Universidade de Brasília
Departamento de Estatística

Regressão Quantílica no Contexto da NBA: Fatores Determinantes nos Salários dos Atletas

Lucas Seiti Yamazaki

Projeto apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2021

Lucas Seiti Yamazaki - 16/0013224

Regressão Quantílica no Contexto da NBA: Fatores Determinantes nos Salários dos Atletas

Orientador(a): Prof. Leandro Tavares Correia
Coorientador(a): Prof. Helton Saulo Bezerra dos Santos

Projeto apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2021

Agradecimentos

- À minha família: minha mãe Cristina, minha irmã Beatriz e meu pai Cícero. Sem vocês nada seria possível.
- Aos professores e servidores da Universidade de Brasília, em especial ao professor Leandro, que me orientou durante todo este trabalho.

Resumo

A *National Basketball Association* (NBA) é a principal liga de basquete do mundo. Diante da imposição de um teto salarial aos times na montagem de seus elencos, faz-se necessário identificar quais são os principais fatores que norteiam a remuneração de jogadores para melhores tomadas de decisão. Utilizando a regressão quantílica com estatísticas da temporada 2020-2021 da NBA como variáveis preditoras e salários da mesma temporada como variável dependente, foram identificados alguns fatores e em que quantis são mais impactantes.

Palavras-chaves: regressão quantílica, fatores determinantes, salário, remuneração

Lista de Tabelas

Lista de Figuras

Sumário

1 Introdução

A *National Basketball Association* (NBA) é a principal liga profissional de basquete masculino dos Estados Unidos. Estabelecida em 1949 (ROSEN, 2008), a liga se tornou um fenômeno não apenas em território americano e hoje atrai espectadores pelo mundo inteiro.

A liga é referência para os melhores jogadores de basquete do mundo. Um bom exemplo disso foi visto nas Olimpíadas de Tóquio de 2020. Na modalidade de basquete masculino, onde cada nação poderia selecionar 12 jogadores para representar seu país, as três seleções que foram ao pódio (Estados Unidos, França e Austrália) tinham 24 jogadores ativos na NBA (WATSON, 2021). A seleção francesa ainda tinha três jogadores que já atuaram na liga americana no passado, fazendo o número ir para 27 do total dos 36 medalhistas olímpicos. Mais ainda, dos 12 países que disputaram a modalidade, 11 tinham pelo menos um representante que atuava pela NBA. Como consequência de se ter os melhores jogadores na mesma liga, a remuneração dos atletas é bastante expressiva.

A liga americana de basquete impõe um teto salarial anual, isto é, um limite que os times podem gastar com seus jogadores. Para a temporada 2021-2022, este valor está fixado em $112,414 milhões de dólares (KASABIAN, 2021). Além disso, a liga também impõe um valor mínimo de folha salarial que cada time deve ter. Este valor corresponde a 90% do teto salarial, isto é, na temporada 2021-22 o mínimo corresponde a $101,173 milhões.

Ao considerar os limites impostos na folha salarial, surge uma questão importante para os times ao construir seus elencos. É de interesse de cada equipe contar com os melhores jogadores possíveis pagando valores baixos em seu processo de aquisição, o que não é um processo fácil. Naturalmente, os melhores jogadores são os mais bem remunerados e vão consumir uma parte grande do teto salarial. Assim, o que acontece fora de quadra se torna tão importante quanto o que acontece dentro do jogo. Como cada franquia consegue avaliar bons jogadores, identificar atletas com um bom custo-benefício e alocar os salários dentro do teto imposto pela liga dita bastante de como um time vai conseguir performar ao longo de uma temporada.

O presente estudo busca explorar este problema ao aprofundar os estudos acerca dos determinantes salariais nos atletas da liga NBA. O intuito é identificar quais razões levam os jogadores a serem mais ou menos remunerados e neste processo trazer informações acerca de como um time poderia montar seu elenco de maneira a otimizar o espaço pre-

sente no teto salarial.

Para alcançar o objetivo proposto, será utilizada a técnica de regressão quantílica, introduzida por Koenker e Bassett (1978). Um estudo similar foi feito por Vincent e Eastman (2009), mas com enfoque na NHL, liga de hóquei no gelo dos Estados Unidos e Canadá. Para este estudo, a variável em análise será os log-salários dos jogadores na temporada 2020-2021 e as variáveis explicativas serão estatísticas pessoais dos atletas na mesma temporada e estatísticas de equipe.

A opção pela regressão quantílica se deve pelo fato de ser uma técnica mais robusta à regressão clássica. Porém, a fim de comparar ambas as técnicas, o estudo também apresentará a regressão por Mínimos Quadrados Ordinários para se ter uma ideia que informação adicional pode ser obtida pelo método quantílico. O trabalho, por conter muitas variáveis, também explora alguns métodos de redução de dimensionalidade, bem como de escolha de modelos por métodos de penalização.

Desta forma, o trabalho tem como objetivo geral identificar fatores determinantes na remuneração dos jogadores da NBA. E, ao longo do seu desenvolvimento, compreender a técnica de regressão quantílica e suas aplicações, bem como explorar suas vantagens em relação à regressão clássica.

O estudo visa implementar a técnica computacionalmente através do *software* R e pretende comparar fatores determinantes entre atletas mais e menos remunerados. E, através dessas ferramentas, analisar a dinâmica do teto salarial imposta pela liga de basquete.

2 Metodologia

Esta seção está dividida em seis partes a começar por uma contextualização da situação econômica que a NBA se encontra atualmente. Em seguida, uma discussão acerca das técnicas de regressão clássica e quantílica é proposta. Depois, a técnica de regressão quantílica é discutida mais afundo e é feito um detalhamento sobre o banco de dados em estudo. Por fim, é apresentado métodos de redução de dimensionalidade e escolha de modelo por regularização utilizados para diminuir o número de variáveis preditoras do modelo.

2.1 Situação Financeira da NBA

As quatro principais ligas esportivas dos Estados Unidos (NBA, MLB, NFL e NHL) são estruturadas no sistema de franquias. Cada liga tem um número fixo de times que não se alteram de ano para ano como ocorre no futebol. Por exemplo, os 20 times que disputam a Série A do Campeonato Brasileiro de futebol em um ano nunca são os mesmos 20 times do ano anterior. Além disso, cada franquia possui um proprietário que é responsável por decisões esportivas e econômicas.

A liga americana é composta por 30 times e cada franquia pode ser negociada para um novo proprietário em transações que chegam a ser bilionárias. A negociação mais cara de um time ocorreu em 2019 quando Joseph Tsai comprou o *Brooklyn Nets* por 2,35 bilhões de dólares (GLEESON, 2019). Mesmo em um ano ruim economicamente por conta da pandemia, o lucro médio de cada time em 2020 foi de 62 milhões de dólares e o valor médio de cada franquia foi de $2,2 bilhões (BADENHAUSEN; OZANIAN, 2021). O time que apresentou menor lucro foi o *Phoenix Suns* com $20 milhões apresentando uma receita de 222 milhões de dólares.

O aspecto lucrativo que a NBA tem não se restringe apenas aos proprietários das franquias, se estendendo aos jogadores que estão entre os mais bem pagos do mundo. Principalmente se comparada com atletas de outras ligas, como NFL, MLB e a já citada *Premier League* (ROA, 2021). A termos de comparação, Jordan Henderson, jogador de futebol pela equipe do *Liverpool*, capitão de seu time, representante da seleção nacional da Inglaterra, com anos de experiência e conquistas expressivas em sua carreira, terá como remuneração salarial aproximadamente o mesmo que Cade Cunningham, do *Detroit Pistons*, jogador de 19 anos que ainda sequer esteve presente em um jogo de NBA. Cade terá 2021-2022 como primeira temporada na liga e mesmo recém-chegado do basquete

universitário, já tem um salário alto com prospectiva de aumento após alguns anos.

Os salários, porém, nem sempre foram tão exorbitantes. Em 2014, foi firmado um acordo televisivo por parte da NBA que renderia à liga $2,6 bilhões por ano a partir de 2016 (MATANGE, 2021). Com isso, uma parte do dinheiro seria destinado aos jogadores como foi estabelecido no *Collective Bargaining Agreement* (CBA) da época. O CBA é um acordo trabalhista coletivo entre jogadores e liga, presente nos os esportes americanos.

