Erasmus University Rotterdam

Erasmus School of Economics

29 – 07 – 2024

International Economics & Business Economics

Bachelor Thesis

The effects on differences in performance on wages in the National Basketball Association

Student name: Daron Arselan
Student number: 605172

Supervisor: prof.dr.ig. (Jan) JC van Ours
Second assessor: dr. (Thomas) TLPR Peeters

Abstract

This thesis researches whether an increase in performance measures in contract years result into an increase in salary. The seasons 2011-2012 through 2022-2023 are sampled, dropping the top and bottom observations to prevent skewness. Determening the effect is done by tying wages of subsequent years to player statistics to judge on performance rewards, and analysing whether a rise in player performance statistics results in higher pay. First, the sample is assessed on whether it is similar to past papers. Afterwards, an analysis solely looking at years prior to transfer is used to answer the hypothesis. In the results, we find that changes in age and the past wages are the only determinant as a prediction for wages. Concerning causal inference, it is hard to determine any causal effect since the research done is observational. unobserved factors and potential reverse causality remain concerns that could affect the interpretation of results.

Table of Contents

1. Introduction and research question

The National Basketball Association (NBA) is a premier professional basketball league globally, renowned for its high level of competition and lucrative player contracts. However, the determination of these contracts and the relationship between a player's performance and their pay remains a topic of considerable interest and debate. This thesis aims to investigate the effect of performance on pay within the NBA, providing an empirical analysis that contributes to the broader discourse on wage determination in professional sports.

Professional sports, including the NBA, offer a unique context to study wage determination due to the availability of quantifiable performance measures and publicly disclosed salaries. Despite this, the correlation between performance and pay is not always straightforward. Factors such as marketability, tenure, potential, and team budget constraints often influence a player's salary, potentially leading to discrepancies between pay and performance.

This study will employ econometric models to analyse the relationship between various performance indicators (such as points per game, assists, rebounds, and player efficiency rating) and player salary increases in contract renewals. By doing so, it aims to shed light on the extent to which performance influences pay in the NBA and identify any potential anomalies or biases in wage determination. The research question is formulated as follows:

"Are wage increases of an NBA player in contract renewals determined by game performance?"

The findings of this research could have significant implications for players, agents, and team managers. For players and agents, understanding the factors that influence pay could inform contract negotiations and career decisions. For team managers, insights into the relationship between performance and pay could aid in budget allocation and team building.

Moreover, this study contributes to the academic literature on wage determination in professional sports, a topic that intersects labour economics, sports management, and behavioural economics. It also has broader societal implications, as professional sports often mirror societal trends and attitudes, including those related to compensation and fairness.

This thesis is structured as follows: Chapter 2 reviews the relevant literature on wage determination in professional sports, Chapter 3 describes the data and methodology used in this study, Chapter 4 presents the empirical results, Chapter 5 concludes with a discussion of the findings and their implications, and Chapter 6 concludes the thesis with the limitations and recommendations for future research.

Hypothesis

The maintained hypothesis is that salary depends on the employer's perception of worker ability. Lots of research has gone into professional sports, and the hypothesis from this thesis might only be unique in very detail. Wage equations have been generated and researched successfully in the papers mentioned. However, the NBA is ever-changing and evolving. Establishing the same conclusion using more recent data is valuable on its own. The goal of this thesis paper is to shed light on what happens in the year prior to contract signing, the contract year. And whether an increase of performance in contract years comes with an increase in pay.

The hypothesis formulated is:

H1: An increase in performance in contract years, relative to previous contract years, has an increasing causal effect on salary.

Inversely, but not a one-to-one opposite, a player might be paid less when performing below the league average. The reason the above hypothesis is not phrased to include both is because teams might be reluctant to decrease pay due to negative reactions from the player. Therefore, the following hypothesis is also formed:

H2: A decrease in performance in contract years, relative to previous contract years, has a decreasing causal effect on salary.

2. Literature review

The goal of this literature review is to analyse the existing work in the field. The reason this is important is that insights can be drawn from research done by professionals. Therefore, the literature review will consist of papers that revolve around wages in professional sports. The structure of this literature review is as follows. First, papers that sample the MLB will be examined and summarized, adding similar relevant studies using the NBA as a sample when possible. This is because researchers started their interest in professional sports with the MLB. Next, we transition to papers that sample the NBA for their research, and note why they are of interest to this thesis paper.

MLB-focused Literature

The credited first paper introducing economic labour analysis is Rottenberg (1956), a paper on the imperfect market of the MLB in the fifties. The very competitive European league, with its relegations and incentives to get into first place, makes the perfect counterfactual according to Rottenberg. He considers alternatives to the reserve clause (Baseball's reserve clause bound a player to his team, and unless released, a player was forbidden to negotiate with any other team. The player either accepted what was offered or sat out the season in hopes that the owner would meet his demands.), such as equal revenue sharing, maximum salary limits, equal market franchise distribution, and roster limits. Each of these is rejected in favour of a free market solution which, on the basis of the invariance principle, he suggests will perform just as well as the reserve clause in allocating talent to where it is most productive.

A paper that goes deeper into the pay of athletes, and more specifically, the exploitation of athletes is by Scully (1974). Exploitation is measured by comparing their salary with the marginal revenue they bring in. The salary calculation includes an array of performance measures. The approach used in finding the relevant measures is particularly interesting for the purpose of this thesis, as the variables were not selected entirely on a priori grounds. Every single performance measure has been included, and the ones that seem relevant remained. The paper concludes that, with statistically significant evidence, players' salary includes about 11% to 20% of their net revenue. It is worth noting that this study was conducted during a time when MLB players were battling for the free contract employment scheme we see today, which was established a year after this paper was published.

Using the same method as Scully (1974), as referenced in the paper itself, to compare team performance effects on their payroll in the MLB and the Premier League (Hall et al., 2002). In the interest of this thesis, the Granger causality test is employed to check for one-way causality. It is shown that since 1995, for both sports, player cost and performance are two-way causal. The reasons given, being tighter trade restrictions than English football, and the national television contract dropping by 60%, could have similar effects in basketball.

The MLB papers show insights into the most researched professional sport, and show the history of statistical analysis on sports. To say this is pivotal for the thesis is an overstatement, but understanding the history of a field is important for critical thinking. Next, research regarding the NBA will be summarized and the importance of the paper regarding this thesis will be stated.

NBA-focused Literature

Lots of examples of early research around the NBA consist of papers that examine the effect and circumstances of black Americans joining the league after desegregation. One of these papers is by Johnson & Marple (1973), where barriers to entry to the NBA are examined. The paper empirically examines hypotheses brought forward by Blalock (1962), one of which stated that blacks would only be hired when they were exemplars in their field. Johnson and Marple found no empirical proof supporting these hypotheses when examining the NBA. The hypothesis that spots for benchwarmers and subpar players in the league were also reserved for whites could also not be confirmed. There was, however, evidence that mediocre black players had a shorter career than mediocre white players. What that means for this thesis paper is that there is no need for variables indicating race in the regression formula.

The assumptions that will be discussed rest heavily on economic market theory, and Kahn (2000) looked past the high wages in professional sports and examined whether market theories hold. The article describes the dynamics of four theories: monopsony, discrimination, Coase theorem, and incentive contracts. Coase theory, which supposes that bargaining over property rights results in the optimal and efficient outcome, can play a role in wage determination in Major League Baseball (Hylan et al., 1996). The monopsony case made in the paper rests on the belief that the sports owners are a small and interconnected group. This, in my opinion, can impact blacklisting and prophesying players, but not wages for an individual, average player.

Rosen & Sanderson (2001) depict the economic dynamics in modern sports. A very important insight into this thesis is that a variety of data suggest that the variation in skill has declined. This stays consistent with the rising supply price of talent. The rise in pay, fewer salary restrictions, and allowing more athletically inclined youth to seriously consider a career in professional sports make it so that weaker competitors are weeded out. This can be an explanation as to why a model like the one built in this thesis may not work the same way it could have in the fifties or sixties.

A paper that goes more into detail regarding actual wage setting, and how to approach it empirically is Yang, C. H., & Lin, H. Y. (2012). The study's findings indicate a potential bias in salary allocation within the NBA, with international players generally earning less than their U.S.-born counterparts, all other factors being equal. This could be interpreted as nationality-based salary discrimination. Interestingly, international players originating from larger economies appear to benefit from preferential treatment in the labour market, underlining the significance of their home country's market in influencing their salaries. Factors such as player positions, height, draft status, and the size of a team's local market are found to be pivotal in determining player salaries. The research also uncovers a trend where U.S.-born white players are typically paid less than U.S.-born black players. This could be attributed to a higher proportion of white players being bench players, or it could potentially indicate racial discrimination.

Regarding the salary cap mentioned in previous papers, it is set in place through a complex set of rules and is calculated as a percentage of league revenue. The salary cap is known as a "soft cap," where teams spending over the cap are subject to reduced privileges. Késenne (2000) writes extensively on the effects of the salary caps on the NBA. In the paper, the effect is seen as a positive. The paper highlights that the implementation of a salary cap can enhance the competitive equilibrium within a sports league by facilitating a more equitable distribution of player salaries. This mechanism effectively curtails the of top-tier player salaries, thereby ensuring a sustainable profit margin for club owners, irrespective of the size of their clubs. Such a financial structure encourages continued investment in the industry. However, a notable drawback of a salary cap is its deviation from the Pareto-optimal point, leading to a potential decrease in total league revenues. In an idealized model of sports, where all stakeholders are perfectly informed and rational, the necessity for salary caps would be rendered moot. Nevertheless, the real-world scenario often deviates from this ideal due to the irrational behaviour of owners and managers. They often overlook the negative externalities of unbalanced competition and engage in aggressive bidding wars in a free agency player market, thereby inflating top player salaries. This behaviour underscores the necessity for regulatory measures such as salary caps.

