DEGREE PROJECT IN TECHNOLOGY,
FIRST CYCLE, 15 CREDITS
STOCKHOLM, SWEDEN 2017

Analysis of Performance Measures That Affect NBA Salaries

SIMON LOUIVION

FELICIA PETTERSSON

KTH ROYAL INSTITUTE OF TECHNOLOGY
SCHOOL OF ENGINEERING SCIENCES

Analysis of Performance Measures That Affect NBA Salaries

SIMON LOUIVION

FELICIA PETTERSSON

Degree Projects in Applied Mathematics and Industrial Economics
Degree Programme in Industrial Engineering and Management
KTH Royal Institute of Technology year 2017
Supervisors at KTH: Henrik Hult, Kristina Nyström
Examiner at KTH: Henrik Hult

TRITA-MAT-K 2017:15
ISRN-KTH/MAT/K--17/15--SE

Royal Institute of Technology
School of Engineering Sciences
KTH SCI
SE-100 44 Stockholm, Sweden
URL: www.kth.se/sci

Abstract

This thesis investigates which factors that affect the salary for basketball players in the NBA and if the salary cap has achieved its purpose. The data for this project was collected from basketball-reference.com and consisted of performance measures from season 2015/2016 and salaries from the beginning of the season 2016/2017.

The study was performed by using multiple linear regression analysis in the software R and the data was handled in Excel. The results from the regression indicates that *position point guard*, if the player has played in *D-league* or not,*Age, Offensive rebounds, Assists, Steals, Two point attempts, Three point attempts, Free throw attempts, Field goal percentage, Usage percentage* and *Defensive rating* are the main factors that affect the salary. The performance measures that had the greatest were *two* and *three point attempts*. The regression model achieved an explanatory level of 57.4%. In complementary to analyze if the salary cap has achieved its purpose, a literature analysis was used and showed that the salary cap systems in North America are neither accurately designed nor do they satisfy the intentions of what they were set to achieve.

Analysering av prestationsmått som påverkar NBA-löner

Sammanfattning

Denna rapport undersöker vilka prestationsfaktorer som påverkar lönen för basketspelare i NBA och om NBA's *salary cap* (lönetak) har uppnått sitt syfte. Datan för projektet hämtades från basketball-reference.com och bestod utav spelarstatistik ifrån säsong 2015/2016 och lön ifrån början av säsong 2016/2017.

Undersökningen utfördes genom linjär regressions analys med hjälp utav mjukvaruprogrammet R och datan hanterades i Excel. Resultatet från regressionen visar att positionen *point guard*, om spelaren spelat i *D-league* eller inte, *ålder, offensiva returer, assists, steals, 2-poängsförsök, 3-poängsförsök, straffkastsförsök, field goal procent, användningsprocent* och *defensiv rating* är faktorer som påverkar lönesättningen. Prestationsmåtten med störst påverkan var *2-poängsförsök* och *3-poängsförsök*. Regressionsmodellen uppnådde en förklaringsgrad på 57.4%. Motsvarande, för att analysera om NBA's salary cap har uppnått sitt syfte gjordes en litteraturstudie som visade att salary cap-systemen i Nordamerika varken är korrekt utformade eller uppfyller sina ursprungliga syften.

Preface

This thesis is written by Felicia Pettersson and Simon Louivion during the spring of 2017 at the Mathematical Institute of the Royal Institute of Technology. We would like to appreciate the guidance from our supervisors Henrik Hult and Kristina Nyström. Lastly, we would like to thank Simon Borgefors who has been supporting us on our path and the legendary Mr. Lavar Ball who funded this essay by launching his five hundred dollar shoes to market.

Contents

1 Introduction **7**
1.1 Background 7
1.2 Aim 8
1.3 Research Questions 8
1.4 Limitations 9

2 Theoretical Framework **10**
2.1 Multiple Linear Regression 10
2.1.1 Assumptions for Linear Regression 10
2.1.2 Ordinary Least Squares 11
2.2 Model Errors 12
2.2.1 Multicollinearity 12
2.2.2 Heteroskedasticity 12
2.2.3 Normal Q-Q 14
2.2.4 Endogeneity 14
2.3 Hypothesis Testing 15
2.3.1 The F-statistic 15
2.3.2 P-value 16
2.3.3 Breusch-Pagan Test 16
2.3.4 Confidence Interval 17
2.3.5 Runs Test 17
2.4 Model Validation 18
2.4.1 Dummy Variable 18
2.4.2 Box-Cox Transformation 18
2.4.3 Log-Transformation 19
2.4.4 AIC - Akaike Information Criterion 19
2.4.5 BIC - Bayesian Information Criterion 20
2.4.6 R^2 and Adjusted R^2 20
2.4.7 Effect Size, η^2 and Cohen's Rule 21
2.4.8 VIF - Variance Inflation Factor 21
2.5 NBA Salary Cap 22
2.6 Literature Review 22
2.6.1 How the Salary Cap Is Supposed to Affect the NBA 22
2.6.2 Salary Cap Differences Between NFL and NBA 23

3 Methodology **24**
3.1 Data 24
3.2 Regression as A Method 24
3.3 Variables 24

3.3.1 Variables of Choice 24
3.3.2 Dependent Variable 24
3.3.3 Covariates 25
3.3.4 Initial Model 26
3.4 Initial Model Validation 27
3.4.1 Possible Transformations 28
3.4.2 Variable Selection - AIC 30
3.4.3 Detecting Multicollinearity - VIF 31
3.4.4 Normal QQ-plot 32
3.4.5 Residuals vs Fitted - Final Model 33
3.4.6 Test for Randomness 33
3.4.7 Breusch-Pagan Test for Final Model 34

4 Results 35
4.1 Final Model 35
4.2 Impact from the Covariates 35
4.3 What Studies Really Have Shown About NBA's Salary Cap 38

5 Discussion 41
5.1 Analysis of Final Model 41
5.2 Adjustment of Data Set 43
5.3 Analysis of Residuals and Outliers 43
5.4 Model Development 45
5.5 Possible Enhancement of the Salary Cap System 45

6 Conclusion 46

7 References 47

A Appendix 49
A.1 Stepwise AIC in R 49

List of Figures

1 Homoskedasticity . . . 13
2 Heteroskedasticity . . . 13
3 Normal Q-Q plot - Normal distributed . . . 14
4 Normal Q-Q plot - Not normal distributed . . . 14
5 Residual vs Fitted Initial Model . . . 28
6 Log-Likelihood . . . 29
7 Residuals vs Fitted Box-Cox Model . . . 29
8 Residuals vs Fitted Log Model . . . 30
9 Normal QQ-plot Final Model . . . 32
10 Residuals vs Fitted Final Model . . . 33
11 NBA Money vs Wins Relationship (Pagels, 2014) . . . 39
12 NFL Money vs Wins Relationship (Pagels, 2014) . . . 40

1 Introduction

1.1 Background

NBA - National Basketball Association is the greatest and most competitive basketball league in the world. Eligible players from all around the world apply to enter the NBA draft to get selected by one of the thirty teams. There are limited spots to the league and only sixty players can enter it through the draft every year. Thirty NBA teams are allowed to have the maximum amount of 15 players on each team so the total league maximum is 450 players. (NBA.com, 2016) That amount could be compared to the National Football League's maximum of 1696 players (NFL.com, 2017), Major League Baseball's maximum of 1280 players (MLB.com, 2017) and the National Hockey League's maximum of 1500 players (NHL.com, 2017). The significantly low amount of players enhances the competition in the NBA and increases the salaries paid to players, which could explain the reason of why NBA players are the best monetarily credited athletes by average annual salary per player. (Gaines, 2015)

NBA uses a salary cap system where the salary cap is set as a percentage level of the leagues total revenue from the previous season. So the salary cap changes every year and has so far increased every year. The cap system is very complex, contains a lot of exceptions and is sometimes refereed to as "Soft Cap" because the are so many loopholes. Each club can use a set percentage of its revenues for their salary expenses. Usually a single player can receive the maximum of 30 percentage of the clubs total salary cap and every club generally has one or two players that earn a significantly greater amount of money in comparison to their teammates. (Coon, 2016)

Basketball is a spectator sport. Every team's income is highly dependent on TV contracts, how many tickets they sell and how popular their club is. Generally it all comes to popularity. For a club to continuously be popular it is essential to win games. The audience expects wins, nobody wants to watch a horrible team that tend to lose their home games. To be a winning team, efficient and great players are needed which is determined by players performances. In summary great performances on the court lead to victory which increases team popularity. This creates revenue for the club and the club will credit their players for these prowess by immense amount of salaries.

As salaries are principally based on performances on the court, commonly but not always the better player will earn more than the less successful player. There exist a lot of different performance measures. The importance here is to investigate and find which of these measures are crucially affecting the NBA salaries. The NBA player contracts are determined before the season starts. Therefore to find the correlation between performance measures and salaries, it is essential to use statistics between current salaries and performance measures from the previous season.

Similar studies analyzing salaries based on performance measures have been performed on the NBA and other sport leagues. One study was performed by Peck on the National Hockey League, NHL (Peck, 2012). Peck did a regression analysis, with salary and performance measures from 710 hockey players. The conclusion was that there is a positive, significant relationship between salary and goals, assists, career games, and All-Star appearances. Another similar study was made by Fullard who also investigated salary in comparison to performance measures in the NHL (Fullard, 2012) and one by Chakravarthy on the National Football League (NFL) (Chakravarthy, 2012). All of the authors used regression analysis as a method.

1.2 Aim

The purpose of the bachelor thesis is to create an assessment tool for benchmarking the salary of NBA players against their current salaries and other similar researches. The project is relevant since it can be used to measure if a player is overpaid or underpaid in relation to his performances on the court. It will therefore be a useful tool when determining if a players salary is accurate and plausible.

The performance measures and qualities that affect the salaries of NBA players are going to be evaluated. This is going to be processed through a regression analysis to identify the most crucial performance measures and enable us to develop a performance based salary model. Further, the thesis also evaluates the salary cap system with the aim of enhancing the system if it turns out to be insufficient.

Since every club wants to win the championship and that is what players are paid for, it would be an appropriate project to find a correlation between these factors. It does not necessarily mean that the performance measures that affect the player salaries also contribute to winning games. The performance based salary model can therefore additionally be developed to identify underpaid players who can contribute to winning games. As salary cap exists it is an smart tool for clubs to efficiently spend their money with the purpose of creating a winning team. This could be associated with the Moneyball strategy used by the Oakland Athletics Baseball in the 2002 season. The general manager Billy Beane used statistical analyzes to acquire new players with a lean budget. (Lewis, 2003)

1.3 Research Questions

The research questions are the following:

- Which performance measures and qualities affect the salaries of NBA players?

- Is NBA's salary cap serving its purpose?

1.4 Limitations

The study will include all players from the NBA season 2015/2016 and their salaries from season 2016/2017. Rookies and players that ended their careers (salary missing) after the season will therefore be excluded. The same applies players that have played less than 100 minutes. Minimum salary and ten day contract players are also removed and will be discussed in the discussion section.

2 Theoretical Framework

2.1 Multiple Linear Regression

Multiple linear regression is a well known method used in mathematical statistics. The method is used in order to investigate the correlation between a dependent response variable y and a set of k independent variables $x_j, j = 0, ..., k$, also called covariates or regressors. The mathematical correlation between the response variable and the regressors can be described in an equation as:

$$y_i = \sum_{j=0}^{k} x_{ij}\beta_j + e_i, \quad i = 1, ..., n, \tag{1}$$

where the β_j variables are called regression coefficients and are unknown until estimated from observed data. The dependent response variable y can therefore be described by the covariates x_j together with the corresponding error term e_i. Since equation (1) consists of n observations and k regressors, it can be expressed in matrix form as the following:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$$

Where

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_m \end{bmatrix}.$$

(Lang, 2015)

2.1.1 Assumptions for Linear Regression

The linear regression model is based on five assumptions.