A partir de 2016, os salários tiveram um salto significativo em seus valores. Na temporada 2015-2016, o teto era de $70 milhões. No ano seguinte, este valor subiu para $94,1 milhões. Desde então, o teto continua a subir a cada ano. Os números são ainda mais significativos ao levar em conta que a NBA adota um sistema de teto salarial flexível. Há diversas exceções em que o time pode ir além do teto salarial. A fim de controlar isso, a liga também impõe outro teto, o chamado *luxury tax*, que se ultrapassado, rende uma multa por cada dólar que excede o teto. Para a temporada de 2021-2022, o *luxury tax* da liga é de $136,606 milhões.

Os elencos da NBA não são grandes. Para a temporada 2021-2022, o máximo de jogadores ativos que cada equipe pode ter é 15 e o máximo de jogadores que podem estar sob contrato ao mesmo tempo é 17 (HELIN, 2021), número mais alto que as temporadas anteriores, devido à preocupação com a pandemia da COVID-19. Dividindo a folha salarial mínima pelo número máximo de jogadores possíveis sob contrato ao mesmo tempo, cada atleta ganharia cerca de 5,9 milhões de dólares por ano. Na prática, esta situação não ocorre, mas consegue elucidar o quão bem pagos os jogadores são.

2.2 Regressão Clássica × Quantílica

Considere o modelo de regressão linear clássica para o problema em questão como foi descrito em Kutner, Nachtsheim e Neter (2004):

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi} + \varepsilon_i, \tag{2.2.1}$$

onde Y_i representa a variável dependente, dita variável resposta. $X_{1i}, X_{2i}, \ldots, X_{pi}$ são variáveis independentes, ditas explicativas, que são constantes conhecidas. O modelo também apresenta $\beta_0, \beta_1, \ldots, \beta_p$ que são parâmetros desconhecidos e precisam ser estimados. Na regressão clássica, a estimação é dada por Mínimos Quadrados Ordinários (MQO). Por fim, há um termo de erro aleatório ε_i que é independente entre as observações $i = 1, \ldots, n$ e tem uma distribuição Normal com média zero e variância σ^2.

No modelo de regressão clássica, o enfoque está na função de regressão, dada pela esperança da variável resposta Y condicionada aos valores das variáveis explicativas $X_1, X_2, \ldots, X_p$ (DAVINO; FURNO; VISTOCCO, 2013). Isto é, dado o vetor de variáveis independentes $\mathbf{X}$, a função de regressão é dada por $E(Y|\mathbf{X})$.

Considerando a variável dependente Y a remuneração de atletas, surge um problema a ser considerado. Na regressão clássica, assume-se que o efeito marginal de uma variável explicativa não muda ao longo da distribuição condicional dos salários (VINCENT; EASTMAN, 2009). Em outras palavras, a função de regressão fica exclusivamente restrita a um local específico da distribuição condicional de Y (DAVINO; FURNO; VISTOCCO, 2013).

Na regressão clássica, a função de regressão é estimada pelo método de Mínimos Quadrados Ordinários, que estima os parâmetros na média condicional. É o estimador mais eficiente considerando o pressuposto de erros independentes com distribuição Normal, porém, muito suscetível a *outliers*, isto é, valores discrepantes (KEEFER, 2013). Em contrapartida, a regressão quantílica, introduzida por Koenker e Bassett (1978), é menos restritiva, pois permite que o efeito marginal das variáveis explicativas possam variar em diferentes pontos da distribuição condicional (VINCENT; EASTMAN, 2009). Desta forma, a Equação (2.2.1) pode ser reescrita da seguinte forma:

$$Y_i = \beta_0(\theta) + \beta_1(\theta)X_{1i} + \beta_2(\theta)X_{2i} + \ldots + \beta_p(\theta)X_{pi} + \varepsilon_{\theta i}, \tag{2.2.2}$$

onde os parâmetros e o erro aleatório são avaliados em um determinado θ-ésimo quantil, com $0 < \theta < 1$. Como os parâmetros $\beta_p(\theta)$ dependem do quantil θ, o efeito marginal das variáveis explicativas na variável resposta varia ao longo da distribuição condicional.

Para o estudo, dado que a variável dependente indica os salários dos jogadores, é natural pensar que os efeitos marginais das estatísticas dos jogadores varie de acordo com a distribuição dos salários. Jogadores mais bem remunerados têm a tendência de ter melhores estatísticas. Logo, utilizar a regressão quantílica à regressão clássica ajuda a melhor elucidar os fatores determinantes na remuneração dos atletas. Com isso, a função de regressão quantílica é dada por:

$$Q_\theta(Y_i|\mathbf{X}) = \beta_0(\theta) + \boldsymbol{\beta}(\theta)\mathbf{X}, \tag{2.2.3}$$

onde $\boldsymbol{\beta}$ é o vetor de parâmetros desconhecidos. Na Equação (2.2.2), Vincent e Eastman (2009) ressalta que nenhuma distribuição é assumida para o termo de erro $\varepsilon_{\theta i}$, apenas que satisfaz $Q_\theta(\varepsilon_{\theta i}|\mathbf{X}) = 0$. Portanto, a regressão quantílica é melhor que à clássica na

presença de distribuições não-normais.

Desta forma, o estudo acerca da remuneração dos atletas da NBA será feita utilizando a regressão quantílica a fim de explorar suas vantagens mais afundo. Com o intuito de comparar as duas técnicas, resultados da regressão clássica com estimativas por MQO também serão apresentadas.

2.3 Regressão Quantílica

Na seção anterior, uma breve discussão foi feita sobre a diferença entre a técnica de regressão clássica e regressão quantílica. A seguir, é feito um aprofundamento acerca do método quantílico, escolhido para o desenvolvimento do trabalho. Para tanto, algumas definições tiradas de Davino, Furno e Vistocco (2013) se fazem úteis e são apresentadas abaixo.

Dado Y uma variável aleatória qualquer, sua média é definida como o ponto c da sua distribuição que minimiza a soma quadrática dos desvios, isto é,

$$\mu = \underset{c}{\operatorname{argmin}}\, E(Y-c)^2. \tag{2.3.1}$$

Já a mediana de Y minimiza a soma absoluta dos desvios:

$$Me = \underset{c}{\operatorname{argmin}}\, E|Y-c|. \tag{2.3.2}$$

Através das observações amostrais, é possível obter os estimadores amostrais $\hat{\mu}$ e $\hat{Me}$ para os pontos c. Considere agora a função de distribuição acumulada (f.d.a.) de Y:

$$F_Y(y) = F(y) = P(Y \leq y), \tag{2.3.3}$$

o θ-ésimo quantil é o valor y tal que $P(Y \leq y) = \theta$. Desta forma, a função quantil é definida como o inverso da f.d.a. de Y:

$$Q_Y(\theta) = Q(\theta) = F_Y^{-1}(\theta) = \inf\{y : F(y) \geq \theta\}, \tag{2.3.4}$$

para $\theta \in [0,1]$. Os quantis também podem ser vistos como um problema de otimização. Desta maneira, o θ-ésimo quantil é dado por:

$$q_\theta = \underset{c}{\operatorname{argmin}}\, E[\rho_\theta(Y-c)], \tag{2.3.5}$$

em que $\rho_\theta(.)$ é a função de perda:

$$\rho_\theta(y) = [\theta - I(y < 0)]y, \tag{2.3.6}$$

onde $I(.)$ é a função indicadora. O objetivo é, portanto, encontrar um c que minimize a perda esperada. Na literatura, há diferentes soluções para este problema. Por exemplo, Hao e Naiman (2007) apresenta a Equação (2.3.6) como uma função de perda absoluta assimétrica:

$$\rho_\theta(y) = [(1-\theta)I(y \leq 0) + \theta I(y > 0)]|y|. \tag{2.3.7}$$

A seguir, é apresentado o desenvolvimento como visto em Koenker (2005). Retornando à Equação (2.3.5), o problema é minimizar a seguinte função:

$$E[\rho_\theta(Y - c)] = (\theta - 1)\int_{-\infty}^{c}(y - c)\,dF(y) + \theta\int_{c}^{\infty}(y - c)\,dF(y). \tag{2.3.8}$$

Igualando a zero e diferenciando em relação a c, tem-se que:

$$0 = (1-\theta)\int_{-\infty}^{c} dF(y) - \theta\int_{c}^{\infty} dF(y) = F(c) - \theta. \tag{2.3.9}$$

Por F ser uma função monótona, qualquer elemento de $y : F(y) = \theta$ minimizará a Equação (2.3.5). Segundo Gilchrist (2000), se $F(.)$ é estritamente crescente e contínua, $F^{-1}(\theta)$ é o único número real y tal que $F(y) = \theta$. Caso contrário, o resultado será um intervalo de θ-ésimos quantis, dos quais o menor elemento deverá ser escolhido (KOENKER, 2005).

