Taste-based Discrimination in the NBA

Matteo Restuccia, Nathan Moonesinghe, Gabe Raffa, Rishabh Raniwala, Antonio Sellemi

Winter Quarter 2024

Abstract

This paper aims to address the question of whether taste-based discrimination exists in today's NBA. This is done by estimating the effect of being Black on wages, holding other factors constant. First, we replicate past papers to certify our data generation process. Following this, we employ p-score matching and an instrumental variable to identify the causal effect of being Black on wages. Then, we explore the possibility of race affecting wages through other avenues, like points per game. Finally, we apply more complex regression models, to observe what features are significant in predicting NBA player wages.

1 Introduction

1.1 Motivation

The National Basketball Association, commonly referred to as the NBA, is widely regarded as a symbol of Black progress. Historically and still today, in the 2023-2024 season, over 70% of NBA players are Black. Moreover, when compared to other prominent American sports leagues like the National Hockey League (NHL) or Major League Baseball (MLB), the demographics within the broader NBA community, including team personnel and coaches, exhibit a notably higher proportion of Black representation. However, while the pervasive presence of the Black community in the NBA seems to be a demonstration of Black success, it fails to capture the entirety of this narrative.

In his *Economics of Discrimination* (1957), Gary Becker demonstrated that customer discrimination could persist under perfect competition and result in long-term inequities in labor markets. A central observation of his publication was that if discrimination depresses the wages of Black workers relative to those of similarly qualified White workers, then a discriminator who does not want to hire Black individuals will have to pay more to hire their White counterparts.

Becker's influential and ground-breaking work was leveraged by civil-rights activists who sought legal remedies for historic injustices against Black people. In 1964, Congress passed a historical piece of federal legislation against discrimination. The Civil Rights Act barred discrimination on the basis of race, ethnicity, or sex in pay, promotion, hiring, firing, and training. With the vitality of the Civil Rights movement persisting to this day, the NBA serves as an excellent case study to assess the nation's adherence to its founding values, as delineated in the U.S. Constitution and Supreme Court precedents.

This prominence of the NBA as significant source of Black wealth drove our investigation into the existence of wage discrimination in the modern NBA, which is indubitably one of the most appropriate settings to perform such an analysis; the NBA and other basketball-affiliated companies have been recording a myriad of comprehensive player and team performance statistics since the 1980s. A natural question, then, is whether Black NBA players are compensated differently than their non Black peers?

1.2 Prior Literature

Whether wage discrimination exists in the NBA and how significant a role it plays in determining Black salaries is a question that has been asked and investigated many times. The inspiration for our analysis stemmed from three investigations.

1.2.1

Kahn and Sherer published *Racial Differences in Professional Basketball Players' Compensation* in 1988 and analyzed player salaries for the 1985-86 season, with the dual objective of identifying the existence of a ”White premium” in the NBA, and providing evidence in support of prevailing theory of customer discrimination. Kahn and Sherer hypothesized that, under free-agency, (a negotiation system rendering salary determination for experienced players to be market determined by enabling players to sign their preferred offer), if customer discrimination exists and if franchise owners are profit maximizers, then a ”White premium” would be established and observable. In other words, if White NBA fans have a preference for spectating White players, given the scarcity of White players in the league, owners desiring to drive up attendance would be willing to pay more to have White players on their rosters.

Kahn and Sherer modeled salary determination as:

$$\ln(S) = a \cdot \ln(V)$$

where S corresponds to salary and V corresponds to a player's marginal revenue product:

$$\ln(V) = c \cdot P + d \cdot R$$

where P corresponds to an overall representation of player performance and R is a race indicator (1 if White, 0 otherwise). However, since P does not exist, Kahn and Sherer approximate salaries as:

$$\ln(S) = \beta' X + \delta \cdot R + \epsilon$$

here X is a vector of player performance variables and local metropolitan characteristics.

Kahn and Sherer performed OLS and 2SLS regression analyses over different specifications of their model and consistently found race effects to significant and indicative of a White premium of 20%. Subsequent investigations found that home attendance was a positive function of White representation on a team, which was consistent with the prevailing theory of customer discrimination driving a ”White premium”.

1.2.2

In 1991, Brown, Spiro, and Keenan published *Wage and Nonwage Discrimination in Professional Basketball: Do Fans Affect It?* and built upon the analysis of Kahn and Sherer, choosing to investigate a different set of performance variables in the event that observed wage gaps were artifacts of mismeasured performance.

Brown, Spiro, and Keenan ran multiple OLS models using per minute career statistics and other metropolitan controls to model 1984-85 season salaries. Ultimately, Brown, Spiro, and Keenan concluded that the observation of a substantial performance-adjusted salary discrepancy was not likely to be an artifact of mismeasured performance on the court. They similarly found a significant coefficient on race that suggested a 15% White premium. Furthermore, a comparable market analysis to that conducted by Kahn and Sherer indicated that NBA teams situated in predominantly White cities exhibited a greater prevalence of White players on the team roster. A race-neutral distribution of NBA players across team rosters had only a 1% likelihood of matching the exact race-roster breakdown. This evidence of heightened demand for White players in regions with predominantly White fan bases aligns with the concept of customer discrimination, which underpins the phenomenon of "White premiums".

1.2.3

Most recently, however, in 1999, Sean D. Johnson published *Wage Discrimination in the National Basketball Association: Is There Discrimination Based on Race*. He considered data from the 1996-97 season and similarly sought to to model the driving forces of NBA player wages. Ultimately, he did not find that race is a significant contributor of player wages and suggested that thus was a result of greater acceptance of Black players among the NBA fan base.

Ultimately, these three papers provided a foundation for our analysis and inspired the development of our model.