- The response variable y is a linear combination of the regressors x_j together with the residual e_i.
- The expected value of the error term, also called the residual, is zero,

$$E[e_i] = 0.$$

- Every error term must be uncorrelated to the others and have the same variance such that:

$$E[e_i^2] = \sigma^2,$$

where σ is unknown.

- The regression model's deterministic component should be a linear function of the separate predictor.
- The amount of observations are greater than the number of regressors and there is no or low mullticollinearity between the regressors.

(Kennedy, 2008)

2.1.2 Ordinary Least Squares

The method of *Ordinary Least Squares, OLS*, can be used to estimate the regression coefficients β and are denoted by $\hat{\beta}$. $\hat{\beta}$ represents the relation between the response variable and the covariates. The OLS estimation $\hat{\beta}$ minimizes the sum of squared residuals $\hat{e}^t\hat{e} = \left|\hat{e}^t\right|^2$, where $\hat{e}$ and $\hat{\beta}$ is defined as

$$\mathbf{\hat{e}} = \mathbf{Y} - \mathbf{X}\hat{\beta}, \quad \hat{\beta} = \begin{bmatrix} \hat{\beta}_\mathbf{0} \\ \hat{\beta}_\mathbf{1} \\ \vdots \\ \hat{\beta}_\mathbf{k} \end{bmatrix}. \tag{2}$$

In order to find the $\hat{\beta}$, the following *normal equations* are solved for $\hat{\beta}$

$$\mathbf{X^t\hat{e}} = \mathbf{0}. \tag{3}$$

By using equation (2) in (3) we get

$$\mathbf{X^t}(\mathbf{Y} - \mathbf{X}\hat{\beta}) = \mathbf{0}.$$

It follows that

$$\hat{\beta} = (\mathbf{X^tX})^{\mathbf{-1}}\mathbf{X^tY}.$$

(Lang, 2016) (Belsley, Kuh and Welsch, 1980)

2.2 Model Errors

2.2.1 Multicollinearity

Multicollinearity occurs when there are near-linear dependencies among the regressors (Montgomery et al., 2012). This means that the OLS estimate does not have a unique solution and occurs when at least one of the covariates can be expressed as a linear combination of the other covariates.

(Lang, 2016)

To detect multicollinearity the estimated standard errors for the regression coefficients must be observed. If the standard errors have high values, problem with multicollinearity probably exists. To eliminate multicollinearity the linearly dependant covariates are removed by identifying their VIF- *Variance Inflation Factor*.

2.2.2 Heteroskedasticity

The linear regression model can be described as the following:

$$y_i = \sum_{j=0}^{k} x_{ij}\beta_j + e_i, \quad i = 1, ..., n.$$

The assumption of Homoskedasticity demonstrates that all the error terms e_i must be uncorrelated to the others and have the same unknown standard deviance σ according to the following:

$$E[e_i] = 0,$$
$$E[e_i^2] = \sigma^2.$$

Since there is a possibility that the error terms are normally distributed it means that the assumption above is not always achieved. Heteroskedasticity implies in violation of this assumption, implying that all error terms do not have the same variance. Then the error terms are defined by the following heteroskedastic assumption:

$$E[e_i] = 0,$$
$$E[e_i^2] = \sigma^2,$$
$$E[e_i^4] < \infty.$$

If a model is assumed to be homoskedasticity when it in fact is heteroskadisticity, problems will occur. (Lang, 2015)

Identify heteroskedastic

It is important to know whether a model is homoskedastic or heteroskedastic. If a model is incorrectly defined problems will occur as mentioned. The parametrization will be inconsistent because of the incorrect assumption that all standard deviations for each error term have the same value. The consequence is that the result of the F-test on the regression will possibly be invalid. It is therefore essential to analyze heteroskedasticity in a model. The easiest way is plot the error term vs the response variable and observe if the behaves constantly.

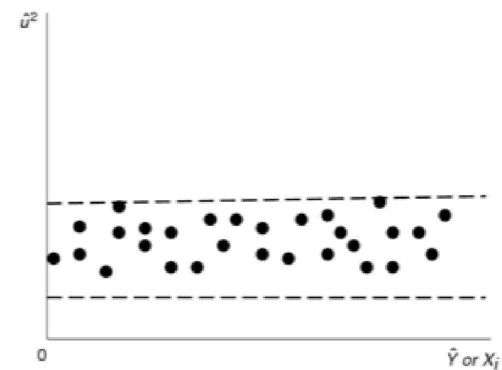

Figure 1: Homoskedasticity

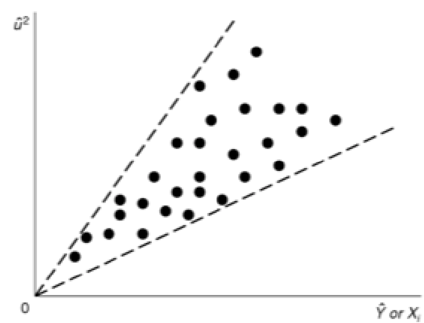

Figure 2: Heteroskedasticity

(Asteriou, 2011)

2.2.3 Normal Q-Q

A *Normal Quantile Quantile plot*, Q-Q plot, could be used to analyze if the residuals are normal distributed. The Q-Q plot represents the standardized resiudals versus the theoretical quantiles.(Ford, 2015) The plots corresponding to the Q-Q plot, should follow a straight line for the model to be classified as normal distributed, illustrated below:

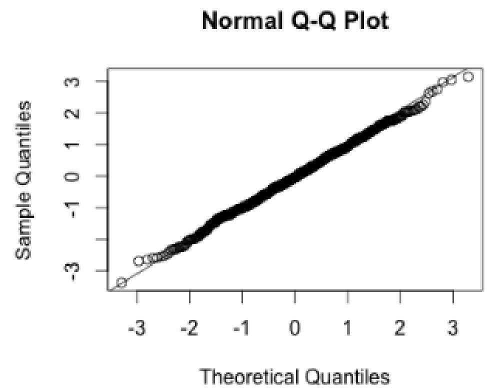

Figure 3: Normal Q-Q plot - Normal distributed

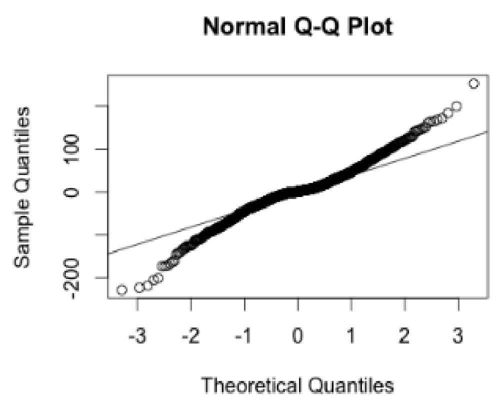

Figure 4: Normal Q-Q plot - Not normal distributed

2.2.4 Endogeneity

Endogeneity is a problem that occurs when the error term $\hat{e}$ is correlated with one or more regressors in the model. The consequences are that the results from the OLS-regression become inconsistent. If there are indications that any regressor in the model conduces endogeneity, it is possible to detect and verify it by plotting the error term $\hat{e}$ on the y-axis versus each of the chosen regressors on the x-axis. If there is a linear outcome in the plot, it demonstrates that endogeneity exists.

(Lang, 2016)

2.3 Hypothesis Testing

To make conclusions from a set of data, a hypothesis test have to be performed. The general process for the test is:

1. Define the null hypothesis H_0 and the alternative hypothesis H_1.
2. Consider the statistical assumptions being made about the data, for example, assumptions about independence or the distributions of the observations.
3. Decide which test statistic is appropriate, state the test statistic and derive the probability distribution.
4. Define the required level of significance α, which is the lower level for H_0 to be rejected. In general a significance level of 5% is used.
5. Define the decision rule.
6. Based on the sample data, calculate the value of the test statistic.
7. Reject or fail to reject the null hypothesis. The decision rule is to reject the null hypothesis H_0 if the observed value is in the critical region, or fail to reject the hypothesis otherwise.

(Investopedia, 2017)

2.3.1 The F-statistic

The F-test is a hypothesis test which makes it possible to test if a number r of the β-estimators should be excluded from the model. The F-statistic is used under the null hypothesis, meaning that the r number of the β:s are all equal to zero. The F-statistic is defined as the relation between two chi-squared distributions. Because of this relation, the F-statistic is shifted to the right. The test statistic for the F-test is the following

$$F(n,p) = \frac{\frac{\chi(n)^2}{n}}{\frac{\chi(p)^2}{p}}.$$

(Lang, 2016)

2.3.2 P-value

The p-value is also used in hypothesis testing. This value represents the probability of the occurrence of a given event. A smaller p-value indicates that the null hypothesis should be rejected.

The p-value is derived from the F-distribution and defined as:

$$P(F(r, n-k-1) > F),$$

where $F(r, n-k-1)$ is the α quantile of the F-distrubution with r number of covariates tested under the null hypothesis, n number of observations, k is the total amount of covariates and F is the F-statistic.

(Lang 2016)

2.3.3 Breusch-Pagan Test

Breusch-Pagan test can be used to identify if a model is heteroskedastic. The test examines if the estimated variance of the error term $Var(\hat{\mathbf{e}}^{\mathbf{2}})$ is dependent of the regressors in the model. If the estimated variance is dependent of the regressors, the conclusion is that the model is heteroskedastic.

The Breush-Pagan test estimates the variance by taking the mean value of all the squared error terms $\hat{\mathbf{e}}^{\mathbf{2}}$. A hypothesis is then created according to the following:

$\mathbf{H_0}$: The model is homoskedastic,

$\mathbf{H_1}$: The model is heteroskedastic.

Afterwards a regression is initiated with $\hat{\mathbf{e}}^{\mathbf{2}}$ as dependent variable together with the other $\mathbf{X}$ regressors such that: $\hat{\mathbf{e}}^{\mathbf{2}} = \mathbf{X}\beta + \mathbf{u}$, $\mathbf{u}$ is the notation for the error term of the regression.

By doing an F-test it is possible to test the hypothesis. If the F-test can confirm that the variables are jointly significant for a certain level of significance, the null hypothesis can be rejected.

(Wooldridge, 2013)

2.3.4 Confidence Interval

When performing a hypothesis test, computing a *confidence interval* is useful. The most common way to analyze the the confidence interval for a single estimation β_i at significance level $1 - \beta$ is with this equation:

$$\beta_i = \hat{\beta}_i \pm \sqrt{F_\alpha(1, n-k-1)} SE(\hat{\beta}_i),$$

where k is the number of coefficients and n the number of observations. The $F_\alpha(1, n-k-1)$ is the cumulative distribution with $n - k - 1$ denominator degrees of freedom and one numerator degrees of freedom. $SE(\hat{\beta}_i)$ is the estimated standard error for β_i.

If the confidence interval is only positive or only negative a conclusion is that the effect of the covariate on the model gives either a positive or a negative result. If the interval contains 0, such a conclusion can not be made. (Montgomery et al., 2012)

2.3.5 Runs Test

Runs test is a statistical test that checks for randomness in a set of data.

The basis of the runs test is formed by the probability that the $(I+1)th$ value is larger/smaller than the Ith value follows a binomial distribution in a set of random data. The run is said to be a series of increasing values or a series of decreasing values. The length of the run is the number of increasing/decreasing values.