Considerando a função de distribuição empírica $F_n(y) = n^{-1}\sum_{i=1}^{n} I(Y_i \leq y)$ na resolução da Equação (2.3.8), o θ-ésimo quantil amostral é obtido. Este é um resultado muito útil, pois transforma o problema de encontrar o quantil amostral, que intuitivamente seria um problema de ordenação, em um problema de otimização.

Agora considerando um modelo de regressão em que Y é a variável resposta e $\mathbf{X}$ o vetor das variáveis explicativas, é possível expandir a ideia da média incondicional como valor que minimiza a Equação (2.3.1) para a estimação da função da média condicional:

$$\hat{\mu}(\mathbf{x}_i, \boldsymbol{\beta}) = E(Y|\mathbf{X} = \mathbf{x}_i) = \underset{\mu}{\operatorname{argmin}}\, E[Y - \mu(\mathbf{x}_i, \boldsymbol{\beta})]^2, \tag{2.3.10}$$

onde $E(Y|\mathbf{X} = \mathbf{x}_i)$ é a função da média condicional. Ao se considerar a função da média

linear $\mu(\mathbf{x}_i, \boldsymbol{\beta}) = \mathbf{x}_i^\top \boldsymbol{\beta}$, a equação anterior se transforma em:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, E[Y - \mathbf{x}_i^\top \boldsymbol{\beta}]^2, \tag{2.3.11}$$

resultando no modelo de regressão linear de mínimos quadrados. O problema de minimização é um problema numérico de álgebra linear.

A mesma abordagem pode ser feita para a Equação (2.3.2) da mediana e para o θ-ésimo quantil, como visto em Equação (2.3.5), resultando em:

$$\hat{q}_Y(\theta, \mathbf{X}) = \underset{Q_Y(\theta, \mathbf{X})}{\operatorname{argmin}}\, E[\rho_\theta(Y - Q_Y(\theta, \mathbf{X}))], \tag{2.3.12}$$

em que $Q_Y(\theta, \mathbf{X}) = Q_\theta(Y|\mathbf{X} = \mathbf{x})$ é a função quantil condicional genérica. Para o caso de modelo linear:

$$\hat{\beta}(\theta) = \underset{\beta}{\operatorname{argmin}}\, E[\rho_\theta(Y - \mathbf{X}\beta)], \tag{2.3.13}$$

onde agora (θ) denota que os parâmetros e seus respectivos estimadores estão avaliados em um θ-ésimo quantil específico.

Como discutido na seção anterior, a regressão quantílica é uma extensão da abordagem clássica permitindo a estimação de quantis condicionais da distribuição da variável resposta Y em função das variáveis explicativas $\mathbf{X}$. As estimativas dos parâmetros têm a mesma interpretação que o modelo clássico.

Quanto à parte inferencial, é importante destacar que há diferentes métodos para conduzir a inferência dos parâmetros da regressão quantílica. Como dito, por não fazer suposições sobre a distribuição do termo de erro, não há uma teoria como existe na regressão clássica, ao assumir erros independentes e identicamente distribuídos Normais.

Para o cálculo do erro padrão dos estimadores, por exemplo, é possível assumir erros independentes e identicamente distribuídos, ocorrendo uma generalização do caso do modelo de regressão clássico. Outra alternativa é considerar erros não independentemente distribuídos. Neste caso, presume-se linearidade local nas funções quantílicas condicionais e o seu comportamento assintótico envolve o estimador "sanduíche" de matriz de covariâncias, apresentado por Huber (1967).

Estas abordagens estão contempladas na biblioteca *quantreg* do R, escolhido para a realização das análises. O pacote inclui também a abordagem pela estimativa de Kernel para cálculo do erro padrão utilizando o estimador "sanduíche" proposto em Powell (1991) e o método de estimação por *bootstrap*. Este último, além de não fazer suposições sobre a distribuição dos erros, oferece outras alternativas de como conduzir a estimação, entre

elas o método tradicional de par "xy" e o como descrito em Parzen, Wei e Ying (1994).

Com os erros padrões dos estimadores obtidos é possível conduzir a inferência de uma maneira mais tradicional com a possibilidade de realizar testes de hipóteses e construção de intervalos de confiança. Para o trabalho, o método de *bootstrap* foi o escolhido com enfoque na realização do teste de significância para coeficientes, onde é necessário o cálculo do erro padrão dos estimadores para cálculo da estatística do teste.

Outro caminho de se conduzir a inferência para a regressão quantílica é a inferência baseada em ranks, explicitada com mais detalhes na seção 3.5 de Koenker (2005). Este método de inferência é baseada na abordagem de *scores* de Rao (1948) e acaba por ser uma classe de testes de rank generalizados para modelos de regressão lineares. Ao longo do trabalho, o cálculo de intervalo de confiança por método de inversão de rank foi explorado, método também disponibilizado no pacote *quantreg*.

2.4 Banco de Dados

O banco de dados utilizado para o trabalho pode ser dividido em duas fontes principais. A primeira corresponde aos salários dos jogadores da NBA durante a temporada 2020-2021 e foi retirada do site HoopsHype (2020). O site é um segmento da *USA Today Sports* que por sua vez mantém uma base de dados online para os salários dos jogadores e a folha salarial dos times das quatro principais ligas esportivas americanas.

A segunda fonte utilizada foi do site Basketball-Reference (2021) que faz parte da *Sports Reference*, um conjunto de sites que contemplam estatísticas das quatro principais ligas esportivas profissionais dos Estados Unidos, futebol americano universitário, basquete universitário e uma base que engloba dados de futebol do mundo todo, abrangendo mais de 100 competições. Esta fonte foi utilizada para obter as estatísticas dos jogadores e dos times da temporada 2020-2021 da NBA que vão corresponder às variáveis explicativas do modelo apresentado na seção anterior.

Quanto à base de dados que reflete a remuneração dos atletas, há 578 observações com nome do jogador, salário anual para a temporada em questão e o salário ajustado de acordo com a inflação, que é irrelevante, dado que 2020-21 é a temporada mais recente na data do estudo e não houve ainda reajuste neste sentido, sendo assim, as observações relativas aos salários são iguais.

Para a base de estatísticas, é possível dividir em cinco partições. A primeira corresponde às estatísticas totais dos jogadores. Nesta base, há as principais estatísticas

observadas no basquete: jogos, jogos como titular, pontuação (incluindo número de arremessos, aproveitamento, cestas de três pontos, lances livres, etc.), rebotes (ofensivos e defensivos), assistências, roubos de bola, bloqueios, perdas de posse e faltas.

A base também tem o nome do jogador, a sua posição, sua idade e o time em qual jogou. No total, a base tem 705 observações. Isto porque, se um atleta jogou por mais que um time, constam as estatísticas por cada time e ainda as totais, isto é, a soma da temporada toda. Desta forma, um jogador que performou por duas equipes, tem três observações na base de dados ao invés de uma.

Há 540 jogadores diferentes na base de dados, que é discrepante em relação aos 578 observados na base de salários. Isto pode ser explicado pelo que foi discutido anteriormente no final da seção 3.1. A liga expandiu o número de jogadores que podem estar sob contrato ao mesmo tempo por precaução com a questão da pandemia da COVID-19. Desta forma, jogadores que não jogaram na temporada e, portanto, não têm estatísticas, estão presentes na base de salários.