Methodology-related Papers

The next papers discussed will be directly relevant to the details of the equation formulated in the following chapter. These papers will cover wage research on the NBA, and will aid me in making choices as to what to include in the wage equation.

What is it we exactly expect from players? This is the question asked in Bodvarsson and Brastow (1998). Consistency was the hypothesis of the paper. The reason a player's performance consistency matters to an employer is as follows. The traditional view is based on a risk aversion model in which a worker's output variability imposes a nonpecuniary cost on the employer arising from his distaste for risk. The authors offer an alternative explanation that is independent of tastes: variance matters because it creates certain pecuniary costs that affect the firm's profitability. This finding can be problematic for this thesis, as a player playing the exact same every match (not improving) should see an increase in wages as the uncertainty of his performance is zero. In the hypothesis formulated, this player should see no increase in wages. This can possibly be further explored in the conclusion.

One of the assumptions made is that players will give the same amount of effort during seasons before signing contracts, as effort in other seasons is not taken into account for this thesis. Akerlof & Yellen (1990) have shown that workers proportionately withdraw effort as their actual wage falls short of their fair wage. There is a possibility that players slack, even in the extreme wage bracket they find themselves in when they feel like their pay is not at what they are worth. This could be the case for players of whom their teammates are paid significantly more, or players who play for teams that have very high total salary caps.

The reason only the year prior to signing a contract is taken, is that there is evidence that performance in years prior to contract year is overlooked in contract determination decisions by NBA general managers, according to a thesis paper written by Fox (2015). Although not a scientific article in a scientific journal, the recency bias is a well-supported phenomenon in social sciences, and has been shown to have an effect in job recommendations, as in Chen et al. (2019).

There are a lot of factors setting players apart, and luckily most of them have been researched. Sadly, skin colour has been shown to influence total career earnings of players (Hoang & Rascher, 1999). There is an argument that the conclusion of blacks earning less could be found due to the data being from the eighties. Although, Johnson & Minuchi (2018) show whites are paid more in the modern day still. According to the authors this is due to a preference of the audience to watch a diverse product as well as interact with those of the same race. With 75% of the players being black, the minority groups of players more valuable based on consumers' preferences are non-blacks. As career earnings are no more than the sum of seasons played, yearly salary is lower by default as well. The Hoang & Rascher results also show black players are more likely to get cut from the team as a bench player.

Superstars, the crème de la crème, with their faces plastered on billboards and sponsors lined up for them, will have wages set differently. Beyond common sense, Humphreys and Johnson (2019) show that star players draw crowds into the stands. This, combined with the assumption that information is complete to the parties in negotiation, will make determining wages for star players a goal that is beyond this thesis.

Conclusion

Finally, to conclude the literature review, two papers that closely resemble the research conducted in this thesis. Stiroh (2006) concludes that the contract year is an incentive for players to play better. Through thorough analysis of NBA players in the eighties and nineties, Stiroh finds that employees optimally vary effort to maximize personal gains, even at the expense of firm gains. Papadaki and Tsagris (2022) show that the relation between pay and performance is non-linear. These works will be further explored in the methodology, as their analysis will be heavily influencing the one in this thesis.

In short, Bodvarsson and Brastow (1998) show that consistency is valued in players. Akerlof & Yellen (1990) have shown that workers proportionately withdraw effort as their actual wage falls short of their fair wage, where in this thesis a fair wage assumption is made. Fox

(2015) shows effort increases in contract years, meaning those will be compared exclusively in the dataset. Humphreys and Johnson (2019) show that star players draw crowds and are compensated for it. To account for this, the top players are removed from the dataset.

3. Data and research methodology

Dataset

The data used for this analysis will come from two esteemed sources in the field of basketball statistics. The data source for the individual and team statistics will be gathered from basketball-reference.com. The reason this website is chosen is that it allows direct importing of tables into Microsoft Excel, and it has been used in NBA-related research by published authors (Papadaki & Tsagris, 2022).

It contains every conceivable statistic from every single NBA player since the Association started tracking them, all of which are gathered are displayed in Table 1. Wages will be gathered from a USA Today Sports initiative called Hoopshype.com. On this website yearly salary is stated for both individual players and team totals per season, adjusted for price inflation.

As contract length is not fixed, every player's statistics will need to be retrieved manually to some degree, where Excel work will come into play. The time constraint for this thesis, and the manual labour that must be put in, will limit the total number of observations that can be considered without compromising the accuracy of the dataset.

After compiling all of the 48 statistics for every active player from seasons 2011/2012 to 2023/2024, some changes are made to the data. First, all of the observations in which players have not been active for more than 10 games are removed. Limiting observations on the bottom is a frequently occurring event, as it limits irrelevant output variability (Bodvarsson & Brastow, 1998). A non-linear effect has been shown between pay and performance (Papadaki and Tsagris, 2022).

All the observations of players earning more than 25 million dollars a year are removed, as these players are paid for more than their performance. This can be seen in a Player efficiency rating per dollar analysis done by Northwestern Sports Analytics Group (Evenson et al., 2024). Here all the players who are allegedly most overpaid for their performance see wages above 25 million dollars a year. The cause could be that at the very peak of performance, the value of performance is non-linear. Or, these players are paid more for their superstar value that drive sales in merchandise and draw in crowds.

This research considers player performance during the NBA regular season, excluding the playoffs. It's crucial to distinguish between the two because the regular season spans from October to April, comprising 82 games. During this period, players may not consistently perform at their best in every game. Additionally, only 8 teams from each conference secure playoff spots, totalling 16 out of the 30 teams. Teams aware they won't make the playoffs may rest their top players in the season's final games, giving more minutes to bench players. Conversely, struggling teams might intentionally lose games (a strategy known as 'tanking') to improve their chances in the upcoming NBA draft (Gong et al., 2021). In the draft, the worst-performing team has the highest chance of landing the number one pick. In contrast, during the playoffs, when the championship is at stake, players elevate their performance (Teramoto & Cross, 2010). Playoff-bound teams focus solely on winning the championship, often relying more on their top players who may even play the entire 48-minute game.

Methodology

The methodology will be influenced by Papadaki, I., & Tsagris, M. (2022) and Stiroh (2007), among others. The first hypothesizes and proves that imperfect information and multi-year contracts create an implicit incentive for workers to strategically alter effort over the contract cycle. The motivating hypothesis in the second is that imperfect information and multi-year contracts create an implicit incentive for workers to strategically alter effort over the contract cycle. The reason these fit well with a hypothesis that states that players wages change according to their change in performance is that both papers use a variance variable, this is either expressed as the difference in statistics across seasons or it is expressed in standard deviations. This will be the driving factor for this thesis paper.

First, to ensure the data collected is comparable to data used in papers of which assumptions, such as a (log)linear relationship between performance measures and pay, and comparable performance metrics, are copied, an analysis of wage rewards on game statistics will be done.

$$\begin{aligned}\operatorname{Log} S_{i\,t+1} = {} & a_0 + a_1 FORWARD_{it} + a_2 CENTER_{it} + a_3 EXP_{it} + a_4 EXP^2{}_{it} \\ & + a_5 AGE_{it} + a_6 AGE^2{}_{it} + a_7 GAME_{it} + a_8 MINUTE_{it} \\ & + a_9 REBOUND_{it} + a_{10} ASSIST_{it} + a_{11} STEAL_{it} \\ & + a_{12} BLOCKED_{it} + a_{13} PTS_{it} + a_{14} CHANGE_{it} + a_{15} STAR_{it} \\ & + e_{it}\end{aligned}$$

Equation 1

Above is the initial wage equation as seen in Yang, C. H., & Lin, H. Y. (2012). The dependent variable ln $Si\ t+1$ denotes the logarithm of yearly salary for individual player i at time t, adjusted for inflation, whereas e_{it} is a white noise error term.

Their research posits that a player's salary for the current year is influenced by his personal attributes and performance on the court in the preceding year, hence the +1 after the t in the dependent variable. The binary variables FORWARD and CENTER are set to one if a player served as a forward or center forward in season t. Although a player's position on the court rarely changes, it is a time-variant variable in their dataset, allowing them to incorporate it into the panel specification to investigate potential salary differences across various positions. Given the conflicting results found in existing literature when different time spans are used, no definitive sign is anticipated a priori. Other personal attributes include the number of NBA seasons (EXP) and age (AGE), both of which are expected to positively affect salary, albeit with diminishing marginal returns. Therefore, the estimated sign for the coefficients of EXP (AGE) and its squared term should be positive and negative, respectively. Performance on the court, being the primary determinant of salary in professional sports, is represented in this study by six widely accepted basketball performance indicators. GAME signifies the number of games a player participates in a season. A player who frequently appears on the court during a season is likely to be physically and mentally stable and a key player for the team. Hence, a significantly positive coefficient for this variable is anticipated. MINUTE represents the average minutes a player spends on the court per game. They expect a positive sign for this variable as starters typically earn more than bench players. Performance variables include rebounds per game (REBOUND), assists per game (ASSIST), steals per game (STEAL), blocked shots per game (BLOCKED), and points per game (PTS). These variables, being the best indicators of performance in basketball, are expected to positively influence salary. For example, studies by Gius and Johnson (1998) and Eschker et al. (2004) found that points per game, rebounds per game, assists per game, and blocked shots per game are significant variables. The dummy variable CHANGE is set to one if a player switched teams in the previous season. They also included a binary variable STAR, set to one if a player was selected for the All-NBA team in the previous season, to account for the possibility that superstars may earn more than what their on-court performance would suggest.