The hypothesis is defined according to the following:

$\mathbf{H_0}$: The sequence was created in a random manner,

$\mathbf{H_1}$: The sequence was not created in a random manner.

The test statistic is defined as:

$$Z = \frac{R - \bar{R}}{S_R},$$

where R is the number of runs, $\bar{R}$ the expected number of runs and S_R the standard deviation of the number of runs. They are calculated as the following:

$$\bar{R} = \frac{2n_1n_2}{n_1 + n_2} + 1,$$

$$S_R^2 = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)},$$

where n_1 and n_2 is the number of positive and negative values in the series.

The null hypothesis is rejected if

$$|Z| > Z_{1-\alpha/2},$$

α is the significance level which in general is 5%. This corresponds to a test statistic where an absolute value greater than 1.96 rejects the null hypothesis.

(Bradley, 1968)

2.4 Model Validation

2.4.1 Dummy Variable

Dummy variables are used when there are data types that are not quantifiable. Dummy variables can be defined as covariates that are qualitative. Utilizing this method is an effective way to make data usable. This qualitative covariate only takes the value of one or zero. One indicates that a certain observation contains a specific quality and zero indicates it does not have the specific quality. (Asteriou, 2011)

For example, we use a dummy variable for the point guard position i basketball. All the players that play on the point guard position receive a one and all other players receive a zero.

2.4.2 Box-Cox Transformation

To rectify heteroskadicity a transformation of the data can be made. The *Box-Cox Method* is one technique that can be applied to help specify an appropriate transformation. If the aim is to transform y to correct non-normality and/or non-constant variance, the power transformation y^λ is a useful class of transformations. To determine λ, the Box-Cox method can be used. This method shows how the parameters of the regression model and λ can be estimated simultaneously through the maximum likelihood method.

The best power transformation that fits a certain data set is found by:

$$y_i^{(\lambda)} = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad y_i^{(\lambda)} = \begin{cases} \frac{y_i^{\lambda}-1}{\lambda} & \lambda \neq 0 \\ \ln(y_i) & \lambda = 0 \end{cases}.$$

(Montgomery et al., 2012)

2.4.3 Log-Transformation

Another transformation that can be used to rectify the problem with heteroskadicity is the *logarithmic transformation*. When the dependent variable is positive by nature, it is often motivated to use log of it. The log-regression is the same as the linear regression, equation (1), except that the dependent variable is transformed to a logarithm:

$$log(y_i) = \sum_{j=0}^{k} x_{ij}\beta_j + e_i, \quad i = 1, ..., n.$$

(Lang, 2016)

2.4.4 AIC - Akaike Information Criterion

The *Akaike Information Criterion* test can be used as a method to evaluate the quality of a model. Where the method mainly assesses how good the model is fit in relation to the complexity of the model. In almost every outcome the ultimate model is the one that generates the lowest AIC value:

$$AIC = n \ln\left(|\hat{e}|^2\right) + 2k,$$

where k is the number of coefficients and n is the number of observations.

This equation identifies models that are overestimated in relation to the optimal which is the reason of choosing the model with the lowest result. The model with the lowest result is the most efficient one and maintains a high coefficient of determination in correlation to other models that can be created by the same data set.

(Lang, 2016)

2.4.5 BIC - Bayesian Information Criterion

Another method to test wich regressors should be included in the model is the *Bayesian Information Criterion* test. The test is performed by comparing the BIC-value for the full model versus the reduced model and then choosing the one with the lowest corresponding value. The BIC-value is expressed as the following:

$$BIC = n \ln(|\hat{e}|^2) + k \ln n.$$

The only difference from AIC is the last term, $k \ln n$. AIC has a $2k$ term. Both methods are derived from the same information's theory and framework but differ in priorities, where BIC mostly reduces the model more than AIC. Which test to use dependencies on the model. (Burnham, 2002)

Always keep in mind that these tests does not provide a completely certain answer on which model is the best, it should only be used as a guidance.

2.4.6 R^2 and Adjusted R^2

R^2 is a statistic measurement of *goodness of fit*. The constant explains how good the model is correlated to the data. R^2 is generally called *Coefficient of Determination*, and is the proportion of variation in the dependent variable y that can be explained by variation in the independent variables x. The goal is to achieve a high value which demonstrates small residuals and to have a model with good fit. (Lang, 2016)

R^2 is defined as following

$$R^2 = \frac{Var(\mathbf{X}\hat{\beta})}{Var(\mathbf{Y})} = 1 - \frac{Var(\hat{\mathbf{e}})}{Var(\mathbf{Y})}.$$

An R^2 equal to 0 means that the dependent variable could not be predicted at all using the independent variables. If R^2 equals 1 instead, it means that the dependent variable could always be predicted by the independent variables. An R^2 between 0 and 1 measures the extent that the dependent variable could be predicted by the regressors. For example, an R^2 of 0.30 means that 30% of the dependent variable is predicted by the regressors.

One problem with R^2 is that by increasing the amount of covariates in the model, the R^2 increases since there is an associated cost in terms of the loss of degrees of freedom. To prevent this, *adjusted* R^2, or $\bar{R}^2$, can be used instead of R^2, as it considers for degrees of freedom.

$\bar{R}^2$ is defined as following

$$\bar{R}^2 = 1 - \frac{(n-1)Var(\hat{\mathbf{e}})}{(n-k)Var(\mathbf{Y})},$$

where k is the number of covariates and n is the number of observations.

(Frost, 2013)

2.4.7 Effect Size, η^2 and Cohen's Rule

Effect size measures how much each covariate affects a model. The effect size is dependent of the number of covariates involved in the regression model. (Becker, 2000) There are different kind of methods to estimate the effect size. One method is through estimating $\boldsymbol{\eta^2}$, according to the following:

$$\boldsymbol{\eta^2} = \frac{\hat{\mathbf{e}}^2_{\mathbf{treatment}}}{\mathbf{e}^2_{\mathbf{total}}},$$

where $\hat{\mathbf{e}}^2_{\mathbf{treatment}}$ is the sum of square of a chosen covariate and $\hat{\mathbf{e}}^2_{\mathbf{total}}$ is the sum of squares of all the covariates in the model. $\boldsymbol{\eta^2}$ calculates how much variance of the response variable can be explained by a single covariate in relation to all of the covariates. Cohen's rule of thumb can be used to determine whether a specific covariate has small, medium or big impact.

Impact	Small	Medium	Big
η^2:	0.02	0.13	0.26

(Cohen, 1988)

2.4.8 VIF - Variance Inflation Factor

Multicollinearity could be detected using *Variance Inflation Factor, VIF*. The VIF-value is defined as

$$VIF = \frac{1}{1-R^2},$$

where R^2 is the coefficient of determination, explained further in detail in section 2.4.6, when running a regression on one specific covariate as dependent variable. There exists one VIF for every coefficient in the multiple regression model. Generally, $VIF > 10$ indicates

a problem with serious multicollinearity requiring correction.
(Lang, 2016)

2.5 NBA Salary Cap

A *salary cap* is an agreement to limit the amount teams can spend on player contracts in professional sports. The idea behind it is to maintain a competitive balance between the teams in the league, so a team with deep pockets can not outcompete other teams. Salary caps are adopted in, among others, the sports leagues National Hockey League (NHL), National Football League (NFL) and National Basketball Association (NBA). The NBA is using a *soft cap*, meaning they allow teams to sign players that exceeds the salary cap under special conditions. An example of soft cap exception is that it allows teams to exceed the cap to re-sign their own players. The soft cap also allows exceeding the cap when teams are signing free agents or signing their first round draft picks to rookie scale contracts.

The *Collective Bargaining Agreement*, CBA, defines the salary cap and rules by which the league operates. The CBA is the legal contract between NBA and the players.

For the season 2016-2017 the NBA salary cap was set at $94.143 million and the luxury tax limit $113.287 million. (NBA.com, 2016) But the amount of the salary cap varies every season. For the 2015–2016 season, it was $70 million and the luxury tax limit was $84.74 million. (NBA.com, 2015)
NBA has initially introduced its salary cap system for the first time in 1946-1947, but the "modern" salary cap was introduced 1984-1985 at $3.6 million.

(Coon, 2016)

2.6 Literature Review

2.6.1 How the Salary Cap Is Supposed to Affect the NBA

The salary cap is a payroll that constraints the amount of salary that each NBA club can pay to a single athlete. In theory this creates the same opportunity for every NBA club to sign a certain player no matter the circumstances. It is not like in European football where the teams with the highest pay roll can buy which ever player they want. Instead salary cap have a huge impact on how teams acquire and retain athletes. (Neiger, 2010)

Through logical reasoning the best performing athletes are the ones that require the highest salary. Because of the salary cap a franchise will not be able to stockpile high performing players. This outcome is supposed to enhance the competitive balance between all of the NBA teams and also emphasize the growth of young talent, as first year players usually consume smaller parts of the cap space. It is believed that competitive balance results in higher attendance which increase the revenues. If a game is competitive people are much more likely to watch it than if the outcome is already certain. For example attendance levels in champions league are much higher than in the domestic leagues. When attendance levels increase, the revenues for the NBA clubs will increase. Further, it is assumed that with competitive balance comes higher media exposure since the games are more interesting. This creates costlier media contracts and lucrative contracts with advertisers which can be seen in the National Football Association's (also uses a salary cap system) super bowl game. The advertisement during that game are one of the worlds most expensive. Lastly a salary cap affects the franchise strategy of spending money. Teams will not invest a large amount of their pay roll to players that have not proved themselves so money will be spent wisely. (Wallace, 2011)

2.6.2 Salary Cap Differences Between NFL and NBA

In a short presentation. NFL uses a hard cap which means that there do not exist too many exceptions in the salary cap model. This could be compared to the numerous amounts of exceptions in NBA's salary cap model which allows clubs to exceed their assigned cap space in many different ways.

3 Methodology

3.1 Data

All of the performance measure data from the NBA season 2015-2016 was collected from basketball-reference.com for the regression analysis. The site uses data that originally comes from the official site NBA.com. Salaries from the basketball season 2016-2017 for the NBA players was collected from ESPN.com, USA's greatest sports television company. A constraint was set to only include players that had played at least 100 minutes during the season. Totally 391 players satisfied this constraint but more players had to be excluded which is explained in the discussion. After the adjustments 335 players and 22 covariates were used in the regression.

3.2 Regression as A Method

The aim of the project was to investigate and identify which performance measures and qualities that affect the salary of NBA player through statistical data. Utilizing regression was therefore a plausible way of approach. It has also been used in previous similar projects such as Chakravarthy's study of salary allocation in the NFL (Chakravarthy, 2012) and Peck's regression analysis of salary determination in the NHL (Peck, 2012).

3.3 Variables

3.3.1 Variables of Choice

The most difficult part of the process was to determine which variables to choose. Almost all of the efficiency measures are linear dependent to each other. Primarily when the project started, 52 variables were included from basketball-reference, causing the regression model to have VIF values around 5000. It was therefore essential to carefully chose which variables to include. Through experience in the sport of basketball and reasoning, an initial model was created with 22 variables of performance measures and qualities in a special formation to counteract excessive multicollinearity.

3.3.2 Dependent Variable

The dependent variable is the salary of the NBA players in dollar.

3.3.3 Covariates

PG - Dummy variable used to determine if a player plays on the position point guard.

SG - Dummy variable used to determine if a player plays on the position shooting guard.

SF - Dummy variable used to determine if a player plays on the position small forward.

C - Dummy variable used to determine if a player plays on the position center.

International - Dummy variable used to determine if a player is from outside the US.