A base em questão corresponde aos totais observados ao longo da temporada. A segunda partição corresponde às mesmas estatísticas, mas com seus números por jogo, pois naturalmente há jogadores que estão presentes em quadra mais vezes que outros.

A terceira partição corresponde a estatísticas ditas avançadas. As duas primeiras bases podem ser registradas apenas observando o jogo. Esta terceira tem medidas que precisam de algum cálculo para ser obtidas. Na base são 22 dessas estatísticas como por exemplo, a *Player Efficiency Rating* (PER), uma medida da produção por minuto de determinado jogador de tal forma que a média da liga seja 15. Ou a *True Shooting Percentage* (TS%) que é uma medida de eficiência de arremessos do jogador, mas levando em conta arremessos de dois e três pontos, assim como lances livres.

Todas as três bases referentes a jogadores são consistentes ao ter 705 observações de 540 jogadores diferentes. Porém, ao comparar os jogadores da base de estatísticas com a base de salários, foram identificados três jogadores que não estavam na base de salários. São eles: Devin Cannady, Frank Mason III e Robert Franks. Foi possível recuperar seus números através da própria Basketball-Reference (2021), que também faz o trabalho de registro de salários. Desta forma, estas observações foram adicionadas depois e todos os jogadores das bases de estatísticas agora constavam na base de salários.

As últimas duas partições são referentes a estatísticas de time, pois a depender da equipe que defende, os jogadores são suscetíveis a terem melhor ou pior desempenho. Um jogador que atua por uma equipe com uma média muito alta de pontos, por exemplo,

tende a pontuar mais também.

Para as estatísticas de time, é possível traçar um paralelo para as bases de jogadores. Há uma base de estatísticas por jogo observáveis e outra base com estatísticas avançadas que tiveram algum cálculo envolvido e tem alguma interpretação. Exemplos de estatísticas avançadas para times incluem a *Strength of Schedule* (SOS) uma medida da força que os times enfrentados durante a temporada tiveram, isto é, maior SOS, maior dificuldade o time teve ao longo da temporada.

Também podem ser citados o *Offensive Rating*, uma estimativa de pontos feitos por 100 posses de bola, assim como o *Defensive Rating*, estimativa de pontos concedidos por 100 posses. Há um total de 10 dessas estatísticas. Ambas as bases de time naturalmente têm 30 observações, uma para cada franquia da NBA.

Assim, serão escolhidas algumas dessas estatísticas para compor as variáveis independentes explicativas presentes na Equação (2.2.2). É possível também que haja uma divisão das análises por posição, uma vez que armadores e pivôs, por exemplo, terão estatísticas muito discrepantes, pois por questão de altura, pivôs agarram mais rebotes do que armadores e isso deve ser considerado.

2.4.1 Correlações

Como descrito anteriormente na subseção de base de dados, há cinco partições para o banco de dados de estatísticas. Após tratamento dos dados, a junção das bases e também da base de salários, assim como a sua transformação em log-salários, totaliza uma base de 626 observações com 106 variáveis. Este número é muito elevado e, portanto, será reduzido.

Para tanto, foram calculadas as correlações de Pearson entre as 102 variáveis de interesse utilizando as observações que têm todas as informações disponíveis. Isto é, jogadores que não apresentam problema de dados faltantes, casos que serão discutidos na próxima subseção. As quatro variáveis não usadas foram a de nome do jogador, time e posição, todas nominais, e a de salários, uma vez que foi incluso a variável já transformada, a de log-salários. Isto porque, para duas variáveis com uma correlação muito grande, é possível deixar uma de fora do modelo a favor da outra.

Após calculadas as correlações, há dois casos de correlação perfeita negativa. Ambas estão presentes nas bases de estatísticas de times e são as correlações entre:

- Número de Vitórias $\times$ Número de Derrotas;

- Número de Vitórias "Pitagóricas" × Número de Derrotas "Pitagóricas".

A definição de uma vitória (ou derrota) pitagórica é, por Dewan, Zminda e STATS (1993), o número de vitórias (ou derrotas) esperadas baseado em pontos marcados e cedidos através da seguinte fórmula:

$$\%\ de\ Vitória = \frac{Pontos\ Marcados^{13,91}}{Pontos\ Marcados^{13,91} + Pontos\ Cedidos^{13,91}} \tag{2.4.1}$$

Após o cálculo, o número de vitórias é obtido ao multiplicar a porcentagem por 72, número de jogos na temporada de 2020-2021. O número de derrotas será, portanto, $72 - Número\ de\ Vitórias$, uma vez que não há empates no basquete. Desta forma, é fácil entender o porquê da correlação perfeita entre os números pitagóricos e os números de vitórias/derrotas normais. Assim, duas variáveis não precisam estar no modelo, apenas as que indicam o número de vitórias ou as duas que indicam o número de derrotas são suficientes.

Foi identificado também outras variáveis com correlações altas. A seguir estão listados os pares de variáveis com correlação maior que 0,9 ou menor que -0,9, bem como o número de suas correlações entre parênteses:

Estatísticas por Jogador

1. Minutos Totais Jogados × Arremessos Totais Convertidos (0,910141)
2. Minutos Totais Jogados × Arremessos Totais Tentados (0,917711)
3. Arremessos Totais Convertidos × Arremessos Totais Tentados (0,985754)
4. Arremessos Totais de 3 Pontos Convertidos × Arremessos de 3 Pontos Tentados (0,990857)
5. Arremessos Totais de 2 Pontos Convertidos × Arremessos Totais Convertidos (0,947403)
6. Arremessos Totais de 2 Pontos Tentados × Arremessos Totais Convertidos (0,956379)
7. Arremessos Totais de 2 Pontos Tentados × Arremessos Totais Tentados (0,928302)
8. Arremessos Totais de 2 Pontos Tentados × Arremessos Totais de 2 Pontos Convertidos (0,988868)
9. Arremessos Livres Totais Tentados × Arremessos Totais de 2 Pontos Convertidos (0,905054)

10. Arremessos Livres Totais Tentados × Arremessos Totais de 2 Pontos Tentados (0,910536)
11. Arremessos Livres Totais Tentados × Arremessos Livres Totais Convertidos (0,989335)
12. Rebotes Totais × Rebotes Totais Defensivos (0,985677)
13. Perdas de Posse Totais × Arremessos Totais Convertidos (0,90385)
14. Perdas de Posse Totais × Arremessos Totais Tentados (0,904274)
15. Perdas de Posse Totais × Assistências Totais (0,903211)
16. Pontos Totais × Minutos Totais Jogados (0,906621)
17. Pontos Totais × Arremessos Totais Convertidos (0,994451)
18. Pontos Totais × Arremessos Totais Tentados (0,989802)
19. Pontos Totais × Arremessos Totais de 2 Pontos Convertidos (0,922292)
20. Pontos Totais × Arremessos Totais de 2 Pontos Tentados (0,93753)
21. Pontos Totais × Arremessos Livres Totais Convertidos (0,910198)
22. Pontos Totais × Arremessos Livres Totais Tentados (0,901912)
23. Pontos Totais × Perdas de Posse Totais (0,906825)
24. Arremessos Tentados por Jogo × Arremessos Convertidos por Jogo (0,975435)
25. Arremessos de 3 Pontos Tentados por Jogo × Arremessos de 3 Pontos Convertidos por Jogo (0,979178)
26. Arremessos de 2 Pontos Convertidos por Jogo × Arremessos Convertidos por Jogo (0,931053)
27. Arremessos de 2 Pontos Tentados por Jogo × Arremessos Convertidos por Jogo (0,935266)
28. Arremessos de 2 Pontos Tentados por Jogo × Arremessos Tentados por Jogo (0,90255)
29. Arremessos de 2 Pontos Tentados por Jogo × Arremessos de 2 Pontos Convertidos por Jogo (0,980076)