As my data is not as extensive as the data in Yang, C. H., & Lin, H. Y. (2012), I will be making slight changes to the regression, trying to keep as close to it as possible. First, I will include dummy variables for all five positions on the basketball court, being power forward, center, point guard, shooting guard, and small forward, where the positions are compared to centers, instead of the two mentioned above. I do not possess the data to see whether a player has changed teams in the past season or how many years of experience he has, which means I will not be including the dummy CHANGE and EXP and its square into my regression. Other than those, the variable STAR is removed, as there will be no stars due to the nature of my dataset. Lastly, dummy variables will be added for every single season included. Not to adjust for price inflation, as this is accounted for in the database, but to adjust for wage inflation and specific seasonal changes. This brings the formula to:

$$\begin{aligned}\mathrm{Log}\, S_{i\,t\,+\,1} = {} & a_0 + a_1 POWERFORWARD_{it} + a_2 CENTER_{it} \\ & + a_3 POINTGUARD_{it} + a_4 SHOOTINGGUARD_{it} \\ & + a_5 SMALLFORWARD_{it} + a_6 AGE_{it} + a_7\, AGE^2{}_{it} + a_8 GAME_{it} \\ & + a_9 MINUTE_{it} + a_{10} REBOUND_{it} + a_{11} ASSIST_{it} \\ & + a_{12} STEAL_{it} + a_{13} BLOCKED_{it} + a_{14} PTS_{it} + \sum_{j=11}^{24} a_{14\,+\,j}\ \mathrm{YEAR}_{ij} \\ & + e_{it}\end{aligned}$$

Equation 2

As this regression is expected to show the results that are predicted in Yang, C. H., & Lin, H. Y. (2012), I move forward towards a model that works towards the hypothesis. The previous formula shows a loglinear relation between wages and performance. What is sought after is whether players wage increases when performance increases. Or, in other words, does the difference in performance equate to a difference in wage? This can be interpreted as looking for the first derivative of the equation.

Issues arise from going from a linear model to its derivative when using linear regression. Taking the first derivative of a linear regression causes the formula to change to a single constant. To make sure do not run into this technicality, the approach will have to be slightly modified. Changing the analysis from a first-difference model to a lagged dependent variable model, which is commonly used when you believe that the current value of your dependent variable is dependent on its previous value and the changes in the other variables.

The performance is attached the yearly wage agreed upon in the contract signed after the season, W-1. For a fair comparison, we compare this performance to the player's next contract year. This is due to the finding that players systematically perform better during contract years (Stiroh, 2007). This is why I take the performance in contract year Y, the last year of contract W-1, and attach the performance reward W to it as a comparison. The explanation of the idea of this research will be aided by the figure below (figure 1). The figure depicts the timeline that will be researched for each player. These players will be examined, and contract years are drawn. They will be examined on whether they have played a sufficient number of games, and whether there are two contract years in the dataset. The initial contract year, Y-1, the player will perform a certain way. After this performance he is assigned a contract based on how well he did, as the first regression should prove performance influences wages.

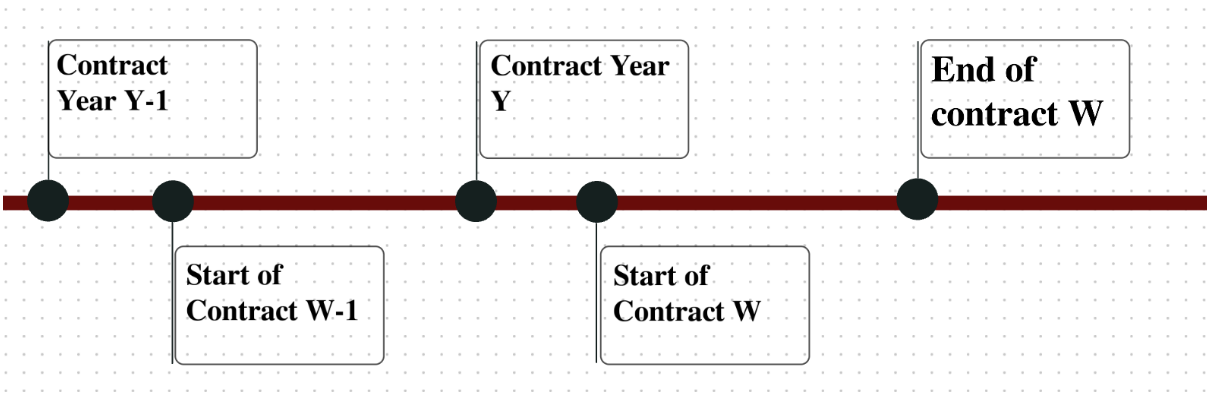

Figure 1: Timeline contract years used in comparison

Using the same methodology used to formulate equation two, the equation will consist of the difference in the variables from the second one.

$$\begin{aligned}\mathrm{Log}\, S_{i\,w} = a_0 &+ POWERFORWARD_{iy} + a_2CENTER_{iy} + a_3POINTGUARD_{iy} \\ &+ a_4SHOOTINGGUARD_{iy} + a_5SMALLFORWARD_{iy} \\ &+ a_6\Delta AGE_{iy} + a_7\,\Delta AGE^2{}_{iy} + a_8\Delta GAME_{iy} + a_9\Delta MINUTE_{iy} \\ &+ a_{10}\Delta REBOUND_{iy} + a_{11}\Delta ASSIST_{iy} + a_{12}\Delta STEAL_{iy} \\ &+ a_{13}\Delta BLOCKED_{iy} + a_{14}\Delta PTS_{iy} + a_{15}\,\mathrm{Log}\, S_{i\,w-1} + e_{it}\end{aligned}$$

Equation 3

Where equation two is used to evaluate whether the sample used for this thesis shows similar results to past papers, equation three is attempting to answer the research question. Where most of the variables are familiar, except for the addition of delta (Δ) in the game statistic variables. The interpretation of delta follows: $a_8\Delta GAME_{iy}$, where i indicates the individual player and y indicates the contract year. Δ would be the difference between this and the previous contract year, for example: $\Delta GAME_{iy} = GAME_{iy} - GAME_{iy-1}$.

Lastly, to differentiate between raises and pay cuts, a variable raise will be introduced, which is 1 if the pay in contract W is higher than the pay in contract W-1, and 0 otherwise. This variable will replace the dependent variable in the regression to determine a difference in pay cuts and rises.

Samples

Since the three samples are not attached to their own hypothesis or equations, an expansion on their use is necessary. As this thesis makes use of three subsamples of the same database, all of them will be explained and their importance will be underlined. The raw data exists of every player and their statistics and wages from seasons 2011/2012 to 2022/2023. For equation 2, the changes made in the methodology regarding the minimum requirements and the maximum wages are implemented. This yields a sample with 3038 observations. This sample is not to evaluate the accuracy of the predictions made with the hypotheses, but to evaluate whether the same results can be yielded using the sample for this thesis as for previous works. This is an integral part for this research as some of the core assumptions made in this thesis are drawn from past papers. This sample not yielding similar results can imply the assumptions are not applicable to the sample.

After the validity of the sample is confirmed with statistically significant results on performance using equation 2, this research attempts to answer the research question using equation 3. For this purpose contract years are used, slimming down the sample to 555 observations. The way contract years are obtained is as follows. All player observations are evaluated on whether it is the first year at a new team, as new contracts often come with a player signing at a new team. This method is not accurate to the highest degree. Contracts in this sample can vary length, which can cause inaccuracy due to the nature of contract lengths. A one-year contract should not be compared to a four-year contract, as they are inherently different from one another. A four-year contract signals belief in the growth of a player and a reservation on the eventual skill he may develop, which often is paired with disproportionally high wages. Another issue is that players are free to extend their contract at the same club, as Koby Bryant famously played his entire career for the Lakers for example. Lastly, players out on loan are included in this sample. This is troublesome as these are usually still under the same contract, playing at a different club.

As the sample used to explore equation 3 has its limitations, a robustness check will need to be implemented to determine whether these limitations are to the detriment of the validity of the sample. This robustness check will exist of contract years of randomly selected players. This sample exists of 75 observations, which is substantially less than the sample used for answering the research question. This is due to the fact every single observation must be manually checked for contract years and team transfers, also making sure contract length offered are similar. The reason this is a valid approach to check for robustness is that this sample does not come with the same limitations that occur when looking for players that changed teams. The argument can be made that this sample is preferable over the one used to interpret results for the conclusion, and this is a subjective subject. A trade-off must be considered between the accuracy of the data, and the number of observations. There is either a small number of observations that is very exact in the identification of contract years, or a larger sample, which looks at whether a player has transferred as an indicating for a contract year.

In short, the sample is first confirmed to be similar to ones used in past papers, then contract years are taken to evaluate the hypothesis. Lastly, a robustness check is implemented to determine if the method of obtaining contract years has been valid.