Dleague - Dummy variable used to determine if a player have played in NBA D-league during the season.

Age - Numeric variable of a players age on February 1 in 2015.

MPPG - Numerical variable of the amount of minutes a player is on the court per game.

DRBPG - Numerical variable of the amount of defensive rebounds a player has per game.

ORBPG - Numerical variable of the amount of offensive rebounds a player has per game.

PFPG - Numerical variable of the amount of personal fouls a player has per game.

ASTPG - Numerical variable of the amount of assists a player has per game.

STLPG - Numerical variable of the amount of steals a player has per game.

BLKPG - Numerical variable of the amount of blocks a player has per game.

TwoPA - Numerical variable of the amount of two point attempts a player has per game.

ThreePA - Numerical variable of the amount of three point attempts a player has per game.

FTAPG - Numerical variable of the amount of free throw attempts a player has per game.

ThreePpr - Numerical variable of the percentage of three pointers made a player has per game.

eFGpr - Numerical variable called effective field goal Percentage. This statistic adjusts for the fact that a 3-point field goal is worth one more point than a 2-point field goal. For example, suppose Player A goes 4 for 10 with 2 threes, while Player B goes 5 for 10 with 0 threes. Each player would have 10 points from field goals, and thus would have the same effective field goal percentage (50 percent).

USGpr - Numerical variable called usage percentage is an estimate of the percentage of team plays used by a player while he was on the floor.

ORtg - Numerical value of the amount of points produced per 100 possessions, called offensive rating.

DRtg - Numerical value of the amount of points allowed per 100 possessions, called defensive rating.

PG, **SG**, **SF**, **C** with comparison reference **PF** - position power forward.

3.3.4 Initial Model

The initial model was defined as

$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \beta_5 x_{i,5} + \beta_6 x_{i,6} + \beta_7 x_{i,7} + \beta_8 x_{i,8} + \beta_9 x_{i,9} + \beta_{10} x_{i,10} + \beta_{11} x_{i,11} + \beta_{12} x_{i,12} + \beta_{13} x_{i,13} + \beta_{14} x_{i,14} + \beta_{15} x_{i,15} + \beta_{16} x_{i,16} + \beta_{17} x_{i,17} + \beta_{18} x_{i,18} + \beta_{19} x_{i,19} + \beta_{20} x_{i,20} + \beta_{21} x_{i,21} + \beta_{22} x_{i,22}$

where i represent the observation and the parameters stated in the following table:

Variable	Covariate	Unit
$x_{i,1}$	PG	Dummy variable, 0 or 1
$x_{i,2}$	SG	Dummy variable, 0 or 1
$x_{i,3}$	SF	Dummy variable, 0 or 1
$x_{i,4}$	C	Dummy variable, 0 or 1
$x_{i,5}$	International	Dummy variable, 0 or 1
$x_{i,6}$	Dleague	Dummy variable, 0 or 1
$x_{i,7}$	Age	Numeric value
$x_{i,8}$	MPPG	Numeric value
$x_{i,9}$	DRBPG	Numeric value
$x_{i,10}$	ORBPG	Numeric value
$x_{i,11}$	PFPG	Numeric value
$x_{i,12}$	ASTPG	Numeric value
$x_{i,13}$	STLPG	Numeric value
$x_{i,14}$	BLKPG	Numeric value
$x_{i,15}$	TwoPA	Numeric value
$x_{i,16}$	ThreePA	Numeric value
$x_{i,17}$	FTAPG	Numeric value
$x_{i,18}$	ThreePpr	Numeric value
$x_{i,19}$	eFGpr	Numeric value
$x_{i,20}$	USGpr	Numeric value
$x_{i,21}$	ORtg	Numeric value
$x_{i,22}$	DRtg	Numeric value

3.4 Initial Model Validation

It is important to validate the initial model and analyze if it satisfies the assumptions in section 2.1.1. Heteroskedaticity must be tested. One possible way to identify this issue is to analyze the data by plotting residuals versus fitted values of the salary.

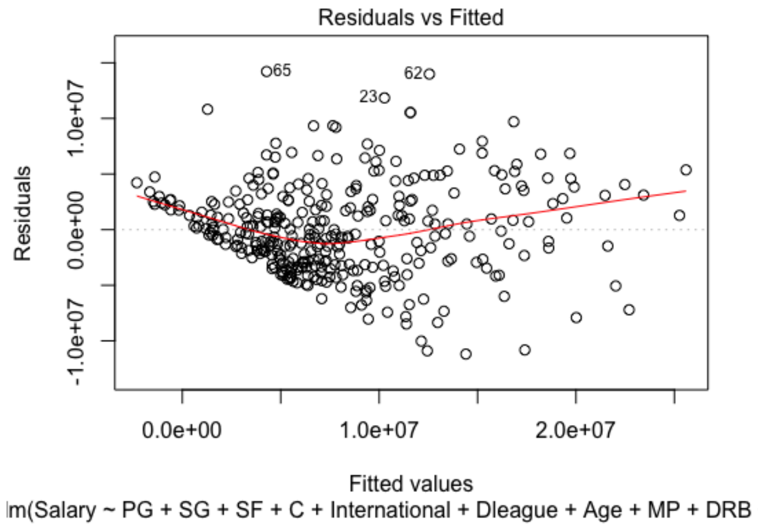

Figure 5: Residual vs Fitted Initial Model

As seen in Figure 5 the red line is not linear so it is concluded that heterskedastic tendencies exist. The conclusion is therefore that the assumption for constant variance is not satisfied. In order to certify that the data is heteroskedastic, a Breusch-Pagan test is performed.

BP	DF	P-value
44.375	22	0.003182

The Breusch-Pagan test generated a P-value below 0.05. Since the hypothesis for homoskedasticity is rejected it can be suggested that the data of the model is heteroskedastic.

3.4.1 Possible Transformations

To handle the issue of heteroskedasticity a box-cox transformation was performed. By processing this method it was suggested to transform the response variable to $\mathbf{y^{0.2323232}}$.

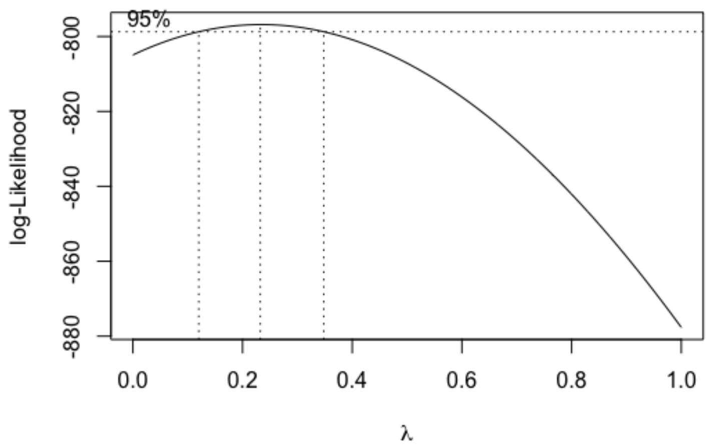

Figure 6: Log-Likelihood

As it can be seen in Figure 6 the lambda value 0.2323232 is where the transformed data has the highest log-likelihood.

As observed in Figure 7, the sizes of the residuals are smaller than in the initial model.

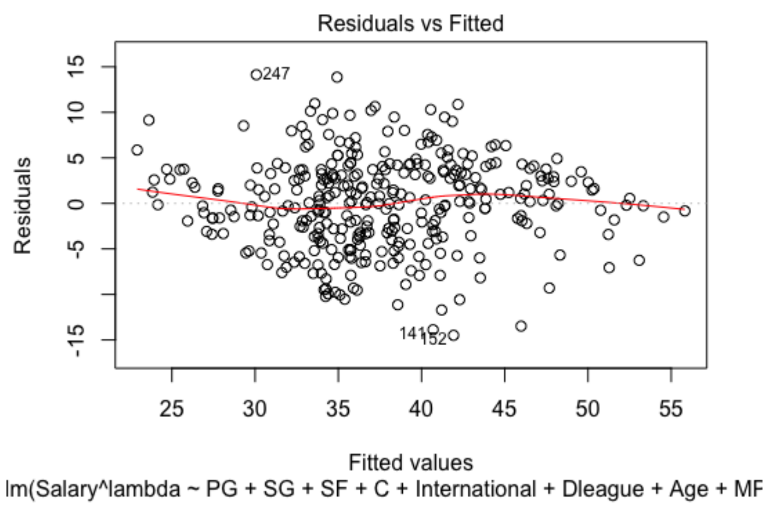

Figure 7: Residuals vs Fitted Box-Cox Model

Also a log-transformation was performed. The residuals from the log transformation can be observed in Figure 8. All residuals have a lower value than |2|, therefore the log-transformation is superior to the Box-Cox transformation. All further calculations and tests will be on the log-transformed model.

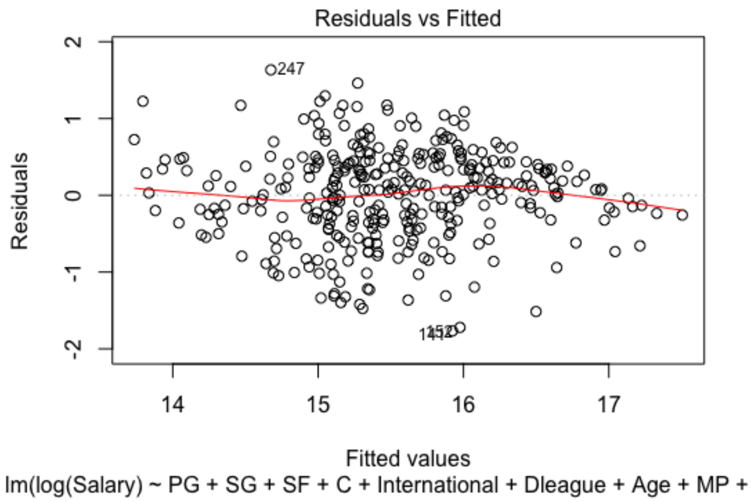

Figure 8: Residuals vs Fitted Log Model

3.4.2 Variable Selection - AIC

To reduce the log-transformed model, stepwise AIC was performed to estimate if a single covariate can be excluded from the model. For each covariate with a high p-value and relatively low η^2, a ΔAIC was calculated. $\Delta AIC = AIC(\textit{Reduced model}) - AIC(\textit{Full model})$ where the *Reduced model* consisted of all variables except one and the *Full model* of all variables. If the ΔAIC was negative the covariate was excluded from the final model. This was repeated stepwise until no ΔAIC was negative. The process can be found in Appendix (8.1).

The covariates that were excluded are stated in the table below.

Covariate	AIC
International	-293.8668
SG	-295.8648
ThreePpr	-297.8520
C	-299.7859
ORtg	-301.6727
DRBPG	-303.5831
MPPG	-305.4380
SF	-307.1782
PFPG	-308.8153
BLKPG	-309.0498

This final model is the model that will be used in the rest of the study:

Dependent variable: log(Salary).
Covariates: PG, Dleague, Age, ORBPG, ASTPG, STLPG, TwoPA, ThreePA, FTAPG, eFGpr, USGpr, DRtg.

3.4.3 Detecting Multicollinearity - VIF

To control multicollinearity between the covariates, a VIF test was performed. The calculations were made with the equation described in section 2.4.8. The result from the test is stated below.

