Charles University

Faculty of Social Sciences
Institute of Economic Studies

BACHELOR THESIS

An Empirical Investigation of Wage Discrimination in Professional Football

Author: **Jakub Blaha**

Supervisor: **Ing. David Kocourek**

Academic Year: **2016/2017**

Declaration of Authorship

The author hereby declares that he compiled this thesis independently, using only the listed resources and literature.

The author grants to Charles University permission to reproduce and to distribute copies of this thesis document in whole or in part.

Prague, July 26, 2017

Signature

Acknowledgments

Firstly, I would like to express my sincere gratitude to my supervisor Ing. David Kocourek for his patience, motivation and expert knowledge. His guidance helped during the whole duration of writing this thesis.

Besides my supervisor, I would like to thank to my parents, Renata and Stanislav, and the rest of my family for the continuous mental and financial support.

Abstract

Salary discrimination is a phenomenon that arises from ineffective behaviour of economic subjects. Even though its presence is incompatible with the theory of profit maximization, salary inequality still persists in the human society. Nevertheless, the investigation of this topic has been largely unheeded in the environment of professional football. In our empirical research, we use the most recent data to investigate the salary gap between white, African American and Hispanic players in the American Major League Soccer. Besides ordinary least squares method that focuses on the impact of ethnicity for the average player, we adopted the method of quantile regression to reveal wage gap between players with below-average pays. Observing each player's performance for 3 seasons, we uncovered salary discrimination against African Americans and Hispanics in the lowest decile of the salary distribution that amounts to 18.9% and 15.3%, respectively. Furthermore, we utilized the difference-in-differences (DID) estimator to find no effect of the increasing level of invested money on the wage gap.

JEL Classification	J30, Z20, Z21, J71 J31 J15
Keywords	discrimination, race inequality, football, quantile regression, OLS, wages, racism

Author's e-mail	kubablaha@seznam.cz
Supervisor's e-mail	kocourek.david@email.cz

Abstrakt

Platová diskriminace je jev, který je výsledkem neefektivního jednání ekonomických subjektů. Ačkoli je její přítomnost neslučitelná s teorií maximalizace zisku, platová nerovnost i přesto nadále přetrvává v lidské společnosti. Nicméně, pokud se jedná o prostředí profesionálního fotbalu, otázce platové diskriminace bývá zřídkakdy věnována pozornost. V naší studii používáme nejnovější data, abychom vyšetřili rozdíly v platech mezi bílými, afroamerickými a latinskoamerickými hráči v americké Major League Soccer. Kromě metody nejmenších čtverců, která se zaměřuje na vliv rasy pro průměrného hráče, jsme použili kvantilovou regresi, abychom odhalili rozdíly v platu pro hráče s podprůměrnou mzdou. Sledováním statistik každého hráče po dobu 3 sezón jsme objevili platovou diskriminaci vůči Afroameričanům a hráčům z Latinské Ameriky v nejnižším decilu platového rozdělení, která vyšplhala k 18.9% a 15.3% ve prospěch bílých hráčů. Mimoto jsme použili metodu difference-in-differences (DID), abychom ověřili, že stoupající výše investovaných peněz neměla na rozdíl v platech napříč rasami žádný vliv.

Klasifikace JEL	J30, Z20, Z21, J71 J31 J15
Klíčová slova	diskriminace, rasová nerovnost, fotbal, kvantilová regrese, OLS, platy, rasismus

E-mail autora	kubablaha@seznam.cz
E-mail vedoucího práce	kocourek.david@email.cz

Contents

List of Tables

List of Figures

Acronyms

FIFA	Fédération Internationale de Football Association
MLS	Major League Soccer
OLS	Ordinary Least Squares
IFAB	International Football Association Board
GSSS	Global Sports Salary Survey
MLB	Major League Baseball
US	United States
QR	Quantile Regression
NHL	National Hockey League
NBA	National Basketball Association
FFP	Financial Fair Play Regulations
DID	Difference-in-differences
GAM	General Allocation Money

Bachelor Thesis Proposal

Author	Jakub Blaha
Supervisor	Ing. David Kocourek
Proposed topic	An Empirical Investigation of Wage Discrimination in Professional Football

Topic characteristics: To eradicate discrimination, intolerance and racism from American soccer, Major League Soccer (MLS) joined Fédération Internationale de Football Association (FIFA) in the campaign called ”Say no to racism” (Mlssoccer.com 2010). On the occasion of Say No to Racism Day, MLS stadiums incorporated field boards with “Say no to racism” signs. The Commissioner of MLS Don Garber commented on this event as follows: *”With players born in 44 different countries, our league reflects the broad range of ethnicities that love the game of soccer and that live in the United States.”* Then he added: *”Major League Soccer has a no-tolerance policy toward racism. The league prohibits and will not tolerate any discrimination or harassment on the basis of race.”* After this statement, one might wonder if club owners and football managers also comprehend the MLS's fight against racial inequality.

In our thesis we would like to examine whether there is a presence of racial discrimination among footballers from top American clubs and therefore to test the effectiveness of MLS's approach. As a measure of discrimination in American soccer, we investigate the relationship between the race and the footballer's income. We also take into account the fact that MLS eased the policy of salary cap and established several policies such as the Designated Player Rule which guarantees that the wages of the most productive players are solely determined by the market forces. We incorporate this system shocks into our examining so that the results obtained from our research would help us to answer the following **research questions**:

1. Do teams in Major League Soccer pay less to African American and Hispanic players?
2. Is there a pay premium for white players in any part of the salary distribution?
3. Did the change in wage policies worsen or improve the position of Hispanics and African Americans in terms of wages?
4. Are players racially discriminated across different positions?

Methodology This thesis uses econometric approach which is called Pooled Cross Sections to estimate the model with factors influencing footballers' income. To estimate Pooled Cross Sections, we utilize the ordinary least squares estimator, which is the most efficient one under given assumptions that need to be satisfied. In addition, we incorporated the quantile regression to be able to obtain estimates at different quantiles of players' salary.

Outline

1. Introduction
2. Theory and Literature Review
3. Methodology and Crucial Terms
4. Dataset Construction and League Overview
5. Model
6. Results
7. Conclusion

Core bibliography

1. Arrow, Kenneth J. (1998): "What Has Economics to Say About Racial Discrimination?." *Journal of Economic Perspectives)* pp. 91-100.
2. Fryer, Roland G (2010): "Racial inequality in the 21^{st} century: The declining significance of discrimination." *National bureau of economic research - Harvard University*
3. KAHN, Lawrence M & K. LANG (2000): "The Sports Business as a Labor Market Laboratory." *Journal of Economic Perspectives. American Economic Association*
4. LEHMANN, Jee-Yeon K. & K. LANG (2011): "Racial Discrimination in the Labor Market: Theory and Empirics." *Journal of Economic Literature* pp. 959- 1006.
5. WOOLDRIDGE, Jeffrey M., (2006): "Introductory econometrics: a modern approach. ." *3rd ed. Mason: Thomson/South-Western*

Author

Supervisor

Chapter 1

Introduction

”The only difference between man and man all the world over is one of degree, and not of kind... Where is the cause for anger, envy or discrimination?”

Mahatma Gandhi

Association football, commonly known as football[1], is a sport where each of both teams consist of 11 players and the main objective of each side is simple - score as many goals as possible to the opponent in order to win the game. Not surprisingly, it is the most popular sport in the entire world (Dunning *et al.* 1999) as there are approximately 265 million registered players (Fifa.com 2007) in 200 dependencies. The rules of the game were determined by International Football Association Board (IFAB), which was established in 1886 and ever since, the money involved in this sport has skyrocketed.

Nonetheless, the labour market of football players is nowadays a highly-competitive one since out of the millions of football players only a fraction can make a living by playing football. The increasing interest of the media and the leading sports companies caused that the money invested in it has increased exponentially as football was more and more professionalised. This phenomenon can be seen for example in terms of wages of footballers which have augmented dramatically. For comparison, in 1901 the Football League in England imposed a restriction that no player would be paid more than £4-a-week[2] for his performance. According to Sportingintelligence.com (2016), though, the average

[1] Even though the name ”football” might signify American football, we decided to use this word throughout this paper as it is more widespread in Europe.

[2] If we adjust for inflation, this value would yield approximately £443.14 in today´s money (May, 2017).

Premier League player earned approximately £49,211 a week.

From an employer's point of view, the market of football players is a specific type of labour market. On the one hand, it differs from other working environments in the possibility to perfectly observe worker's performance using statistical data. On the other hand, it presumably shares some similar characteristics with other types of markets with employees, such as observable elements of supply and demand (Rosen & Sanderson 2001), which need to be taken into consideration. Owing to the publication of the Civil Rights Act of 1964, the most discussed issue in neoclassical era is the fact that different groups of employees, be they skilled or not, receive different wages (Wells *et al.* 2001). This treatment of making a distinction against a person based on their age, race or gender is called discrimination and its presence provoked response in all kinds of social affairs (Arrow 1998). Since it represents a specific sort of economic inefficiencies, it is often a subject of economic academic literature. Being perceived as socially unacceptable and hostile, racial discrimination has been probably one of the most considered forms due to the fact that various policies have been made to remove the difference in dissimilar treatment through races. The motivation for these policies has been here since the colonial era, because European Americans have been given exclusive privileges in terms of voting rights, education, citizenship and most importantly for our study, better working conditions including higher wages (Sears 1988). Nevertheless, the definition of discrimination states that there is no presence of disriminatory behaviour if the persons observed do not have the same productivity. However, productivity arises from such factors as education or training that have been provided in a different way to people based on their race. However, the professional football is a very unique case of labour market. First of all, this sports industry is under general public scrutiny and players should be treated the same in terms of career development. Second, clubs are disclosed to the performance of their employees as they are provided with detailed game statistics. Hence, football clubs have complete knowledge about the productivity of their employees. In connection with this fact, wage discrimination should not be enabled and wage differences should only reflect the importance of a player for the team. Thus, if salary discrimination could be detected in such a competitive, visible and scrutinized industry, the results of the research might encourage academics to study discrimination also in other areas.

One of the reasons for clubs to have nondiscriminatory behaviour is the fact that they can be considered as profit maximisers (Zimbalist 2003), and therefore they should set the marginal revenue equal to the marginal cost in order to attain the highest profit. Speaking in general terms, the worker should receive the very same amount of money as he is capable of earning for his employer. The question that might arise is whether the clubs remunerate their players justly based on their perfomance or they use prejudices against, or in favour of players with non-white origins. There have been many studies examining the relationship between race and compensation in sport, namely in professional basketball and baseball, but only a few were devoted to football due to the unavailability of data. Szymanski (2000) uses wage expenditure and performance statistics in order to test the presence of racial inequality among top English clubs. Discovering positive ceteris paribus effect of the proportion of black players employed on the team performance, he proposed a discriminatory behaviour from some team owners who preferred purchasing white players in spite of worse results. A direct effect of race on player's compensation in National Basketball Association was investigated by Hamilton (1997). Once controlled for performance, he devoted his research to salary differential and compared the situation in mid-1980s and 1990s. Controlling for players' characteristics, he used OLS and Tobit regressions to come to the conclusion that 20% premium paid to white players in 1985 were removed after one decade. However, once examining the same dataset by censored quantile regression, he discovered two curiosities. In the lower end of the dataset, the whites earned substantially less than their black counterparts. But, the premium of 18% was received by whites in the upper end of the distribution. The outcome of his study is very unexpected as discriminating better and expensive players tends to be more costly for the employer. His results were reconfirmed by the paper of Kahn & Shah (2005) who also found wage discrimination against non-white players. These findings lead to a question, whether wages of white and non-white players also differ in football - the most followed and commercialized sport in the world (Mueller *et al.* 1996). Even though Major League Soccer would appear as an improbable place to detect racial discrimination, no study has been focused on this topic in recent years.

Our thesis aims to clarify two ambiguities. First of all, we investigate the relationship between players' income and race in Major League Soccer, which is the only national league that completely discloses the information about players' salaries[3] and at the same time we utilize the American data that was used in the majority of researches on this topic. We are highly inspired by the most recent studies and besides classical OLS method, we also adopt the quantile regression analysis to be able to investigate racial discrimination in different quantiles of the salary. Aditionally, we aim to find out the effect of several salary policies that were introduced to augment the quality of the league. As an example we mention the Designated Player Rule that allowed MLS clubs to sign players who would otherwise be considered above their salary cap. This approach will provide us with a knowledge, whether or not such a system shock caused a wage gap between Hispanic, African American and white players. The structure of the thesis looks as follows:

Chapter 2 summarizes the literature written about the racial inequality in sport and other labour markets. In addition, we provide the theory standing behind the inequity and intolerance in working process - crucial conception for our thesis. The forms of discrimination are mentioned as well. Chapter 3 provides an introduction to pooled independently cross-sections data estimation and explains the purpose of the Difference-in-Differences estimator. Last but not least, we present the quantile regression to be able to examine individual quantiles of the salary distribution. Chapter 4 deals with the dataset construction and also introduces plenty of descriptive statistics so that the reader has better notion about the distribution of the data. Moreover, it covers the description of the Major League Soccer which is the field of our exploration. Chapter 5 describes the model used for answering the research questions. The discussion of results from our regressions is presented in Chapter 6. We also find it necessary to point out some shortcomings of the investigation. Finally, Chapter 7 summarizes our findings and acts as a spur for further research.

[3] As of June 2017, the data was available at: mlsplayers.org

Chapter 2

Theory and Literature Review

This chapter explains the theory standing behind the economics of discrimination. In the first part, we define what discrimination is, we also discuss its forms and consequences. In the second part we introduce the principle of the wage gap and two types of wage discrimination. In the third part, the reader is informed about the literature written about this phenomenon and we also present the results of previous studies in sports industry.

2.1 Discrimination

The word discrimination comes from the latin verb *discrimire* which signifies "to separate, to make a distinction" (Oxford University 2017). It can therefore be assumed that discrimination is a treatment of making a distinction against, or in favor based on the group where the given person pertains. This act significantly contradicts essential principles of human rights and represents one of the strongest economic inefficiencies. There are different types of discrimination, though. Unequal treatment can be aimed at persons of different age, race, religion or nationality. Since the main objective of the study is dissimilar treatment based on the person's ethinicity, this thesis primarily concerns racial discrimination and its direct impacts.

The key assumption when considering a different treatment by employers based on the employees' race is that we should observe people of equal productivity (Becker 1971). This condition is hardly measured through labour market as employers seldom have perfect information about the performance of their employees and this limitation markedly reduces the possibilities for potential investigations. There are still some researches dealing with employment discrimination which can be measured, for example, at the moment of initial hire. One of them was introduced by Pager *et al.* (2009) who found patterns between the probability of being hired and demographic characteristics in the US. According to the research he made, Hispanic and black applicants with clean backgrounds had approximately the same chances of being hired as white applicants recently released from prison. Interpersonal skills of applicants are arguable but for the sake of justifiability, the candidates for job offers were given identical résumés.