Robustness checks

On top of the use of the sample, multiple robustness checks will be employed to establish internal validity. First, the correlation between variables will be examined. It is assumed in multiple regression that the variables are independent. When this assumption is violated, it can lead to issues such as inflated standard errors, misleading coefficient estimates, and reduced predictive accuracy of the model. Therefore, correlation will be examined to judge the robustness.

A test for outliers will be done with the IQR method. Here a boxplot will be generated, where the outliers are defined as the twenty-fifth percentile minus one and a half times the interquartile range and the seventy-fifth percentile plus one and a half times the interquartile range. In the case outliers are present, a decision will have to be made whether to remove them or not. Removing them likely increases the statistical significance, but it takes away from the general applicability of the model. As outliers are very common in the NBA, with large differences in wages being observed, outliers will not be removed.

Lastly, sensitivity analysis will be done to see how sensitive the results are to changes. This will consist of adding different game statistics to see what happens to the statistics that are in the regression. First a completely random, low-discrepancy sequence, whereafter more variables are added. The idea being that when a theoretically irrelevant statistic is disrupting the entire regression, it is not a robust model.

Checks that test for the time series related robustness, such as the Augmented Dickey-Fuller test for stationarity, are excluded as the time series in this study is not as you would see it in the typical sense. Therefore, I find it unnecessary to include. As for the use of an instrumental variable (IV), it is very hard to think of and find a variable that does not influence income, but does influence performance. It is also complex to use an IV for a multiple linear regression model.

Potential issues

The most prominent potential issue is the assembly of the dataset, which includes observations that can skew the regression results. For example, a loaned player will not be in a contract year, and decreased effort as shown in Stiroh (2007), is not the greatest problem with this. It is that the year played is not rewarded with a wage for his performance. What it means is that playing poorly in the second year of a four year contract will not decrease the wage in the subsequent year.

What is also noteworthy is that these player's statistics come from the games they play against one another, or together as well. This means that the independence of observations assumption of linear regression is violated. Autocorrelation can cause issues with the model. In the real world this can be seen as two players, as teammates on the floor, playing together. There can only be a limited number of attacking plays, shot chances, and rebounds for a single team within the time limit of the game. A chance made by one player, is a chance that cannot be made by the next. The sample is too small to adjust for every team in the league to combat this problem.

Finally, multicollinearity can be a major problem for the regression. Pre-emptively, statistics like the player efficiency rating (PER) are left out. Although proven to be significant, PER is calculated from multiple significant factors as blocks, steals, field goals etc. This does not mean that other statistics are not corelated with one another. It makes sense that a player that scores more, has less assists. This means that these can be inversely correlated. On the other hand, it can be that the player that makes a lot of points, gets the ball more, and thus has more chances to make assists. The very dynamic game of basketball causes problems for the regression.
Over- and underfitting of the model does not pose much of a problem in this analysis, as the variables used are credited to past peer reviewed papers.

4. Results

table 2: summary table of statistics of NBA players from 2011 to 2023.

Variable	Obs	Mean	Std. dev.	Min	Max
log (Salary)	3038	15	0.758	14	17
Position numeric	3038	3	1	1	5
Age	3038	26	4	19	42
Age Squared	3038	682	237	361	1764
Games	3038	57	19	10	83
Minutes Played	3038	1268	694	33	3167
Points	3038	508	350	10	2375
Assists	3038	110	104	0	704
Steals	3038	40	28	0	191
Blocks	3038	26	27	0	241
Rebounds	3038	220	154	2	1114

Table 2 presents summary statistics for variables used in Regression 1, focusing on 3038 observations related to basketball players. The logarithm of salary (log (Salary)) has a mean of 15 and a standard deviation of 0.758, ranging from 14 to 17, making it skewed to the right. This finding suggests that a minority of players earns a large amount of money. Player statistics, such as minutes played, points, assist, steals, blocks, and total rebounds follow the same trend. Position numeric seems evenly distributed throughout the sample. These observations underscore the influential role of exceptional performance in both on-court statistics and player salaries within the dataset. Age averages 26 years with a standard deviation of 4, ranging from 19 to 42, while Age Squared averages 680 with a standard deviation of 237, ranging from 361 to 1.764.

Table 3: Regression between Log(Salary(t+1)) and performance variables as in equation 2

Dependent Variable	**Log (Salary (t+1))**
Nr. of Obs: **3038**	**R-squared** 0.39

Independent Variable	**Coefficient**
Power Forward	-0.027 (0.037)
Point Guard	-0.228 (0.053)***
Small Forward	-0.132 (0.045)***
Shooting Guard	-0.135 (0.046)***
Age	0.249 (0.025)***
Age^2 †	-3.83 (0.467)***
Games †	-11.51 (1.188)***
Minutes Played †	0.566 (0.061)***
Points †	0.271 (0.075)***
Assists †	0.699 (0.191)***
Total Rebounds †	0.264 (0.151)*
Steals †	-0.021 (0.646)
Blocks †	1.319 (0.670)**
Season 2011-12	-0.256 (0.050)***
Season 2012-13	-0.333 (0.050)***
Season 2013-14	-0.397 (0.050)***
Season 2014-15	-0.320 (0.050)***
Season 2015-16	-0.150 (0.054)***
Season 2016-17	-0.074 (0.051)
Season 2017-18	0.093 (0.059)
Season 2018-19	-0.121 (0.053)**
Season 2019-20	-0.030 (0.051)
Season 2020-21	-0.072 (0.049)
Season 2021-22	-0.101 (0.048)**
Season 2022-23	Omitted
Constant term	11.308 (0.340)***

Note: † implies the statistic is divided by 1000, which means the coefficient is exaggerated by a factor of 3. * Indicates a statistical significance at the 10% level, ** indicates a statistical significance at the 5% level. *** indicates a statistical significance at the 1% level. The positions are a categorical variable, compared to center players. Season 2022-23 is shown as being omitted, due to this being the reference.

Table 3 analyses the log-transformed future salaries of 3038 basketball players, the regression model yields an R-squared value of 0.39, indicating moderately low explanatory power. Most findings are in line with what is hypothesized in the methodology. Remarkable regarding this regression is that all positions seem to earn less than the position they are compared to, which are centers, except for power forwards, who do not see a significant difference of any kind compared to center players. Also, the total number steals is not significant statistically in this model. Despite this finding, they will be included in further analysis as theory shows they should have an impact on wage. This is a satisfactory result, as most variables seem to have a statistically significant effect on wages. This allows us to conclude that the same assumptions made in Yang, C. H., & Lin, H. Y.(2012) apply to this data.

Table 4: Summary table of the statistics

Variable	Obs	Mean	Std dev	Min	max
Log(salary(W))	555	15.46	0.83	13.81	16.80
Log(salary(W-1))	555	15.53	0.78	13.81	16.81
Δage	555	1.60	1	1	7
Δage squared	555	3.60	6	1	49
Δgames	555	1	20	-63	66
Δminutes played	555	-24	664	-1912	1909
Δrebounds	555	-1	147	-589	617
Δassist	555	-2	77	-378	319
Δpoints	555	-12	312	-1260	1098
Position	555	3	1,47	1	5

Table 4 comprises 555 observations detailing changes (Δ) in various player metrics and salaries over a specific period. The change in age seems to be skewed right, as players cannot age less than one year between contracts, but can have up to seven years between two contract years. Relative to their spreads, the means of game statistics such as the changes in games, minutes played, rebounds, assists, and points are relatively close to zero. This can be attributed to the fact that these factors are limited throughout the league, as there is only a finite number of plays to be made within a season. Looking at the logarithms of wages, we find an increase of roughly 0.07 between the average wage in previous contract years, and current ones. The position variables spreads roughly evenly between all positions.

Table 5: The effect of difference in performance statistics on salary with added sensitivity analysis variables.

Number of obs	555	555	555
F	(14, 540): 22.62	(16, 538): 20.84	(20, 534): 17.29
Prob > F	0	0	0
R-squared	0.34	0.34	0.34
Root MSE	0.64	0.64	0.64
Dependent Variable	Log (Salary(W))		
	Coefficient (t)	Coefficient (t)	Coefficient (t)
Power Forward	0.06(0.62)	0.06(0.63)	0.05(0.60)
Point Guard	0.10(1.08)	0.10(1.06)	0.10(1.05)
Small Forward	0.01(0.05)	0.00(0.02)	0.00(0.05)
Shooting Guard	0.16(1.96)**	0.16(1.92)*	0.15(1.87)*
Log (Salary(W-1))	0.53(15.06)***	0.53(14.59)***	0.53(14.69)***
Δage	0.30(3.43)***	0.30(3.45)***	0.29(3.33)***
Δage squared	-0.04(-2.89)***	-0.04(-2.91)***	-0.04(-2.81)***
Δgames†	-2.25(-0.89)	-2.15(-0.85)	-2.50(-0.98)
Δminutes played†	-0.20(-1.30)	-0.27(-1.48)	-0.42(-1.88)*
Δpoints†	0.52(2.44)**	0.63(2.48)***	0.11(0.29)
Δrebounds†	0.09(0.25)	0.15(0.41)	0.68(1.25)
Δassist†	0.55(1.13)	0.61(1.23)	0.89(2.51)**
ΔSteals†	1.13(0.55)	1.37(0.66)	1.54(0.70)
ΔBlocks†	4.09(2.21)**	4.27(2.29)***	4.24(2.21)**
filler		0.10(1.00)	0.10(1.06)
ΔPER†		-10.25(-0.82)	-16.42(-1.07)
Δturnover%†			0.69(0.08)
Δusage%†			-12.57(-0.91)
ΔWinshares			0.03(0.73)
ΔVORP			-0.02(-0.26)
Constant	6.90(12.42)***	6.91(12.24)***	6.87(12.16)***

Note: † implies the statistic is divided by 1000, which means the coefficient is exaggerated by a factor of 3. * Indicates a statistical significance at the 10% level, ** indicates a statistical significance at the 5% level. *** indicates a statistical significance at the 1% level. The positions are a categorical variable, compared to center players.