Covariate	VIF
PG	2.042455
Dleague	1.366553
Age	1.166231
ORBPG	2.857945
ASTPG	3.796701
STLPG	2.920548
TwoPA	7.117036
ThreePA	2.758344
FTAPG	3.376217
eFGpr	1.268387
USGpr	5.074694
DRtg	1.544452

As mentioned earlier in section 2.4.8, a VIF-value>10 indicates a problem with multicollinearity. Since no variables exceeds 10, no serious multicollinearity exists. Although TwoPA and USGpr are close to 10, 7.117036 and 5.074694 respectively, which was expected since TwoPA is a part of the equation that creates USGpr. This will not affect the interpretations of the statistics and thereby not a problem.

3.4.4 Normal QQ-plot

A normal QQ-plot has been done to investigate if the residuals are normally distributed. This is the result:

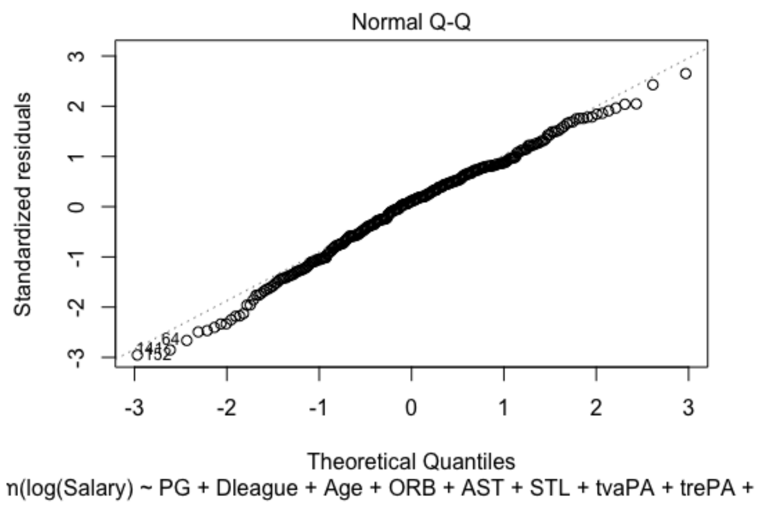

Figure 9: Normal QQ-plot Final Model

It can be observed from the plot that the model is approximately normally distributed. A log transformation was used to make the model better.

3.4.5 Residuals vs Fitted - Final Model

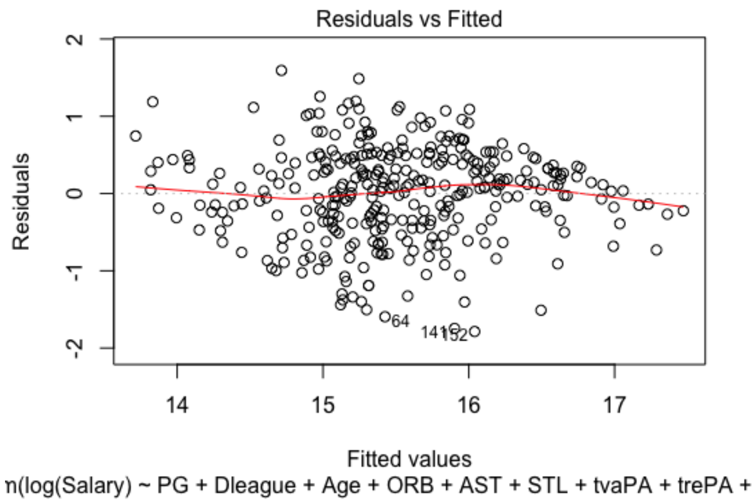

Figure 10: Residuals vs Fitted Final Model

In the final model we can see that the residual is much better than before. However the "S" formed curve is a bit problematic. Primarily most of the residuals are above the zero line, then below and above again. There is a chance that the residuals are not independent. To test this we will use runs test for randomness and observe if the variables are random. Also a Breusch-Pagan test will be used to observe if heteroskedasticity still is present.

3.4.6 Test for Randomness

Runs Test was performed to check for randomness of the model.

Statistic	Runs	n1	n2	n	P-value
0.2192	170	167	167	334	0.8265

As stated in section 2.3.5, the hypothesis that the sequence was produced in a random manner should be rejected if the absolute value of the test statistic exceeds 1.96. Since the statistic is $0.2192 < 1.96$, the null hypothesis can not be rejected so the data is assumed to be from a random process.

3.4.7 Breusch-Pagan Test for Final Model

BP	DF	P-value
13.821	12	0.3123

As it can be observed the P-value>0.05 so accordingly the null hypothesis homoskedasticity is not rejected.

4 Results

4.1 Final Model

Covariate	β-estimate	Standard Error	Eta Squared	P-value	Lower 2.5%	Upper 97.5%
(Intercept)	17.140256	1.467591	-	< 2e-16	14.252977479	20.02753425
PG	-0.226607	0.120756	0.010818032	0.061482	-0.464177073	0.01096368
Dleague	-0.280104	0.136343	0.012937830	0.040743	-0.548339604	-0.01186838
Age	0.050562	0.008711	0.094717966	1.55e-08	0.033423993	0.06769944
ORBPG	0.134817	0.070039	0.011375822	0.055125	-0.002975142	0.27260846
ASTPG	0.101447	0.034970	0.025470135	0.003976	0.032648743	0.17024458
STLPG	-0.472973	0.132638	0.037989466	0.000418	-0.733919207	-0.21202736
TwoPA	0.153087	0.025528	0.100460392	5.40e-09	0.102863684	0.20331074
ThreePA	0.175680	0.028678	0.104378366	2.63e-09	0.119259880	0.23210048
FTAPG	0.052903	0.035810	0.006732326	0.140565	-0.017547952	0.12335433
eFGpr	1.132419	0.709629	0.007846474	0.111517	-0.263675690	2.52851301
USGpr	-0.048861	0.015232	0.030968531	0.001472	-0.078827038	-0.01889520
DRtg	-0.037192	0.011323	0.032420399	0.001133	-0.059467624	-0.01491577

Final model values:

- Multiple R-squared: $R^2 = 0.574$
- Adjusted R-squared: $\bar{R}^2 = 0.5582$
- F-statistic for all covariates to be equal to zero: 36.16 on 12 and 322 DF
- P-value for all covariates to be equal to zero: <2.2e-16

The final regression rendered the following equation:

$log(Salary) = 17.140256 - 0.226607 \times [PG] - 0.280104 \times [Dleague] + 0.050562 \times [Age] + 0.134817 \times [ORBPG] + 0.101447 \times [ASTPG] - 0.472973 \times [STLPG] + 0.153087 \times [TwoPA] + 0.175680 \times [ThreePA] + 0.052903 \times [FTAPG] + 1.132419 \times [eFGpr] - 0.048861 \times [USGpr] - 0.037192 \times [DRtg]$

4.2 Impact from the Covariates

After reduction of the model, the final model was calculated as stated above in section 4.1. 12 covariates were left after reduction and included in the final model.

PG

According to Cohen's rule, the position point guard has a small impact on the model since the $\eta^2 = 0.010818032$. But since the p-value is close to 0.05 the significance is quite high.

	Standard Error	Eta Squared	P-value
PG	0.120756	0.010818032	0.061482

Dleague

The result from the dummy variable Dleague is a low η^2 and a relatively low p-value.

	Standard Error	Eta Squared	P-value
Dleague	0.136343	0.012937830	0.040743

Age

The age of the player have a small to medium impact on the model according to Cohen's rule. The η^2 is the third highest in the model and the p-value is extremely low, which means that the covariate's existens can not be questioned.

	Standard Error	Eta Squared	P-value
Age	0.008711	0.094717966	1.55e-08

ORBPG

Offensive rebounds has a low η^2 and a p-value close to the significance level 0.05, meaning that the covariate should be included in the model.

	Standard Error	Eta Squared	P-value
ORBPG	0.070039	0.011375822	0.055125

ASTPG

Assists has a small impact on the model, according to a low η^2-value. The p-value is also low, meaning it is still is significant in the model.

	Standard Error	Eta Squared	P-value
ASTPG	0.034970	0.025470135	0.003976

STLPG

This covariate also have a low η^2 and a low p-value.

	Standard Error	Eta Squared	P-value
STLPG	0.132638	0.037989466	0.000418

TwoPA

Two point attempts has the largest impact on the Salary, together with three point attempts. The η^2 is relatively high, and the p-value is extremely low.

	Standard Error	Eta Squared	P-value
TwoPA	0.025528	0.100460392	5.40e-09

ThreePA

Three point attempts is also significant in the model since it has one of the largest η^2 and lowest p-value. According to Cohen's rule, it has close to medium impact on the model.

	Standard Error	Eta Squared	P-value
ThreePA	0.028678	0.104378366	2.63e-09

FTAPG

Free throw attempts has an $\eta^2 = 0.006732326$ and a p-value of 0.140565, meaning it has quite low impact.

	Standard Error	Eta Squared	P-value
FTAPG	0.035810	0.006732326	0.140565

eFGpr

The next covariate, effective field goal percentage, has a high p-value and a low η^2.

	Standard Error	Eta Squared	P-value
eFGpr	0.709629	0.007846474	0.111517

USGpr

Usage percentage has a small impact on the model according to Cohen's rule, and a low p-value.

	Standard Error	Eta Squared	P-value
USGpr	0.015232	0.030968531	0.001472

DRtg

Defensive rating is also a covariate with small impact, since the η^2 is 0.032420399. The p-value is low, 0.001133.

	Standard Error	Eta Squared	P-value
DRtg	0.011323	0.032420399	0.001133

4.3 What Studies Really Have Shown About NBA's Salary Cap

This is an example of how the teams pay rolls (cap space) are correlated with the win percentage in the NBA and NFL from Pagels' essay.

In the figures 11 and 12 below we can see a simple linear regression between the win percentage of all the games as the response variable and each teams pay roll (salary cap) as a regressor during 2014-2015. The R^2 value for this relation shows that 21 percentage of the variation in salary cap explains the variation of the win percentage. For the same relation in the NFL the R^2 shows that 7 percentage of the variation in salary cap explains the variation of the win percentage.

As observed, NBA with its soft cap system has a higher spending variation compared to NFL's hard cap system. NBA also has a much higher correlation between cap space versus win percentage. This significant outcome shows that the lack of spending regulations, mainly because of all of salary cap exceptions, might create an opportunity for wealthy owned teams to outcompete other teams. It would mean that the whole leagues revenue and player salaries could decrease as teams with wealthy owners would dominate (Pagels, 2014).

