaign is surely a positive step, the true harmful effects of racism may still be felt by the players themselves. One way to value a player is their transfer market value. This leads to the question: “Are players valued differently based on their race in the Premier League?”. It is suspected that, *ceteris paribus*, a Black player will be valued less on the open market. In European football players, are purchased by clubs for a set price and then sign personal wage agreements with the club. While wage agreements are often kept private, a player’s value on the market can be found through numerous sources like transfermarkt.com. By using this data and combining with player performance statistics, this paper seeks to understand if players are valued differently based on their Race.

Data for this paper was collected from two main sources. All player statistics were gathered from fbref.com (517 observations). This website gathers its own data as well as some harder to compute statistics like expected goals from StatsBomb. The transfer value data was obtained from transfermarkt.com along with the players race (510 observations). The data set originally had 517 observations. We then had to remove Goalkeepers due to the fact that common player statistics would not have produced reliable results for value. Once Goalkeepers and players with no value were removed, there were 474 observations in the data set.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Table 1: Summary Statistics | | | | | | |
| Variable | Obs | Mean | Std. Dev. | Min | Max | Skewness |
| Player Value(Millions of USD) | 473 | 18.453 | 21.951 | 0.275 | 140.800 | 2.597 |
| Shot Creating Actions (Per 90 Minutes) | 440 | 2.013 | 1.633 | 0.060 | 16.360 | 3.399 |
| Tackles (Per 90 Minutes) | 444 | 1.905 | 1.362 | 0.328 | 20.000 | 6.387 |
| Expected Assists | 396 | 1.701 | 1.923 | 0.100 | 18.400 | 2.858 |
| Expected Goals | 409 | 2.483 | 3.415 | 0.100 | 20.300 | 2.656 |
| Age | 472 | 25.294 | 4.084 | 16.000 | 36.000 | -0.001 |
| Race: |  |  |  |  |  |  |
| White | 473 | 0.013 |  | 0 | 1 |  |
| Black | 473 | 0.385 |  | 0 | 1 |  |
| Asian | 473 | 0.603 |  | 0 | 1 |  |
| Position: |  |  |  |  |  |  |
| Defender | 473 | 0.374 |  | 0 | 1 |  |
| Defender, Forward | 473 | 0.004 |  | 0 | 1 |  |
| Defender, Midfielder | 473 | 0.015 |  | 0 | 1 |  |
| Forward | 473 | 0.167 |  | 0 | 1 |  |
| Forward, Defender | 473 | 0.002 |  | 0 | 1 |  |
| Forward, Midfielder | 473 | 0.080 |  | 0 | 1 |  |
| Midfielder | 473 | 0.233 |  | 0 | 1 |  |
| Midfielder, Defender | 473 | 0.038 |  | 0 | 1 |  |
| Midfielder, Forward | 473 | 0.091 |  | 0 | 1 |  |
| Caption: In addition to the variables here there are two variables, Country and Club which have 61 an 20 variables respectively, and are not included for brevity. | | | | | | |

Player value is listed in millions of USD. If a player was sold within the current season, the listed sale price was used. Shot creating actions and tackles are adjusted for 90 minutes. Expected assists and expected goals are a calculated value based on factors that include shot distance, angle, and players in front of the ball. For more information about these measures, links will be listed at the end of this paper. Race is listed as White, Black, or Asian. This was manually added to the data set based on the picture associated with the player on tranfermarkt.com. There are more nuanced race classifications than such, but would prove difficult to categorize for analysis. Primary positions are listed before secondary positions, ordered by minutes played. If there is no secondary position listed, the player only played one role in the season. The Central Limit Theorem is satisfied by the size of this sample. Although our dependent variable value is right skewed, it becomes more normalized once transformed into log form, which satisfies Central Limit Theorem as well.