However, the crucial statement that should be noted is that the environment of professional sport has one special feature which distinguishes it from the rest of the labour market. The performance and productivity of its participants are measurable since the statistics of each individual are observed and can be subsequently used for research studies. Thanks to this, we are able to obtain all data needed to determine whether the difference in terms of wages also exists between African American, Hispanic and white footballers. But first, we find it necessary to provide some key facts about the situation in the American market with workers as racism has been rooted in the American society since the establishment of the country (Takaki 1979). Hence, we find it very important for reader's perception about the meaning of our research.

2.2 Racial Discrimination

Racial discrimination remains present in every aspect of human society (Arrow 1998). Fortunately, in the course of 20^{th} century, life of the blacks improved dramatically. The difference in unequal treatment in United States was intended to be abolished in 1964 when Civil Rights Act was enforced. It ended racial separation in schools, prohibition on voting and most importantly, worse conditions at the workplace. Besides, the US government implemented a policy which is known as affirmative action to assure that groups like African Americans, who had been a subject to discrimination for ages, would be accepted more often when applying for a job (Feinberg 2003). Not surprisingly, these political acts motivated plenty of researchers to scrutinize racism and its consequences in all areas of labour market (Cain 1986).

Even though racial inequality is not as prevalent nowadays as it was earlier in the past, its continous existence still can be proved in many aspects of social affairs. Most importantly, racial discrimination can be also found in economic aspects, namely, in terms of income, life expectancy or hiring standards. In the United States, blacks earn 24% less compared to whites and live on average 5 fewer years. Concerning Hispanics, they earn 25% less than whites holding other factors fixed (Fryer Jr 2010). This empirical fact inspired many to examine conditions in professional team sports, especially in North America. Although racial inequality produces a certain opportunity cost, it still exists in many sports industries (see section 2.4). However, before we introduce the findings of racial discrimination in sport, the section 2.3 explains two behaviours that can lead to the race pay gap.

2.3 Wage Differential

Wage discrimination in the labour market occurs if, after controlling for special characteristics such as experience or education, the coefficient on race, gender or age is negative and statistically significant. One of the first studies that took wage gap into consideration was introduced by Oaxaca (1973). He inquired into the wage differential between genders and he clarified that male-female wage difference is not caused by discrimination, but it rather stems from the variance in productivity. The most important limitation of his study lies in the fact that he did not have an adequate amount of data at his disposal. He was consequently followed by researchers who introduced various extensions in order to provide results of higher importance. Subsequent studies included also time dimension into their models allowing them to measure returns to skills, and how they have been changing the wage gap over time (Blau & Kahn 2016). This technique identifies whether wage difference among people of different ethnicity has widened or narrowed, and if the factors that are hidden in the racial dummy variable, have augmented or lowered their effects over time.

As the studies dealing with racial discrimination have been accumulating, literature has been divided into two branches. The first is statistical and it relies on the assumption that the employer would lean his decision on the group averages if they did not have perfect information about employees. Taste-based discrimination, by contrast, is based on the prejudice of decision-maker. The second appears to be more likely our case since we are able to observe the performance of each player. Nevertheless, we explain the concepts of both of them.

2.3.1 Stastistical Discrimination

The pioneers of this form of discrimination were Edmund Phelps and Kenneth Arrow (Fang & Moro 2010). The key assumption of this type of discrimination is that the employer evaluates their workers without having complete knowledge about their productivity. The only characterictics that can be observed by the employer are gender, age, experience, ethnicity or race. Therefore, workers can disclose their skills only by the form of résumé and partly during the process of interview. Otherwise, unobservable characterictics are presumed to be correlated with observable features and workers are classified accordingly (Arrow 1971). Statistical discrimination occurs when employers treat potential employees based on the group where they belong (Lang *et al.* 2012). The question which might arise is whether the workforce productivity varies across races. Altonji & Pierret (2001) came to the conclusion that race is a significant variable influencing productivity of workers. The evidence of statistical discrimination was found and moreover, they clarified that the more information the employer obtains about their workers, the less significant easily observable variables become as determinants of workers' earnings. A model of statistical discrimination has been used by Lang & Manove (2011) who examined a wage differential between whites and blacks. As workers' ability was unobserved, they used the score on Armed Forces Qualification Test as a proxy for their ability. Their study vindicated that the wage difference between white and black workers is equal to the return of additional year of education, holding other factors fixed.

2.3.2 Taste-based Discrimination

Principles of this theory were presented by Becker (1971). The crucial assumption for his model is that there is perfect competitiveness among firms. Furthermore, we need to assume that employers are perfectly informed about workers' race. While the latter condition is undoubtedly satisfied in professional football environment, the first mentioned might be debatable in the case of Major League Soccer, since clubs following discriminatory behaviour, and therefore accruing negative profits created by discriminating, cannot so easily leave the market in the long-run.

According to this theory, the utility of the employer is dependent on profit (π) and on the amount of non-white workers (L_b). The former mentioned influences the utility function positively whereas the latter negatively. As each employer has different taste for prejudices, coefficient d_p states the relationship between utility and number of prejudiced workers.

$$u_e = \pi - d_p \cdot L_b \tag{2.1}$$

Equation 2.1 presumes that both whites and blacks have equal marginal productivity implying that employers avoid cooperation with a specific group of employees with the same characteristics no matter how productive they are. The profit function therefore looks as follows:

$$\pi = f(L_w + L_b) - w_w \cdot L_w - w_b \cdot L_b \tag{2.2}$$

where w_w and w_b stand for wages for black and white workers, L_w is an amount of white workers and f is a production function. Microeconomics suggests that employers would choose the number of white workers L_w, and the number of black workers L_b that maximize the utility function:

$$u_e = f(L_w + L_b) - w_w \cdot L_w - w_b \cdot L_b - d_p \cdot L_b$$

These values L_b^* and L_w^* must satisfy the conditions given below, so-called first-order conditions.

$$f'(L_w^* + L_b^*) - w_w = 0, L_w^* > 0 \tag{2.3}$$

$$f'(L_w^* + L_b^*) - w_b - d_p = 0, L_b^* > 0 \tag{2.4}$$

These two conditions imply that employers are willing to hire new labour force until the moment, when the marginal cost and the marginal product are in equality. In case of white workers, the marginal cost is only w_w. However, in case of black workers, the marginal cost for employer is their wage w_b, added to the discrimination coefficient d_p. Given the assumptions we have stated earlier, solely white workers would be hired if $w_w - w_b < d_p$. On the other hand, only blacks would be employed if $w_w - w_b > d_p$. We can easily deduce that the market equilibrium occurs if

$$w_w^* = w_b^* + d_p^* \tag{2.5}$$

where w_w^* and w_b^* are optimum wages. In the equilibrium, discrimination coefficient d_p^* of the marginal discriminator would be equal to the wage differential between black and white workers. We can therefore conclude that the wage differential is positively correlated with the prejudice of the marginal employer and its magnitude also depends on his taste for discrimination.

2.4 Literature Review

The problem of racial discrimination has motivated plenty of economists to investigate this social discrepancy. Gary Becker (1971) wrote his paper *The Economics of Discrimination* where he divided discrimination into three main types - employer, co-worker and customer. Each one has its name after the group which tends to pursue discriminatory behaviour. He asserted that firms that consider their workers as equal, and treat them so, have competitive advantage against the firms that discriminate, as they do not have to cope with additional costs caused by unequal treatment. Discriminating firms are therefore forced to leave the market because of inefficiency. On the other hand, avoiding discrimination among co-workers should result in equally segregated teams that are chiefly highly-competitive. Lastly, customer discriminatory behaviour in the sports industry can be observed if supporters had trouble attending matches where the opponent team woud be composed of more players of a different race than usual.

As far as professional football is concerned, very few researches have been written about and thus, we provide also results from another sports sectors. One of the first studies that engaged in the topic of racism in sport was executed by Gwartney & Haworth (1974), who, in their work focused on the Major League Baseball (MLB), discovered patterns between the success rate and participation of black players in a team. Their sample was compiled from games played between 1947 and 1956. Statistics proved that five teams with highest number of African Americans were among six teams with highest win per game ratio. Another finding was introduced by Eitzen *et al.* (1982) who alleged that professional sport can be viewed as a field of equal economic opportunity for minorities since teams desire to maximize profit. Their conjecture that sports provide an exceptional opportunity for minorities was bolstered by the fact that representation of non-white players in major teams is much higher than in the rest of labor force. According to Kahn (1991), 74.3% of all players were black in NBA during the season 1985-86. In MLB, the proportion was only 27.8%. Still relevant, though, if we take into account the fact that the very first black player who was selected to play alongside his white team-mates made his major league debut on April 15, 1947. His name was Jackie Robinson and his story even inspired the movie director Brian Helgeland to film a documentary about him[1]. Thanks to his perseverance and bravery, he won recognition for his performance even though he faced many types of discrimination. First of all, some of his team-mates refused to play along his side. One member of his team even required to be traded somewhere else rather than play with him in the same team. In addition to that, he was constantly being humiliated by his opponents. But, the management of Brooklyn Dodgers were convinced about him being involved in the team and decided to penalize others rather than play against him.

In the paragraphs above, we covered examples of co-worker and employer discrimination, customer discriminatory behaviour is much more usual than these two, though. Unlike co-worker and employer prejudice, inequal treatment from customers cannot be eliminated by market forces. According to Scully (1974), presence of African Americans in MLB significantly descreased team revenue, all else being equal. This finding probably reflects preferences of white fans who might have preferred attending matches with non-black players. Nevertheless, there was no evidence of lower revenues based on the racial composition of playing teams in the season 1976-77 (Sommers & Quinton 1982). Brown *et al.* (1991) discovered a relation between fans and wage discrimination. According to their study, wage differential in Major League Baseball was for the most part caused by the decision-making of spectators[2]. As a proof of discriminatory behaviour we can also consider the finding of Nardinelli & Simon (1990)[3] who explored the market with players' baseball cards. Cards with white players were sold for higher prices than for equally qualified blacks. This might suggest racial prejudice among fans, but the demand for baseball cards can be completely different from that for matches.

Another form of discrimination was found by Jiobu (1988) who came to the conclusion that, if controlled for performance, black players had a signifi-

[1] The name of the film is "*42*", which points to the number on his shirt.

[2] In their study, they revealed an insignificant, negative effect on attendance. Replacing one white player for black player resulted in 8,400 fans lost.

[3] If we control for performance, black and Hispanic players had significantly lower prices of baseball cards, the magnitude was -14.2% and -9.8%, respectively.

cantly higher exit rate from Major League Baseball than whites. On the other hand, the exit rate of Hispanics was not significantly different from zero during the seasons 1971-1985. There exists also researches that investigate the game statistics. Scully (1973) unreveals that African Americans have significantly higher points per minute ratio in NBA. In his research through sport environment he also spotted that blacks outperform whites in American football. Their dominance was caught in all performance aspects and the difference was found to be significant at 10 out of 12 variables. Kahn & Sherer (1988) revisited the case of wage segregation. They compared players with same salaries and concluded that at the same level of wage, blacks outperform white players[4].

As was said earlier, race segregation can have many forms. One of the biggest motivation for our thesis was the discovery of Holmes (2011). He used quite nontraditional method that is called quantile regression to show significant differences in individual points of the salary distribution of MLB players. Even though he was inspired by Hamilton (1997), his study came to the exactly opposite outcome. While Hamilton examined NBA and discovered 18% premium for whites[5] in the upper part of the distribution, Holmes' investigation of MLB uncovered 25% pay gap against blacks in the lower quartile. His paper serves as a proof that even though racial discrimination causes economic losses and conflicts in society, it is still accepted by the general public, even in such a scrutinized industry of professional sport.

In comparison with North America, research studies on the topic of racial discrimination in European professional sports are extensively limited primarily due to the lack of data. Wilson & Ying (2003) suggest that team performance in Europe could be considerably improved by hiring players from Latin America. Their work was based on the data from the period after the Bosman[6] ruling was enforced. According to Frick (2006) who surveyed the German Bundesliga,

[4] Nevertheless, there was no race difference concerning players being drafted by NBA teams.

[5] In his model, he did not include a covariate implying the player's height. The results of his study might be biased as whites are on average higher than their black counterparts.

[6] Jean-Marc Bosman was a Belgian playing for RFC Liège. After his contract expired in 1990, he wanted to change teams but his potential new team refused to accept his transfer fee. His wage was afterwards reduced and he no longer played for the first team. Which resulted in taking his case to the European Court of Justice. He ended up suing for restraint of trade. Afterwards, Bosman and all players in EU were enabled to transfer for free at the end of their existing contracts. Formerly, footballer were the property of their clubs. But now, they become employers like any other workers in European Union.

players from Western Europe, Eastern Europe and Latin America receive higher bonuses relative to their German peers. The differences were found to be significant with values of 15% in favour of Western Europeans, 30% for players from Eastern Europe and a premium of 50% for "football artists" from Latin America.

Chapter 3

Methodology and Crucial Terms

The main task of this chapter is to present econometric concepts that we have been using during the analysis for every step of the practical part to become easier to comprehend.

3.1 Independently pooled cross section

Independently pooled cross section sample is obtained if we randomly draw observations from each time period (Wooldridge 2015). This approach of data collection has many reasons. Primarily, we have a larger sample which results in more precise estimators. Furthermore, test statistics become more valid. The other reason standing behind this approach is the fact, that the dependent variable might differ just because of time. As the distributions of population could be different across time periods, we include $t-1$ dummy variables, where t is the number of time periods we observe. This measure avoids dummy variable trap and also allows the intercept to vary across periods. If we interact a period dummy variable with one of the explanatory variables included into the model, we can assess if the impact of this covariate has changed over time.