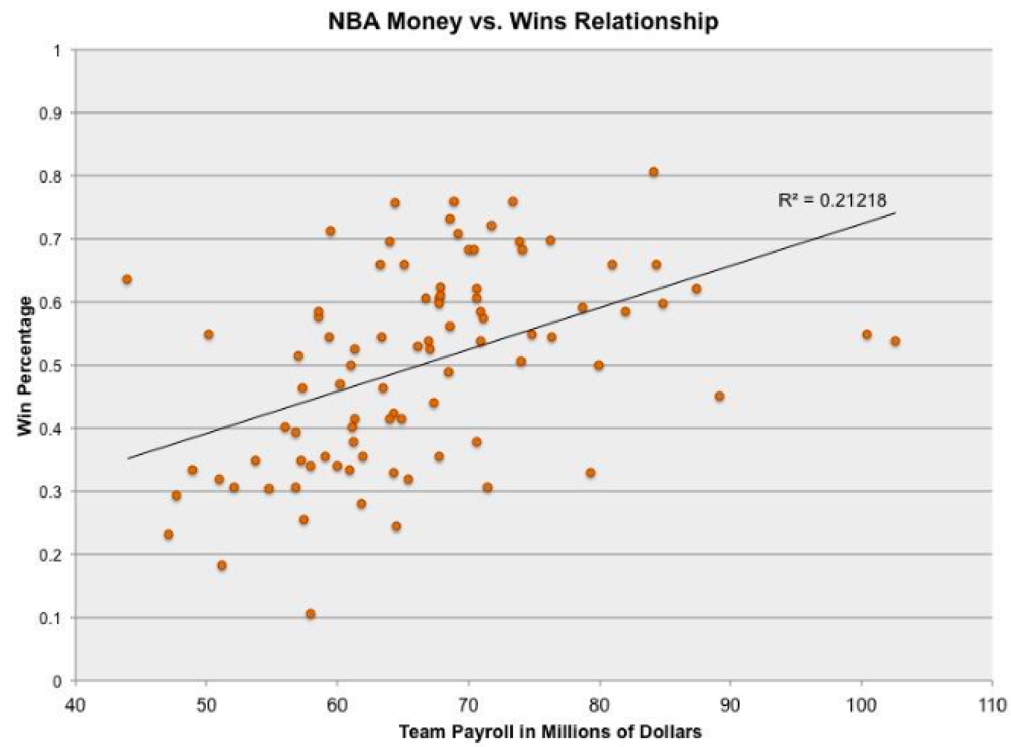

Figure 11: NBA Money vs Wins Relationship (Pagels, 2014)

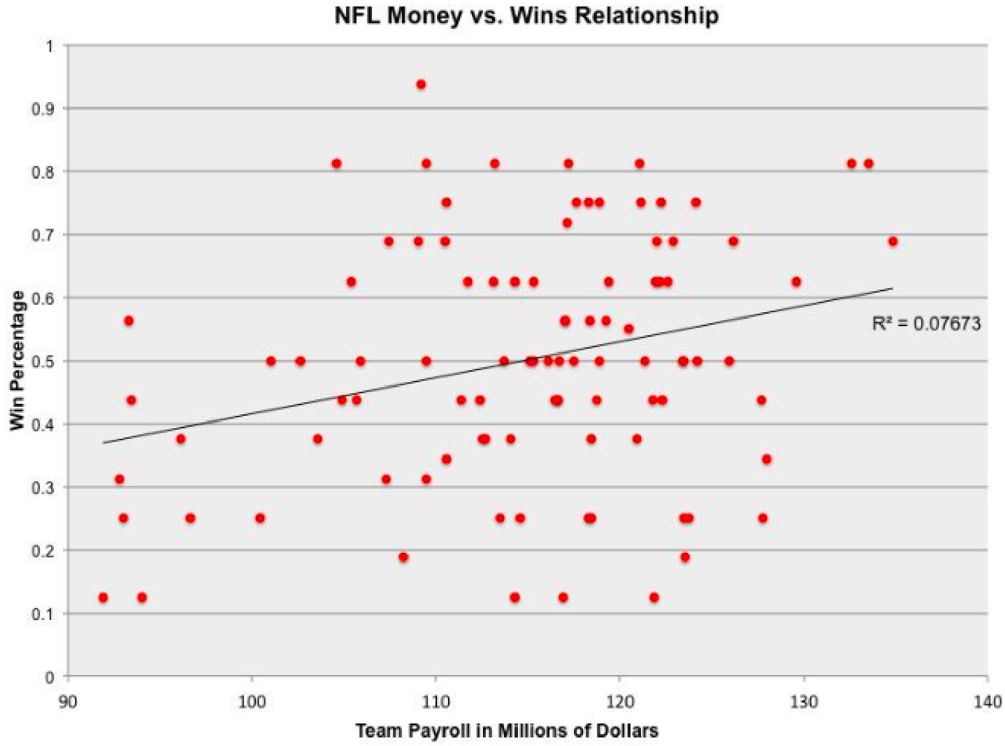

Figure 12: NFL Money vs Wins Relationship (Pagels, 2014)

Another study called *Salary Caps and Competitive Balance in Professional Sports Leagues*, by Evan S. Totty and Mark F. Owens, has analyzed how the salary caps in NBA, NHL and NFL affect the competitive balance through a regression analysis. The results show that salary caps in NHL and NFL did not really affect the competitive balance but affecting the competitive balance in the NBA negatively. Also the results showed that revenue sharing arrangements promote competitive balance in a manner that is consistent with economic theory. (Totty et al., 2014)

There was no evidence in their analysis that salary caps would increase the competitive balance in any of the leagues. The negative impact was most clear in the NBA case, the authors assume that the reason for that are the exemptions that limit player movement by allowing teams to spend over the salary cap. The conclusion from the essay is that the salary cap systems in North America are neither properly designed nor do they satisfy the intentions of what they were set to achieve. (Totty et al., 2014)

5 Discussion

5.1 Analysis of Final Model

PG

According to the η^2 value corresponding to the variable PG, point guard had a low effect on the final model. Point guard was the only position that the stepwise AIC process chose to remain in the model. The coefficient is negative. This implies that if you are an NBA player and a point guard you will earn less than if you are not a point guard. Mainly there are a few PG's among the best paid players. Although there are a lot more low paid PG's than on the other positions so the average point guard is therefore less paid.

Dleague

According to the η^2 value corresponding to the variable Dleague, it had a low effect on the final model. Development league is the NBA's minor league. This is where players get sent if they do not perform good enough. As observed and expected the coefficient in front of Dleague in the regression model is negative. This implies that players who have spent some time in the D-league during the previous season are likely to obtain a lower salary in the next season. Further enhancements and additions could also be done to find

Age

According to the η^2 value corresponding to the variable Age, it had a low effect on the final model. The regression model implies that as a player get older and still plays in the league, he will increase his salary from one season to the next.

ORBPG

According to the η^2 value corresponding to the variable ORBPG, offensive rebounds per game had a low effect on the final model. There is a positive coefficient in front of offensive rebounds per game which implies that the more offensive rebounds a player average during on season the higher will his salary be the succeeding season. In the NBA players average much more defensive rebounds than offensive rebounds, these are rarer and more valuable which probably is the reason of why defensive rebounds per game is reduced from the final model.

ASTPG

According to the η^2 value corresponding to the variable ASTPG, assists per game had a low effect on the final model. The coefficient in front of the variable is positive. This implies that the more assists a player averages during one season the more will he earn the upcoming season. This is a bit surprising. Assist is usually one of the most important and common measure used and talked about. It was believed to have a greater impact on the salary and thereby a higher η^2 value.

STLPG

According to the η^2 value corresponding to the variable STLPG, steals per game had a low effect on the final model. The coefficient in front of the variable is negative. This implies that the more steals a player averages during one season the less will he earn the upcoming season. Steals per game is a performance measure that has a very small range. Usually small and quick players are the best performing in this measure which is probably the reason of why the coefficient is negative.

TwoPA

According to the η^2 value corresponding to the variable TwoPA, two point attempts per game had a medium effect on the final model. The coefficient in front of the variable is positive. This implies that the more two point shot attempts a player averages during one season the more will he earn the upcoming season. This variable had a higher effect on the salary than the other variables, the explanation behind this is that the more points a player score the more shot attempts will he take. High scoring players are always attractive for basketball teams so their salaries increase accordingly.

ThreePA

According to the η^2 value corresponding to the variable ThreePA, three point attempts per game had a medium effect on the final model. The coefficient in front of the variable is positive. This implies that the more two point shot attempts a player averages during one season the more will he earn the upcoming season. This variable had a greater effect on the salary than the other variables, the explanation behind this is that the more points a player score the more shot attempts will he take. High scoring players are always attractive for basketball teams so their salaries increase accordingly.

FTAPG

According to the η^2 value corresponding to the variable FTAPG, free throw attempts per game had a low effect on the final model. The coefficient in front of the variable is positive. This implies that the more free throw shot attempts a player averages during one season the more will he earn the upcoming season.

eFGpr

According to the η^2 value corresponding to the variable eFGpr, effective field goal percentage per game had a low effect on the final model. The coefficient in front of the variable is positive. This implies that the higher percentage effective field goals mede(two and three pointers made/two and three pointers attempted) during one season the greater will his salary be the succeeding season.

USGpr

According to the η^2 value corresponding to the variable USGpr, usage percentage had a low effect on the final model. Usage percentage is defined as an estimate of the percentage of team plays used by a player while he was on the floor. The coefficient in front of the variable is negative. This implies that the lower percentage usage during one season the higher will his salary be the succeeding season. Usage percentage is a mix of free throw attempts, field goal attempts and turnovers a player takes in relation to the amount of these the whole team takes.

DRtg

According to the η^2 value corresponding to the variable DRtg, defensive rating had a low effect on the final model. Defensive rating is defined as an estimate of how many points a player allows the opponent to score per 100 possessions. The coefficient in front of the variable is negative. This implies that a high defensive rating average during one season will decrease the salary for the succeeding season. Even though defense is important people attend games to see offense and in general offense is more valued than defense and the main event of every basketball game. This is probably the reason of why the defensive rating has a negative impact on the salary and offensive rating was reduced from the model.

5.2 Adjustment of Data Set

Initially 476 data points for performance and qualities 2015-2016 and salaries for 2016-2017 was collected. Rookie players who had not played in the NBA before 2015-2016 were removed because no salary information existed. A boundary of at least 100 minutes played was in forced so that only players with legitimate performance values were included. Players with minimum salaries and ten day contracts were removed from the data. Minimum salary players were removed because they all had the same Y values and different values of the independent variables which would result in different coefficients for each minimum salary. This would have a negative effect on the regression. Ten day contract players' performances did not correlate at all with the rest of the players'. Their salaries where much lower but had decent stats since the performance variables are measured per game. Therefore it was important to remove these. Also players that retired in 2016 were removed. After the adjustment 335 data existed and was used in the OLS.

5.3 Analysis of Residuals and Outliers

Initially before removing ten day contract and minimum players, these were outlier data points. After the removal the R^2 value increased significantly and the final model did not have any outliers according to Cook's distance. Also the standardized residuals from the final model are smaller than 3, see Figure 9, which indicates that outliers do not exist. To investigate if the players with the largest standardized residuals have anything in common, the following players are analyzed.

Jarett Jack - In the 2015-2016 season the player had above average stats. His salary the same season was 6.3m and end of a long term contract. In mid season he got badly injured and later waived by the Brooklyn Nets. In his next season when he was 32 years old he signed with for 1.5m. Because of his previous injury and age he was worth more. In summary the player is an above average player during 2015-2016 but earns a very small amount in relation to his performance mainly due to injury.

Rodney Hood - 2015-2016 was his second year and he put up above average stats. It looks like he has signed a bad multi-year contract with the Utah Jazz. In 2016-2017 he only earned 1.35m which is really rare in relation to his stats. A normal player with his stats would normally earn around 6m.

Robert Covington - Is a similar case. 2013-2014 He played one season in Houston rockets with very low stats. In 2014-2015 he signed a multi-year deal with Philadelphia Sixers. In all three seasons with the Sixers he has had above average stats. His salary during 2016-2017 was only 1m which is really low for his stats. The reason is probably that he had a bad season in the Houston Rockets before signing a multi-year contract.

C.J. McCullum - He went from averaging 6.8 points in 2014-2015 to averaging 20.6 points in 2015-2016. This is one of the greatest player developments of all time. During the 2016-2017 he only earned 3.2m because the Portland Trailblazers had to free cap space so from the 2017-2018 he has signed an 4 year extension where he will earn around 25m every year.

Nikola Pekovic - Signed a multi-year contract in 2013 when he was a above average player for 12m a year for five years. In 2014-2015 he got injured and only played 12 games during the season where he did not perform nearly as well as he used to.