Each point on this plot represents one player. From the above graph we should see that Black players are on average valued more than White or Asian players. This paper will focus on the analysis of differences between Black and White players since there is not enough Asian players to make assumptions. For this paper it is assumed that Black players are valued differently such that:

Value = 𝛽0 + 𝛽1RaceBlack + µ

There are two hypotheses that could be relevant for the data:

H0: 𝛽1 = 0 or H1: 𝛽1≠0

Because our nature of dependent variable, it is hard to tell whether there is a bivariate relationship immediately. However players values are spread out, and they

appear to be normally distributed once log adjusted. When a simple regression is run. 𝛽1 is 0.2923 with a standard error of 0.4839 which refutes the idea that Black players are valued less by clubs, but is in line with observations about the boxplots. Looking at the P-value as well we can tell this is a statistically insignificant estimate. If we translate this back to non-log form, this equates to a boost of roughly 1.4 million dollars in value based on the players race being Black. The simple regression gives an R2 of 0.0201 and the estimate itself is considered insignificant. OLS is likely an overestimate in this case. One key omitted variable that is likely causing bias is Clubs. This is most likely causing an upward bias since players from specific clubs are valued more. More competitive clubs may care less about a players race in order to succeed. It is also likely that since there is lower number of Black players in the Premier League, they have already passed some sort of barrier in lower level football. Controlling for performance metrics that typically drive value will help reduce the omitted variable bias as well. Some basic controls for performance are expected goals, expected assists, shot creating actions, and tackles. These omitted variables are likely causing an upward omitted variable bias as well since if you perform better a players value is likely to be higher. There are additional controls for player characteristics such as age and position. These controls compare players in similar points in the career. By adding all aforementioned controls, this should lead to a more accurate model.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table 2: Regression Results | | | | | |
|  | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
| Coefficient – Black (%Change in Value) | 0.2923  (0.4839) | 0.0487  (0.4339) | -0.3853  (0.5257) | -.0042102  (.42583) | -.0423  (.4161) |
| Robust S.E. | 0.601 | 0.415 | 0.529 | 0.524 | 0.334 |
| Coefficient – Asian  (%Change in Value) | -0.0858  (0.4811) | -0.2906  (0.4311) | -0.7277  (0.5237) | -0.2595  (0.426) | -.3368  (.416) |
| Controls | None | xGoals | xGoals, xAssists | xGoals, xAssists,  Club | All |
| Sample Size | 473 | 409 | 371 | 371 | 367 |
| Adjusted R-Squared | 0.0201 | 0.2419 | 0.2548 | 0.5299 | 0.5547 |
| -Controls for “All” are xGoals, xAssists, SCA, Tackles, Nation, Club, and Position  -Value, xGoals, xAssists, sca90, and tackles are all log adjusted  -Value originally given in millions of USD  -The base variable is “White” for Race | | | | | |

Since the results from the simple linear regression were inconclusive, it is important to add the aforementioned controls in. From table 2, R2 increases as more controls are added to the simple regression. When expected goals and expected assists were added to the regression, 𝛽1 gets much closer to zero. This suggests that there was actually a downward bias for these variables. This makes the hypothesis that there is no difference in player value based on race look more likely since the difference is quite small. The biggest jump in R2 occurs when Clubs are added as a control. 𝛽1 remained negative, but increased in value closer to zero, confirming the upward bias of this estimator. This also serves as a proxy for performance, since a player for 2019-2020 Premier League champions Liverpool, is likely a world class player compared to a player on Norwich city who were relegated the same season. When we compare R2 between model 4 and model 5 with all controls, there is little difference. 𝛽1 with all controls was -.0423, representing a 4.23% decrease in value based on Race. Once all controls were added, R^2 were only reached 0.5547, giving room for a high amount of omitted variable error. This is to be expected, however. Football is a largely fluid game. It is referred to as “The Beautiful Game” for this reason. Creating performance measures are much harder when their isn’t defined actions like in a sport like baseball. Also for any given interaction on the pitch there are at a maximum two participants. This leaves 20 other players doing effectively a positioning battle making it hard to account for the value of these actions. Also for this data set there is only one performance measure for defenders. Defending is considered a much more subjective position so it likely predictions for defenders will underperform.