3.2 Difference-in-differences

If we are interested in studying an impact of a policy, a law, or the effect of a treatment, the method which is used in quantitative researches is called difference-in-differences (Ashenfelter & Card 1984). To be able to apply difference-in-differences (DID) estimator, we require data both from a control and treatment group[1] at least at two time periods. In a simplified form, we can write the model as follows:

$$y = \beta_0 + \delta_0 T_2 + \beta_1 GoI + \delta_1 T_2 \cdot GoI + u \tag{3.1}$$

Where y is the response variable, T_2 is a dummy implying that the data was collected from the second time period and GoI is a dummy variable equal to 1, if the observation belongs to the group of interest. The coefficient of interest is in our case δ_1, since it captures the effect of interaction term $T_2 \cdot GoI$. The logic behind the difference-in-differences estimator is described below.

$$\mathbf{E}[y \mid GoI, T_2] = \beta_0 + \delta_0 + \beta_1 + \delta_1$$

$$\mathbf{E}[y \mid OO, T_2] = \beta_0 + \delta_0$$

Where the abbreviation *"OO"* stands for other observations that are not included in the group of interest. T_1 signifies the base year. If we subtract these two equation from each other, we obtain the first difference:

$$\mathbf{E}[\Delta y_{T_2}] = \mathbf{E}[y_{GoI} - y_{OO} \mid T_2] = \beta_1 + \delta_1 \tag{3.2}$$

The other difference is obtained if we subtract expected values from the base year.

$$\mathbf{E}[y \mid GoI, T_1] = \beta_0 + \beta 1$$

$$\mathbf{E}[y \mid OO, T_1] = \beta_0$$

And the difference is therefore:

$$\mathbf{E}[\Delta y_{T_1}] = \mathbf{E}[y_{GoI} - y_{OO} \mid T_1] = \beta_1 \tag{3.3}$$

The final step of obtaining DID estimator looks as follows:

$$\mathbf{E}[\Delta y_{T_2} - \Delta y_{T_1}] = \delta_1 \tag{3.4}$$

The coefficient δ_1 will in our case reflect, how much the salaries of African Americans and Hispanics have changed with the increasing commercialization of MLS together with the introduction of new salary policies. As the first signif-

[1] So-called control group is the baseline measure for the experiment. On the other hand, the treatment group is the one that receives experimental manipulation.

icant increase in wages occured in 2010[2] (see Model 1 in the table 6.2), we interact this time variable with race dummies to obtain estimates for *y2010*Black* and *y2010*Hisp.* These dummies are intended to uncover whether an increase in investment has had any discriminatory elements or not. If we observed a change in the salaries of black and Latin players only, before and after the easing of the salary cap, we would fail to control for some omitted variables, such as economic situation in the US or popularity of soccer there. Once we also use white players as a control in the DID model, we eliminate a possible bias of our estimator, even though some of the variables influencing players' salary are unobserved.

3.3 Quantile Regression

Quantile regression was introduced by Koenker & Bassett Jr (1978) and offers another approach for data analysis. While OLS gives us estimates at the average, quantile regression accounts for outliers and can estimate any quantile q for $q \in (0, 1)$ of the explained variable. This treatment ensures higher robustness of our estimates. Thanks to this tool, we are able to measure the effect of race on compensation for less paid players that are more susceptible due to their relative lower expenses for employers. One of the advantages might be also that analysis of characteristics between explanatory and explained variables becomes richer and more understandable as we are not focused only on the conditional mean. The estimator of QR for quantile q minimizes the function:

$$Q(\beta_q) = \sum_{i:y_i \geq x_i^{'}\beta}^{N} q \mid y_i - x_i^{'}\beta_q \mid + \sum_{i:y_i < x_i^{'}\beta}^{N} (1-q) \mid y_i - x_i^{'}\beta_q \mid \tag{3.5}$$

where q varies depending on the quantile we want to observe. The coefficient interpretation remains the same as in the case of OLS, except, instead of the average, we refer to the corresponding quantile. Another advantage seems to be the fact that quantile regression does not require homoskedasticity and normal distribution of the error terms. These two assumptions are rarely satistified in the case of the salary distribution of the entire population because of high extreme values. For more detailed description of this method, see Koenker & Hallock (2001).

[2] In our own interest, we tried to interact all of the season dummies, but none of them has been found significant for the model.

Chapter 4

League Overview and Dataset Construction

This part of our thesis is dedicated to a detailed description of the data used in our survey. First of all, we decided to provide some basic information about the MLS as well as to highlight its financial standing. The Major League Soccer was chosen for our analysis due to its uniqueness in terms of publishing players' salaries. Another reason was the fact that it reflects the typical elements of North American sports to which most of the studies concerning discrimination in professional sports have been devoted. The next step of this chapter is a description of some policies that helped MLS to increase its quality and to adjust the salary system that was introduced with establishment of the league. The most famous one is the Designated Player Rule that was established in 2007 in order to allow clubs to pay transfers and wages exceeding the salary cap. Last part of this chapter focuses on the detailed description of the dataset and to the justification of each variable used for estimating the determinants of player's compensation.

4.1 American Major League Soccer

Major League Soccer is the top professional football league in the USA and Canada. Soccer, as the European football is called in American English, has always lagged behind other popular sports such as baseball, American football or basketball. But lately, thanks to the increasing popularity of soccer in the US, and to the policies that have been implemented, it has lured some of the football superstars[1]. These new acquisitions appear to have very beneficial effect on the clubs' financial standing. When we look at the calendar 2016, all 22 teams recorded exponential growth in attendance, viewership, merchandise sales or social media (Forbes.com 2017). As far as attendance is concerned, it even exceeds NBA and NHL with average number of 21,692 spectators per game (Statista.com 2017). MLS's establishment dates back to the year 1995 as a condition for United States Soccer Federation to FIFA, which awarded United States by organizing FIFA World Cup in 1994. Though, the first season was not played until 1996, with the presence of 10 teams (Mlssoccer.com 1996). In the first few seasons, the league was not profitable at all and matches were played on half-empty stadiums. In 2002, two of the founding clubs decided to cease operations in the league after having financial problems[2]. The turning point occured in 2007 when David Beckham, an English superstar and the first league's designated player, left Real Madrid to join Los Angeles Galaxy. This season was also affected by expansion beyond the United States borders. The first team from Canada that entered the league was Toronto FC. Ever since, the money pumped into the system has risen dramatically and the league attracted other teams to participate. As of 2017, the competition consists of 22 teams and the most successful of them is Los Angeles Galaxy that has won 5 trophies.

As salaries in MLS are limited by a salary cap, clubowners are prohibited to spend excessive money on compensations. This measure prevents also competitive imbalance among teams. With the introduction of the Designated Player Rule, though, the wages paid to footballers increased substantially. In 2017, and for the first time in MLS history, the total guaranteed compensation of all players exceeded $200 million. The median salary of the whole competition increased by 15% relative to the season 2016 up to $135,000 and as a result of the developing salary system, the number of millionaires among MLS players grew to 28 (Espnfc.com 2017). If we take a look at the table 4.1 which depicts racial composition of the highest paid players in 2016, we can notice, that two out of ten players with highest salaries were Hispanics and only one of them was African American. This ratio appears to be surprising as in the random

[1] David Beckham started this trend and was subsequently followed by such footballers as Steven Gerrard or Frank Lampard. These two top level British football players transferred from English Premier League in 2015 and were signed as designated players. Another famous designated player was the Spaniard David Villa who signed in 2014 a contract, that ensured him a yearly salary of $5,610,000.

[2] Both Floridian teams, Tampa Bay Mutiny and Miami Fusion FC, ended their presence in MLS.

sample that we have collected, black players represent 26.4%.

Table 4.1: The highest paid MLS footballers in 2016

Player	Salary	Race	Position	Mins	Goals	Assists
Ricardo Kaká	$7,167,500	Hispanic	Midfielder	1955	9	10
Sebastian Giovinco	$7,115,556	White	Striker	2418	17	15
Michael Bradley	$6,500,000	White	Midfielder	2160	1	5
Steven Gerrard	$6,132,500	White	Midfielder	1491	3	11
Frank Lampard	$6,000,000	White	Midfielder	1280	12	3
Andrea Pirlo	$5,915,690	White	Midfielder	2770	1	11
David Villa	$5,610,000	White	Striker	2869	23	4
Jozy Altidore	$4,825,000	Black	Striker	1487	10	5
Clint Dempsey	$4,605,942	White	Midfielder	1429	8	2
Giovani dos Santos	$4,250,000	Hispanic	Striker	2350	14	12

4.2 Salary system

Since the establishment of the league, MLS salaries has been constrained by a salary cap. Its main purpose was to ensure a competitive balance among teams. This prevented MLS from attracting players of the highest quality as the most talented players sought contracts in more generous European clubs. To ease the strict wage policy and therefore increase quality of players in the competition, MLS introduced several rules while still sticking to the salary cap.

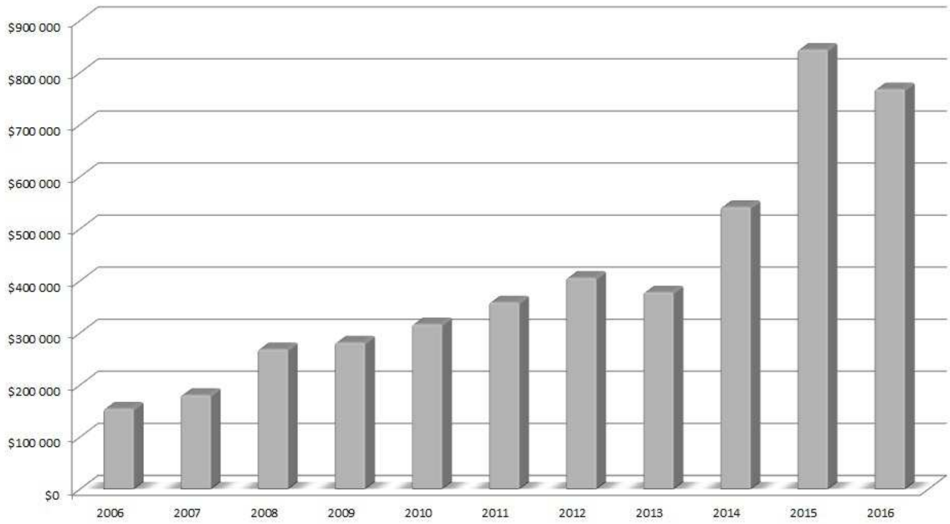
Designated Player Rule - Also known as Beckham Rule (Reuters 2017) since David Beckham was the first who used this rule to transfer to Los Angeles Galaxy. It was adopted in 2007 and it allows MLS teams to sign players outside their club's salary cap. The main goal of the policy was to enable MLS clubs to compete for international players in the football labour market. Besides higher wages, clubs are also allowed to pay significant transfer money which would be impossible if there was not for this rule.

Allocation Money - is additional money that is available to football clubs and therefore each MLS team is entitled to receive this money in addition to its salary budget. Moreover, for those teams that do not have a third Designated Player, General Allocation Money (GAM) uses its funds collected by MLS to enable them to purchase more expensive players of their desire. In 2017, MLS also introduced the so-called Additional targeted allocation money that allocated additional $1,200,000 for each team (Mlssoccer.com 2017).

Generation Adidas - This rule was created thanks to the cooperation between MLS and US soccer. It was aimed to improve the quality of young football talent in the US. This programme sponsored by German company Adidas, covers the salaries of young supertalents who after that do not have intentions to leave the US. For these players, salary cap does not count.

The long-term effect of these policies is obvious. Once the restrictions about salary cap were eased, the trend in terms of wages was mostly upward-sloping. While in 2006 the sample average compensation of 100 most productive players was mere $152,845, one decade later the value increased by 401% to $766,481. The evolution of salaries with top 100 MLS players is depicted in the figures 4.1 and 4.2.

Figure 4.1: MLS average salary of 100 best players

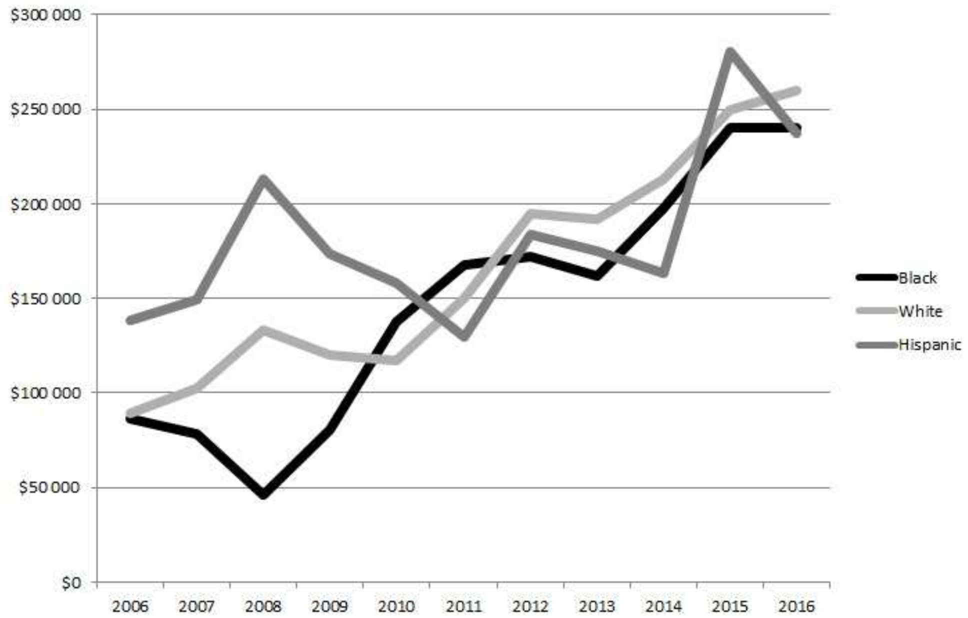


Source: MLS Players Union (2016), author's computations.

4.3 Dataset Construction

In order to answer our research questions, we have compiled a vast dataset consisting of 1,100 observations. This large number of observed players will enable us to focus also on different quantiles of the distribution. From each

Figure 4.2: Development of players' median salaries



Source: MLS Players Union (2016), author's computations.
Note: Mind that the seasonal fluctuations might be influenced by the dataset structure. What is important is the long-term trend.

season 2006-2016, we took 100 most productive players in terms of goals and assists because their figures provide reliable information about their performance. Otherwise, the data generated would not be meaningful. It would be optimal to include random effect for each individual and arrange the data into a panel. This method would enable us to account for unique and unobserved characteristics of each player. Nevertheless, as the majority of the players appear in the dataset at most twice (see table 4.2), random effects technique would not be appropriate.

Table 4.2: Occurrence in the dataset

Occurrence	1×	2×	3×	4×	5×	6×	7×	8×	9×	10×	11×
No. of players	382	163	42	18	11	8	4	3	2	1	1

Note: Each number implies the occurence of the players in the sample. Once we sum up the row with the numbers of players, we obtain 635 individuals. If we multiply the number of players by their occurence in the dataset, we get to the 1,100 observations.

Thus, we created a matrix that contains 18 variables out of which 17 are explanatory. In addition to that, we added 10 dummy variables accounting for individual seasons. To examine a possible wage discrimination against non-white players, we needed to collect three types of data.

Performance data - The first one deals with players' performance that we need to control for in order to discover wage inequality given the same productivity. This data should have the highest effect on player's compensation and it is available on the official webpage of MLS. To the best of our knowledge, no model explaining footballer's salary has been constructed and hence, we needed to select the most important variables by intuition. These statistics imply player's productivity, stamina, activity or indiscipline.