Allen Crabbe - Improved his stats a lot in 2015-2016. In 2016-2017 he was offered a generous 4 year 75m dollar deal around 18m per year. This was really high compared to his performance the previous year. Although he is only 24 years old so maybe the Portland Trailblazers saw this as an future investment.

Through analyzing these players it can be concluded that multi-year contracts affect the outcome of the model negatively. As observed injuries on players with multi-year contracts or young players who have a breakthrough after signing a multi-year contract have residuals that stand out. By removing the players above the R^2 value increases by three percentage which implies that long term contracts have a grand impact on the results. The conclusion is that if a point corresponding to a certain player has a big residual, there is a great chance that the player has signed a multi-year contract.

5.4 Model Development

Primarily another set up of initial covariates could have been used. It is not obvious that the covariates that we chose to include in the model is the optimal set. Other covariates could have had a greater significance. There is an issue of multi-year contracts. Future projects could develop our model by handle the issue of long-term contracts, perhaps by using data from several seasons. Futher, the performance based salary model can be developed to identify underpaid players who can contribute to winning games. This is what the Oakland Athletics baseball management did and won the league. By enhancing our model a similar method could be used to try to win the NBA.

5.5 Possible Enhancement of the Salary Cap System

In the literature analysis it was concluded that salary caps really do not achieve what they are supposed to. However a first step would be to remove the exceptions so that all of the teams would have the same amounts to spend on player salary. This would decrease the impact that each team's payroll has on the chance of winning and emphasizes the competitive balance.

6 Conclusion

The conclusion from the regression analysis is that 12 different covariates have an impact on the salary. The regression model achieved explanatory level of 57.4%. By comparing the covariates in the final model, it can be concluded that the performance measures that had the greatest impact on the salaries were *two* and *three point attempts*. All the other covariates *point guard* position, if the player has played in *D-league* or not, *Age, Offensive rebounds, Assists, Steals, Free throw attempts, Field goal percentage, Usage percentage* and *Defensive rating* in the final model had and impact but not as major as the two mentioned ones. This model can now be used as benchmark to other studies and when valuating which salary should be set for free agents or new contract deals. The model can also be developed to handle the issue of long-term contracts and to identify underpaid players who can contribute to winning games. The conclusion from the literature analysis shows that the salary cap lacks to achieve everything it is supposed to fulfill.

7 References

D. Asteriou, S. G. Hall, *Applied econometrics*, Palgrave Macmillan (2011)

L. A. Becker, *Effect Size (ES)*, UCCS.EDU (2000) `http://www.uccs.edu/lbecker/effect-size.html` 2017-04-21

D. A. Belsley, E. Kuh, R. E. Welsch, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, Wiley Series in Probability and Statistics (1980)

J. V. Bradley, *Distribution-Free Statistical Tests* (1968)

K. P. Burnham, D. R. Anderson, *Model selection and multimodel inference Pectical Information-Theoretic approach*, 2nd edidtion (2002)

P. Chakravarthy, *Salary Allocation and Risk Preferences in the National Football League: The Implications of Salary Allocation in Understanding the Preferences of NFL Owners*, Harvard College (2012)

J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition (1988)

L. Coon, *NBA Salary Cap FAQ*, Collective Bargaining Agreement (2016) `http://www.cbafaq.com/salarycap.htm#Q5` 2017-05-03

C. Ford, *Understanding Q-Q Plots*, University of Virginia Library (2015) `http://data.library.virginia.edu/understanding-q-q-plots/` 2017-04-19

J. Frost, *Regression analysis: How do I interpret r-squared and asses the goodness-of-fit?* (2013)

J. Fullard, *Investigating Player Salaries and Performance in the National Hockey League*, Brock University (2012)

C. Gaines, *The NBA is the highest-paying sports league in the world*, Business Insider (2015) `http://www.businessinsider.com/sports-leagues-top-salaries-2015-5?r=US&IR=T&IR=T` 2017-05-09

Investopedia, *Hypothesis-testing*, `http://www.investopedia.com/exam-guide/cfa-level-1/quantitative-methods/hypothesis-testing.asp#ixzz4ecUhPYTa` 2017-04-20

P. Kennedy, *A guide to econometrics. 6. ed.*, Oxford: Blackwell (2008)

H. Lang, *Elements of Regression Analysis*, KTH Mathematics (2015)

H. Lang, *Elements of Regression Analysis*, KTH Mathematics (2016)

M. Lewis, *Moneyball: The Art of Winning an Unfair Game*, W. W. Norton and Company (2003)

MLB.com, (2017) https://www.mlb.com 2017-05-05

D. Montgomery, E. Peck, G. Vining: *Introduction to Linear Regression Analysis*, Wiley-Interscience, 5th Edition (2012)

NBA.com, (2015) http://www.nba.com/2015/news/07/08/nba-salary-cap-2016-official-release/ 2017-05-04

NBA.com, *FAQ* (2016) http://www.nba.com/news/faq 2017-05-05

NBA.com, (2016) http://www.nba.com/2016/news/07/02/nba-salary-cap-set/ 2017-05-04

C. Neiger, *How Salary Caps Changed Sports*, Investopedia (2010) http://www.investopedia.com/financial-edge/0910/how-salary-caps-changed-sports.aspx 2017-05-09

NFL.com, (2017) https://www.nfl.com 2017-05-05

NHL.com, (2017) https://www.nhl.com 2017-05-05

J. Pagels, *Are Salary Caps for Professional Athletes Fair?*, Priceonomics (2014) https://priceonomics.com/are-salary-caps-for-professional-athletes-fair/ 2017-05-09

K. Peck, *Salary Determination in the National Hockey League: Restricted, Unrestricted, Forwards, and Defensemen*, Western Michigan University (2012)

E. S. Totty, M. F. Owens, *Salary Caps and Competitive Balance in Professional Sports Leagues*, Journal for Economic Educators (2011) https://www.researchgate.net/publication/227458677_Salary_Caps_and_Competitive_Balance_in_Professional_Sports_Leagues 2017-05-09

M. Wallace, *The Advantages of Salary Caps*, Small Business - Chron.com (2011) http://smallbusiness.chron.com/advantages-salary-caps-18682.html 2017-05-09

J. M. Wooldridge, *Introductory Econometrics: A Modern Approach*, South-Western, 5th Edition (2013)

A Appendix

A.1 Stepwise AIC in R

Start: AIC=-291.87

log(Salary) ~ PG + SG + SF + C + International + Dleague + Age +
MPPG + DRBPG + ORBPGPG + PFPG + ASTPG + STLPG + BLKPG + TwoPA + ThreePA
+ FTAPG + ThreePpr + eFGpr + USGpr + ORtg + DRtg

		Df	Sum of Sq	RSS	AIC
-	International	1	0.0009	122.19	-293.87
-	SG	1	0.0009	122.19	-293.87
-	trePpr	1	0.0040	122.19	-293.86
-	C	1	0.0242	122.21	-293.80
-	ORtg	1	0.0428	122.23	-293.75
-	DRB	1	0.0479	122.24	-293.74
-	MP	1	0.0484	122.24	-293.74
-	SF	1	0.0776	122.27	-293.66
-	PF	1	0.1251	122.31	-293.53
-	eFGpr	1	0.2363	122.42	-293.22
-	BLK	1	0.3622	122.55	-292.88
-	FTA	1	0.5540	122.74	-292.35
	<none>			122.19	-291.87
-	ORB	1	0.7958	122.98	-291.69
-	PG	1	0.9739	123.16	-291.21
-	Dleague	1	1.6266	123.81	-289.44
-	USGpr	1	2.4625	124.65	-287.19
-	AST	1	2.5478	124.74	-286.96
-	DRtg	1	3.2725	125.46	-285.01
-	STL	1	3.5213	125.71	-284.35
-	trePA	1	7.2262	129.41	-274.62
-	tvaPA	1	8.8192	131.01	-270.52
-	Age	1	11.7691	133.96	-263.06

Step: AIC=-293.87

log(Salary) ~ PG + SG + SF + C + Dleague + Age + MPPG + DRBPG + ORBPGPG +
PFPG + ASTPG + STLPG + BLKPG + TwoPA + ThreePA + FTAPG + ThreePpr + eFGpr +
USGpr + ORtg + DRtg

		Df	Sum of Sq	RSS	AIC
-	SG	1	0.0007	122.19	-295.87
-	trePpr	1	0.0044	122.19	-295.86
-	C	1	0.0286	122.22	-295.79
-	ORtg	1	0.0434	122.23	-295.75
-	MP	1	0.0475	122.24	-295.74
-	DRB	1	0.0482	122.24	-295.74
-	SF	1	0.0768	122.27	-295.66
-	PF	1	0.1254	122.31	-295.52
-	eFGpr	1	0.2355	122.42	-295.22

-	BLK	1	0.3621	122.55	-294.88
-	FTA	1	0.5531	122.74	-294.35
	<none>			122.19	-293.87
-	ORB	1	0.7977	122.99	-293.69
-	PG	1	0.9737	123.16	-293.21
+	International	1	0.0009	122.19	-291.87
-	Dleague	1	1.6280	123.82	-291.43
-	USGpr	1	2.4639	124.65	-289.18
-	AST	1	2.5563	124.75	-288.93
-	DRtg	1	3.2875	125.48	-286.97
-	STL	1	3.5817	125.77	-286.19
-	trePA	1	7.2391	129.43	-276.58
-	tvaPA	1	8.8641	131.05	-272.41
-	Age	1	11.7769	133.97	-265.04

Step: AIC=-295.86
log(Salary) ~ PG + SF + C + Dleague + Age + MPPG + DRBPG + ORBPGPG +
PFPG + ASTPG + STLPG + BLKPG + TwoPA + ThreePA + FTAPG + ThreePpr + eFGpr +
USGpr + ORtg + DRtg

		Df	Sum of Sq	RSS	AIC
-	trePpr	1	0.0047	122.19	-297.85
-	C	1	0.0280	122.22	-297.79
-	ORtg	1	0.0463	122.24	-297.74
-	MP	1	0.0467	122.24	-297.74
-	DRB	1	0.0531	122.24	-297.72
-	SF	1	0.1121	122.30	-297.56
-	PF	1	0.1251	122.31	-297.52
-	eFGpr	1	0.2349	122.42	-297.22
-	BLK	1	0.3615	122.55	-296.88
-	FTA	1	0.5628	122.75	-296.32
	<none>			122.19	-295.87
-	ORB	1	0.8373	123.03	-295.58
+	SG	1	0.0007	122.19	-293.87
+	International	1	0.0007	122.19	-293.87
-	PG	1	1.4959	123.69	-293.79
-	Dleague	1	1.6465	123.84	-293.38
-	USGpr	1	2.4887	124.68	-291.11
-	AST	1	2.5571	124.75	-290.93
-	DRtg	1	3.3055	125.50	-288.92
-	STL	1	3.6886	125.88	-287.90
-	trePA	1	7.2965	129.49	-278.44
-	tvaPA	1	8.8974	131.09	-274.32
-	Age	1	11.8010	133.99	-266.98

Step: AIC=-297.85

log(Salary) ~ PG + SF + C + Dleague + Age + MPPG + DRBPG + ORBPGPG + PFPG + ASTPG + STLPG + BLKPG + TwoPA + ThreePA + FTAPG + eFGpr + USGpr + ORtg + DRtg