1.3 Replication

In order to verify our data sources and confirm the validity of our source material's results, we replicated the regression conducted by Johnson. We did not have access to Johnson's data set as he used a physical copy of *The Sporting News Official NBA Register: 1996-1997 Edition*; this compendium was not available online or at any nearby library. Therefore, we chose to pull our data from other sources (hence the need to validate them). We used Basketball Reference and the NBA's officially reported statistics in order to generate the controls of interest that Johnson used. As Johnson used career averages up to the year prior to the salary year being analyzed, we needed to figure out how to calculate these averages ourselves for the relevant players in our dataset. Since neither source provided career averages up to a certain point in a player's career, we wrote a web scraper in Python to accomplish the task. The web scraper makes use of the fact that Basketball Reference URLs with regard to a player's career statistics follow a predictable pattern based on a player's first and last name. Using this pattern, we were able to scrape data from relevant player careers from the year they were drafted up until the 1995-1996 season, or the season before the salary year analyzed. We averaged the statistics for each year by the number of games the player played in the season, as all game statistics used by Johnson were in a "per game" format.

Using this compilation, we performed an OLS regression using the same controls as Johnson and the same variable of interest, the race of a player. We were able to match Johnson's results quite closely, as shown below, with a couple of noticeable differences. For one, while Johnson found the "Blocks per Game" variable to be statistically significant when determining a player's salary, our regression did not. Our hypothesis as to why this occurred is the omission of certain players with high block rates from our dataset that were included in Johnson's dataset; this would be due to errors thrown by our web scraper, causing certain players to be skipped either due to unique URL patterns or due to missing data in their reported career statistics. We further hypothesize that "Blocks per Game" tends to be heavily right-skewed as only a few players consistently put up multiple blocks per game; therefore, the omission of players with a high number of blocks per game could have driven our coefficient in the replication regression to be statistically insignificant.

Another key difference between our replication and Johnson's regression is that Johnson found both "Disqualification Percentage" and "Personal Fouls per Game" to be statistically significant; our regression does not include a control variable for "Disqualification Percentage" and did not find "Personal Fouls per Game" to be statistically significant. In the case of the former, we were unable to find reliable disqualification statistics and so had to omit the control from our regression; this is the only control that we omit that Johnson includes. In the case of the latter, we hypothesize that because our model omits "Disqualification Percentage", we are unable to separate the effects of "Disqualification Percentage" from "Personal Fouls per Game" and so their combined effect ends up being statistically insignificant. It makes sense that these controls would be intertwined; as Johnson hypothesizes, the coefficient on "Personal Fouls per Game" is positive in his regression because it is an indicator for how tough a defensive player a certain player is, while the coefficient on "Disqualification Percentage" is negative as too many fouls / getting removed from the game means a player is on the court less and thus negatively impacts salary (Johnson, 40).

Putting these two discrepancies aside, our regression agrees with Johnson on which constants are significant and which are not for all other constants used by Johnson. Thus, we claim that our dataset is similarly constructed and distributed as Johnson's and thus our data sources should be valid for further analysis. A side-by-side comparison of the two regressions is included below.

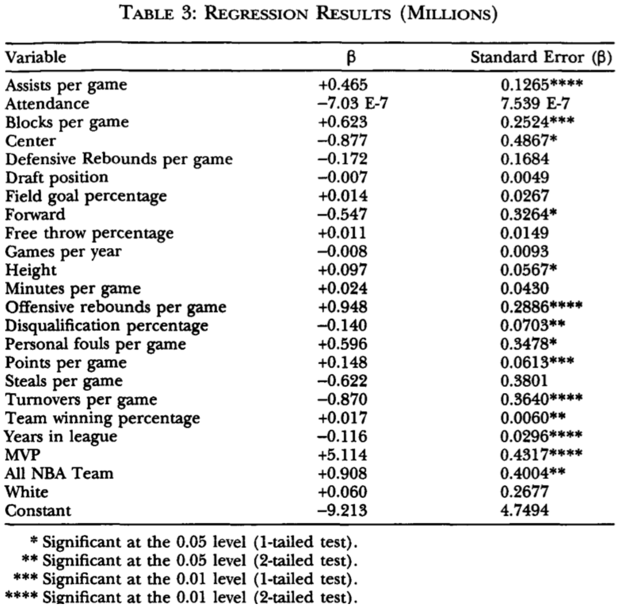
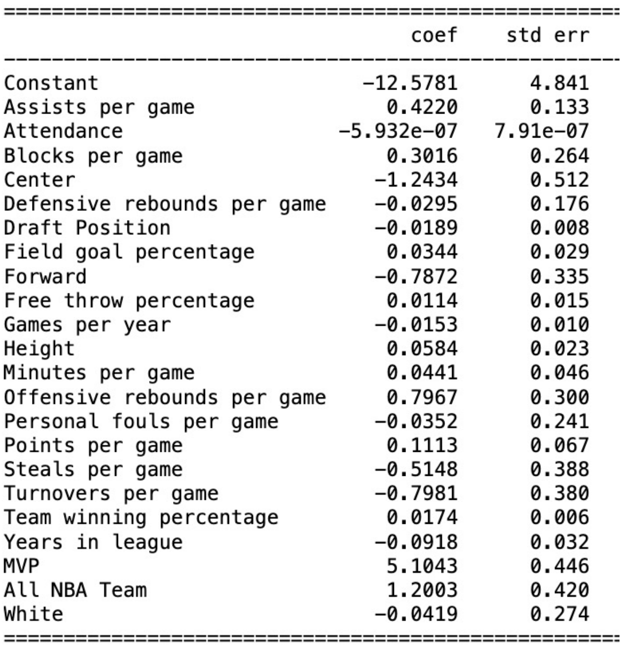


TABLE 3: REGRESSION RESULTS (MILLIONS)