Salary data - The next type was the information about player's salary. American sports leagues are unique due to their transparency and they provide, contrary to the leagues in Europe, information about players' yearly salaries. This data is at disposal on the webpage of MLS Players Union. This association serves as the collective bargaining representative and as a protection of all players' rights. For each individual player, they list his yearly compensation divided into two components - base salary and guaranteed compensation. The guaranteed compensation is base salary added to the annualized bonuses and options of players that are included in the conctract. It is more appropriate for the study as it also includes the bonuses obtained when signing the contract. Thus, salaries at the end of each season vary accordingly. Performance bonuses are not included because there is uncertainty about obtaining these benefits and therefore, a possible correlation between salaries and performance statistics is ruled out.

Demographic data - The ethnicity data were observed either from the US sports channel ESPN or Mlssoccer.com as both of them offer individual pictures and also information about places of birth as well as the data about players' age, which can be used for determining the peak or the bottom in terms of compensation in their carreers. Since the main question to be answered is wage discrimination of African Americans and Hispanics, we were obliged to exclude Asian footballers from the sample[3].

We obtained the ultimate dataset that consists of 290 (26.36%) black, 537 (48.82%) white and 273 (24.82%) Latin footballers. These proportions are

[3] Since there were only 3 Asian players in our dataset, their exlusion could not cause a violation of random sampling assumption.

sufficiently high and makes the investigation of any pay gaps possible. While analysing our data, we found some interesting facts that occured in the seasons 2006-2016:

- Between 2006-2016, 6,175 goals were scored. From that amount, 2817 (45.62%) were recorded by whites, 1727 by blacks (27.97%) and 1631 (26.41%) by Hispanics.
- 9 different players won MLS trophy for the top scorer; only Jeff Cunningham (black) and Chris Wondolowski (white) managed to win this award twice.
- The best player in terms of productivity was Sebastian Giovinco. Despite his high guaranteed compensation of $7,115,556 in the season 2015, he was very contributive for his team Toronto FC as he was able to record 38 points (22 goals and 16 assists).
- In the season 2010, David Beckham played only 466 minutes and scored 2 goals. Given his salary of $6,500,000, one goal cost his club $3,250,000 which is the highest ratio Salary/Goal observed.
- The cheapest goals were scored in the same season by Chris Wondolowski whose each goal cost only $2,667.
- One minute played of an average black player cost his employer $208.5, for Hispanic players the amount was $204.2 and for whites $234.1.
- 3 most often fouling players were African Americans; in the season 2013, Quincy Amarikwa fouled every 17.6 minutes which made him the most indisciplined player in the league.

The table 4.3 lists all explanatory variables that were used for the research.

Table 4.3: Definitions of variables

Variable	Definition	Exp. sign
STRIKER	1 if striker, 0 otherwise	?
MIDFIELD	1 if midfielder, 0 otherwise	?
GS	Games started per season	+
MINS	Minutes played per season	+
GOALS	Goals per season	+
ASSISTS	Assists per season	+
SHTS	Shots per season	+
SOG	Shots on goal per season	+
FOCO	Fouls commited per season	-
FOSU	Fouls suffered per season	+
YCARD	Yellow cards per season	-
OFFS	Offsides per season	-
STAR	1 if star, 0 otherwise	+
AGE	Age of the player	+/-
AGE^2	Age of the player squared	-/+
BLACK	1 if black, 0 otherwise	?
HISP	1 if Hispanic, 0 otherwise	?
YEAR_t	1 if taken from $year_t$, 0 otherwise	+

The dependent variable is salary in logarithmic form and it is the key element for the thesis. Log-level model moderates the effect of outliers emerging from extreme values in the upper part of the distribution. We expect the compensations to be determined *ex post* and on that account, we collected salaries after each season. The table 4.4 represents the distribution of salaries from the sample.

Table 4.4: Salary - Descriptive statistics

Variable	Mean	Min	10%	25%	50%	75%	90%	Max
Salary ($)	407,315	12,900	48,400	83,562	156,417	258,469	650,000	7,167,500
i-th player		1^{st}	110^{th}	275^{th}	550^{th}	825^{th}	990^{th}	1100^{th}

The percentages denote corresponding quantiles

Note: The players are arranged in ascending order.

The reader can notice that the median is significantly lower than the mean. It is caused by the dramatic difference between 90^{th} and 100^{th} percentile and we can deduce that at least 10% of the highest paid players have unproportionally higher salaries than the rest of the sample. This fact urges us to use quantile regression in addition to OLS and to estimate the conditional quantiles of the dependent variable. Moreover, we will be allowed to investigate discrimination in the lower part of the wage distribution. This approach is reasonable as discriminating better and more paid players would incur much higher relative costs for the employer. The table 4.5 offers means of the lower quartile compared to the averages of whole sample. At the first glance we can notice that while performance statistics do not differ so much, the percentage difference of salary is immense between these two groups.

Table 4.5: Comparison between whole sample and lower quartile

	Salary	GS	MINS	Goals	Assists	SHTS	SOG	FoCo	FoSu	Offs	Ycard	Age
25^{th} percentile	54,356	16.08	1,456	3.69	2.43	28.38	12.29	22.46	22.85	7.20	2.53	23.94
Whole sample	407,315	19.87	1,765	5.25	3.65	39.04	16.36	25.79	29.18	9.45	2.98	26.55
% difference	649.35%	23.56%	21.19%	42.29%	50.49%	37.56%	33.17%	14.83%	27.72%	31.24%	17.94%	10.89%

To the best of our knowledge, there is no study which would examine the relation between footballer's race and wage and hence, the data processing needs to be done according to studies from another field of sports environment. For National Basketball League, Brown *et al.* (1991) introduced the method of all-career statistics. Their work was extended by Rehnstrom (2009), who used all-career statistics as well, and found 24.5% premium for white players in NBA in the season 2008-09. This approach might seem reasonable as employers are more interested in observing how workers have been doing in a long-term period. But, professional football is a special case, and in comparison with basketball, there are many competitions on the top level and statistical comparison between them might lead us to false results. The approach of all-career statistics would not allow us to observe no other players than Americans which would significantly lower the extent of our work. In addition, some older former stars towards the end of their career receive relatively lower wages even though their all-career statistics exceed numbers of other players. For this reason, we adopted the method of Yang & Lin (2012) who examined the impact of nationality on salaries in NBA. According to them, a player's current salary is determined by performance in the previous year. Nevertheless, as compensations might not be determined on the basis of one season, our thesis is composed of two separate regressions - the first examines performance statistics from previous season and the second deals with average statistics from three previous seasons. The second technique was used by Hill (2004) in the examination of racial wage differential in basketball. This method eschews sudden fluctuations in players' performance across individual seasons. Furthermore, employers tend to remunerate players according to their longer consistent performance[4].

4.3.1 Performance statistics

After specifying the method of data collection and data transformation, we can proceed to the description of the data and to the reasoning standing behind its inclusion into the dataset. Goals and assists are likely to be the most important determinants of the player's salary since they reflect the most his importance for the team. Similarly, games started and minutes played per season imply how much trust the coach has in the player. In addition, the first mentioned shows player's readiness and medical condition during the season. Consequently, we expect all of them to have a positive impact on the salary. We decided to omit a number of games played because each club is provided for only 3 substitutions during the game and it is very usual that players are substituted in the last 15 minutes of the game. The correlation between the time spent on the pitch and games would therefore decline and it would also decrease the trustworthiness of the model.

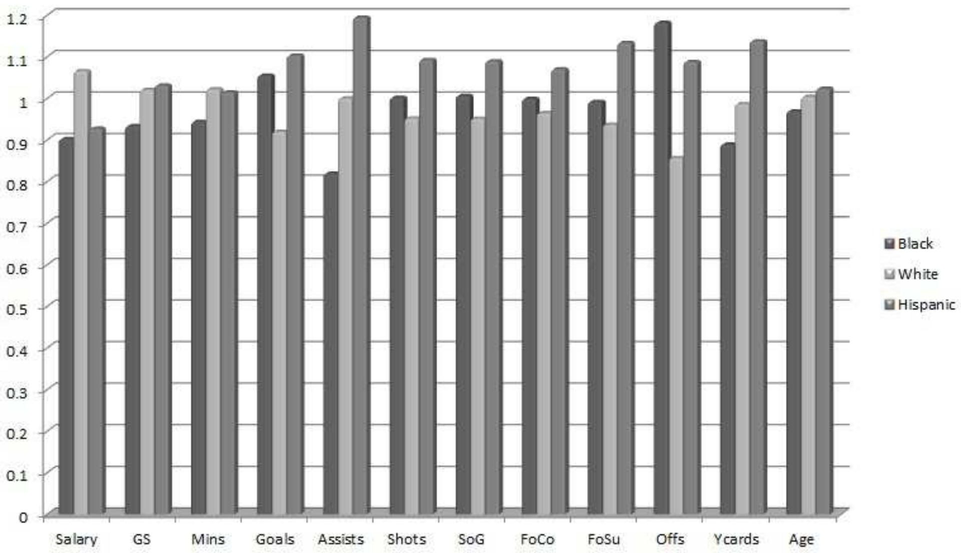
To avoid dummy variable trap, we omitted the position of the defender from the model[5] and this position is therefore included in the intercept. The other two positions, striker and midfielder, are included as dummy variables and their effect on wage is difficult to be estimated. The indicators of quality for players on these two positions differ from indicators for defenders. The other two variables included in the study are shots and shots on goal per season. Basically, both of them imply the player's activity on the pitch and eagerness for scoring goals. For this reason, we expect them to influence the salary positively as well. The variable *offsides* is likely to have a negative impact on

[4] While compiling the dataset, we noticed, though, that there was no player, whose salary would remain constant for 3 years and more. Player's remuneration is therefore often adjusted and changes over time based on previous productivity.

[5] The position of a goalkeeper was completely excluded from our study as their statistics are imcomparable with statistics of other players on the pitch.

the player's compensation since it rather reflects his inattentiveness. On the other hand, the variable standing for fouls suffered per season should serve as a proxy for belligerence and the coefficient is very likely to be positive. If we move to the position of a defender, there are two statistics that need to be emphasized. Namely, yellow cards (YCARD) and fouls commited (FOCO). Both of them indicate indiscipline and their presence should have a negative effect on the wage.

Figure 4.3: 3-seasons average statistics



Note: We weighted players' performance based on overall average, which is equal to 1. If the column is higher than 1, it signifies that players of this race had above-average performance.

As can be seen from the figure 4.3, both Hispanics and African Americans earn on average less than their white peers despite their better performance in terms of goals and shots. From the table 4.6, which captures statistics based on 1-season observation, we can spot that the salary of whites and blacks in the middle of the distribution is almost the same, though. Median salaries of both are significantly exceeded by salaries of Hispanics. Both blacks and Hispanics appear to be more undisciplined which can be seen in higher number of offsides as well as fouls commited. Blacks were also given 7.4% less time on the pitch than whites. However, based on these results, we cannot conclude anything about discriminatory behaviour. If we want to detect any premium for players of different ethnicity, we need to control for the player's performance, and therefore descriptive statistics are not sufficient.

Table 4.6: Sample means, medians & standard deviations - 1 season

1-season statistics	Black		White		Hispanics		Overall	
Variable	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Salary ($)	367,631	149,667	434,152	150,000	377,723	175,000	407,315	156,417
	(212229)		(1123149)		(719240)		(974237)	
Games started	19.73	20	21.36	23	21.15	23	20.89	22
	(7.68)		(7.52)		(6.75)		(7.41)	
Minutes	1,763	1,796	1,894	1,980	1,850	1,955	1,849	1,924
	(640.77)		(638.71)		(565.19)		(624.24)	
Goals	5.96	5	5.21	4	5.97	5	5.61	4
	(4.27)		(3.89)		(3.60)		(3.94)	
Assists	3.21	3	3.78	3	4.38	4	3.78	3
	(2.57)		(3.31)		(3.42)		(3.19)	
Shots	41.51	36	38.79	34	44.32	42	40.99	37
	(25.71)		(24.54)		(22.44)		(24.45)	
Shots on goal	17.49	15	16.16	14	18.3	17	17.09	15
	(10.87)		(10.93)		(9.75)		(10.67)	
Fouls commited	26.73	25	25.42	24	28.26	27	26.49	25
	(13.59)		(12.22)		(13.11)		(12.87)	
Fouls suffered	30.32	28	27.86	25	33.38	29	29.88	27
	(16.96)		(14.90)		(18.51)		(16.56)	
Offsides	11.99	10	8.09	5	10.41	7	9.72	6
	(10.47)		(9.99)		(10.00)		(9.99)	
Yellow cards	2.79	2	3.13	3	3.43	3	3.12	3
	(2.08)		(2.28)		(2.08)		(2.19)	
Age	25.73	25	26.65	26	27.18	27	26.55	26
	(4.17)		(4.16)		(4.71)		(4.34)	
N	290		537		273		1100	

standard deviations are noted in parentheses

Table 4.7 illustrates descriptive statistics where players' performance was observed and averaged for the 3 previous seasons. As we can see from the lower standard deviations, this treatment serves as a robustness check. Compared to the table 4.6, most of the values are slighly lower. Otherwise, both approaches appear to have very similar descriptive statistics. Players of black skin played on average 138 minutes less than whites and 125 minutes less than players from Latin America. Hispanics were more active than the rest of the dataset as their average player scored 19.88% more goals than his white counterpart. They also recorded better results concerning assists, shots or shots on goal. On the contrary, Southerner temperament is manifested in terms of commited fouls and yellow cards where they outdid other footballers.