30. Arremessos Livres Tentados por Jogo × Arremessos Livres Convertidos por Jogo (0,98356)
31. Rebotes por Jogo × Rebotes Defensivos por Jogo (0,975719)
32. Pontos por Jogo × Arremessos Convertidos por Jogo (0,990328)
33. Pontos por Jogo × Arremessos Tentados por Jogo (0,980314)
34. Pontos por Jogo × Arremessos de 2 Pontos Tentados por Jogo (0,906043)
35. *True Shooting Percentage* × *Effective Field Goal Percentage* (0,960019)
36. Porcentagem de Rebotes Agarrados × Porcentagem de Rebotes Defensivos Agarrados (0,94654)
37. Porcentagem de Assistências × Assistências por Jogo (0,903312)
38. Contribuição de Vitórias × Contribuição de Vitórias via Ataque (0,957635)
39. *Box Plus/Minus* × *Offensive Box Plus/Minus* (0,929571)

Estatísticas por Time

40. Arremessos de 3 Pontos Tentados × Arremessos de 3 Pontos Convertidos (0,923078)
41. Lances Livres Tentados × Lances Livres Convertidos (0,901023)
42. Pontos Marcados × Arremessos Convertidos (0,902714)
43. Vitórias Pitagóricas × Vitórias (0,952204)
44. Vitórias Pitagóricas × Derrotas (-0,9522)
45. Derrotas Pitagóricas × Vitórias (-0,9522)
46. Derrotas Pitagóricas × Derrotas (0,952204)
47. Margem de Vitória × Vitórias (0,946073)
48. Margem de Vitória × Derrotas (-0,94607)
49. Margem de Vitória × Vitórias Pitagóricas (0,999202)
50. Margem de Vitória × Derrotas Pitagóricas (-0,9992)

51. Dificuldade de Jogos × Vitórias Pitagóricas (-0,91719)
52. Dificuldade de Jogos × Derrotas Pitagóricas (0,917195)
53. Dificuldade de Jogos × Margem de Vitória (-0,92091)
54. Sistema de Avaliação Simples × Vitórias (0,946682)
55. Sistema de Avaliação Simples × Derrotas (-0,94668)
56. Sistema de Avaliação Simples × Vitórias Pitagóricas (0,99919)
57. Sistema de Avaliação Simples × Derrotas Pitagóricas (-0,99919)
58. Sistema de Avaliação Simples × Margem de Vitória (0,999848)
59. Sistema de Avaliação Simples × Dificuldade de Jogos (-0,91398)
60. Estimativa Ofensiva × Vitórias Pitagóricas (0,900034)
61. Estimativa Ofensiva × Derrotas Pitagóricas (-0,90003)
62. Estimativa Ofensiva × Margem de Vitória (0,900341)
63. Estimativa Ofensiva × Sistema de Avaliação Simples (0,901629)
64. Estimativa de Diferença de Pontos × Vitórias (0,946369)
65. Estimativa de Diferença de Pontos × Derrotas (-0,94637)
66. Estimativa de Diferença de Pontos × Vitórias Pitagóricas (0,999379)
67. Estimativa de Diferença de Pontos × Derrotas Pitagóricas (-0,99938)
68. Estimativa de Diferença de Pontos × Margem de Vitória (0,999882)
69. Estimativa de Diferença de Pontos × Dificuldade de Jogos (-0,91801)
70. Estimativa de Diferença de Pontos × Sistema de Avaliação Simples (0,999853)
71. Estimativa de Diferença de Pontos × Estimativa Ofensiva (0,901567)
72. Taxa de Tentativas de Lances Livres × Lances Livres Tentados (0,970654)
73. Taxa de Tentativas de Arremessos de 3 Pontos × Arremessos de 3 Pontos Convertidos (0,9029811)

74. Taxa de Tentativas de Arremessos de 3 Pontos × Arremessos de 3 Pontos Tentados (0,9794014)

75. Taxa de Tentativas de Arremessos de 3 Pontos × Arremessos de 2 Pontos Tentados (-0,957667)

76. *True Shooting Percentage* × Estimativa Ofensiva (0,91812)

As duas primeiras correlações são fáceis de ser explicadas. Quanto mais tempo em quadra, maior é o número de arremessos. A correlação 3 também segue a mesma linha de raciocínio. Quanto mais arremessos são tentados, mais são convertidos. Portanto, destas três estatísticas, minutos totais jogados, arremessos totais convertidos e arremessos totais tentados, apenas uma dessas variáveis bastaria estar no modelo. A variável escolhida será a de arremessos totais convertidos.

Da mesma maneira, particionando as maneiras de se pontuar em um jogo de basquete, temos as correlações 4, 8 e 11. Quanto mais arremessos de 2 pontos, de 3 pontos e lances livres são tentados, mais são convertidos. Desta forma, não há necessidade de se manter ambos os números de tentativas e conversão de cada categoria no modelo. Como no último caso, só as variáveis de conversão serão consideradas.

A correlação 5 é explicada porque, predominantemente, arremessos no basquete são cestas de 2 pontos. Assim, quanto maior o número de arremessos convertidos, maior é o número de arremessos de 2 pontos convertidos. Como não houve correlação muito alta entre arremessos de 2 pontos e de 3 pontos, o número de arremessos convertidos total será eliminado, uma vez que temos variáveis que indicam o número de cestas de 2 e 3 pontos e juntas somam o número total de cestas.

Uma vez que já foi decidido que as variáveis de arremessos tentados serão deixadas de fora a favor das variáveis de arremessos convertidos, as análises para as correlações 6, 7, 9, 10 e 14 serão omitidas. A correlação 12 se explica como a 5. Os rebotes são, em sua maioria, defensivos. Portanto, a variável de rebotes totais será omitida, pois na base também constam os rebotes ofensivos, que somados aos defensivos resultam no total.

Para 13 e 15, a explicação é por volume de jogo. Jogadores que participam mais do jogo, como um grande número de cestas (13) ou grande número de assistências (15), tendem a errar mais, isto é, a perder a posse de bola mais vezes. Como na base também constam os números de perda de posse por jogo, a variável que indica as perdas de posse totais será desconsiderada.

As correlações entre 16 e 23 todas se referem a pontos totais e alguma outra

variável. O banco de dados contém a subdivisão da pontuação entre cestas de 2 e 3 pontos, assim como lances livres. Manter a variável de pontos totais é, portanto, desnecessária, uma vez que seu efeito pode ser explicado por outras variáveis.

Para as correlações de 24 a 34, as interpretações são similares às já apresentadas nos parágrafos anteriores, mas agora as explicações se estendem para os números por jogo. Então, da mesma maneira, os números de tentativas total e para cestas de 2 pontos, 3 pontos e lances livres são desconsiderados a favor de suas conversões. O número de rebotes totais por jogo também sai do modelo, bem como o número de pontos por jogo.

As correlações de 35 a 39 correspondem a estatísticas avançadas, aquelas que precisam de algum cálculo para serem obtidas, diferentes das anteriores, que são meramente observações de ocorrências dentro do jogo. Para isto se fazem necessárias algumas definições, a começar pela correlação 35. O *Effective Field Goal Percentage* é a porcentagem de arremessos convertidos ajustada, em que leva em conta que um arremesso de 3 pontos vale mais do que um de 2 pontos. Já a *True Shooting Percentage* é uma medida de eficiência de arremesso que leva em conta cestas de 2 e 3 pontos, além de lances livres. Como ambas as estatísticas levam em conta a diferença de pesos em arremessos de 2 e 3 pontos, é fácil ver de onde surge a correlação. Por também levar em conta lances livres, a *True Shooting Percentage* será a variável mantida.

Para a correlação 36, a porcentagem de rebotes agarrados é uma estimativa dos rebotes agarrados dos rebotes disponíveis enquanto o jogador estava em quadra. Novamente é identificada uma relação entre o total e os rebotes defensivos. Seguindo as análises anteriores, o número total é desconsiderado a favor do registro de rebotes defensivos. Da mesma forma, a porcentagem de assistências é uma estimativa dos arremessos em que o jogador deu assistência sobre o número de arremessos totais enquanto o jogador estava em quadra. É natural pensar que este número está relacionado com o número de assistências por jogo, como pode ser visto na correlação 37. Neste caso, o número de assistências será a variável mantida.

A contribuição de vitórias é o número estimado de vitórias adicionadas pelo jogador. Esta estatística pode ser dividida em contribuição via ataque e defesa, isto é, o número de vitórias adicionadas devido à performance ofensiva ou defensiva. Para a correlação 38, a contribuição via ataque está correlacionada com a contribuição total porque é mais fácil um jogador impactar o jogo pelo ataque do que pela defesa. O número total pode ser desconsiderado a favor do número da contribuição pelo ataque.