Variable	β	Standard Error (β)
Assists per game	+0.465	0.1265****
Attendance	–7.03 E-7	7.539 E-7
Blocks per game	+0.623	0.2524***
Center	–0.877	0.4867*
Defensive Rebounds per game	–0.172	0.1684
Draft position	–0.007	0.0049
Field goal percentage	+0.014	0.0267
Forward	–0.547	0.3264*
Free throw percentage	+0.011	0.0149
Games per year	–0.008	0.0093
Height	+0.097	0.0567*
Minutes per game	+0.024	0.0430
Offensive rebounds per game	+0.948	0.2886****
Disqualification percentage	–0.140	0.0703**
Personal fouls per game	+0.596	0.3478*
Points per game	+0.148	0.0613***
Steals per game	–0.622	0.3801
Turnovers per game	–0.870	0.3640****
Team winning percentage	+0.017	0.0060**
Years in league	–0.116	0.0296****
MVP	+5.114	0.4317****
All NBA Team	+0.908	0.4004**
White	+0.060	0.2677
Constant	–9.213	4.7494

* Significant at the 0.05 level (1-tailed test).
** Significant at the 0.05 level (2-tailed test).
*** Significant at the 0.01 level (1-tailed test).
**** Significant at the 0.01 level (2-tailed test).

Figure 1: Johnson's Regression



```
====================================================
                                  coef    std err
----------------------------------------------------
Constant                      -12.5781      4.841
Assists per game                0.4220      0.133
Attendance                  -5.932e-07   7.91e-07
Blocks per game                 0.3016      0.264
Center                         -1.2434      0.512
Defensive rebounds per game    -0.0295      0.176
Draft Position                 -0.0189      0.008
Field goal percentage           0.0344      0.029
Forward                        -0.7872      0.335
Free throw percentage           0.0114      0.015
Games per year                 -0.0153      0.010
Height                          0.0584      0.023
Minutes per game                0.0441      0.046
Offensive rebounds per game     0.7967      0.300
Personal fouls per game        -0.0352      0.241
Points per game                 0.1113      0.067
Steals per game                -0.5148      0.388
Turnovers per game             -0.7981      0.380
Team winning percentage         0.0174      0.006
Years in league                -0.0918      0.032
MVP                             5.1043      0.446
All NBA Team                    1.2003      0.420
White                          -0.0419      0.274
====================================================
```

Figure 2: Replication Regression

1.4 Data Generation

In generating the data, we made a few important choices, in order to address some shortcomings of the aforementioned papers. First, we chose to include the last 4 years of play statistics for each player, averaged out with equal importance. We argue that this is the ideal subset of data to consider.

First, we want to filter for only "experienced" players, which we define as players who have been in the league for 4 or more years. This is because all rookie contracts expire after 2 or 4 years, so by including players who have been in the league for 4 or more, we remove outlier contracts. This is important in the context of evaluating taste-based discrimination, as rookie contracts have much less flexibility and cannot be negotiated. Therefore, if there is any discrimination, it won't be directly through salary, but will be on draft choice (higher draft picks get paid more). This is almost impossible to fairly evaluate, since statistics depend heavily on the level of competition, and each college team plays different competition. By including players with experience of at least 4 years, we guarantee everyone in our dataset has had the opportunity to negotiate a contract. As can be seen in Figure 3, keeping players with 4 or more years of experience also makes the salary data much more evenly distributed, as we get rid of rookies and bench players who drag the salary data down.

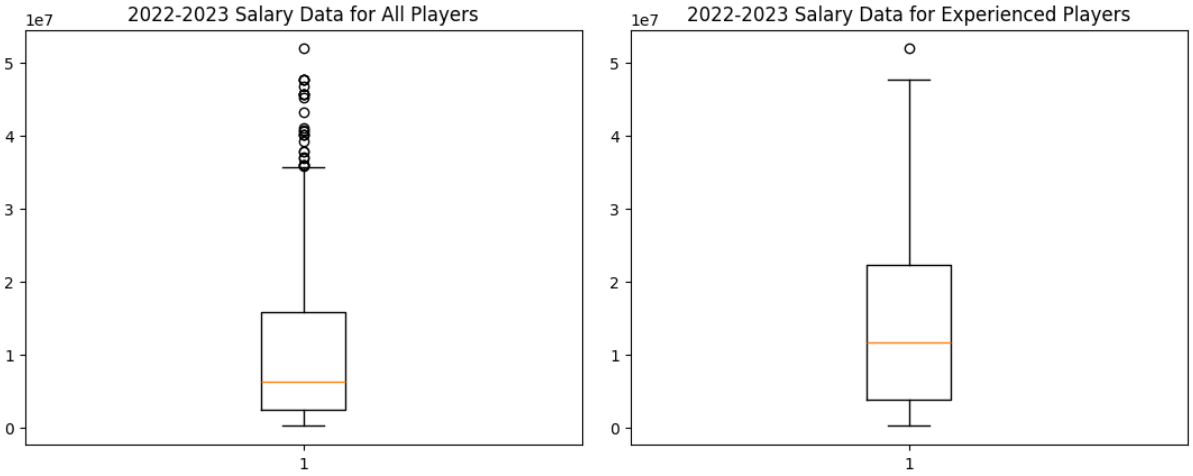


Figure 3: Outlier Comparison when including all, or only experienced, players

Second, by keeping the last 4 years of playing data, we are guaranteed to include most players' contract year. This is important, as the year in which a player is about to sign a contract is the most important (playing statistics-wise) in determining the size of that contract. The only exception to this is Devin Booker, who is still on a contract signed in 2019. For him, we checked his stats manually and noticed he has been very consistent in his play since then. Therefore, we chose to keep him in the study. Further, including multiple years before the contract signings (e.g. for players who signed their contracts in 2020) is ideal when we can, as those are certain to play into the new contract. Finally, in our models we make the key assumption that player salary does not affect play (more on this below), so we feel it is okay to include a few years after the contract, as players should play roughly as did before.

Finally, keeping the last 4-years allows us to look into a fixed window of time, during which players likely played fairly consistently. Including entire player's careers, as other papers did, fails to take into account that there may be year bringing averages down or up, but that those might be deep in the past.