Table 5 shows the effect of difference in performance statistics on salary in the first column, the second column adds a filler and the difference in player efficiency rating. In the third column, differences in the turnover and usage percentage are added, together with the difference in winshares and the value over replacement player (VORP). The r-squared for all the regressions are similar.

In the original regression, it shows that past salary, paired with a change in age, age squared, points scored and blocks show significant results. On the other hand, a difference in minutes, games, rebounds, or steals do not seem to affect the change in wage much. Also noteworthy, it seems shooting guards see growth compared to any other position.

The second column depicts the addition of the first robustness check, comprising of a randomly assigned, normally distributed filler, together with the player efficiency rating. Player efficiency should be able to disturb an unstable model, as it is an advanced statistic that encompasses all statistics used in the prior model. Despite it theoretically influencing the regression, we see only modest differences between column 1 and 2. More specifically, the raise in pay shooting guards see diminishes in statistical significance from 5 percent to 10 percent.

Then, further disturbing the model, the difference in turnover percentage, usage percentage, winshares, and the VORP are added. Whilst not being significant themselves, these highly complex variables cause some disturbances in the previously significant variables. First, points do no longer seem to be a statistically significant. The reason for this could be a high corelation with some of the added variables. A significance appears in the difference in minutes played, at the 10 percent level, as well as the 5 percent level, assists seem to have become significant. The difference in blocks becomes less significant, dropping to the 5 percent level, from the 1 percent level in the previous two columns. A correlation table will be examined between the difference in points, assists, blocks, and the additional control variables.

Table 6: correlation matrix for various statistics

	ΔPoints	**ΔAssists**	**ΔBlocks**	**ΔTurnover%**	**ΔUsage%**	**ΔWinshares**	**ΔVorp**
ΔPoints	1						
ΔAssists	0.71	1					
ΔBlocks	0.54	0.34	1				
ΔTurnover%	-0.21	-0.00	-0.16	1			
ΔUsage%	0.36	0.16	0.05	-0.08	1		
ΔWinshares	0.73	0.56	0.55	-0.32	0.03	1	
ΔVorp	0.62	0.52	0.46	-0.25	0.16	0.88	1

Table 6 is the correlation table added in an attempt to explain the reason the difference in points is not significant in the third column of table 5. A possible explanation is that it could be directly corelated to one of the added variables. As we can see there is a high corelation of 0.73 between the difference in winshares and the difference in points. Winshares, which is an estimate of how many wins can be attributed to a certain player. This, like the player efficiency rating, tallies up and averages out a large number of statistics, points being one of them. This correlation is likely the reason the points has become statistically insignificant. Another regression excluding the highly correlated difference in winshares is seen in table 7 (appendices), where indeed the difference in points is once again seen as significant at the 5 percent level.

Table 8: the effect of differences in statistics on binary variable raise.

Number of obs	**555**	
F(14, 540)	21.67	
Prob > F	0.0000	
R-squared	0.2493	
Root MSE	0.43216	
Dependent Variable		Raise
Variable	Coefficient (t)	P>|t|
Log (Salary(W-1))	-0.25(-12.73)***	0.000
Position		
Power Forward	0.01(0.25)	0.806
Point Guard	0.08(1.26)	0.210
Small Forward	0.06(0.94)	0.350
Shooting Guard	0.17(3.06)***	0.002
ΔAge	0.12(2.33)**	0.020
Δage^2	-0.02(-2.63)***	0.009
ΔGames†	-1.13(-0.67)	0.500
ΔMinutes†	0.17(1.59)	0.112
ΔPoints†	-0.09(-0.60)	0.548
ΔRebounds†	-0.06(-0.25)	0.805
ΔAssists†	-0.02(-0.05)	0.957
ΔSteals†	-1.68(-1.23)	0.220
ΔBlocks†	1.74(1.44)	0.149
Constant term	4.32(13.92)***	0.000

Note: † implies the statistic is divided by 1000, which means the coefficient is exaggerated by a factor of 3. * Indicates a statistical significance at the 10% level, ** indicates a statistical significance at the 5% level. *** indicates a statistical significance at the 1% level. The positions are a categorical variable, compared to center players. The dependent variable wage is binary, taking form as 0 and 1.

The final results table has the dependent variable changed to a binary that is zero for pay cuts and one for pay raises. We can see that shooting guards experience a significantly higher number of raises, compared to other positions. An increase in age shows a higher percentage in raises. Interestingly, previous salary shows a decrease in the raise percentage. This could be due to there being no growing past opportunities after reaching the high paying contracts. None of the changes in game statistics show any effect on whether a player gets a raise.

Robustness checks

Table 9: summary table of statistics

Variable	Obs	Mean	Std. dev.	Min	Max
ΔAge	75	3.6	2	1	10
ΔAge^2	75	190	130	41	600
ΔGames	75	-2	25	-55	56
ΔMinutes Played	75	2	1010	-2284	2296
ΔRebounds	75	6	214	-912	416
ΔAssists	75	20	98	-199	328
ΔPoints	75	78	507	-981	1173
log_salary (W)	75	15.51	0.70	13.99	16.55
log_salary (W-1)	75	15.05	0.74	13.82	16.55

Table 9 comprises 75 observations detailing changes (Δ) in various player metrics and salaries over a specific period. This table will be compared to table 4, which looks at the same statistics of a different sample, as the data is assumed to yield similar results. A large similarity to table 4 is concluded, as the means of the changes in player statistics are also close to zero, with the spreads in statistics being slightly larger in the sample used in table 4. This is likely due to the fact more observations are included in table 4. The changes in mean of the logarithm of salary between the periods is larger, being 0,46 in this sample, compared to 0,07 in table 4.

Table 10: The effect of difference in performance statistics on salary

Number of obs	**75**
F(14. 60)	9.69
Prob > F	0.0000
R-squared	0.4368
Root MSE	0.58513
Dependent variable:	
Log (Salary(w))	Coefficient (T-value)
Position	
Power Forward	0.37 (0.19)*
Point Guard	0.2 (0.27)
Small Forward	0.11 (0.16)
Shooting Guard	0.18 (0.26)
Log(salary(w-1))	0.21 (0.15)
Δage	0.28 (0.2)
ΔAge Squared†	-4.53 (3.48)
Δgames†	-1.13 (5.93)
Δminutes played†	-0.14 (0.36)
ΔSteals†	-2.19 (4.84)
ΔBlocks†	-4.61 (3.37)
Δrebounds†	1.36 (0.77)*
Δassist†	2.46 (1.19)**
Δpoints†	0.6 (0.48)
Constant term	11.9 (2,26)***

Note: † implies the statistic is divided by 1000, which means the coefficient is exaggerated by a factor of 3. * Indicates a statistical significance at the 10% level, ** indicates a statistical significance at the 5% level. *** indicates a statistical significance at the 1% level. The positions are a categorical variable, compared to center players.

Table 10 examines the impact of changes in performance metrics on the log-transformed salary of 75 basketball players, compared to their previous contract year. Interestingly, the change in most game statistics, including previous salary, do not seem to be significant when determining salary. The change in the total number of rebounds a player gathered, is statistically significant on a 10 percent level. This is noteworthy, as the total number of rebounds does not show a significant effect on wage in table 3. Compared to centers, power forwards are the only position to see a wage increase on average over contract years. The change in assists seems to be the only significant predictor at a 5 percent level. Possibly signifying the importance of playmaking and basketball IQ in wage setting.

Table 11: correlation matrix for significant variables used in regression analysis

	Log(salary(W))	**Δage**	**Δage^2**	**Δgames**	**Δminutes played**	**Δrebounds**	**Δassists**	**Δpoints**
Log (Salary(W))	1							
Δage	0.0945	1						
Δage^2	0.0845	0.949	1					
Δgames	-0.236	-0.131	-0.140	1				
Δminutes played	-0.301	-0.0933	-0.112	0.840	1			
Δrebounds	-0.284	-0.0552	-0.0796	0.719	0.839	1		
Δassists	-0.211	-0.0374	-0.0322	0.560	0.735	0.576	1	
Δpoints	-0.297	-0.0616	-0.0802	0.728	0.914	0.787	0.710	1

Table 11 presents the correlation matrix for the variables used in the regression analysis, focusing on their interrelationships. Log(salary(W)) shows positive correlations with the difference in (Δ) Age and Δage^2, indicating slight positive associations with changes in age and its square. Δage and Δage^2 are highly correlated (0.9492), naturally reflecting that age changes and their squared values move together. Δgames, Δminutes played, Δrebounds, Δassists, and Δpoints exhibit varying negative correlations with Log(salary(W)), ranging from -0.2119 to -0.3014, suggesting no substantial corelation between the variables. Strong positive correlations are observed between Δgames and Δminutes played (0.8404), Δgames and Δrebounds (0.7197), Δminutes played and Δrebounds (0.8398), Δminutes played and Δpoints (0.9141), Δrebounds and Δpoints (0.7872), and Δpoints and Δassists (0.7109), indicating that these performance metrics often increase together. This interdependence among performance metrics suggests that improvements in one area are likely associated with improvements in others, which could influence their collective impact on salary changes.