Em um time médio, comparado ao jogador médio da liga, o *Box Plus/Minus* é

uma estimativa de pontos que o jogador contribuiu a mais que um jogador médio em 100 posses. Ou seja, um valor negativo desta estatística indica que o jogador está performando pior do que o jogador médio da NBA e um valor positivo, o contrário. O *Offensive Box Plus/Minus* é a estimativa ofensiva. E, novamente, como é mais fácil impactar o jogo ofensivamente, maiores valores de *Offensive Box Plus/Minus* implicam em maiores *Box Plus/Minus*. Assim, o *Box Plus/Minus* será retirado.

A partir da correlação 40, as variáveis correspondem a estatísticas por time. Até a 42, são estatísticas simples e podem ser interpretadas de maneira similar às medidas por jogador. Assim, em 40 os arremessos de 3 pontos tentados e em 41 os lances livres tentados são desconsiderados. Para 42, a variável que será mantida será a de pontos marcados, eliminando a variável de arremessos convertidos.

Depois da correlação 43, todas as estatísticas são ditas avançadas. Nas correlações 43 a 46 é visto que as vitórias e derrotas registradas têm uma alta correlação com as vitórias e derrotas pitagóricas. Assim, como também foi visto na parte de correlações perfeitas, é possível manter apenas uma dessas variáveis. Dentre as quatro variáveis, apenas o número de vitórias será mantido.

A margem de vitória é a média de saldo de pontos que o time obteve em seus jogos. Uma maior margem indica um maior número de vitórias como visto em 47 e, sendo assim, a variável de margem de vitória pode ser desconsiderada a favor da variável que indica o número de vitórias. Para as correlações que seguem e que têm alguma variável que já foi eliminada, a análise será omitida.

Nas correlações 51 a 53, a dificuldade de jogos leva em conta os oponentes que um time enfrentou ao longo da temporada e o quão difíceis foram. Como em 53, esta variável se mostra correlacionada com a margem de vitória, que por sua vez, foi eliminada por estar altamente correlacionada com o número de vitórias, a dificuldade de jogos será eliminada.

Entre 54 e 59, o sistema de avaliação simples é um medidor que leva em conta a margem de vitória e a dificuldade de jogos. Por sua alta correlação com o número de vitórias, também será desconsiderado. Para as correlações 60 a 63, a estimativa ofensiva traz uma estimativa de pontos marcados por 100 posses de bola. Como todas as variáveis que são apresentadas já foram eliminadas, a estimativa ofensiva será mantida.

Para 64 a 71, a estimativa de diferença de pontos é uma diferença da margem de pontos por 100 posses de bola. Uma maior diferença, indica um maior número de vitórias. Além disso, como a estimativa ofensiva foi mantida, não há necessidade de também manter

a estimativa de diferença total, sendo eliminada.

Para 72 a 75, a taxa de tentativas de lances livres indicam o número de lances livres tentados divididos pelo número de arremessos tentados e a taxa de tentativas de arremessos de 3 pontos indicam a porcentagem de arremessos de 3 pontos tentados. É fácil entender a correlação entre a taxa de tentativas e o número total de tentativas e, portanto, apenas as tentativas serão mantidas. Mas, como as tentativas foram desconsideradas a favor dos números de arremessos convertidos, apenas estes últimos serão mantidos.

Por fim, na correlação 76, como a *True Shooting Percentage* é uma medida de eficiência que leva em conta o peso dos lances livres e arremessos de 2 e 3 pontos, maiores valores de *True Shooting Percentage* resultam em uma maior estimativa ofensiva. Assim, a estimativa ofensiva será a variável escolhida para ser mantida.

2.4.2 Valores *Missing* (NA)

A base com todas as estatísticas conta com observações que têm valores *missing* (NA). Ao todo, das 626 observações, 55 apresentam pelo menos um valor faltante em uma das 106 variáveis. Essas observações correspondem a 48 jogadores distintos e os valores faltantes estão divididos nas seguintes variáveis:

Tabela 1: Descrição das variáveis que apresentam NA

Variável	**Descrição**
FG..x	Aproveitamento de arremessos
X3P..x	Aproveitamento de arremessos de 3 pontos
X2P..x	Aproveitamento de arremessos de 2 pontos
eFG.	*Effective Field Goal Percentage:* Esta estatística ajusta pelo fato de que um arremesso de 3 pontos vale um ponto a mais do que um arremesso de 2 pontos
FT..x	Aproveitamento de lances livres
TS..x	Uma medida de eficiência de arremesso que leva em conta arremessos de 2 e 3 pontos e lances livres
X3PAr.x	Porcentagem de tentativas de arremessos de 3 pontos
FTr.x	Número de lances livres por tentativa de arremesso
TOV.	Uma estimativa do número de posses perdidas em 100 jogadas

Após a análise de correlações feita na Subsubseção 2.4.1 e redução para 75 variáveis, a base final continua a apresentar as mesmas 55 observações que apresentam pelo menos um NA em suas variáveis. Porém, das variáveis apresentadas na Tabela 1, a

única variável que foi retirada por ter equivalência com outra foi a de *Effective Field Goal Percentage* (eFG.), justamente por ser explicada pela *True Shooting Percentage* (TS..x), mas que também apresenta o problema de informação faltante.

Ao investigar mais afundo estas variáveis, nota-se que em sua maioria correspondem a estatísticas expressas em porcentagens. O seu cálculo se dá pela divisão de eventos de interesse por ocasiões totais. Desta forma, se um jogador não registrou o evento de interesse na temporada toda a estatística não é contabilizada. Sem o quantitativo total, a razão fica impossibilitada de ser calculada, pois seria como dividir por zero. E isto acaba resultando em um valor faltante.

As duas variáveis com maior número de NAs foram aproveitamento de arremessos de 3 pontos e aproveitamento de lances livres. Ou seja, jogadores que não tentaram nenhum arremesso de 3 pontos e nenhum lance livre durante a temporada. Então, é natural levantar a hipótese de que estes casos foram de jogadores que atuaram pouco durante a temporada.

Entre as 55 observações faltantes, ao verificar a variável G.x que corresponde às aparições em jogos na temporada, independente de ser como titular ou reserva, tem-se que:

Tabela 2: Observações com NA por número de jogos

Jogos	**Observações**
1	6
2	5
3	6
4	5
5 ou 6	5
7 a 10	6
11 a 15	8
16 a 25	8
>25	6
Total	**55**

Então, das 55 observações identificadas com algum valor faltante, 27 correspondem a jogadores com seis jogos ou menos. O baixo número de jogos acarreta também outro problema que é a distorção de algumas estatísticas, o que pode prejudicar a análise de regressão. Um exemplo é a estatística de *Player Efficiency Rating* (PER), uma medida de produção do jogador por minuto padronizada tal que a média da liga é 15.

Um dos jogadores que atuaram apenas uma vez foi Udonis Haslem que atua pelo Miami Heat. O PER registrado foi 54,6 e foi o mais alto da base. O segundo maior foi de Nikola Jokic do Denver Nuggets que esteve presente em todos os 72 jogos da temporada e teve PER de 31,3. É uma diferença muito grande para uma estatística que por si só é padronizada para ter média 15.

Com o intuito de evitar casos de estatísticas muito discrepantes devido a poucas atuações na temporada, uma solução é estabelecer um limite mínimo de jogos para inclusão de jogadores na base de dados do estudo. É importante ressaltar que, ao optar por este limite, outros jogadores que, apesar de atuarem pouco, registraram todas as estatísticas e não tem dados *missing* serão retirados da base também.

Para ajudar na tomada de decisão, a tabela abaixo foi feita levando em conta diferentes números de jogos mínimos a serem impostos e o tamanho da base decorrente destes mínimos, bem como o percentual da redução da base e o número de NAs.