In terms of the actual play-statistics, we debated between using statistics per season, per game, per minute or per 100-possessions. The past papers kept statistics per game or per minute. However, this does not take into account that different players play different amounts, both within a season and within each game. Therefore, it would be hard to compare play-statistics between players if we kept the statistics in those denominations. Then, in choosing between stats per minute and per 100 possessions, we considered pacing. Some teams may play slower, leading to fewer action (e.g. shooting, passing, etc..) per minute. This may not necessarily be a bad thing, as one can still win games that way, as the other team is kept to lower scores as well. Is is thus unfair to judge players based on their per-minute performance, as this may be a byproduct of their team's pacing, and may not actually indicate how productive a player is. The per 100-possessions data controls for all the above concerns, as is therefore how we generate our data.

2 Causal Models

2.1 Causal Setting and Assumptions

If the following two assumptions hold, we are in the causal setting and can build homogeneous treatment estimators.

$$CIA: Y(0)_i, Y(1)_i \perp D_i \,|\, x_i = x \forall\, x \tag{1}$$

$$COC: 0\, <\, P(D_i = 1 \,|\, x_i = x)\, < 1 \,\forall x \tag{2}$$

Lemma 1 says that there exist a set of variables such that treatment is unconfounded after conditioning on such variables. In practice, this only holds if we observe *all* of these covariates, which is unlikely and is a very strong assumption. Unfortunately, this assumption also cannot be tested, so we can't know for sure if this holds or not. We attempt to loosen this assumption in the rest of the paper, so as to make identifying the true causal effect more likely. Lemma 2 states that, for each group sharing the same values of controlling variables, there needs to be a non-zero probability of being Black or non Black. This is a much more likely assumption, and can easily be tested. Note that this assumption breaks down when we condition on too many covariates, but we address this concern in the matching section below.

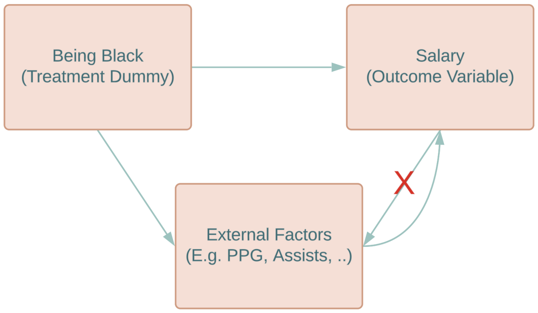


Figure 4: Causal Setting DAG

In order to understand the causal setting we are in, we visualized it using the directed acyclic graph (DAG) above. This graph visualizes what variables we believe have an effect on what other factors. Specifically, being Black may have a direct effect on salary, but likely has an effect on some external factors (e.g. interactions with coaches), which in turn may have an effect on salary. As a side note, this specific phenomenon may lead us to over/under estimate the effect of being Black on Salary, as this is Omitted Variable Bias (OVB). We address this concern in the instrumental variable section.

2.2 Variable Selection

Our initial feature set was quite large, including a myriad of player performance metrics, indicators on team, position, even whether a player is a key player for his team. While we initially believed a lot of these would be significant, we suspected the data would be inherently low dimensional. To test this hypothesis, we did some dimensional analysis on the data. If the data were indeed high dimensional with low correlation between the various features, we would expect each of the principal components to explain significant amount of the variance in the data (i.e. all of the principal components would roughly explain similar amounts of the variance in the data). However, as can be seen in Figure 5, the first two principal components combined to explain almost 80% of the variance in the features, and the first 3 explained almost 90%. There is a steep drop in variance explained by each additional feature, suggesting that the data is highly multicollinear. This is likely due to the fact that better players will have high points per 100 possessions, along with much better stats all around. Therefore, we had to take steps to address this.

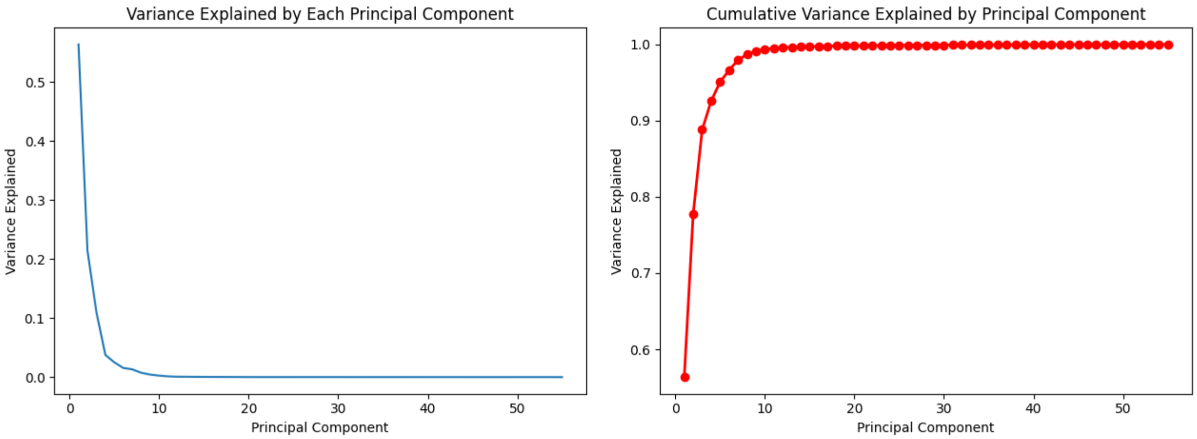


Figure 5: Dimensionality Analysis

Naturally, feature selection with Lasso arose as a possible choice, given our expectation of a fairly sparse distribution of predictors that actually are determinants of salary from our original feature set. We know that cross validation choice of lambda performs very poorly for feature selection, leading us to try an alternative method of choosing our regularization parameter, namely the Belloni-Chen-Chernozhukov-Hansen rule. In implementation, this is an iterative method that uses the residual from an initial fitting to calculate a final regularization parameter. Using this parameter, we ran Lasso, and we found only *Minutes Played Per Game* and *Points per 100 possessions* to be significant predictors of Salary at the 95% confidence level. Note that for this feature importance analysis we did not include the Black indicator, as the goal here was to find the most important important controls according to the data.