The effect of Race on value may be different depending on a players age. As players age, clubs may care more about a players race, as there is less potential for them to develop into a better player. Conversely they may be more willing to sign young Black players because there is a higher chance they can become world class. If this were true, the interaction coefficient would increase as a player ages. To test this, the base regression was run with the interaction between Race and Age.

|  |  |  |  |
| --- | --- | --- | --- |
| Table 3: Interaction Between Race(Black) and Age | | | |
| Age | Interaction Coefficient | Marginal Effect | T-Statistic |
| 18 | -1.444  (.952) | 2.852 (.444) | -1.520 |
| 19 | -.966  (.873) | 3.093 (.276) | -1.110 |
| 20 | -1.411  (.899) | 3.116 ( .344) | -1.570 |
| 21 | -1.412  (1.249) | 2.942 (.193 ) | -1.140 |
| 22 | -1.332  (.819) | 3.212 (.199) | -1.630 |
| 23 | -1.490  (.829) | 2.915 (.258) | -1.800 |
| 24 | -1.072  (.814) | 2.922 (.200) | -1.320 |
| 25 | -1.17  (.819) | 2.977 (.213) | -1.430 |
| 26 | -.998  (.814) | 2.869 (.206) | -1.230 |
| 27 | -1.455  (.842) | 2.400 (.292) | -1.730 |
| 28 | -.932  (.819) | 2.619 (.223) | -1.140 |
| 29 | -.7995136  (.835) | 2.435 (.271) | -0.960 |
| 30 | -.860  (.883) | 2.456 (.387) | -0.970 |
| 31 | -1.723  (.881) | 1.479 (.349) | -1.960 |
| 32 | -1.847  (1.123) | 1.009 (.770) | -1.650 |
| 33 | -2.123  (1.148) | 0.361 (.777) | -1.850 |

Looking at the interaction table above, this seems to be the case. The interaction coefficient decreased until a players prime playing age at 26, increasing again until age 33. Comparing the T-statistics to the normal regression as well, this interaction seems to better explain a players value than either Race or Age by themselves. However only two of the interactions are considered significant at any age. Marginal effects are listed next to the interaction coefficient for comparison. The marginal effects support the same trend between Race and age.

To check for Heteroscedasticity B-P and White tests were run. The B-P test produced a high p-value of 0.4941 and the White test produced a low p-value of 0.0004. Although the tests contradict each other, looking at the Leverage-Residuals plot below, the data does not appear heteroscedastic. Because there was some inconclusiveness, robust errors were added. Once robust errors were added, the dataset passed the RESET test. Robust errors were then added to table 2.

Chart, scatter chart

Description automatically generated**Chart 2: Leverage - Residual Plot**

One outlier was removed from the data set for extremely high leverage. This change was incorporated into the calculations. The player that was removed was Raheem Sterling, who is an outlier in nearly every statistical category. He is also a Black player for Manchester City, which would raise his leverage on the data set further.

This paper used regression and other econometric techniques to analyze differences in player value in the Premier League based on Race. While the individual effect for Race was not statistically significant, some interaction levels between age and Race were. This suggests that race may have differing levels of impact on value at different places in a players career. Although Race by itself was not significant, there is evidence of a downward trend. There are three problems with the dataset that prevent an accurate effect of Race from shining through. The first is that the dataset only includes statistics from one Premier League season. Clearly a players performance in previous seasons impact their value so using a one season dataset won’t full capture a players value. Another key problem is the pure subjectivity and randomness of the sport. Every manager plays different styles. Some players won’t be valued the same by some clubs because they won’t fit into their preferred style of play. There is also the lack of measurements for off the ball movements and other subjective parts of the game. There is simply no way to capture all of this. Likewise among the collected data provided, defensive statistics are much less sophisticated than offensive ones. Until this part of measurement catches up it will be hard to truly compare defenders. The third reason is that differences between races would be more clear in wage pay than in transfer value, since they are specific to each club. Going forward to get more accurate, League officials should surrender all wage data to private group of researchers to investigate this claim. By doing so, Leagues could live up to their own promotions like “No Room for Racism”.

Works Cited

2019-2020 Premier League Stats. (n.d.). Retrieved December 06, 2020, from https://fbref.com/en/comps/9/3232/2019-2020-Premier-League-Stats

Premier League - Top market values 20/21. (n.d.). Retrieved December 06, 2020, from https://www.transfermarkt.us/premier-league/marktwerte/wettbewerb/GB1

Premier League risks Kick It Out confusion with 'No Room for Racism' campaign. (2019, March 25). Retrieved December 06, 2020, from <https://www.theguardian.com/football/2019/mar/25/premier-league-kick-it-out-no-room-for-racism>

XGoals Explanation. (n.d.). Retrieved December 06, 2020, from https://www.americansocceranalysis.com/explanation