		Df	Sum of Sq	RSS	AIC
-	C	1	0.0241	122.22	-299.79
-	ORtg	1	0.0444	122.24	-299.73
-	MP	1	0.0479	122.24	-299.72
-	DRB	1	0.0486	122.24	-299.72
-	SF	1	0.1133	122.31	-299.54
-	PF	1	0.1239	122.32	-299.51
-	eFGpr	1	0.2317	122.43	-299.22
-	BLK	1	0.3588	122.55	-298.87
-	FTA	1	0.6072	122.80	-298.19
	<none>			122.19	-297.85
-	ORB	1	0.9424	123.14	-297.28
+	trePpr	1	0.0047	122.19	-295.87
+	International	1	0.0010	122.19	-295.86
+	SG	1	0.0010	122.19	-295.86
-	PG	1	1.4991	123.69	-295.77
-	Dleague	1	1.6422	123.84	-295.38
-	USGpr	1	2.5039	124.70	-293.06
-	AST	1	2.5775	124.77	-292.86
-	DRtg	1	3.3701	125.56	-290.74
-	STL	1	3.7651	125.96	-289.69
-	trePA	1	7.3745	129.57	-280.22
-	tvaPA	1	8.9043	131.10	-276.29
-	Age	1	11.8908	134.09	-268.74

Step: AIC=-299.79

log(Salary) ~ PG + SF + Dleague + Age + MPPG + DRBPG + ORBPGPG + PFPG + ASTPG + STLPG + BLKPG + TwoPA + ThreePA + FTAPG + eFGpr + USGpr + ORtg + DRtg

		Df	Sum of Sq	RSS	AIC
-	ORtg	1	0.0413	122.26	-301.67
-	DRB	1	0.0498	122.27	-301.65
-	MP	1	0.0513	122.27	-301.64
-	SF	1	0.1021	122.32	-301.51
-	PF	1	0.1432	122.36	-301.39
-	eFGpr	1	0.2273	122.45	-301.16
-	BLK	1	0.4224	122.64	-300.63
-	FTA	1	0.6027	122.82	-300.14
	<none>			122.22	-299.79

-	ORB	1	0.9216	123.14	-299.27
+	C	1	0.0241	122.19	-297.85
+	International	1	0.0049	122.21	-297.80
+	trePpr	1	0.0008	122.22	-297.79
+	SG	1	0.0003	122.22	-297.79
-	PG	1	1.4839	123.70	-297.74
-	Dleague	1	1.6588	123.88	-297.27
-	USGpr	1	2.5565	124.78	-294.85
-	AST	1	2.5788	124.80	-294.79
-	DRtg	1	3.3521	125.57	-292.72
-	STL	1	3.7412	125.96	-291.69
-	trePA	1	7.7887	130.01	-281.09
-	tvaPA	1	8.9728	131.19	-278.05
-	Age	1	11.8669	134.09	-270.74

Step: AIC=-301.67

log(Salary) ~ PG + SF + Dleague + Age + MPPG + DRBPG + ORBPGPG + PFPG +
ASTPG + STLPG + BLKPG + TwoPA + ThreePA + FTAPG + eFGpr + USGpr + DRtg

		Df	Sum of Sq	RSS	AIC
-	DRB	1	0.0327	122.29	-303.58
-	MP	1	0.0487	122.31	-303.54
-	SF	1	0.1128	122.37	-303.36
-	PF	1	0.1651	122.42	-303.22
-	BLK	1	0.4988	122.76	-302.31
	<none>			122.26	-301.67
-	FTA	1	0.9225	123.18	-301.15
-	eFGpr	1	1.0715	123.33	-300.75
-	ORB	1	1.1561	123.42	-300.52
+	ORtg	1	0.0413	122.22	-299.79
+	C	1	0.0210	122.24	-299.73
+	International	1	0.0050	122.25	-299.69
+	SG	1	0.0022	122.26	-299.68
+	trePpr	1	0.0002	122.26	-299.67
-	PG	1	1.4790	123.74	-299.64
-	Dleague	1	1.6347	123.89	-299.22
-	AST	1	2.7427	125.00	-296.24
-	USGpr	1	2.9667	125.23	-295.64
-	DRtg	1	3.6774	125.94	-293.75
-	STL	1	4.0498	126.31	-292.76
-	trePA	1	8.2280	130.49	-281.85
-	tvaPA	1	9.2660	131.53	-279.20
-	Age	1	12.0837	134.34	-272.10

Step: AIC=-303.58

log(Salary) ~ PG + SF + Dleague + Age + MPPG + ORBPGPG + PFPG + ASTPG + STLPG + BLKPG + TwoPA + ThreePA + FTAPG + eFGpr + USGpr + DRtg

		Df	Sum of Sq	RSS	AIC
-	MP	1	0.0530	122.35	-305.44
-	SF	1	0.0985	122.39	-305.31
-	PF	1	0.1486	122.44	-305.18
-	BLK	1	0.4682	122.76	-304.30
	<none>			122.29	-303.58
-	FTA	1	1.0435	123.34	-302.74
-	eFGpr	1	1.0459	123.34	-302.73
+	DRB	1	0.0327	122.26	-301.67
+	ORtg	1	0.0242	122.27	-301.65
+	C	1	0.0226	122.27	-301.64
+	SG	1	0.0058	122.29	-301.60
+	International	1	0.0044	122.29	-301.60
+	trePpr	1	0.0007	122.29	-301.58
-	PG	1	1.5470	123.84	-301.37
-	Dleague	1	1.6259	123.92	-301.16
-	ORB	1	1.7702	124.06	-300.77
-	AST	1	2.9433	125.24	-297.62
-	USGpr	1	3.1826	125.48	-296.98
-	STL	1	4.2933	126.59	-294.02
-	DRtg	1	4.4224	126.72	-293.68
-	trePA	1	8.9724	131.26	-281.87
-	tvaPA	1	9.9498	132.24	-279.38
-	Age	1	12.1447	134.44	-273.87

Step: AIC=-305.44

log(Salary) ~ PG + SF + Dleague + Age + ORBPGPG + PFPG + ASTPG + STLPG + BLKPG + TwoPA + ThreePA + FTAPG + eFGpr + USGpr + DRtg

		Df	Sum of Sq	RSS	AIC
-	SF	1	0.0949	122.44	-307.18
-	PF	1	0.1402	122.49	-307.05
-	BLK	1	0.4693	122.81	-306.16
	<none>			122.35	-305.44
-	eFGpr	1	1.0653	123.41	-304.53
-	FTA	1	1.0716	123.42	-304.52
+	MP	1	0.0530	122.29	-303.58
+	DRB	1	0.0370	122.31	-303.54
+	C	1	0.0262	122.32	-303.51
+	ORtg	1	0.0213	122.33	-303.50
+	SG	1	0.0021	122.34	-303.44
+	International	1	0.0018	122.34	-303.44

+	trePpr	1	0.0006	122.34	-303.44
-	PG	1	1.5275	123.87	-303.28
-	Dleague	1	1.8190	124.17	-302.49
-	ORB	1	1.9211	124.27	-302.22
-	AST	1	3.0179	125.36	-299.27
-	STL	1	4.2706	126.62	-295.94
-	DRtg	1	4.4857	126.83	-295.38
-	USGpr	1	4.4860	126.83	-295.38
-	Age	1	12.1013	134.45	-275.84
-	tvaPA	1	14.5453	136.89	-269.81
-	trePA	1	14.6513	137.00	-269.55

Step: AIC=-307.18

log(Salary) ~ PG + SF + Dleague + Age + ORBPGPG + PFPG + ASTPG + STLPG + BLKPG + TwoPA + ThreePA + FTAPG + eFGpr + USGpr + DRtg

		Df	Sum of Sq	RSS	AIC
-	PF	1	0.1327	122.57	-308.81
-	BLK	1	0.4368	122.88	-307.99
	<none>			122.44	-307.18
-	FTA	1	1.0142	123.45	-306.42
-	eFGpr	1	1.0743	123.52	-306.25
+	SF	1	0.0949	122.35	-305.44
+	MP	1	0.0494	122.39	-305.31
-	PG	1	1.4355	123.88	-305.27
+	ORtg	1	0.0318	122.41	-305.26
+	SG	1	0.0240	122.42	-305.24
+	DRB	1	0.0218	122.42	-305.24
+	C	1	0.0140	122.43	-305.22
+	International	1	0.0010	122.44	-305.18
+	trePpr	1	0.0000	122.44	-305.18
-	Dleague	1	1.8319	124.27	-304.20
-	ORB	1	2.0391	124.48	-303.64
-	AST	1	3.1377	125.58	-300.70
-	USGpr	1	4.3925	126.83	-297.37
-	STL	1	4.4256	126.87	-297.28
-	DRtg	1	4.4599	126.90	-297.19
-	Age	1	12.2444	134.69	-277.25
-	tvaPA	1	14.4582	136.90	-271.79
-	trePA	1	14.6149	137.06	-271.40

Step: AIC=-308.82

log(Salary) ~ PG + Dleague + Age + ORBPGPG + ASTPG + STLPG + BLKPG + TwoPA + ThreePA + FTAPG + eFGpr + USGpr + DRtg

		Df	Sum of Sq	RSS	AIC
-	BLK	1	0.6477	123.22	-309.05
	<none>			122.57	-308.81
-	FTA	1	1.0601	123.63	-307.93
-	eFGpr	1	1.1338	123.71	-307.73
+	PF	1	0.1327	122.44	-307.18
-	PG	1	1.3747	123.95	-307.08
+	SF	1	0.0874	122.49	-307.05
+	ORtg	1	0.0531	122.52	-306.96
+	MP	1	0.0417	122.53	-306.93
+	SG	1	0.0355	122.54	-306.91
+	C	1	0.0281	122.55	-306.89
+	DRB	1	0.0097	122.56	-306.84
+	International	1	0.0035	122.57	-306.82
+	trePpr	1	0.0002	122.57	-306.82
-	Dleague	1	1.6999	124.27	-306.20
-	ORB	1	1.9149	124.49	-305.62
-	AST	1	3.0363	125.61	-302.62
-	USGpr	1	4.3229	126.90	-299.20
-	DRtg	1	4.7751	127.35	-298.01
-	STL	1	5.2684	127.84	-296.72
-	Age	1	12.3315	134.91	-278.70
-	tvaPA	1	14.3967	136.97	-273.61
-	trePA	1	14.4868	137.06	-273.39

Step: AIC=-309.05
log(Salary) ~ PG + Dleague + Age + ORBPGPG + ASTPG + STLPG + TwoPA + ThreePA + FTAPG + eFGpr + USGpr + DRtg

		Df	Sum of Sq	RSS	AIC
	<none>			123.22	-309.05
+	BLK	1	0.6477	122.57	-308.81
-	FTA	1	0.8352	124.06	-308.79
-	eFGpr	1	0.9745	124.20	-308.41
+	PF	1	0.3436	122.88	-307.99
+	ORtg	1	0.2025	123.02	-307.60
+	C	1	0.1446	123.08	-307.44
-	PG	1	1.3476	124.57	-307.41
-	ORB	1	1.4179	124.64	-307.22
+	SF	1	0.0455	123.18	-307.17
+	SG	1	0.0403	123.18	-307.16
+	MP	1	0.0387	123.18	-307.15
+	DRB	1	0.0072	123.21	-307.07
+	trePpr	1	0.0071	123.21	-307.07
+	International	1	0.0028	123.22	-307.06

-	Dleague	1	1.6151	124.84	-306.69
-	AST	1	3.2205	126.44	-302.41
-	USGpr	1	3.9379	127.16	-300.51
-	DRtg	1	4.1287	127.35	-300.01
-	STL	1	4.8660	128.09	-298.07
-	Age	1	12.8924	136.11	-277.71
-	tvaPA	1	13.7613	136.98	-275.58
-	trePA	1	14.3606	137.58	-274.12

TRITA -MAT-K 2017:15
ISRN -KTH/MAT/K--17/15--SE

www.kth.se