To complete our variable selection, we chose a few additional covariates which we believed might affect salary. Specifically, we included: net rating, age, assists, block, player height, an indicator for first round draft picks, and interactions between forwards and points as well as guards and points. These last two were chosen due to the fact that these two positions are expected to score more points. We wanted to identify first rounders to quantify ”hype” for a player, as both fans and coaches are usually more excited about first rounders, and that may affect salary. We included net rating, a measure of their team’s point differential per 100 possessions with them on the court, to capture the overall performance of a player, including intangibles. Finally, we included blocks, assists and age - these may obviously affect salary in some way (though they might be correlated with the rest, still).

2.3 Matching

The motivation for matching stems from the fact that our $x's$ are continuous and high-dimensional. The COC assumption does not hold in an OLS specification because we do not have Black and non Black players for each level of x

to compare salaries. Matching on p-score reduces the dimensionality of the covariates to a single number, avoiding biases generated by the curse of dimensionality, and upholding the COC assumption.

Under CIA and COC, the following assumptions hold, and the ATT is identified[1]:

$$Balancing\ Condition:\ D_i \perp x_i \,|\, p(x_i) = p(x) \forall\, x \tag{3}$$

$$p - score\ based\ CIA:\ Y(0)_i, Y(1)_i \perp D_i \,|\, p(x_i) = p(x) \forall\, x \tag{4}$$

The TE of interest:

$$ATT: E[Y(1)_i - Y(0)_i \,|\, D_i = 1] \tag{5}$$

The estimator:

$$A\hat{T}T = \frac{1}{N_1} \sum_{i=1}^{N_1} (Y_i - Y_{j(i)}) \tag{6}$$

$$(i, j(i))\, s.t. D_i = 1 - D_{j(i)} and\ j(i) = argmin_{k=N_1+1,\ldots,N} \left(p(x_i) - p(x_k)\right)^2$$

and N_1 is the sample size of Black players

Since our covariates potentially suffer from multi-collinearity, we chose to use Lasso with Cross-Validation to select covariates to estimate the propensity scores. The results are in Table 1.

Net rating	Age
Points	Assists
Blocks	Total rebounds
Minutes per game	Player height
First rounder	Points interacted with Forward
Points interacted with Guard	Total rebounds interacted with Forward

Table 1: Selected Covariates

We estimated the propensity scores using a logistic regression model with the selected covariates in Table 1.

$$P(Y \hat{=} 1|X) = \frac{e^{X'\hat{\beta}}}{1 + e^{X'\hat{\beta}}} \tag{7}$$

We binned the estimated propensity scores to determine the distribution of propensity scores. The results are in Table 2. The distribution of propensity scores is highly left skewed for Black players, so we consider trimming the sample in order to make Black and non Black players more comparable for matching.

The trimming procedure is as follows:

$$\max(\min(P_B, P_{NB})) \leq p - score \leq \min(\max(P_B, P_{NB}))$$

[1] Proven in Imbens paper cited below

Stratum	Black	Non_Black
(0,0.1]	0	4
(0.1,0.2]	0	4
(0.2,0.3]	2	5
(0.3,0.4]	2	4
(0.4,0.5]	3	3
(0.5,0.6]	4	5
(0.6,0.7]	13	0
(0.7,0.8]	13	9
(0.8,0.9]	29	5
(0.9,1]	71	3

Table 2: Full Sample P-score Binning

The trimmed sample propensity score distribution is in Table 3. After trimming, we lost 47 Black players, 10 non Black players, and propensity scores below 0.2. The propensity score distribution for Black players is still left-skewed, but not as extreme as before. We work with this sample to conduct matching based on propensity score.

Stratum	Black	Non-Black
(0.2,0.3]	2	3
(0.3,0.4]	2	4
(0.4,0.5]	3	3
(0.5,0.6]	4	5
(0.6,0.7]	13	0
(0.7,0.8]	13	9
(0.8,0.9]	29	5
(0.9,1]	24	3

Table 3: Trimmed P-score Binning

Using the 5 nearest neighbors, we matched based on propensity score, which resulted in the estimates in Table 4. Since the ATT is insignificant, we fail to reject the null hypothesis that there is no difference in the average salary between Black players compared to what their salary would have been had they been non Black.

ATT	0.034286
T-stat	0.01167
p.val	0.99069

Table 4: Matching Results

Matching on propensity score works if we believe that our covariates, or x's, include all observable confounders between race and salary. However, one example of an unobservable confounder between race and salary that we do not observe is player-coach or player-owner interactions. The attitude of a player can be assessed from such an interaction, which can later influence a player's salary. However, there could also exist an innate biases among owners or coaches that could weight the effect of such interactions on salary differently across race. In such a situation, we need to use a method to deal with potentially unobservable confounders.