Table 4.7: Sample means, medians & standard deviations - 3 seasons

3-seasons statistics	Black		White		Hispanics		Overall	
Variable	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Salary ($)	367,631	149,667	434,152	150,000	377,723	175,000	407,315	156,417
	(212229)		(1123149)		(719240)		(974237)	
Games started	18.55	19.33	20.28	21.33	20.49	21.50	19.87	20.83
	(6.79)		(6.74)		(6.05)		(6.64)	
Minutes	1,666	1,711	1,804	1,873	1,791	1,879	1,765	1,824
	(567.53)		(571.08)		(512.17)		(559.24)	
Goals	5.54	4.33	4.83	4	5.79	5	5.25	4.33
	(3.77)		(3.46)		(3.45)		(3.57)	
Assists	2.99	2.67	3.65	3	4.36	4	3.65	3
	(2.04)		(2.76)		(3.04)		(2.71)	
Shots	39.10	34.75	37.18	33.67	42.66	40	39.04	36
	(22.99)		(22.30)		(20.60)		(22.19)	
Shots on goal	16.46	15	15.57	14	17.82	16.67	16.36	14.83
	(9.52)		(9.98)		(9.29)		(9.73)	
Fouls commited	25.77	24.25	24.88	23.50	27.60	26.67	25.79	24.58
	(11.89)		(10.68)		(11.74)		(11.33)	
Fouls suffered	28.95	28	27.34	25	33.07	29	29.18	26.50
	(1.68)		(1.90)		(1.84)		(1.85)	
Offsides	11.17	10	8.10	5	10.28	7	9.45	6.67
	(9.02)		(8.73)		(9.56)		(9.12)	
Yellow cards	2.65	2.50	2.94	3	3.39	3	2.98	3
	(1.68)		(1.90)		(1.84)		(1.85)	
Age	25.73	25	26.65	26	27.18	27	26.55	26
	(4.17)		(4.16)		(4.71)		(4.34)	
N	290		537		273		1100	

standard deviations are noted in parentheses

4.3.2 Other Factors

Undoubtedly, performance statistics are relevant in determining player's wage, but there are other factors to be considered. Age is one of them and it definitely should not be missed out. With increasing age, players should obtain more experience and be rewarded by higher wages. This effect must be, though, accompanied with a decreasing marginal effect. At one point of a career, players tend to lose their physical strength and stamina, which leads to an overall decrease in performace. Holmes (2011) uncovered an increasing effect of age on the salary of MLB players. The turning point in his study occurred at the age of 17.6 years, the estimates turned out to be statistically insignificant, though. On the other hand, Yang & Lin (2012) found a decreasing effect of age on the salary in American NBA. As another covariate in their salary equation, they included the variable *star* that was equal to 1 if the player was chosen into NBA All-Star in the previous season. We adopted the same approach since this variable control for a possible fact that football superstars might be remunerated more than appropriate for their performance on the pitch.

The last set of variables concerns the years from which the data were taken. These dummy variables include besides other things the economic situation in the US as well as the popularity of soccer there. As the statistics were collected in seasons 2006-2016, our model contains ten dummies and the base year 2006 is therefore included in the intercept. These variables are expected to be of a high magnitude and significance for the reasons stated earlier.

Chapter 5

Model

In the first part of this chapter we justify the aproach that has been chosen. After justifying the methodologies, we will move on to present our model that estimates the determinants of a player's salary.

5.1 General Model

To answer the research questions, we have collected a vast set of data from several seasons. To be able to estimate the development of salaries in football enviroment, we monitor data of 1,100 different observations during eleven seasons, which means that we face a combination of cross-sectional and time series data. According to Wooldridge (2012), the data with cross-sectional and time series aspects can be arranged in two kinds of data sets. If we obtain information from a random sample at different points in time, this data is called an independently pooled cross section. On the other hand, collecting data from the same individuals leads to a panel dataset. The assumption of observing same units in time is not satisfied in our data and hence, we arrange it into independently pooled cross section.

$$log(Salary) = \alpha + \mathbf{X}\vec{\beta} + \mathbf{R}\vec{\gamma} + \mathbf{T}\vec{\delta} + \vec{u} \tag{5.1}$$

In this thesis, we utilize the explained variable (salary) in a logarithmic functional form as we want to know the percentage change in salary. At the same time, once we utilize the logarithm transformation, we take into consideration the heteroskedasticity that arises from the large variance in salaries. Using log-level model means that the interpretation must be changed and each coefficient β_i, γ_i and δ_i multiplied by 100, describes a percentage increase or decrease in player's wage. Vector **X** is a vector of size $k \times n$ as there are k explanatory variables standing for performance, position or age, and n observations. **R** is a vector implying the race of each individual player and finally **T** is a vector of time dummies which is equal to 1 if observed in the particular year.

Even though the OLS method was used by the majority of researches dealing with racial discrimination in sport (Kahn 1991), we test the satisfaction of the assumptions in appendices and also point to possible problems.

Before we move to our model, there is one issue that needs to be considered - wage is not in real, but in nominal dollars. As wages in our dataset increase also due to inflation, we should adjust them in order to examine solely the effect on real wages. This procedure requires deflating wages from 2007-2016 to 2006 dollars. Fortunately, this turns out to be unnecessary as far as we use log-linear model. The difference between real and nominal wage can only be seen in the form of different coefficients on the season dummies. We provide an explanation with a simplified model using only two seasons.

$$log(wage_i/PS_2) = log(wage_i) - log(PS_2) \tag{5.2}$$

Where the constant PS_2 describes the price level change in the season 2 relative to the season 1. While $wage_i$ varies across players, PS_2 remains the same and therefore, $log(PS_2)$ is basically absorbed into the intercept. The bottom line is that, nominal wages do not have to be turned into real ones for studying the determinants of footballers' compensation.

5.2 Our Model

As we noted earlier, some of the research questions about pay discrimination will be answered using Ordinary Least Squares method. Our model has following form:

$$
\begin{aligned}
log(Salary) =& \beta_0 + \beta_1 Goals + \beta_2 Assists + \beta_3 SoG + \beta_4 Shots + \beta_5 Fouls + \\
& \beta_6 YCard + \beta_7 Offs + \beta_8 Mins + \beta_9 GS + \beta_{10} FoCo + \beta_{11} Age + \\
& \beta_{12} Age^2 + \beta_{13} Striker + \beta_{14} Midfield + \gamma_1 Black + \gamma_2 Hisp + \\
& \delta_0 year2007 + \delta_1 year2008 + \delta_2 year2009 + \delta_3 year2010 + \\
& \delta_4 year2011 + \delta_5 year2012 + \delta_6 year2013 + \delta_7 year2014 + \\
& \delta_8 year2015 + \delta_9 year2016 + u_{it}
\end{aligned}
\tag{5.3}
$$

The dependent variable $log(Salary)$ denotes logarithm of individual salaries. This form was chosen since we are interested in percentage rather than absolute change in player's wage. For instance, one more goal scored increases or decreases salary by $\beta_1 100\%$. The error term u_{it} stands for speciffic, unobserved effect of each player.

The main concern of this study is expressed by dummy variables $Black$ and $Hisp$. If these two variables turn out to be significant, this study will disclose either positive or negative pay discrimination relative to white players. A positive coefficient would therefore imply pay premium either for black or Hispanic players. In case of negative coefficient, we will speak about racial wage discrimination against non-whites. Besides racial dummies, we also included dummy variables that control for individual years. As we mentioned earlier, these variable contain information about time trends. Omitting them would cause a serious problem for our model, since for example, the value of dollar in 2006 is not equal to its value in 2016. Presumably, years will also be of statistical significance. The increasing importance and popularity of professional football in the latest years attracted many investors who pumped money into the system.

Chapter 6

Results

As was said earlier, we used two different regression analyses[1] to detect racial discrimination in professional football. The most important lesson is that the MLS clubs turned out to have discriminatory behaviour. The table 6.4 summarizes wage gaps at the individual parts of the salary distribution and the estimates indicate that Latinos and African Americans are deprived of a significant part of their salary due to their origins. Tables 6.3 and 6.2 represent the results of OLS estimates. For both methods of data transformation, we run 4 individual regressions[2] to answer the research question we posed in the thesis proposal. The results from both approaches are slightly different due to a different nature of the data. Since we detected heteroskedacity and OLS assumption of homoskedasticity was violated (Appendix A), we obtained also heteroskedasticity-robust standard errors (see table B.1 in Appendix B) to show that most of the variables have very similar t-statistics and statistical significance. To account for non-normality of error terms, we also ran a robust regression that dampens the impact of excessively high values in the upper part of the salary distribution. The results are introduced in Appendix B and demonstrate that the inference about racial discrimination remained unchanged. We will now focus on the description of OLS results, though, since it is the common method in the existing literature when examining pay discrimination between races.

[1] OLS and quantile regressions.

[2] The Model 1 is the core regression that helps us to answer the question about discrimination of the average player. Models 2 and 3 deal with the effect of increasing wage level on the pay gap and finally model 4 clarifies whether or not the players are discriminated on different positions.

6.1 Discussion of the results

As the principle of 3 seasons has slightly higher adjusted R^2 and therefore, the variance in salary is better explained by the explanatory variables, we will be describing results of the table 6.2 in the following text. Dummy variables $BLACK$ and $HISP$ have both negative coefficient once we examine the average player in the horizon of three seasons. However, this is not where we should look for wage discrimination because better and more competitive players would easily transfer to another club if they felt like being discriminated against. Hence, the coefficients for racial dummies are not statistically significant and we would not be able to reject the null hypothesis $H_0 : \gamma_1 = \gamma_2 = 0$.

As a very interesting outcome we can consider coefficients for positions. Given the same performance, midfielders earn on average 16.1 % less than defenders at 5% significance level. This benefit might emerge from the fact that we could not control for other variables that reflect quality of defenders. Such factors could be for example a number of tackles or a number of won aerial duels. Unfortunately, this data is not provided and it is inaccessible for our study. Model 4 takes into account possible interactions between the player's race and position and it consists of four multiplications of two dummy variables. Their presence should account for different physical aspects of each race. As we see in the table 6.1, physically stronger African Americans are more likely to be centre forwards whereas relatively weeker Latinos are more often to be found as wingers or centre midfielders. The effect of these interactions is that for each position and race the intercept differs. For example, for a midfielder from Latin America, the intercept would be: $12.126 - 0.164 + 0.160 - 0.322 = 11.800$. The reader can notice that none of the interactions is significicant[3] and the value

Table 6.1: Relationship between race and position

Race/Position	Defender	Midfielder	Striker
Black	12.46%	29.75%	57.79%
White	14.37%	44.96%	40.67%
Hispanic	5.88%	44.86%	49.26%

[3] As the assumptions of normality is violated and t-statistics using OLS are no longer valid, we checked the significance using a robust regression. The inference about insignificance remained the same.

of adjusted R^2 indicates that the explained variance of the players' income remained the same after adding these four predictors into the model. Not surprisingly, the more games a player is included in the starting line-up, the more money he is rewarded. The high correlation of 83.28% between *games* and *minutes* results in a very surprising outcome.

Figure 6.1: Correlation between GS and MINS

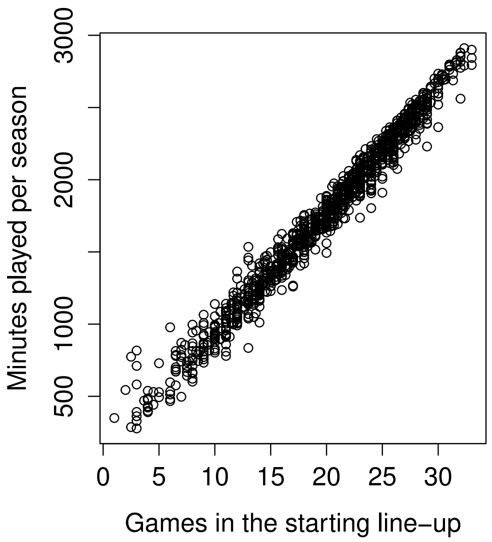


Table 6.2 shows, that if all other predictors are kept constant, *minutes* played have a negative impact on wage. In case of real-life football, other explanatory variables tend to change as well when the predictor *minutes* changes. Consequently, estimating changes just for this variable might be unnatural and we need to be very careful about the interpretation. A false inference would be to assume that players who play less, earn more money. The coefficient rather says that players who manage to perform the same work during less time, and are more time-efficient, are likely to obtain higher compensation.

Goals and assists were both found to be positive and highly statistically significant. This is not surprising because these two variables imply the player's contribution for the team. Similarly, ten more shots per season are supposed to increase salary by 7%. On the contrary, despite the positive sign, shots on goal are insignificant for the model. Adverse effect was expected from the variables *fouls commited* and *offsides* as they both reflect indiscipline and incaution. In the case of fouls, we would even be able to reject the null hypothesis at 1%

level. The impact of *fouls suffered* is rather negligible as being fouled once more increases the salary only by 0.4%. However, since this coefficient is significant, we can deduce that clubs appreciate combative and dribbling players who do not hesitate to get their teams into a dangerous situation at the cost of being fouled. Surprisingly, a positive sign was also obtained for yellow cards. The premium of 6.82% is not directly related to the amount of yellow cards, but this variable might correlate with the player's activity. As far as star players are concerned, we came to the similar conclusion as Yang & Lin (2012) who also found large magnitude and significance in American NBA.

The estimated coefficients imply that *age* has an increasing effect on wage. To find the turning point, we need to use first derivative to find an extreme:

$$0 = \widehat{\beta_1} + 2\widehat{\beta_2} age \Rightarrow age^* = -(\widehat{\beta_1}/2\widehat{\beta_2})$$

.

After filling in values from our model, we get $age^* = -\frac{-0.15968}{2\times 0.00461} = 17.32$ years which is considered to be the lowest bottom of player's career. Predictable outcomes were received for dummy variables that controlled for time trend. Since 2010, the increasing popularity of MLS jointly with new salary policies has caused year dummies to have a very large magnitude and significance. Nevertheless, increased money in the system did not cause any wage gap across races, which can be seen from models 2 and 3. We used the DID estimator to study the differential effect between the treatment group (non-whites) and the control group (whites) and how this effect changes for the average player after the salary cap easing. As a system shock, we considered the season 2010 because this year appears to be the break-even point for the MLS in terms of money spent on wages. Even though the level of wages raised substantially and brought an average premium of 34.58%, the easing of salary cap turned out to be non-discriminatory. Nevertheless, if we wanted to examine the establishment of the Designated Player Rule before and after this policy was introduced, we would have to interact the season 2007. This year, though, does not seem to have any influence on the guaranteed compensation of an average player. Furthermore, we only controlled for one season before this adjustment of salary system was presented. If one would like to examine the exact effect of Beckham's rule on the wage gap, one season is not enough.

Unfortunately, we were limited by the data unavailability for salaries which are not listed before 2006.

In total, there are 7 explanatory variables[4] that are not significant for the equation. To check for their joint significance, we ran a F-test, where

$$F = [(0.5319 - 0.5297)/(1 - 0.5319)] \times (1,072/7) = 0.7179$$

This result is below the 5% critical value and we cannot reject the null hypothesis of statistical insignificance. In other words, these variables are redundant for the model and the appropriate technical approach would be to exclude them from our research. However, their presence is crucial as racial dummies are of the main interest for the thesis and without them, our hypothesis would lose its meaning. Furthermore, season dummy variables include inflation and other both economic and non-economic factors. Their exclusion would lead to a comparison of nominal instead of real values of dollar. We provide at least the table C.1 (see Appendix C) that compares results between the restricted and unrestricted models. The restricted model could be used, if we were not interested in investigating the wage differential between races, but only in the determinants of the player's salary. We can see that the changes in estimates are negligible and including insignificant predictors does not cause any larger biases. Consequently, remaining two covariates were not omitted as they provide a valuable information about player's position, activity and accuracy.