Tabela 3: Informações da base de dados conforme mínimo de jogos

Mín. de Jogos	**Tam. da Base**	**Redução**	**NAs**
1	626	1,28%	55
2	618	2,72%	49
3	609	4,31%	44
4	599	6,39%	38
5	586	7,99%	33
6	576	9,11%	29
7	569	10,54%	28
8	560	12,14%	24
9	550	13,26%	23
10	543	14,38%	22

Após análise, foi decidido que o limite mínimo de oito jogos cumpriria os objetivos propostos para o problema. Desta forma, de 626 observações, a base passa a ter 560 registros. Porém, o problema de dados faltantes continua a persistir. Mas se antes as 55 observações com NA correspondiam a 8,79% da base, agora, estando presente em 24 das 560 observações, os dados *missing* compõem aproximadamente 4,29% da base.

Ao investigar mais afundo estas observações, nota-se que o problema de dados faltantes agora restringe-se a apenas duas variáveis: aproveitamento de cestas de 3 pontos e aproveitamento de lances livres. Das 24 observações, seis delas são atribuídas ao acerto do lance livre. Estes são jogadores que não atuaram muito como um todo e, portanto,

não tiveram a oportunidade de tentar um lance livre. Dentre os seis, nenhum atuou em um jogo como titular, todos foram saindo do banco de reservas.

Nenhuma das 24 observações têm NA nas duas variáveis ao mesmo tempo. Assim, as outras 18 observações restantes são relacionadas ao aproveitamento em arremessos de 3 pontos que podem ser explicadas pela característica dos jogadores envolvidos. Dentre os 18 jogadores, 13 são pivôs e cinco são ala-pivôs.

No basquete, as duas posições historicamente têm como característica jogadores mais altos, fortes e que jogam mais próximos à cesta. E, apesar da NBA estar constantemente evoluindo e cada vez mais o arremesso de 3 pontos estar sendo praticado independentemente da posição, ainda há jogadores que apenas não possuem essa característica. É o caso, por exemplo, de Jakob Poeltl e Clint Capela que jogaram em 69 e 63 partidas, respectivamente, e não arremessaram nenhuma vez para 3 pontos.

Com o intuito de tentar eliminar os casos de dados faltantes na base, uma maneira de contornar o problema é reintroduzir algumas variáveis na base. Ao fazer a análise de correlações, as variáveis referentes às tentativas de cestas de 3 e tentativas de lances livres, correspondendo às correlações 4 e 11, foram eliminadas por terem correlações muito altas. Porém, estas duas variáveis em conjunto às variáveis que indicam o número de conversões tanto de arremessos de 3 pontos, como de lances livres, trazem toda a informação contida nas variáveis X3P..x e FT..x, que têm NAs.

Desta forma, ao reintroduzir as variáveis X3PA.x, que indica as tentativas de arremessos de 3 pontos, e FTA.x, que indica o número de tentativas de lances livres, a informação das conversões estará presente de uma maneira indireta. Para obter os dados referentes aos aproveitamentos, basta dividir X3P.x por X3PA.x para arremessos de 3 pontos e dividir FT.x por FTA.x para lances livres.

Assim, a base que será trabalhada ao longo do restante do trabalho continua com 75 variáveis, mas agora com 560 observações devido ao limite mínimo de oito jogos para cada jogador. Apesar de algumas observações sem dados faltantes saírem da base, esta foi a solução adotada para viabilizar o ajuste do modelo de regressão. E com isso, o problema de dados faltantes é resolvido.

2.5 Redução de Dimensionalidade

A base com todas as estatísticas tinha 106 variáveis ao todo. E mesmo com a análise de correlações reduzindo este número em cerca de 25%, a base continua com o

número elevado de 75 variáveis. Este número dificulta o enfoque do trabalho, que busca ajustar um modelo de regressão. Assim, para a redução de dimensionalidade algumas soluções foram propostas e serão apresentadas ao longo da seção a seguir.

2.5.1 Análise de Componentes Principais

Para a redução de dimensionalidade, uma das soluções testadas foi a da técnica de *Principal Component Analysis* (PCA), ou análise de componentes principais. O PCA é uma técnica multivariada que busca condensar as variáveis preditoras em uma combinação não correlacionada delas (componentes) de tal forma a capturar o máximo de informação contidas nestas variáveis.

Por meio do *software* R, a análise de componentes principais foi feita a partir da função *prcomp* e retornou as informações que podem ser vistas na tabela abaixo: o desvio padrão da componente, a proporção da variância explicada pela componente e a proporção acumulada da variância conforme mais componentes vão sendo levadas em conta.

Tabela 4: Importância das Componentes Principais

Componente	**Desvio Padrão**	**Prop. da Variância**	**Prop. Acumulada**
1	4,438	28,13%	28,13%
2	2,734	10,68%	38,81%
3	2,644	9,99%	48,79%
4	2,029	5,88%	54,68%
5	1,737	4,31%	58,99%
6	1,727	4,26%	63,25%
7	1,629	3,79%	67,04%
8	1,576	3,55%	70,59%
9	1,487	3,16%	73,75%
10	1,410	2,84%	76,59%

A primeira componente consegue condensar 28,13% da informação contida nas variáveis explicativas. Em seguida, a segunda e terceira componentes explicam por volta de 10% cada uma. Analisando a proporção acumulada, 70% da variância consegue ser explicada por oito componentes, o que seria um número válido para o problema comparando com o número anterior de 75 variáveis preditoras.

Porém, o problema que surge ao utilizar esta solução, é ao atribuir significado às componentes. Embora o PCA seja uma técnica que consiga condensar uma porcentagem

da informação dentro de uma combinação de variáveis, se esse conjunto de variáveis não têm significado, ele não contribui para a análise. É o que acontece neste caso, pois perde a capacidade de interpretação do modelo e a análise do problema fica inviabilizada.

Ao tentar o PCA no problema, esperava-se que as componentes apresentassem um significado mais claro, como por exemplo, variáveis ofensivas ou defensivas, que condensassem algum aspecto do jogo. Mas ao analisar a primeira componente, tem-se uma combinação de mais de 30 variáveis, entre elas: pontuação, rebotes, número de jogos, número de faltas cometidas e roubos de bola.

Ou seja, não há uma clara distinção para a primeira componente e, portanto, não há um significado claro. Assim, não é possível reduzir a dimensionalidade por meio da análise de componentes principais e esta solução será descartada para o restante do trabalho.

2.6 Regularização

Com a solução de Análise de Componentes Principais não sendo benéfica para o problema do trabalho, outras opções surgem como alternativas. A regularização de modelos, apesar de não ser uma técnica de redução de dimensionalidade, é uma dessas alternativas e será explorada a seguir.

Ao lidar com um modelo com muitas variáveis preditoras, como é o caso deste problema, é preciso ter um método para seleção das variáveis que entrarão no modelo. A mais comum, de estimação por Mínimos Quadrados Ordinários, apesar de em certas condições ser o estimador BLUE (*Best Linear Unbiased Estimator*), isto é, o estimador linear sem viés e de variância mínima, geralmente retornará um grande número de variáveis preditoras.

Apesar do estimador BLUE ser preferível, é possível fazer opção por outros estimadores. Se, ao introduzir um pouco de viés, o estimador é substancialmente mais preciso que o não viesado, é preferível escolher o estimador viesado, pois ele terá uma maior probabilidade de estar mais próximo do valor do parâmetro verdadeiro (KUTNER; NACHTSHEIM; NETER, 2004).

A ideia dos métodos de regularização baseia-se justamente nesta troca de viés por precisão. E, no processo, muitos parâmetros ao serem estimados, acabam sendo zerados. Mais ainda do que por MQO. Logo, ao optar por um modelo feito por regularização, trabalha-se com menos variáveis e seus estimadores, apesar de um pouco viesados, serão

mais precisos. Esta característica torna este método especialmente atraente para áreas de *Machine* e *Deep Learning*.

Além disso, os métodos de regularização lidam bem com bases de dados com mais parâmetros do que observações e na presença de variáveis altamente correlacionadas. Nas subsubseções que se seguem, serão apresentadas a abordagem geral do método, a regressão Ridge, a regressão Lasso e outras duas abordagens também presentes na literatura.