2.4 Instrumental Variable

In order to loosen the CIA and address OVB at once, we employ an instrumental variable (IV). Specifically, we decided to use whether a player was born in a majority Black county as a proxy for being Black. Note that we define majority Black as greater than the national average of 13%. For international players, we use their nation's Black percentage. This is not problematic, as other nations (and especially those the NBA players come from) are much more homogeneous. In order for our IV to be valid, it needs to satisfy the 3 conditions below:

Exogeneity (an IA for the IV):

$$(y_i(0,0), y_i(0,1), y_i(1,0), y_i(1,1), D_{Oi}, D_{1i}) \perp z_i \tag{8}$$

Exclusion:

$$y_i(d,0) = y_i(d,1) \quad \forall d \in \{0,1\} \tag{9}$$

Relevance:

$$E[D_{1i} - D_{0i}] \neq 0 \tag{10}$$

Since being born in a majority Black county is clearly random, it can be argued that it satisfies the first two conditions:

- Exogeneity: Since the majority Black variable is random and does not determine a player's physical attributes, it is unlikely that it is correlated with any non-included variable. Since we control for performance metrics, any non-included variables are less tangible. One example could be player attitude, as this may both be correlated with being born in a majority Black town and affect player salary. While this might be the case, any effect someone's county might have had on their attitude will likely have been flatted by the time they get to the NBA: to get to the NBA, you have to have a certain attitude (e.g. being hard working, resilient, personality ...), so within the NBA these attributes are likely fairly random. Another excluded variable someone's may be correlated with is player popularity. This variable can clearly also affect salary. However, we argue that a player's county does not affect their popularity much. Since we argued that player's personalities are fairly random within the NBA, a player's county can't be correlated with popularity through this avenue. Thus, the only avenue is fans supporting a player only due to them being born in their county. While there are certainly a few fans for which this is the case, this likely does not impact player popularity in any significant way. Therefore, we don't have to worry about the correlation between a player's county and their popularity. Note that both the above arguments also easily extend to player countries, for the international players.
- Exclusion: Being born in a majority Black county likely only affects salary through being Black, and does not affect salary directly. This is fairly easy to argue, as it is very unlikely that owners choose salary by considering where players were born.

To test relevance, we simply run a Pearson correlation test, and test the significance of the Black variable in the first stage of the regression. As can be seen in Table 5, the Pearson correlation coefficient, which tests the null hypothesis that two variables have no linear relationship between each other, is found to be very significant. This is a good sign, so we continued with the first stage regression. As can be seen in Table 6, the model is found to be highly significant in explaining the majority Black variable, which is another strong signal for the relevance of our instrument.

IV and Black Pearson Correlation	P-Value
0.2139	0.0037

Table 5: Correlation Between IV and Black

Variable	Coefficient
const	0.000240
Black	0.003728

Table 6: First Stage Results

Parameter	Value
P-Value	0.3085
Black Coefficient	-9.9643

Table 7: Second Stage Results

The two stage regression was then ran, using the controls chosen in the variable selection, and the majority Black variable as the instrumental variable. Looking at the results of the two stage linear regression in Table 6 (with the controls omitted for conciseness) we find the P-value to be insignificant. Note that 2 stage least squares has the following asymptotic variance:

$$\hat{\theta}_n^{IV} \rightarrow_p \theta$$
$$\sqrt{n}(\hat{\theta}_n^{IV} - \theta) \rightarrow_d N(0, V)$$

And therefore, this p-value is a theoretically sound one. Given these results, we thus cannot reject the null that being Black does not affect salary in the NBA.

2.5 Binning

As the instrumental variable still found the race coefficient to be insignificant, we explored the possibility that race might affect salary through other avenues. To do so, we ran a regression interacting the Black dummy with the covariates chosen in the variable selection section. In addition, we added 2 point shooting percentage and 3 point shooting percentage: there is a well know stereotype that White people are better shooters, so we wanted to see if that had any effect on salary. Below are the results that were found significant:

Variable	P Value
MPG	2.83×10^{-8}
PTS	7.89×10^{-5}
Black 3P%	0.015

Variable	Coefficient
MPG	1.056
PTS	0.777
Black 3P%	-0.364

Table 8: Regression results

Unsurprisingly, MPG and PTS per 100 possessions remain very significant. Much more interestingly, the interaction term between 3 point shooting percentage and the Black dummy was found to be significant. This result cannot be interepreted as causal, but is an interesting effect. An increase of 1% in 3 point shooting percentage is associated with a smaller effect on wage for Black people, as compared to non Black people (specifically a decrease of $340,000). This might be due to the stereotype: owners notice and reward non Black people more than Black people, for good shooting. Nonetheless, this result is purely observational, since we are not doing causal inference.

3 Prediction Models

3.1 Models

After our finding of insignificance on our race beta, our focus shifted toward prediction - what actually are determinants of salary, and consequently, what are the mechanisms through which race would likely affect salary? We decided on three models, each with their own justification.

1. An OLS Regression with Greedy Forward Variable Selection : This serves as our baseline model. It is perhaps the simplest model we could have chosen, which we will compare features selected, predictive performance, and overall fit with the other two models specified.
2. A Lasso Regression with BCCH Chosen Lambda: This model actually came first in our analysis. Lasso was a very intuitive model choice from the start. For one thing, Lasso is very robust to multicollinearity, which we absolutely expected when dealing with multiple variables pertaining to the same player. Ridge was another option that we decided against given that it does not really select for features, and our ultimate goal was to understand which features were most important.
3. A Random Forest Regression: This was a natural decision after fitting an OLS and Lasso. Although our feature set initially began with included interaction terms, we felt that perhaps all the nonlinearities that could explain the model were not captured, and therefore we wanted to try a non-parametric model. A random forest was a natural choice - splines and kernels were out of the question given the many features and relatively small dataset.

We encountered many other possible ideas to explore but eventually settled on these as a sort of gradient from simple and interpretable to more complex. Another model we considered were some gradient boosting trees. However, after fitting the data on a single tree, we found high variance and low bias, so Random Forests to reduce variance seemed like the obvious choice.

3.2 Results

3.2.1 OLS with Greedy Forward

In the same order as above, we will go through, show results, and interpret them. First, our OLS Regression with Greedy Forward Variable Selection.

From our aforementioned variable set, exactly two variables were chosen with the Greedy Forward Algorithm: Points per 100 Possessions (PTS) and Minutes per Game (MPG). On the front of predictive accuracy, we can turn to our inference statistics.