As we move to the 1-season statistics OLS results, some changes were recorded as the adjusted R^2 dropped by two percentage points. While the majority of predictors have slightly low t-statistics, the variable implying how many shots a player had per season gained on its significance. On the other hand, the variable *midfielder* was no longer found to be significant at 10% significance level and its magnitude also dropped compared to the 3-seasons records. Since other estimates, including racial dummies, remained very similar, we proceed to the outcome of quantile regression which offers much more interesting discoveries.

[4] *Striker*, *Shots on goal*, *Black*, *Hispanic*, *year 2009*, *year 2008* and *year 2007*.

Table 6.2: OLS regressions of 3-seasons statistics

Variable	Model 1 Estimate (t-stats)	Model 2 Estimate (t-stats)	Model 3 Estimate (t-stats)	Model 4 Estimate (t-stats)
Constant	12.127*** (15.240)	12.125*** (15.229)	12.140*** (15.226)	12.126*** (15.166)
STRIKER	-0.152 (-1.545)	-0.152 (-1.545)	-0.153 (-1.549)	-0.204* (-1.678)
MIDFIELD	-0.176** (-2.053)	-0.176** (-2.053)	-0.176** (-2.057)	-0.164 (-1.536)
GS	0.063*** (2.653)	0.063*** (2.656)	0.063*** (2.665)	0.063*** (2.697)
MINS	-0.001*** (-3.141)	-0.001*** (-3.144)	-0.001*** (-3.152)	-0.001*** (-3.245)
GOALS	0.059*** (3.733)	0.058*** (3.734)	0.059*** (3.736)	0.057*** (3.648)
ASSISTS	0.045*** (3.975)	0.045*** (3.969)	0.045*** (3.953)	0.048*** (4.162)
SHTS	0.007* (1.850)	0.007* (1.835)	0.007* (1.820)	0.007* (1.842)
SOG	0.005 (0.459)	0.005 (0.466)	0.005 (0.478)	0.006 (0.572)
FOCO	-0.008*** (-2.691)	-0.008*** (-2.686)	0.008*** (-2.672)	-0.008*** (-2.705)
FOSU	0.004* (1.905)	0.004* (1.907)	0.004* (1.910)	0.004** (1.965)
YCARD	0.066*** (3.842)	0.066*** (3.839)	0.066*** (3.840)	0.067*** (3.909)
OFFS	-0.009** (-2.150)	-0.009*** (-2.154)	-0.009** (-2.150)	-0.009** (-2.259)
STAR	0.571*** (8.226)	0.571*** (8.224)	0.571*** (8.224)	0.561*** (8.000)
AGE	-0.160*** (-2.687)	-0.160*** (-2.683)	-0.161*** (-2.698)	-0.158*** (-2.663)
AGE^2	0.005*** (4.241)	0.005*** (4.236)	0.005*** (4.249)	0.004*** (4.221)
BLACK	-0.024 (-0.411)	-0.021 (-0.350)	-0.024 (-0.409)	-0.137 (-0.885)
HISP	-0.055 (-0.905)	-0.055 (-0.908)	-0.060 (-0.950)	0.160 (0.774)
year2016	0.774*** (6.677)	0.782*** (6.119)	0.761*** (6.126)	0.788*** (6.775)
year2015	0.854*** (7.408)	0.854*** (7.404)	0.856*** (7.410)	0.865*** (7.471)
year2014	0.581*** (5.015)	0.581*** (5.014)	0.583*** (5.021)	0.596*** (5.117)
year2013	0.551*** (4.785)	0.551*** (4.784)	0.553*** (4.791)	0.564*** (4.875)
year2012	0.549*** (4.888)	0.549*** (4.887)	0.550*** (4.893)	0.564*** (5.001)
year2011	0.366*** (3.279)	0.366*** (3.280)	0.367*** (3.286)	0.378*** (3.375)
year2010	0.297*** (2.689)	0.297*** (2.686)	0.297*** (2.694)	0.310*** (2.805)
year2009	0.136 (1.238)	0.136 (1.237)	0.136 (1.244)	0.143 (1.306)
year2008	0.035 (0.322)	0.035 (0.321)	0.035 (0.324)	0.039 (0.364)
year2007	0.001 (0.005)	0.001 (0.005)	0.001 (0.007)	0.003 (0.031)
BLACK*year2010	-	0.220 (1.256)	-	-
HISP*year2010	-	-	-0.214 (-1.109)	-
BLACK*STRIKER	-	-	-	0.162 (0.931)
BLACK*MIDFIELD	-	-	-	0.101 (0.561)
HISP*STRIKER	-	-	-	-0.149 (-0.663)
HISP*MIDFIELD	-	-	-	-0.322 (-1.441)
R^2	0.532	0.532	0.532	0.534
Adjusted R^2	0.520	0.520	0.520	0.520
N	1100	1100	1100	1100
F-statistic (DF)	45.11 (27; 1072)	43.46 (28; 1071)	43.47 (28; 1071)	39.45 (31; 1068)

*Significant at 10% level, **Significant at 5% level, ***Significant at 1% level

Table 6.3: OLS regressions of 1-season statistics

	Model 1	Model 2	Model 3	Model 4
Variable	Estimate (t-stats)	Estimate (t-stats)	Estimate (t-stats)	Estimate (t-stats)
Constant	12.255*** (15.109)	12.255*** (15.100)	12.252*** (15.077)	12.236*** (15.011)
STRIKER	-0.087 (-0.891)	-0.087 (-0.891)	-0.087 (-0.889)	-0.126 (-1.041)
MIDFIELD	-0.125 (-1.473)	-0.126 (-1.472)	-0.125 (-1.469)	-0.105 (-0.982)
GS	0.035* (1.789)	0.035* (1.786)	0.035* (1.781)	0.035* (1.793)
MINS	-0.001** (-2.319)	-0.001** (-2.314)	-0.001** (-2.308)	-0.001** (-2.374)
GOALS	0.043*** (3.546)	0.043*** (3.544)	0.043*** (3.545)	0.043*** (3.508)
ASSISTS	0.032*** (3.609)	0.032*** (3.606)	0.032*** (3.608)	0.033*** (3.712)
SHTS	0.009*** (3.018)	0.009*** (3.012)	0.009*** (3.016)	0.009*** (3.054)
SOG	-0.007 (-0.812)	-0.007 (-0.811)	-0.007 (-0.814)	0.006 (-0.760)
FOCO	-0.006** (-2.168)	-0.006** (-2.166)	-0.006** (-2.168)	-0.006*** (-2.239)
FOSU	0.006*** (2.934)	0.006*** (2.933)	0.006*** (2.931)	0.006*** (2.992)
YCARD	0.028** (2.049)	0.028*** (2.048)	0.028** (2.049)	0.029** (2.142)
OFFS	-0.005 (-1.509)	-0.005 (-1.508)	-0.005 (-1.510)	-0.006 (-1.601)
STAR	0.730*** (9.951)	0.730*** (9.945)	0.730*** (9.945)	0.724*** (9.800)
AGE	-0.169*** (-2.803)	-0.169*** (-2.801)	-0.169*** (-2.793)	-0.167*** (-2.763)
AGE^2	0.005*** (4.413)	0.005*** (4.410)	0.005*** (4.398)	0.005*** (4.377)
BLACK	-0.045 (-0.766)	-0.045 (-0.733)	-0.045 (-0.766)	-0.122 (-0.778)
HISP	-0.029 (-0.475)	-0.029 (-0.475)	-0.027 (-0.429)	0.145 (0.692)
year2016	0.770*** (6.570)	0.770*** (5.955)	0.773*** (6.157)	0.778*** (6.625)
year2015	0.887*** (7.730)	0.887*** (7.727)	0.887*** (7.717)	0.893 (7.759)
year2014	0.557*** (4.814)	0.557*** (4.812)	0.556*** (4.803)	0.565*** (4.869)
year2013	0.495*** (4.283)	0.495*** (4.784)	0.494*** (4.275)	0.502*** (4.335)
year2012	0.499*** (4.400)	0.549*** (4.281)	0.499*** (4.392)	0.508*** (4.468)
year2011	0.302*** (2.652)	0.302*** (2.650)	0.302*** (2.646)	0.309*** (2.708)
year2010	0.287** (2.544)	0.286** (2.543)	0.287** (2.540)	0.296*** (2.615)
year2009	0.168 (1.499)	0.168 (1.498)	0.167 (1.495)	0.173 (1.542)
year2008	0.021 (0.187)	0.021 (0.187)	0.021 (0.186)	0.023 (0.206)
year2007	0.062 (0.563)	0.062 (0.563)	0.062 (0.562)	0.061 (0.557)
BLACK*year2010	-	0.220 (1.256)	-	-
HISP*year2010	-	-	-0.214 (-1.109)	-
BLACK*STRIKER	-	-	-	0.122 (0.687)
BLACK*MIDFIELD	-	-	-	0.056 (0.302)
HISP*STRIKER	-	-	-	-0.113 (-0.497)
HISP*MIDFIELD	-	-	-	-0.265 (-1.168)
R^2	0.514	0.514	0.514	0.515
Adjusted R^2	0.502	0.501	0.501	0.501
N	1100	1100	1100	1100
F-statistic (DF)	41.98 (27; 1072)	40.44 (28; 1071)	40.44 (28; 1071)	36.61 (31; 1068)

*Significant at 10% level, **Significant at 5% level, ***Significant at 1% level

Even though the race has not played an important role for the OLS equation, we cannot speak about non-discriminatory behaviour in MLS because the problem of wage differentials might be rather rooted in the bottom part of the salary distribution as was uncovered by Holmes (2011) in Major League Baseball. He argued that premia for whites in the lower subset of the population is caused by relatively small costs for the employer because the performance of these players is not likely to be important for the team. Consequently, adopting the same method, we uncovered that salary differences for the players in the bottom decile and quartile are much higher and they are significant at the significance level where $\alpha = 0.05$. The final pay premia for whites were calculated using the transformation $100(exp(\widehat{\gamma_{1,2}}) - 1)\%$ that gives us the exact percentage change in predicted salaries (Wooldridge 2015). Once players' performance was observed for three seasons, blacks in the bottom decile received 18.86% less than their white counterparts. We can see that this difference falls as we get to the upper part of the salary distribution and for $\tau = 0.9$, there is even pay premium equal to 8.87% in favour of black players. However, the coefficients lose statistical significance once we exceed $\tau = 0.25$. A similar situation occur as far as Hispanic players are concerned. While the poorest 10% of them are deprived of 15.3% from their salaries compared to whites, the wage differential for the median player is only 6.85% and not statistically significant. This coefficient even goes to positive values once we examine the upper quartile.

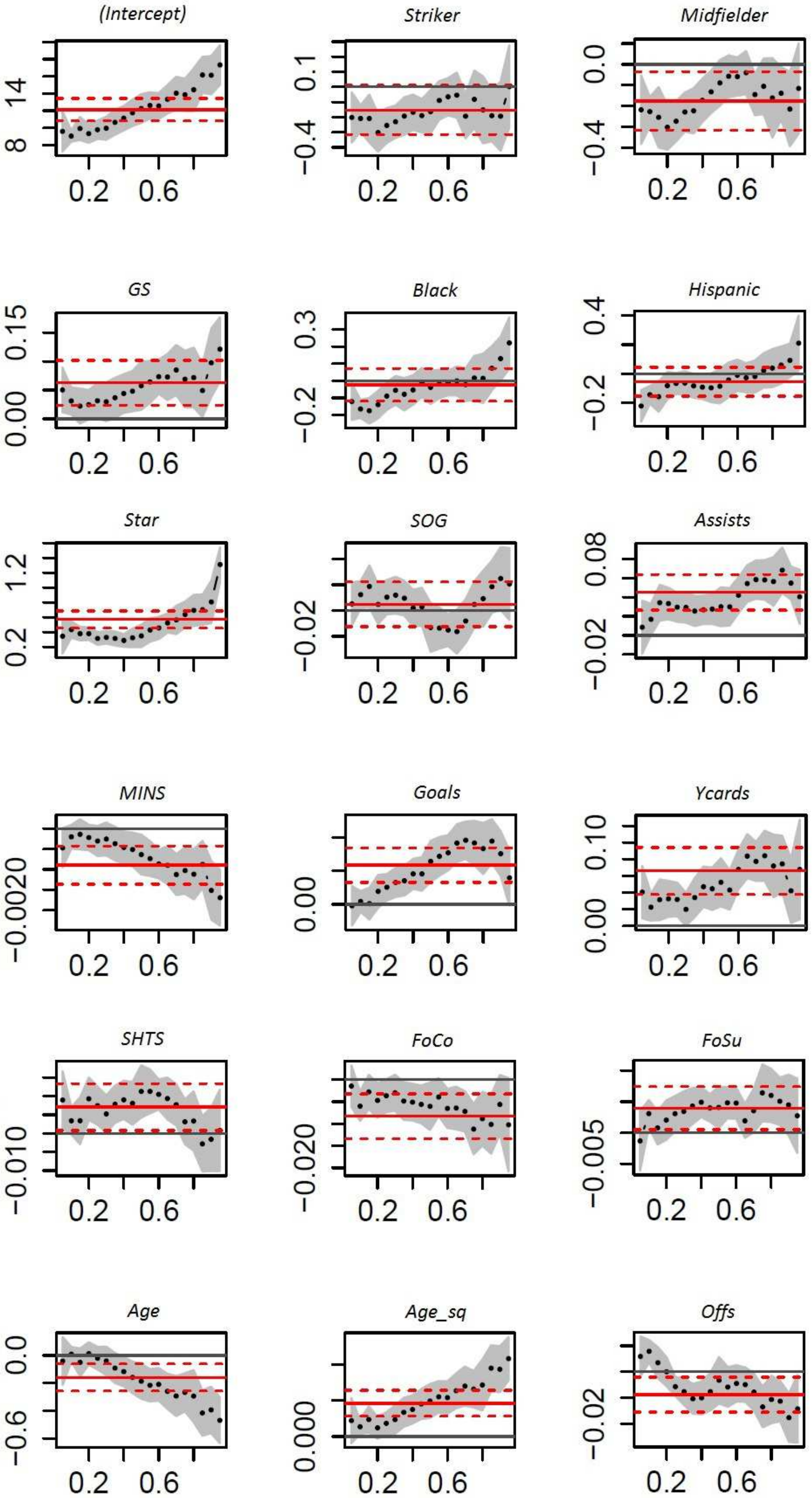
Table 6.4: Results of Quantile Regression

	1-season statistics				3-seasons statistics			
	Black		Hispanics		Black		Hispanics	
Quantile	Estimate	Wage gap	Estimate	Wage gap	Estimate	Wage gap	Estimate	Wage gap
10%	-0.090*	-8.61%	-0.142**	-13.24%	-0.209***	-18.86%	-0.166***	-15.30%
	(-1.953)		(-2.422)		(-4.792)		(-3.021)	
25%	-0.078*	-7.50%	-0.010	-1.00%	-0.120**	-11.31%	-0.101**	-9.61%
	(-1.727)		(-0.179)		(-2.570)		(-2.214)	
50%	-0.023	-2.27%	0.086	8.98%	-0.031	-3.05%	-0.071	-6.85%
	(-0.360)		(1.322)		(-0.547)		(-1.090)	
75%	-0.023	-2.27%	0.043	4.39%	0.008	0.80%	0.019	1.92%
	(-0.356)		(0.527)		(0.103)		(0.253)	
90%	0.056	5.76%	0.041	4.19%	0.085	8.87%	0.071	7.36%
	(0.643)		(0.546)		(0.915)		(1.063)	

t-statistics are noted in parentheses

*Significant at 10% level, **Significant at 5% level, ***Significant at 1% level

Figure 6.2: Quantile regression coefficients



Note: The axis x describes the corresponding quantile while the axis y shows the magnitude of coefficient. The grey area represents a 90% confidence interval (CI), so if the interval does not include zero, the coefficient is in a given quantile significantly different from zero. The line is from OLS and the dashed lines represent 90% CI. All graphs perfectly demonstrate that the coefficients change as we examine different parts of the distribution.