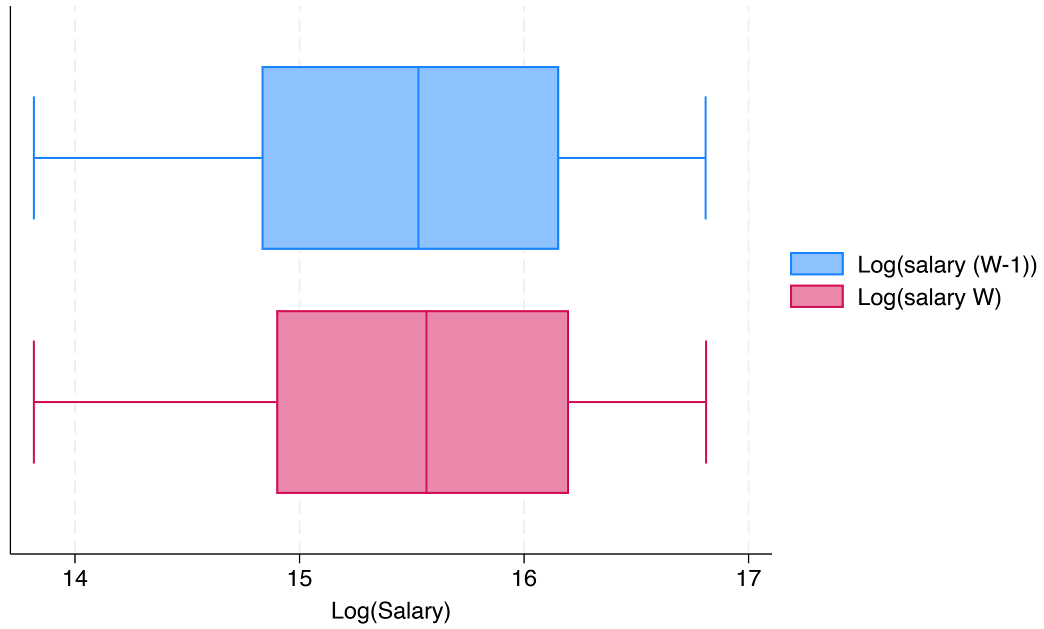

Figure 2: Boxplots of variables used for regression analysis.

The box plot compares the distributions of the logarithm of salary for the previous year (Log(salary(W-1)), in blue) and the current year (Log(salary(W)), in red). The medians are similar, with Log(salary(W-1)) slightly lower, indicating a slight increase in median salary. The interquartile ranges (IQRs) and overall ranges are comparable, suggesting consistent variability and spread in salary distributions across both periods. The absence of outliers and the symmetry of both distributions support the normality assumption and indicate no significant skewness or extreme values. These consistent patterns across the two time periods enhance the internal validity of the regression analysis, suggesting that the relationship between past and current salaries is stable and reliable.

The box plot for delta minutes played in contract years (graph 4, appendices) shows a wide distribution of changes in playing time. The median is close to zero, indicating that the typical change in minutes played is minimal. The interquartile range (IQR) spans from approximately -500 to 500 minutes, capturing the middle 50% of the data. The minimum and maximum values, excluding outliers, are around -1500 and 1500 minutes, respectively. Notable outliers are present beyond -1750 and 1750 minutes, indicating extreme changes in playing time for some players. These outliers may result from significant factors like exceptional performance improvements or reductions due to injuries. Overall, the plot highlights that while most players see minor changes in playing time during contract years, a few experience substantial increases or decreases.

The box plot for delta age, shown in graph 5 in appendices, illustrates the difference in age between contract years. The minimum value starts at 1 year, reflecting the minimum time between contracts. The central tendency is represented by the median, around 2 years, indicating that the typical age difference between contract renewals is modest. The interquartile range (IQR) spans from approximately 1 to 2.5 years, capturing the middle 50% of the data. Notable outliers extend beyond 3 years, with values up to 7 years, suggesting some players experience significantly longer gaps between contracts. These outliers may result from unique career circumstances, such as extended recovery periods or personal decisions. Overall, the plot highlights that while most players renew their contracts within a few years, a few experience substantially longer intervals.

The box plots for delta points (Graph 6, in appendices) and delta assists (Graph 7, in appendices) illustrate the changes in these performance metrics during contract years. In both plots, the central tendency, marked by the median, is close to zero, indicating minimal typical changes in points scored and assists. The interquartile range (IQR) for delta points spans from approximately -750 to 750 points, while for delta assists it ranges from -170 to 170 assists, capturing the middle 50% of the data for each metric. The minimum and maximum values, excluding outliers, are around -1000 to 1000 points for delta points and -250 to 250 assists for delta assists. Notable outliers in both graphs extend beyond these ranges, with delta points reaching approximately -1500 to 1000 and delta assists extending from -400 to 400. These outliers suggest significant deviations in performance, possibly due to changes in roles, team dynamics, or individual improvements. Overall, while most players experience modest changes in points scored and assists during contract years, the presence of substantial outliers indicates that external factors such as team context and player development significantly influence these metrics.

5. Conclusions

This thesis aimed to investigate the impact of performance statistics on professional athletes' salaries, with a focus on assessing the robustness of the regression models used. The hypotheses were stated as:

H1: An increase in performance in contract years, relative to previous contract years, has an increasing causal effect on salary.

And

H2: A decrease in performance in contract years, relative to previous contract years, has a decreasing causal effect on salary.

Regressions show that certain performance measures are statistically significant, showing an increase in these statistics increase the pay non-linearly. According to this research, answering the hypothesis, increasing age, points, and blocks, will increase wages in the next year.

Aging happens naturally, of course, but it does increase wage. This is most likely due to minimum guarantees increasing with experience. A worthy mention is that when sensitivity analysis is done, difference in minutes show to be statistically significant at a 10% level, with a negative coefficient. Players probably wish they would get as many minutes as possible, yet the regression shows a negative coefficient for players who had more minutes on the court compared to the previous season. A possible explanation for this in unintuitive result is that the relation between the change in minutes played and the wage is that minutes played does not have a loglinear relation with wages, similar to age in Yang, C. H., & Lin, H. Y. (2012). Another explanation could be that a player would want to maximise his efficiency, and maximise points and assists in as little time as possible. Increasing points is a good way to get eyes on you, which is most-likely why the regression shows a significant positive coefficient. Steals are relatively rare, and require tremendous effort on the defensive, where perceived effort could be a cause of increased wage.

Differentiating the first and second hypothesis from one another, table 9 highlights what causes raises for players to occur. No in-game statistics seem to be statistically significant, which implies there is no difference in interpreting the results for raises and cuts.

Internal validity

To address internal validity, we must look back on the methodology and the database. The data origins from a highly reliable source. Most of the advanced statistics are estimated, which is why I opted out of using them as inaccuracies can occur whilst capturing the statistic. The sample including over 500 observations seen in Table 7 is also accrued through looking at whether the player played at a different club the season prior. This can draw in trades and does not include contract extensions; this inaccuracy was chosen over the hand-sourced model due to the vast increase in observations. There is also no say into whether there are confounding variables that are not included into the regressions. Ones that come to mind are effort, chemistry, and the possibility that different, undisclosed wage categories get judged differently.

The study's design inherently limits the ability to draw definitive causal inferences due to its observational nature. While the significant relationship between past salary and current salary is evident, causality cannot be conclusively established. I attempted to strengthen causal inference by controlling for lagged variables and examining changes over time (Δ variables). Nevertheless, unobserved factors and potential reverse causality remain concerns that could affect the interpretation of results.

As the sample included all observations, no selection bias could have been taken place. I found that the performance variables are correlated, which jeopardises the isolation and therefore the interpretability of the coefficients. It can also obscure any effect that the statistical insignificant coefficient potentially could show, as collinearity can increase the standard errors. Luckily, the coefficients are stable.

The sample of handpicked observations seems to show different results to the main sample used for the research. In this sample, only the change in assists and the change in rebounds seem to be significant in the determination of the change in wages. The different results can be caused due to two reasons. One being the method used to gather observations for the main sample of 555 observations is faulty and the results cannot be interpreted to reject a hypothesis as stated in this thesis. The other being that the sample used as a robustness check is experiencing turbulence due to the small number of observations, being 75.

External validity

The robustness of the findings was tested through sensitivity analyses, which showed that key predictors remained significant even with the inclusion of additional variables. This suggests that the model is relatively stable and supports its potential applicability in other contexts. However, replication of this study in different settings and with different populations is necessary to confirm these results.

As the timeframe of this thesis encapsulates the latest years of the NBA, it is hard to tell whether the pay schemes players were exposed to decades back are comparable to today's game with all its evolutions. What is safe to say, is that the future years are most-likely covered in this model as there is no reason to believe the way contracts are set up will change in the coming years.