Coef	Value	Std Err	P-Value
const	-26.8289	4.259	0.000
MPG	1.0015	0.223	0.000
PTS	0.7595	0.202	0.001

Table 9: Coefficient Information: OLS

We have a fairly high R^2 value of 0.777 with an adjusted R-squared of 0.764. We include this as our main descriptor because MSE, AIC, and BIC do not carry as much interpretive power. For now, this can be our information statistic measure of fit. All the chosen coefficients are, as expected, significant, with relatively tame standard errors, giving us confidence that these results are actually worth comparing to. Note the actual values of the coefficients, which we will compare in passing when we look at the results of our Lasso regression.

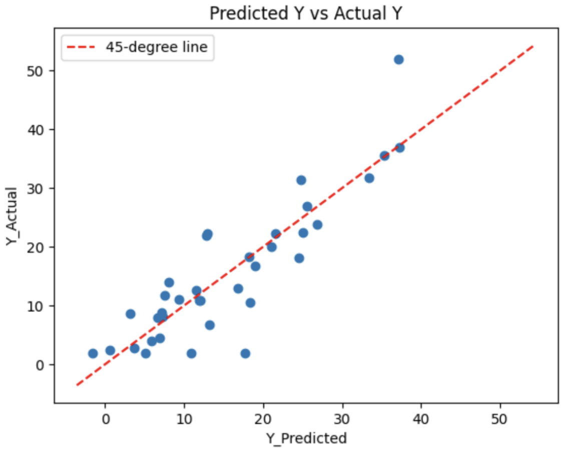


Figure 6: OLS Predictive Plot Out of Sample

Overall, it looks like our fit performs quite well out of sample. Visualizing this will help us compare our out of sample performance of the three models.

3.2.2 Lasso with BCCH Regularization

In a very interesting revelation, we get that the Lasso with BCCH regularization ends up choosing the exact same two predictive covariates - PTS and MPG. Intuitively, we were expecting that more features would be included in the OLS with greedy forward than our Lasso. However, getting the same feature subset serves as a corroboration that these are the predictors that will significantly explain the variance of our data, at least under a linear regression framework with its assumption. However, the coefficients themselves are quite different - after finding that our regularization parameter chosen was relatively high given the scaling of the data, we found that the Lasso regularization is likely biasing our coefficients down, leading to poorer out of sample performance as we will see. Instead, we turn to the Lasso as the main way to feature select, rather than necessarily as a predictive model. Since we scaled our data, there is no constant term as well in this Lasso regression.

Coef	Value	P-Value
PTS	0.197333	0.000
MPG	0.215505	0.000

Table 10: Coefficient Information: Lasso

We get an adjusted R^2 of 0.6001, which is significantly lower than its OLS counterpart. This model's strictly worse predictive power is even more pronounced when considering a plot of fitted vs actual.

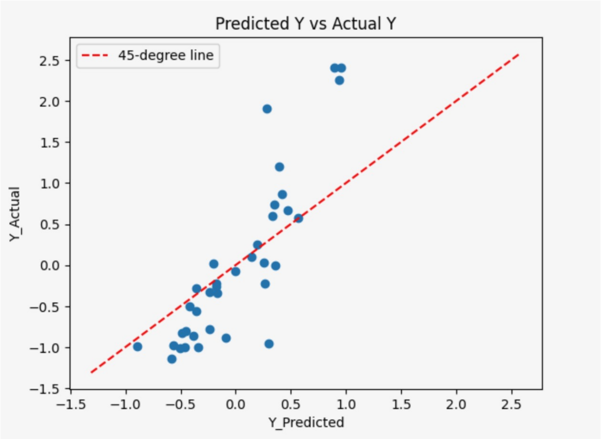


Figure 7: Lasso Predictive Plot Out of Sample

3.2.3 Random Forest

Now, our most granular model, which we hoped would capture the non-linearities that are not immediately obvious or included in our intuitive additions of interaction terms. The Random Forest did in fact select a larger subset of features, while PTS and MPG still remained part of that set. The total set of features includes:

PTS, BLK, TOV, lnExperience, MPG

where PTS, MPG are the same as before, but BLK represents blocks, TOV represents turnovers, and lnExperience is a log transformation of experience (which is a variable we created with the idea that subsequent years become less and less consequential in building salary). It seems that the random forest sees some potential nonlinearities that could add to the predictive power of our model. We say these are nonlinearities purely because they are added variables outside of Lasso and OLS, which means if these variables are selected they likely do not have a strict linear relationship with our variable of interest, salary. Now, to see how this Random Forest actually does:

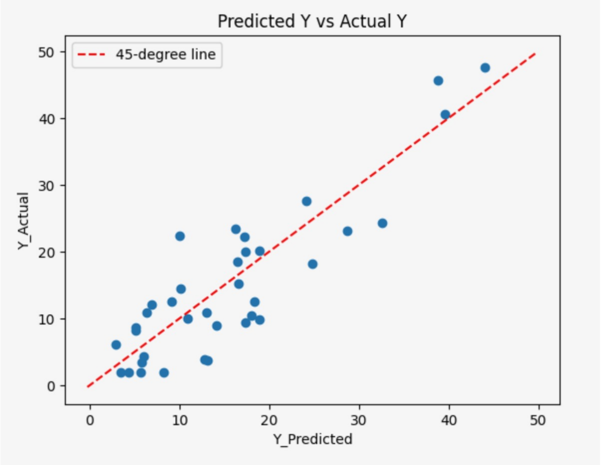


Figure 8: RF Predictive Plot Out of Sample

We can see that it looks quite good visually. In terms of R^2, we get an adjusted value of 0.778 out of sample, and 0.965 in sample. This fairly large gap in R^2 may be attributed to an overfitting problem. However, the NBA has salary caps and minimums, as well as a rough outline for salary progression set by precedent, which means our out of sample should generally be quite similar to our in sample. Overall, we see that a RF does not perform significantly better than our OLS out of sample, leading us to question the additional covariates that it found to be significant.