Chapter 7

Conclusion

This bachelor thesis investigated wage discrimination in professional football using two different econometric methods. The first one is called Ordinary Least Squares and it estimates the effect of race for an average player in the population. On the other hand, quantile regression offers different and more complex data analysis as it enables to examine the effect for individual points of the salary distribution. For both methods, we used the same variables about players' performance, age, race and also the season from which given salary was observed. Salaries were observed *ex post* and thanks to the explanatory variables included into our model, we were able to explain more than 50% of the variance in the dependent variable.

To investigate a potential racial salary discrimination, we have collected seasonal MLS data from 2004-2016. Our large dataset, composed of a rich set of explanatary variables, was focused on the most productive players from individual seasons. We discovered a strong economical and statistical significance of wage discrimination in the lower part of the salary distribution which is aimed against Hispanic and African American players. Using the method of ordinary least squares we found a wage differential for the average player of 2.37% and 5.35% that is accrued against black and Latin players. These values were far from being statistically significant, though. Quantile regression reveals that the premium is much higher for poorer players. When the lowest quartile was the main focus of our observation, the wage gap against players from Latin America reached 9.61%. Similarly, blacks suffer from salary inequality that amounts to 11.31%. Once we looked at the bottom decile, the situation appeared to be more dramatical. Despite the same performance, black players receive 18.86% less. By implication, our thesis revealed that poorer non-white players are more vulnerable as the employers pay them less despite the economic inefficiency. The crucial lesson is that salary discrimination in professional football exists in spite of the ample amount of supporters and performance analysts. These findings were enabled especially thanks to the large amount of observations that we have collected. Lately, significant results have been rarely found, even though they exist, albeit only for a subset of the entire population. Quantile regression is therefore a very powerful tool for detecting discrimination, but the sample needs to be sufficiently large to observe differentials at individual quantiles.

In the proposal, we stated 2 other research questions, which, given the dataset we worked with, we were able to answer. With regards to a possible racial discrimination on different positions, we did not find any statistical evidence. Even though the race plays an important role in determining which position each footballer will play on, salaries are distributed justly according to the player's performance. The last question we wanted to clarify was whether or not the increasing interest of fans and investors in MLS, and introduction of new salary policies, have caused any wage gap. We used DID estimator to find quite high, but insignificant effect of the increasing wage level on compensations of black and Hispanic players.

The results of our thesis bring very surprising findings. The evidence of racial discrimination in MLS might prompt others to examine and search for racial discrimination in other sports areas. With regards to professional football, the improvement of our analysis could be done, if the European best competitions decided to publish yearly data about players' salaries. In spite of the increasing quality level in MLS, this league still lags behind the top leagues from England, Spain or Germany where the salaries and differences between them are incomparably higher (BBC 2015). However, for now, the reveal of salary fees from European clubs is hardly expected as both clubs and players are afraid of potential conflicts with tax authorities.

Bibliography

Altonji, J. G. & C. R. Pierret (2001): "Employer learning and statistical discrimination." *The Quarterly Journal of Economics* **116(1)**: pp. 313–350.

Arrow, K. (1971): "The theory of discrimination. Princeton university, department of economics." *Industrial Relations Section* pp. 313–350.

Arrow, K. J. (1998): "What has economics to say about racial discrimination?" *The Journal of Economic Perspectives* **12(2)**: pp. 91–100.

Ashenfelter, O. C. & D. Card (1984): "Using the longitudinal structure of earnings to estimate the effect of training programs."

BBC (2015): "How long would it take you to earn a top footballer's salary?" [online], Available at: `http://www.bbc.com/news/world-31110113`, [Accessed: July 12, 2017].

Becker, G. (1971): "The economics of discrimination 2nd ed. (University of Chicago Press, Chicago)."

Blau, F. D. & L. M. Kahn (2016): "The gender wage gap: Extent, trends, and explanations." *Technical report*, National Bureau of Economic Research.

Brown, E., R. Spiro, & D. Keenan (1991): "Wage and nonwage discrimination in professional basketball: do fans affect it?" *American Journal of Economics and Sociology* **50(3)**: pp. 333–345.

Cain, G. G. (1986): "The economic analysis of labor market discrimination: A survey." *Handbook of Labor Economics* **1**: pp. 693–785.

Dunning, E. *et al.* (1999): "The development of soccer as a world game." *Sport Matters: Sociological Studies of Sport, Violence and Civilisation.* pp. 80–105.

EITZEN, D. S., G. H. SAGE *et al.* (1982): *Sociology of American sport.* Wm. C. Brown Company Publishers.

ESPNFC.COM (2017): "MLS pumping money into salaries causes ripple effect throughout league." [online], Available at: `http://www.espnfc.com/major-league-soccer/19/blog/post/3112526/major-league-soccer-pumping-money-into-salaries-causes-ripple-effect-throughout-league`, [Accessed: June 12, 2017].

FANG, H. & A. MORO (2010): "Theories of statistical discrimination and affirmative action: A survey." *Technical report*, National Bureau of Economic Research.

FEINBERG, W. (2003): "Affirmative action." *The Oxford Handbook of Practical Ethics* .

FIFA.COM (2007): "265 million playing football." [online], Available at: `https://www.fifa.com/mm/document/fifafacts/bcoffsurv/emaga_9384_10704.pdf`, [Accessed: March 25, 2017].

FORBES.COM (2017): "MLS records banner year in 2016, cements position among top u.s. pro sports leagues." [online], Available at: `https://www.forbes.com/sites/markjburns/2016/10/26/mls-records-banner-year-in-2016-cements-position-among-top-us-pro-sports-leagues/`, [Accessed: June 12, 2017].

FRICK, B. (2006): "Salary determination and the pay-performance relationship in professional soccer: Evidence from Germany." *Sports Economics After Fifty Years: Essays in Honour of Simon Rottenberg. Oviedo: Ediciones de la Universidad de Oviedo* pp. 125–146.

FRYER JR, R. G. (2010): "Racial inequality in the 21st century: The declining significance of discrimination." *Technical report*, National Bureau of Economic Research.

GWARTNEY, J. & C. HAWORTH (1974): "Employer costs and discrimination: The case of baseball." *Journal of Political Economy* **82(4)**: pp. 873–881.

HAMILTON, B. H. (1997): "Racial discrimination and professional basketball salaries in the 1990s." *Applied Economics* **29(3)**: pp. 287–296.

HILL, J. R. (2004): "Pay discrimination in the NBA revisited." *Quarterly Journal of Business and Economics* **43(1/2)**: pp. 81–92.

HOLMES, P. (2011): "New evidence of salary discrimination in Major League Baseball." *Labour Economics* **18(3)**: pp. 320–331.

JIOBU, R. M. (1988): "Racial inequality in a public arena: The case of professional baseball." *Social Forces* **67(2)**: pp. 524–534.

KAHN, L. M. (1991): "Discrimination in professional sports: A survey of the literature." *ILR Review* **44(3)**: pp. 395–418.

KAHN, L. M. & M. SHAH (2005): "Race, compensation and contract length in the NBA: 2001–2002." *Industrial Relations: A Journal of Economy and Society* **44(3)**: pp. 444–462.

KAHN, L. M. & P. D. SHERER (1988): "Racial differences in professional basketball players' compensation." *Journal of Labor Economics* **6(1)**: pp. 40–61.

KOENKER, R. & G. BASSETT JR (1978): "Regression quantiles." *Econometrica: Journal of the Econometric Society* pp. 33–50.

KOENKER, R. & K. HALLOCK (2001): "Quantile regression: An introduction." *Journal of Economic Perspectives* **15(4)**: pp. 43–56.

LANG, K., J. LEHMANN, & K. YEON (2012): "Racial discrimination in the labor market: Theory and empirics." *Journal of Economic Literature* **50(4)**: pp. 959–1006.

LANG, K. & M. MANOVE (2011): "Education and labor market discrimination." *The American Economic Review* **101(4)**: pp. 1467–1496.

LAPCHICK (2016): "The 2016 racial and gender report card: Major League Soccer." [online], Available at: `https://www.mlssoccer.com/glossary/targeted-allocation-money`, [Accessed: July 24, 2017].

MLSSOCCER.COM (1996): "1996 season recap." [online], Available at: `https://www.mlssoccer.com/history/season/1996`, [Accessed: May 30, 2017].

MLSSOCCER.COM (2010): "MLS joins in FIFA Say No to Racism Day." [online], Available at: `https://www.mlssoccer.com/post/2010/01/23/mls-joins-fifa-say-no-racism-day`, [Accessed: May 16, 2017].

MLSSOCCER.COM (2017): “Targeted allocation money.” [online], Available at: https://www.mlssoccer.com/glossary/targeted-allocation-money, [Accessed: May 24, 2017].

MUELLER, F., R. CANTU, & S. VAN CAMP (1996): “Team sports.” *Catastrophic Injuries in High School and College Sports* p. 57.

NARDINELLI, C. & C. SIMON (1990): “Customer racial discrimination in the market for memorabilia: The case of baseball.” *The Quarterly Journal of Economics* **105(3)**: pp. 575–595.

OAXACA, R. (1973): “Male-female wage differentials in urban labor markets.” *International Economic Review* pp. 693–709.

OXFORD UNIVERSITY (2017): “Definition of discrimination; origin.” [online], Available at: https://en.oxforddictionaries.com/definition/discriminate, [Accessed: March 11, 2017].

PAGER, D., B. BONIKOWSKI, & B. WESTERN (2009): “Discrimination in a low-wage labor market: A field experiment.” *American Sociological Review* **74(5)**: pp. 777–799.

REHNSTROM, K. (2009): “Racial salary discrimination in the NBA: 2008-2009.” *Major Themes in Economics* **11**: pp. 1–16.

REUTERS (2017): “Beckham put MLS on fast track to respectability.” [online], Available at: http://www.reuters.com/article/us-soccer-usa-beckham-idUSKBN14U2E5, [Accessed: May 12, 2017].

ROSEN, S. & A. SANDERSON (2001): “Labour markets in professional sports.” *The Economic Journal* **111(469)**: pp. 47–68.

SCULLY, G. W. (1973): “Economic discrimination in professional sports.” *Law and Contemporary Problems* **38(1)**: pp. 67–84.

SCULLY, G. W. (1974): “Pay and performance in Major League Baseball.” *The American Economic Review* **64(6)**: pp. 915–930.

SEARS, D. O. (1988): “Symbolic racism.” In “Eliminating racism,” pp. 53–84. Springer.

SOMMERS, P. M. & N. QUINTON (1982): "Pay and performance in Major League Baseball: The case of the first family of free agents." *The Journal of Human Resources* **17(3)**: pp. 426–436.

SPORTINGINTELLIGENCE.COM (2016): "Global sports salaries survey 2016." [online], Available at: `https://www.globalsportssalaries.com/GSSS%202016.pdf`, [Accessed: March 14, 2017].

STATISTA.COM (2017): "Average per game attendance of the five major sports leagues in North America 2016/17." [online], Available at: `https://www.statista.com/statistics/207458/per-game-attendance-of-major-us-sports-leagues/`, [Accessed: June 18, 2017].

SZYMANSKI, S. (2000): "A market test for discrimination in the English professional soccer leagues." *Journal of Political Economy* **108(3)**: pp. 590–603.

TAKAKI, R. T. (1979): *Iron Cages: Race and Culture in 19th-century America.* Oxford University Press, USA.

WELLS, L. T., N. ALLEN, J. MORISSET, & N. PIRNIA (2001): *Using Tax Incentives to Compete for Foreign Investment: Are They Worth the Cost?* Washington, DC: FIAS.

WILSON, D. P. & Y.-H. YING (2003): "Nationality preferences for labour in the international football industry." *Applied Economics* **35(14)**: pp. 1551–1559.

WOOLDRIDGE, J. M. (2015): *Introductory Econometrics: A Modern Approach.* Nelson Education.

YANG, C.-H. & H.-Y. LIN (2012): "Is there salary discrimination by nationality in the NBA? foreign talent or foreign market." *Journal of Sports Economics* **13(1)**: pp. 53–75.

ZIMBALIST, A. (2003): "Sport as business." *Oxford Review of Economic Policy* **19(4)**: pp. 503–511.

Appendix A

OLS Assumptions

1. **Linear in Parameters:** In the population model, the relation between the dependent variable and the independent variables is assumed to be linear. Therefore, $\vec{\alpha}$, $\vec{\beta}$ and $\vec{\gamma}$ are linear vectors.

2. **Random Sampling:** Our sample consists of all players who satisfy given condition. Once this condition was satisfied, we measured characteristics of all players to use this information and make an inference about the entire population. To prove the randomness of our sample, we compared the racial composition with the data reported by the Institute for Diversity and Ethnics in Sport. Values of 48.8% (whites) and 24.8% (Latinos) from our dataset were almost perfectly matched by 48% and 24.8% (Lapchick 2016). The proportion of black players is split into national and international, but as the percentage of Asians is only 0.7%, their representation in our sample will be very similar as well.

3. **No Perfect Collinearity:** In our sample, there is no variable that is constant, and there are no perfect linear relationships among explanatory variables included into the model. If it was so, our matrix **X** of explanatory variables would be singular and it would have perfect multilinear dependence. In this case, we would not be able to compute OLS estimator since matrix (**X'X**) could not be inverted. Table A.1 depicts mutual correlations between independent variables and most of the values show no signs of perfect collinearity. Very strong collinearity is between *Age* and Age^2, but the inclusion of both of them was necessary due to the non-linear relationship between the player's age and salary.