6. Limitations and future research

This research has its limitations, first is its sample. Although it might encompass the entire NBA over almost a decade, excluding the top and bottom players. the NBA is such a dynamic and complex league. This complexity makes it so that extrapolating these results to different professional sports or sectors is irresponsible. Other than that, as discussed in the external validity, I do believe this research is usable from whenever the latest NBA rules were instated to when they get changed. These are the boundaries I set, as rule changes can possibly affect the integrity of the research. There is no doubt the measurements by the NBA are accurate, but, as discussed in the internal validity, the method of extracting the contract years may jeopardise the research. While the study included a comprehensive set of performance variables, there may be other relevant factors that were not considered. Variables such as psychological factors, off-field behaviours, or specific contractual clauses might also influence salary but were not included in the analysis.

For future research, this methodology can be extrapolated to different sports such as the MLB and NFL. The research can be replicated with performances in the 80's and 90's to see whether the trends found apply across a broader timeline. Also, as this is a thesis written with limited time, a more grounded and comprehensive dataset can be set up to address possible measurement errors used to conduct this research. Also, if possible, the factors such as psychological factors, off-field behaviours, or specific contractual clauses can also be included to draw conclusions regarding true causal effects.

Bibliography

Akerlof, G. A., & Yellen, J. L. (1990). The fair wage-effort hypothesis and unemployment. *The Quarterly Journal of Economics, 105*(2), 255-283. https://doi.org/10.2307/2937787

Blalock, H. M. (1962). Occupational discrimination: Some theoretical propositions. *Social Problems, 9*(3), 240-247. https://doi.org/10.2307/799234

Bodvarsson, Ö. B., & Brastow, R. T. (1998). Do employers pay for consistent performance? Evidence from the NBA. *Economic Inquiry, 36*(1), 145-160. https://doi.org/10.1111/j.1465-7295.1998.tb01702.x

Chen, R., Ai, Q., Jayasinghe, G., & Croft, W. B. (2019). Correcting for recency bias in job recommendation. *ACM International Conference on Information and Knowledge Management, 2185-2188*. https://doi.org/10.1145/3357384.3358131

Evenson, I., Immediately, D., & Madhu, A. (2024, April 8). Who are the most overpaid and underpaid players in the NBA? *Northwestern Sports Analytics Group*. https://sites.northwestern.edu/nusportsanalytics/2024/04/08/who-are-the-most-overpaid-and-underpaid-players-in-the-nba/

Fox, C. (2015). NBA contracts and recency bias: An investigation into irrationality in performance pay markets. *Haverford College, Department of Economics*. https://osf.io/ncg4v/#!

Gong, H., Watanabe, N. M., Soebbing, B. P., Brown, M. T., & Nagel, M. S. (2021). Exploring tanking strategies in the NBA: An empirical analysis of resting healthy players. *Sport Management Review, 25*(3), 546-566. https://doi.org/10.1080/14413523.2021.1970972

Hall, S., Szymanski, S., & Zimbalist, A. S. (2002). Testing causality between team performance and payroll. *Journal of Sports Economics, 3*(2), 149-168. https://doi.org/10.1177/152700250200300204

Hoang, H., & Rascher, D. (1999). The NBA, exit discrimination, and career earnings. *Industrial Relations, 38*(1), 69-91. https://doi.org/10.1111/0019-8676.00110

Hylan, T. R., Lage, M. J., & Treglia, M. (1996). The Coase theorem, free agency, and Major League Baseball: A panel study of pitcher mobility from 1961 to 1992. *Southern Economic Journal, 62*(4), 1029-1042. https://doi.org/10.2307/1060946

Johnson, C., & Minuci, E. (2020). Wage discrimination in the NBA: Evidence using free agent signings. *Southern Economic Journal, 87*(2), 517-539. https://doi.org/10.1002/soej.12461

Johnson, N. R., & Marple, D. P. (1973). Racial discrimination in professional basketball: An empirical test. *Sociological Focus, 6*(4), 6-18. https://doi.org/10.1080/00380237.1973.10570857

Kahn, L. M. (2000). The sports business as a labour market laboratory. *The Journal of Economic Perspectives, 14*(3), 75-94. https://doi.org/10.1257/jep.14.3.75

Késenne, S. (2000). The impact of salary caps in professional team sports. *Scottish Journal of Political Economy, 47*(4), 422-430. https://doi.org/10.1111/1467-9485.00171

Papadaki, I., & Tsagris, M. (2022). Are NBA players' salaries in accordance with their performance on court? In M. Tsagris (Ed.), *Contributions to Economics*, 405-428. Springer. https://doi.org/10.1007/978-3-030-85254-2_25

Rottenberg, S. (1956). The baseball players' labour market. *Journal of Political Economy, 64*(3), 242-258. https://doi.org/10.1086/257790

Rosen, S., & Sanderson, A. (2001). Labour markets in professional sports. *The Economic Journal, 111*(469), F47-F68. https://doi.org/10.1111/1468-0297.00598

Scully, G. W. (1974). Pay and performance in major league baseball. *The American Economic Review, 64*(6), 915-930. https://www.jstor.org/stable/1815242

Teramoto, M., & Cross, C. L. (2010). Relative importance of performance factors in winning NBA games in regular season versus playoffs. *Journal of Quantitative Analysis in Sports, 6*(3). https://doi.org/10.2202/1559-0410.1260

Yang, C. H., & Lin, H. Y. (2012). Is there salary discrimination by nationality in the NBA? *Journal of Sports Economics, 13*(1), 53-75. https://doi.org/10.1177/1527002510391617

Appendices

Table 1: Statistics available

Position	Position of a player, dual position players (e.g./PF) are seen as their primary position
Age	The player's age on Feb 1rst of a season
Wage reward	Wage one season after the statistics are gathered
Team	
Games	The number of games a player participated in
Minutes Played	The total number of minutes spent on the court
Total Rebounds	The number of times the ball is caught after a missed shot
Assists	Total number of times a pass is immediately causal to a field goal
Steals	The number of times the ball is taken from opposition
Blocks	The number of times the ball gets stopped during a shot attempt
Turnovers	The number of times the ball is lost to opposition
Points	The total number of points accrued by a player
Robustness variables	
Value over Replacement Player	A box score estimate of the points per 100 TEAM possessions that a player contributed above a replacement-level (-2.0) player, translated to an average team, and prorated to an 82-game season.
Player efficiency rating*	A measure of per-minute production standardized such that the league average is 15.
Usage percentage	An estimate of the percentage of team plays used by a player while they were on the floor.
Win shares	An estimate of the number of wins contributed by a player.
Turnover percentage	An estimate of turnovers committed per 100 plays.

Note: Multiple statistics are displayed as estimates. Estimates are implied when part of the equations is not accurately measurable. This includes comparisons to average players, and immeasurable statistics (for example, total rebounds available)

*PER formula: uPER = (1 / Minutes Played) *[3 Pointers+ (2/3) * Assists+ (2 − factor * (team assists / team field goals)) * field goals+ (free throws *0.5 * (1 + (1 − (team assists / team field goals)) + (2/3) * (team assists / team field goals)))- VOP * Turnovers − VOP * Dribble% * (field goal attempts − field goals)- VOP * 0.44 * (0.44 + (0.56 * Dribble%)) * (free throws attempted − free throws)+ VOP * (1 − dribble%) * (total rebounds − offensive rebounds)+ VOP * Dribble% * offensive rebounds + VOP * steals + VOP * Dribble% * Blocks- personal fouls * ((log (Free throws) / (log (Personal Fouls))) − 0.44 * (log Free Throws Attempted / log (Personal Fouls)) * VOP)] where factor = (2 / 3) - (0.5 * (log (Assists) / log field goals)) / (2 (log_Field Goals / log free throws))

VOP = (log (points) / log (Field Goals Attempted)) − log (offensive Rebounds) + log (Turnover) + 0.44 * log Free throw attempted)

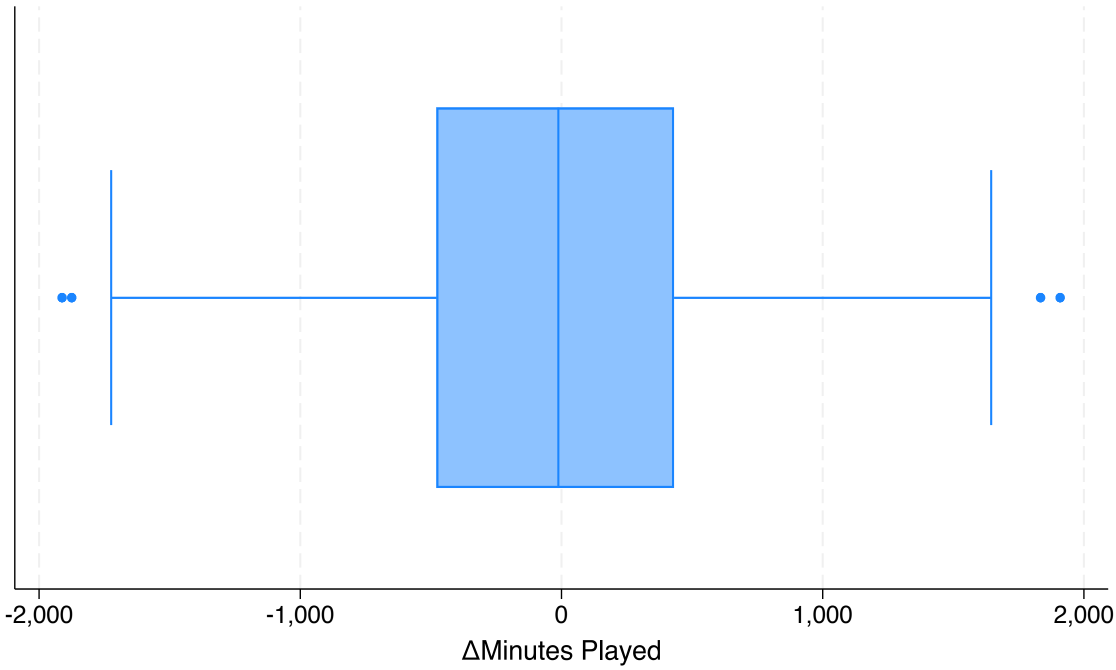

Figure 3: boxplot of differences in minutes played.