Our results here with the three models served as a basis for our exploration of binning. Our results also suggest that the actual determinants of salary are a much smaller subset than much of the prior literature found. Is the NBA really a pure meritocracy, basing salary purely off of your points and your time on the court?

4 Conclusion

We reached two conclusions throughout the course of the study. The first conclusion was in opposition to our initial thoughts coming into the study. We initially thought that previous studies such as the one conducted by Johnson had too few controls in their regressions. For one example, we thought that interaction terms, such as ”Rebounds per 100 Possessions” × ”Center”, would be significant in a regression against player salary as we hypothesized that players in certain positions are paid to do certain things. However, this did not end up being the case. Across different methods such as Double LASSO with Belloni-Chen-Chernozhukov-Hansen rule and Lasso with greedy-forward variable selection, our regressions consistently landed on ”Minutes per Game” and ”Points per 100 Possessions” as the only significant controls toward salary after regularization. This is a much smaller set of controls than what Johnson or other previous papers used, potentially indicating that other authors overfit their salary models and reached biased conclusions as a result. Despite only using these two variables, our models were able to give salary predictions with high accuracy. While this is surprising it makes sense: basketball is a points based game. Players who stay on the court for longer (i.e. can be relied on for long periods of time) and score more, will simply be more valuable players. While there are more facets to evaluating how ”good” a player is, points and time on the court should be but the most important ones.

Our second conclusion was in relation to our null hypothesis. In our null, we hypothesized that race does not play a factor in NBA salaries in the current day. Through our causal inference methods, we are not able to reject this null. When we assumed the conditional independence assumption (CIA)and estimated CATE with p-score matching, we did not find the result to be significant. Further, even when we tried to address omitted variable bias as well as get rid of the CIA assumption through an instrumental variable, we still found an insignificant effect of race on wage. Therefore, we are unable to reject the null. In searching for the effect of race through other avenues, we found the effect of a 3 point shooting percentage to be different between Black and non Black individuals. Specifically, Black individuals seemed to benefit less from an increase in 3 point shooting percentage. Due to the observational nature of the study, this can’t be said to be a causal effect, but this is an interesting finding nonetheless. Further, by regressing the subsets of Black and non Black people separately in the data, we found through a binned analysis of the two significant controls indicated by BCCH LASSO that we could not reject the null hypothesis as the confidence intervals for the coefficients on the controls for each binned group (Black and White) overlapped. However, the coefficients themselves on the ”Points per 100 Possessions” control were separated by more than a standard deviation, indicating that more analysis into wage discrimination in the NBA could be fruitful.

5 Next Steps

There are a few next steps that this research lends itself to. First and foremost, we would like to verify our own work on a larger dataset. To do so, we might utilize each player’s salary in different years as different data points, with a corresponding control vector of that player’s career statistics up to the year that the player’s salary is being pulled from. Concretely, this would mean that Lebron James’s 2022-23 salary and career statistics up to 2021-22 would be treated as a separate data point as James’s 2018-19 salary and career statistics up to 2017-18. By doing so, we would hope to increase the amount of data at our disposal and thus decrease the variance in our estimates. We would also like to see if certain trends have developed over time; specifically, if certain statistics have gained more relevance as time has passed than others.

Secondly, we would like to explore what exactly has led to the discrepancy in the coefficient for ”Points per 100 Possessions” between Black and White players. As Johnson brings up in his paper, biases of this nature can usually be traced back to a reason embedded in consumer preferences. Thus, one potential avenue of exploration may be to see if consumer preferences or demographics have changed dramatically for NBA viewers and fans since the 1995-1996 season, or up to when Johnson took data for his paper. Another avenue would be to see if the difference in coefficients is made up for via a different statistic. It is a common stereotype that White players in the NBA are ”shooters”; this stereotype may lead to other statistics such as ”3 Point Percentage” contributing in a significantly different way to White player salaries than Black player salaries.

Lastly, we’d like to see how different our regressions look after accounting for player popularity. In today’s age of social media, likes, follows, and retweets, player popularity is more commoditized than ever. Therefore, we would like to see how this shifting trend may have impacted salaries, if at all.

6 Works Cited

Basketball-Reference.com. (2023, November 25). NBA ABA League Index. [Online]. Available: https://www.basketball-reference.com/leagues/

CIA World Factbook. (2021). Ethnic groups. [Online]. Available: https://www.cia.gov/the-world-factbook/field/ethnic-groups/

Imbens, G. W. (2004). Matching Methods in Practice: Three Examples. The Journal of Human Resources.

The Sport Journal. (2010). Determinants of NBA Player Salaries by Evan Plous Kresch and Daniel A. Rascher. [Online]. Available: https://thesportjournal.org/article/determinants-of-nba-player-salaries/

Wikipedia contributors. (2024, March 8). National Basketball Association. In Wikipedia, The Free Encyclopedia. Retrieved March 8, 2024, from https://en.wikipedia.org/wiki/National_Basketball_Association

University of Chicago Press Journals. (1992). The Journal of Political Economy, Volume 100, Issue 6. [Online]. Available: https://www.journals.uchicago.edu/doi/abs/10.1086/298174

Villanova University Charles Widger School of Law. (1991). The Villanova Sports and Entertainment Law Journal, Volume 1, Issue 2. [Online]. Available: https://digitalcommons.law.villanova.edu/cgi/viewcontent.cgi?article=1219context=mslj

Wiley Online Library. (1991). Journal of Applied Corporate Finance, Volume 4, Issue 4. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1536-7150.1991.tb02300.x

Kaggle. (2023, October 13). NBA Players. [Online]. Available: https://www.kaggle.com/datasets/justinas/nba-players-data

Kaggle. (2023, December 15). US County Data by evangambit - Kaggle Dataset [Online]. Available: https://www.kaggle.com/datasets/evangambit/us-county-data/data