4. **No serial correlation:** This assumption rules out the correlation of

Table A.1: Correlation of explanatory variables

Variable	Str	Midf	Star	SOG	GS	Black	Hisp	MINS	G	A	SHTS	FoCo	FoSu	Offs	Yc	Age	Age^2
Str	1	-0.79	-0.05	0.38	-0.28	0.13	0.02	-0.28	0.44	-0.13	0.31	-0.11	-0.04	0.56	-0.28	-0.01	-0.01
Midf	-0.79	1	0.02	-0.15	0.2	-0.14	0.05	0.19	-0.25	0.28	-0.08	0.09	0.15	-0.39	0.18	0.02	0.02
Star	-0.05	0.02	1	0.16	0.22	-0.01	-0.12	0.24	0.17	0.2	0.14	0.13	0.21	0.04	0.13	0.18	0.18
SOG	0.38	-0.15	0.16	1	0.38	0.01	0.09	0.38	0.88	0.41	0.95	0.23	0.41	0.68	0.02	0.13	0.13
GS	-0.28	0.2	0.22	0.38	1	-0.12	0.05	0.99	0.29	0.45	0.44	0.51	0.52	0.1	0.41	0.2	0.19
Black	0.13	-0.14	-0.01	0.01	-0.12	1	-0.34	-0.11	0.05	-0.15	0	0	-0.01	0.11	-0.11	-0.11	-0.11
Hisp	0.02	0.05	-0.12	0.09	0.05	-0.34	1	0.03	0.09	0.15	0.09	0.09	0.15	0.05	0.13	0.08	0.09
MINS	-0.28	0.19	0.24	0.38	0.99	-0.11	0.03	1	0.3	0.45	0.44	0.52	0.52	0.1	0.42	0.19	0.17
G	0.44	-0.25	0.17	0.88	0.29	0.05	0.09	0.3	1	0.32	0.82	0.16	0.31	0.67	-0.04	0.2	0.20
A	-0.13	0.28	0.2	0.41	0.45	-0.15	0.15	0.45	0.32	1	0.43	0.18	0.48	0.11	0.13	0.26	0.26
SHTS	0.31	-0.08	0.14	0.95	0.44	0	0.09	0.44	0.82	0.43	1	0.27	0.44	0.62	0.07	0.12	0.12
FoCo	-0.11	0.09	0.13	0.23	0.51	0	0.09	0.52	0.16	0.18	0.27	1	0.5	0.12	0.63	0.07	0.05
FoSu	-0.04	0.15	0.21	0.41	0.52	-0.01	0.15	0.52	0.31	0.48	0.44	0.5	1	0.21	0.31	0.11	0.10
Offs	0.56	-0.39	0.04	0.68	0.1	0.11	0.05	0.1	0.67	0.11	0.62	0.12	0.21	1	-0.06	0.17	0.17
Yc	-0.28	0.18	0.13	0.02	0.41	-0.11	0.13	0.42	-0.04	0.13	0.07	0.63	0.31	-0.06	1	0.14	0.127
Age	-0.01	0.02	0.18	0.13	0.2	-0.11	0.08	0.19	0.2	0.26	0.12	0.07	0.11	0.17	0.14	1	0.996
Age^2	-0.01	0.02	0.18	0.13	0.19	-0.11	0.09	0.17	0.2	0.26	0.12	0.05	0.1	0.17	0.13	0.996	1

errors from two different time periods. This is not a problem for our data as we did not observe the same individuals for each time period.

5. **Homoskedasticity:** This assumption ensures that the error u has constant variance for all values of independent variables. The most common tests for homoskedasticity are White's test and Breusch-Pagan. The B-P test uncovers heteroskedasticity and hence, we also present White's standard errors that are heteroscedasticity-consistent. The results show that the majority of explanatory variables preserve their statistical significance (see Appendix B). We reject the null hypothesis of homoskedasticity at a very low p-value.

Studentized Breusch-Pagan test: H_0: Homoskedasticity

$$\text{BP} = 127.0637,\ \text{df} = 27,\ \text{p-value} = 6.396 \times 10^{-7}$$

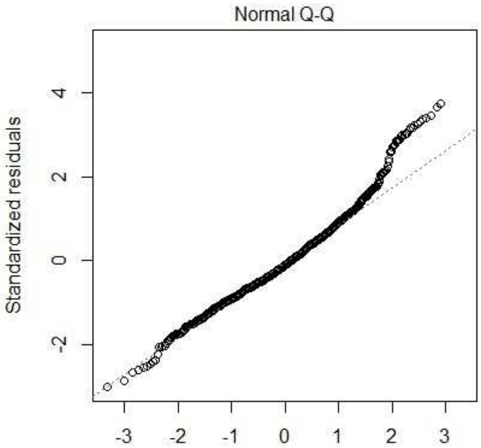
6. **Normality:** The errors u are not dependent on explanatory variables and are normally distributed with zero mean and variance equal to σ^2 for the whole population: $u \sim$N(0,σ^2). There are two common hypotheses about how to check this assumption to be valid - The Anderson-Darling test and the Shapiro-Wilk test. The null hypothesis H_0 is that the data is normally distributed for both of them. Running the S-W test we reject the null hypothesis of normality.

Shapiro-Wilk normality test: H_0: Normality of errors

$$\mathrm{W} = 0.9719, \text{ p-value} = 8.868 \times 10^{-6}$$

From the figure A.1 we can observe that the errors are right skewed, meaning that most of the error terms are with a long tail on the left side of the distribution. The shape of the Q-Q plot is heavily influenced by the extreme values in the highest decile of the salary distribution. Removing these outliers would cause a loss of a valuable information, though. Even though our sample consists of 1,100 observations, relying on Central Limit Theorem may not be appropriate due to long-tailed errors. Consequently, we also run a robust regression (see Appendix B) that weights and dampens the effect of outliers to find out that both results and statistical significances[1] are very similar and inference about discrimination is not influenced. As both assumptions are not required for the quantile regression, we can rely on the discriminating behaviour we have uncovered in the lower part of the salary distribution.

Figure A.1: Q-Q plot of the regression



[1] The only variable that differs across different method for 3-seasons statistics is the position of striker. Its negative coefficient is different from zero using robust regression.

Appendix B

Robust methods

Table B.1: Comparison between OLS and White's standard errors

	1-season statistics				3-seasons statistics			
Variable	Estimate	Robust SE	OLS SE	Stat. Significance	Estimate	Robust SE	OLS SE	Stat. Significance
Constant	12.255	0.984	0.811	***	12.127	0.949	0.796	***
STRIKER	-0.087	0.090	0.098	-	-0.152	0.090	0.099	*/-
MIDFIELD	-0.125	0.072	0.085	*	-0.176	0.072	0.085	**
GS	0.035	0.019	0.020	*	0.063	0.024	0.024	**
MINS	-0.001	0.0002	0.0002	**	-0.001	0.0003	0.0003	***
GOALS	0.043	0.014	0.012	***	0.059	0.018	0.015	***
ASSISTS	0.032	0.011	0.009	***	0.045	0.013	0.011	***
SHTS	0.009	0.003	0.003	***	0.007	0.004	0.004	*
SOG	-0.007	0.008	0.008	-	0.005	0.011	0.011	-
FOCO	-0.006	0.003	0.002	*/**	-0.008	0.003	0.003	**/***
FOSU	0.006	0.002	0.001	***	0.004	0.002	0.002	*
YCARD	0.028	0.015	0.014	*/**	0.066	0.018	0.017	***
OFFS	-0.005	0.003	0.003	-	-0.009	0.004	0.004	**
STAR	0.730	0.091	0.073	***	0.571	0.084	0.069	***
AGE	-0.169	0.076	0.060	**/***	-0.160	0.073	0.059	**/***
AGESQ	0.005	0.001	0.001	***	0.005	0.001	0.001	***
BLACK	-0.045	0.060	0.059	-	-0.024	0.059	0.058	-
HISP	-0.029	0.066	0.061	-	-0.055	0.064	0.060	-
year2016	0.770	0.112	0.117	***	0.774	0.117	0.116	***
year2015	0.887	0.118	0.115	***	0.854	0.124	0.115	***
year2014	0.557	0.110	0.116	***	0.581	0.114	0.116	***
year2013	0.495	0.101	0.116	***	0.551	0.105	0.115	***
year2012	0.499	0.109	0.113	***	0.549	0.110	0.112	***
year2011	0.302	0.111	0.114	***	0.366	0.111	0.111	***
year2010	0.287	0.110	0.113	***/**	0.297	0.106	0.110	***
year2009	0.168	0.103	0.112	-	0.135	0.102	0.109	-
year2008	0.021	0.106	0.111	-	0.035	0.106	0.108	-
year2007	0.062	0.103	0.110	-	0.001	0.103	0.107	-

*Significant at 10% level, **Significant at 5% level, ***Significant at 1% level

Table B.2: Comparison between OLS and Robust regressions

	1-season statistics					3-seasons statistics				
Variable	OLS	OLS SE	Robust	Robust SE	Stat. Significance	OLS	OLS SE	Robust	Robust SE	Stat. Significance
Constant	12.255	0.811	12.051	0.749	***	12.127	0.796	12.071	0.735	***
STRIKER	-0.087	0.098	-0.09	0.09	-	-0.152	0.099	-0.189	0.091	-/**
MIDFIELD	-0.125	0.085	-0.111	0.079	*/-	-0.176	0.085	-0.191	0.079	**
GS	0.035	0.02	0.024	0.018	*/-	0.063	0.024	0.051	0.022	**
MINS	-0.0007	0.0003	-0.0004	0.0002	**/*	-0.001	0.0003	-0.0007	0.0003	***
GOALS	0.043	0.012	0.041	0.011	***	0.059	0.015	0.061	0.015	***
ASSISTS	0.032	0.009	0.021	0.008	***/**	0.045	0.011	0.044	0.011	***
SHTS	0.009	0.003	0.01	0.003	***	0.007	0.004	0.008	0.004	*/**
SOG	-0.007	0.008	-0.010	0.008	-	0.005	0.011	0.0002	0.01	-
FOCO	-0.006	0.002	-0.005	0.002	**	-0.008	0.003	-0.007	0.003	***/**
FOSU	0.006	0.001	0.005	0.002	***	0.004	0.002	0.004	0.002	*/**
YCARD	0.028	0.014	0.022	0.013	**/*	0.066	0.017	0.059	0.016	***
OFFS	-0.005	0.003	-0.004	0.003	-	-0.009	0.004	-0.007	0.004	**/*
STAR	0.730	0.073	0.719	0.068	***	0.571	0.069	0.478	0.064	***
AGE	-0.169	0.06	-0.158	0.056	***	-0.16	0.059	-0.163	0.055	***
AGESQ	0.005	0.001	0.005	0.001	***	0.005	0.001	0.005	0.001	***
BLACK	-0.045	0.059	-0.054	0.055	-	-0.024	0.058	-0.040	0.054	-
HISP	-0.029	0.061	0.007	0.056	-	-0.055	0.060	-0.053	0.056	-
year2016	0.770	0.117	0.790	0.108	***	0.774	0.116	0.776	0.107	***
year2015	0.887	0.115	0.844	0.106	***	0.854	0.115	0.800	0.107	***
year2014	0.557	0.116	0.576	0.107	***	0.581	0.116	0.578	0.107	***
year2013	0.495	0.116	0.506	0.107	***	0.551	0.115	0.542	0.106	***
year2012	0.499	0.113	0.536	0.105	***	0.549	0.112	0.554	0.104	***
year2011	0.302	0.114	0.306	0.105	***	0.366	0.111	0.394	0.103	***
year2010	0.287	0.113	0.322	0.104	**/***	0.297	0.110	0.342	0.102	***
year2009	0.168	0.112	0.218	0.103	-/**	0.135	0.109	0.187	0.101	-
year2008	0.021	0.111	0.075	0.102	-	0.035	0.108	0.081	0.100	-
year2007	0.062	0.110	0.101	0.102	-	0.001	0.107	0.039	0.099	-

*Significant at 10% level, **Significant at 5% level, ***Significant at 1% level

Note: The method we have used is called Huber's M-estimator and it deadens the effect of outliers.

Appendix C

Unrestricted vs. restricted model

Table C.1: Results after the exclusion of insignificant variables

Variable	Before (t-stats)	After (t-stats)	Difference	Percentage change
Constant	12.127*** (-15.24)	11.884*** (15.231)	0.243	27.51%
MIDFIELD	-0.176** (-2.053)	-0.077 (-1.385)	-0.099	-9.43%
GS	0.063*** (-2.653)	0.062*** (2.662)	0.001	0.1%
MINS	-0.001*** (-3.141)	-0.001*** (-2.98)	0	0%
GOALS	0.059*** (-3.733)	0.06*** (4.645)	-0.001	-0.1%
ASSISTS	0.045*** (-3.975)	0.044*** (3.953)	0.001	0.1%
SHTS	0.007* (-1.85)	0.008*** (3.757)	-0.001	-0.1%
FOCO	-0.008*** (-2.691)	-0.009** (-3.088)	0.001	0.1%
FOSU	0.004* (-1.905)	0.003 (1.546)	0.001	0.1%
YCARD	0.066*** (-3.842)	0.069*** (4.144)	-0.003	-0.3%
OFFS	-0.009** (-2.150)	-0.01** (-2.521)	0.001	0.1%
STAR	0.571*** (-8.226)	0.583*** (8.512)	-0.012	-1.19%
AGE	-0.160*** (-2.687)	-0.148** (-2.517)	-0.012	-1.19%
AGE^2	0.005*** (-4.241)	0.004*** (4.092)	0.001	0.1%
year2016	0.774*** (-6.677)	0.708*** (7.96)	0.066	6.82%
year2015	0.854*** (-7.408)	0.783*** (8.844)	0.071	7.36%
year2014	0.581*** (-5.015)	0.507*** (5.773)	0.074	7.68%
year2013	0.551*** (-4.785)	0.48*** (5.562)	0.071	7.36%
year2012	0.549*** (-4.888)	0.484*** (5.618)	0.065	6.72%
year2011	0.366*** (-3.279)	0.301*** (3.51)	0.065	6.72%
year2010	0.297*** (-2.689)	0.239** (2.788)	0.058	5.97%
R^2	0.532	0.530		
Adjusted R^2	0.520	0.521		
N	1100	1100		
F-statistic (DF)	45.11 (27; 1072)	60.76 (20; 1079)		

*Significant at 10% level, **Significant at 5% level, ***Significant at 1% level