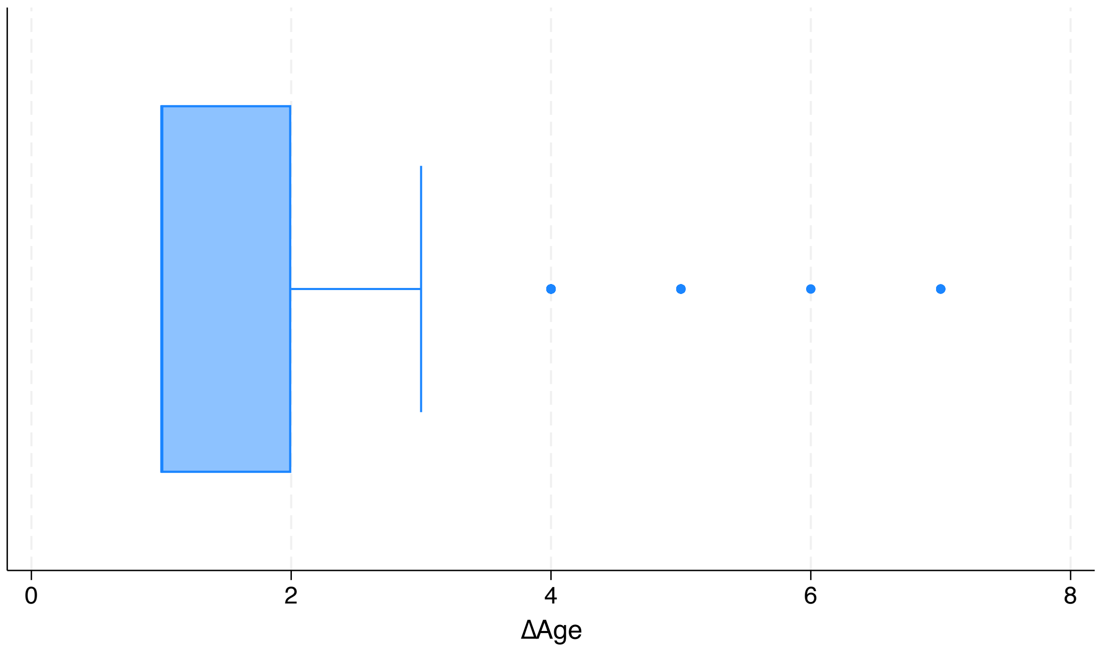

Figure 4: boxplot of differences in age

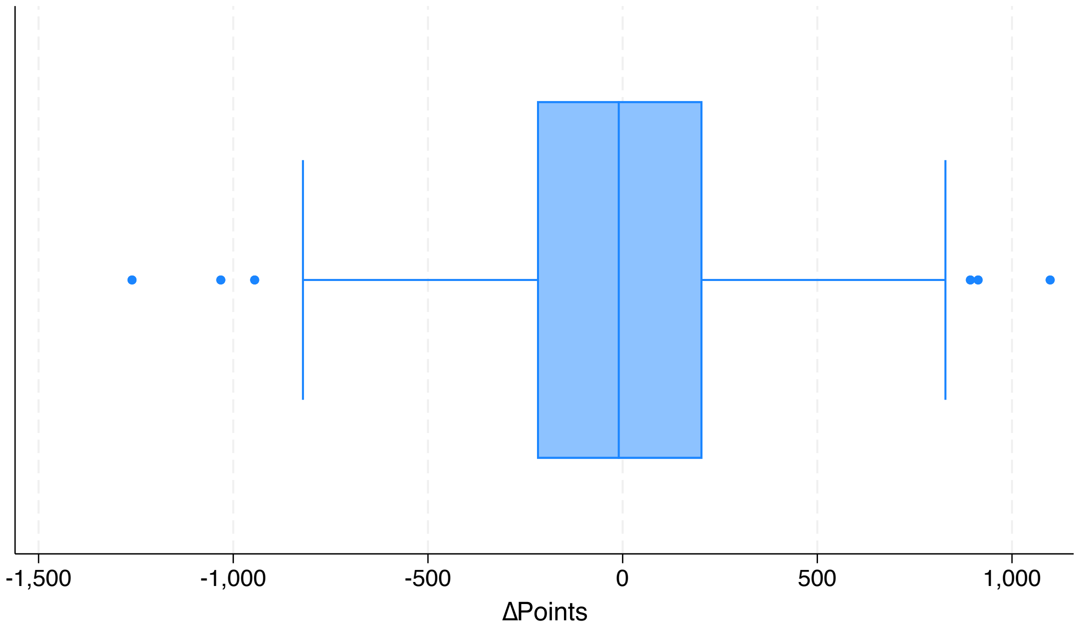

Figure 5: boxplot of the differences in points

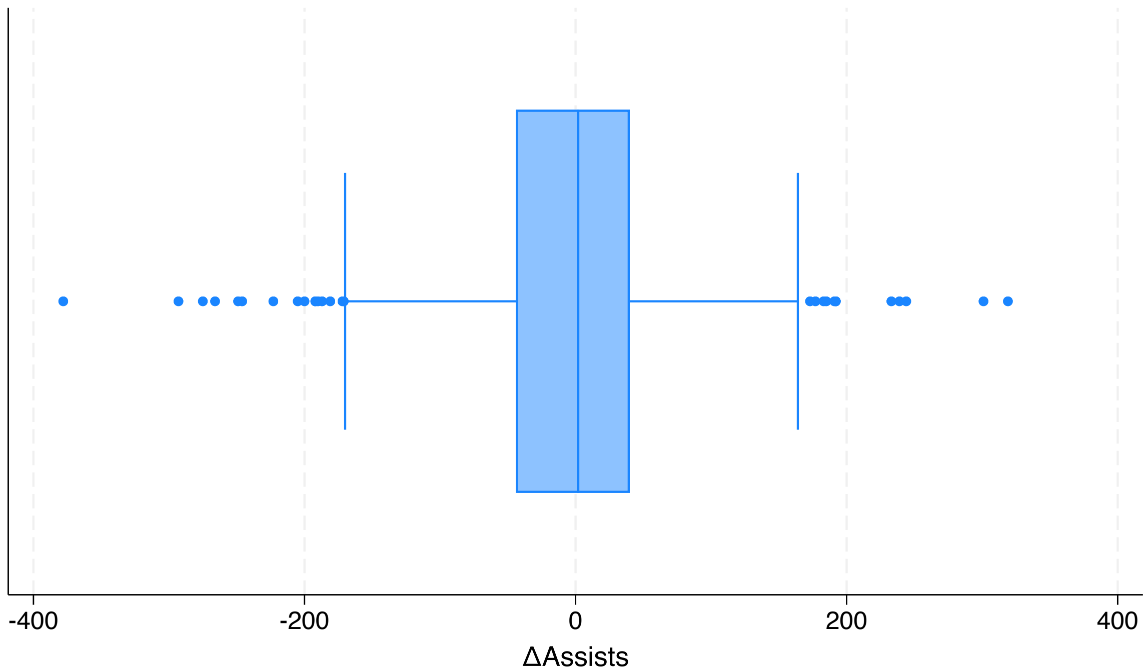

Figure 6: Boxplot of the differences in assists

Table 6: identical table to table 5 column 3, leaving out WS%

Number of obs	**555**
F(19, 535)	17.98
Prob > F	0.0000
R-squared	0.3391
Root MSE	0.64152

Log (Salary (W))	Coefficient (t)
Log(salary(W-1)	0.53 (14.63)***
Position	
Power Forward	0.05 (0.59)
Point Guard	0.10 (1.07)
Small Forward	0.01 (0.06)
Shooting Guard	0.15 (1.87)*
Δage	0.29 (3.35)***
Δage^2	-0.04 (-2.83)**
Δgames†	-2.11 (-0.85)
Δminutes†	-0.37 (-1.78)
Δpoints†	0.88 (2.50)**
Δrebounds†	0.17 (0.46)
Δassists†	0.59 (1.09)
Δsteals†	1.13 (0.54)
Δblocks†	4.06 (2.15)**
filler	0.10 (1.07)
ΔPER	-0.01 (-0.90)
ΔTOV%	-0.00 (0.00)
ΔUSG%	-0.02 (-1.21)
ΔVORP	0.03 (0.59)
Constant term	6.88 (12.15)***

Note: † implies the statistic is divided by 1000, which means the coefficient is exaggerated by a factor of 3. * Indicates a statistical significance at the 10% level, ** indicates a statistical significance at the 5% level. *** indicates a statistical significance at the 1% level. The positions are a categorical variable, compared to center players.