2.6.1 Abordagem Geral

Os métodos de regularização se baseiam em penalizações no cálculo das estimativas dos parâmetros. Consistem em reduzir a soma de quadrados dos resíduos sujeita a soma do valor absoluto dos coeficientes à potência de q a menos de uma constante λ. Sua forma geral pode ser vista a seguir:

$$\hat{\boldsymbol{\beta}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q, \tag{2.6.1}$$

onde y_i representa a variável dependente, $x_{i,j}$ as variáveis independentes, β_j o parâmetro a ser estimado e λ é o chamado *tuning parameter*, um parâmetro entre zero e infinito que controla o tamanho da redução a ser aplicada às estimativas. A potência q define o tipo de regularização: Ridge ($q = 2$), Lasso ($q = 1$) ou Horseshoe ($q < 1$).

Estes métodos são mais úteis justamente pela soma do valor absoluto dos coeficientes elevado à q. O Critério de Informação de Akaike (AIC), por exemplo, leva em conta no seu cálculo o número de coeficientes estimados. Porém, para o caso de um dos coeficientes ser muito pequeno, próximo de zero, a regularização será superior ao AIC por levar em conta a magnitude dos coeficientes.

É importante notar que diferentes λ levam a um conjunto de parâmetros estimados diferentes. Como citado, estes métodos introduzem viés no modelo, mas diminuem a variância. Para $\lambda = 0$ seria o caso da estimação por Mínimos Quadrados Ordinários. Para um λ grande, a variância seria próxima de zero, mas teria um viés muito alto.

2.6.2 Penalização Ridge

O primeiro método de regularização a ser explorado será o da regressão Ridge, introduzido por Hoerl e Kennard (1970). Surgiu como uma alternativa para solucionar problemas em que as covariáveis originais eram não-ortogonais ($\mathbf{X^T X} \neq \mathbf{I}$). É o caso de

$q = 2$ na Equação (2.6.1), ou seja, toma a seguinte forma:

$$\hat{\boldsymbol{\beta}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2. \tag{2.6.2}$$

A solução encontrada pelos autores foi justamente a de inserir viés em troca do ganho de precisão ao diminuir o erro quadrático médio. Com a penalização imposta, a regressão Ridge busca "encolher" (do termo em inglês *shrink*) ou reduzir os coeficientes β estimados. Porém, essa redução, apesar de ser muito próxima de zero, não chega a zerar de fato as estimações dos parâmetros.

Por este motivo, a regressão Ridge acaba não sendo a mais adequada para o problema em questão. Isto porque, além de um modelo preciso, um dos objetivos é obter um modelo parcimonioso, com poucas variáveis explicativas. E, em comparação com outras alternativas, a Ridge acaba não sendo a melhor neste quesito.

2.6.3 Regressão Lasso

A regressão Lasso, por sua vez, é um modelo que resulta em muitas estimativas sendo zeradas, o que ajuda na diminuição de variáveis preditoras. Lasso, introduzida por Tibshirani (1996), é um acrônimo do inglês *least absolute shrinkage and selection operator*. Um operador absoluto para seleção e redução ("encolhimento") de variáveis para um modelo linear de regressão, da seguinte forma:

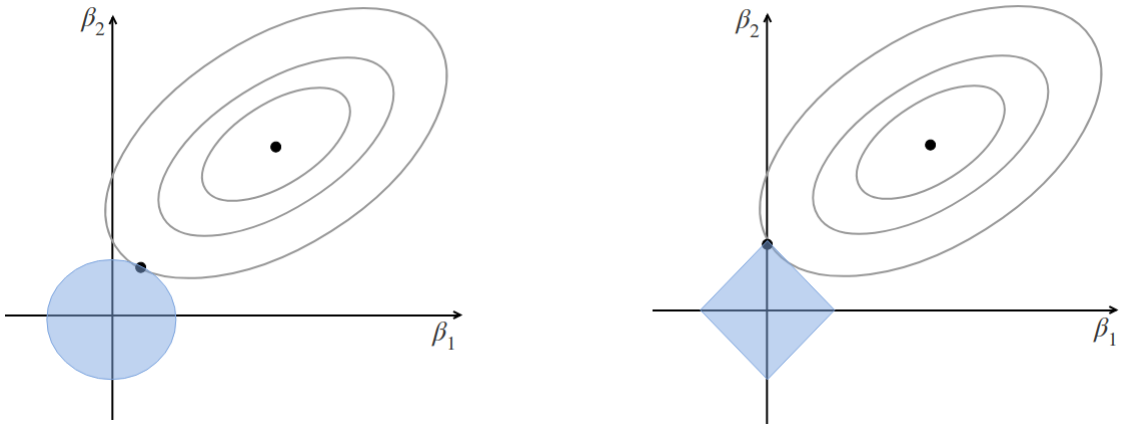
$$\hat{\boldsymbol{\beta}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{2.6.3}$$

Para entender a diferença entre a regressão Lasso e a Ridge é possível ver abaixo como suas equações funcionam de maneira gráfica. O ponto preto denota o custo mínimo de β. O objetivo, tanto da Equação (2.6.2), como da Equação (2.6.3), é minimizar este custo através da minimização dos βs, respeitando as restrições. Esta restrição, é justamente a penalização que está sendo imposta pelo termo $\lambda \sum_{j=1}^{p} |\beta_j|^q, q = 1, 2$.

Pela Figura 1, fica claro o que foi mencionado na subsubseção anterior, de que na restrição Ridge, apesar das estimativas dos parâmetros chegarem próximos de zero, nunca são de fato zeradas. Desta forma, todas as variáveis independentes entram no modelo.

Já para o caso do Lasso, os βs são de fato zerados. No caso da Figura 1, o β_1 é zero e a variável associada a este parâmetro não entraria no modelo. A penalização, portanto, funciona como um selecionador de variáveis. Diante disso, por ser vantajoso

Figura 1: Estimativas via penalizações Ridge (esquerda) e Lasso (direita)



Fonte: LEG/UFPR (2018)

para o problema do trabalho, a regressão Lasso será utilizada.

2.6.4 Outras Abordagens

Além das penalizações Ridge e Lasso, outras alternativas aparecem na literatura e serão apresentadas a seguir. Uma delas é o estimador Horseshoe, caso em que $q < 1$ na Equação (2.6.1). Este estimador foi introduzido por Carvalho, Polson e Scott (2010) e surge de um método bayesiano. É uma regularização mais robusta e que introduz ainda mais zeros na seleção de variáveis. Por lidar com aspectos bayesianos, que não são enfoque do trabalho, esta penalização não foi escolhida para ser trabalhada.

Outra abordagem é a penalização Elastic Net, introduzida por Zou e Hastie (2005). Sua principal característica é uma ponderação entre as abordagens Ridge e Lasso. Este método de regularização introduz um novo parâmetro α de *tuning*, que a depender de sua escolha, pode favorecer o método Ridge, o Lasso ou fazer um meio-termo entre os dois e pode ser visto a seguir:

$$\hat{\boldsymbol{\beta}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} \left[(1-\alpha)|\beta_j| + \alpha \beta_j^2 \right]. \tag{2.6.4}$$

É fácil ver que na Equação (2.6.4) se $\alpha = 1$, a equação é a mesma da regressão Ridge e se $\alpha = 0$, a equação se torna a mesma da regressão Lasso. No artigo, é apresentado que o Elastic Net é particularmente útil no caso de mais parâmetros do que observações e que, em contraste, o Lasso não é muito satisfatório na seleção de variáveis para este caso. Como no trabalho há mais observações que parâmetros, este método não foi o utilizado.

3 Resultados

Esta seção se divide em uma análise descritiva dos dados explorando a variável dependente do modelo, algoritmos de seleção automática de modelos, a escolha de um modelo mais parcimonioso por regressão Lasso e, por fim, os resultados de regressão quantílica deste modelo escolhido.

3.1 Análise Descritiva: Salários

Ao longo desta subseção, é feita uma análise descritiva acerca do comportamento da variável de salários de jogadores da NBA, bem como dos log-salários, que corresponde à variável resposta do modelo a ser ajustado posteriormente. Para as análises a seguir, como o objeto em análise é o salário da liga como um todo, a base original foi a considerada. A base tratada descrita na Subsubseção 2.4.2 sem as observações com dados faltantes foi usada a partir da subseção seguinte sobre escolha do modelo.

Como explicitado na Subseção 2.4, a base original de salários totaliza 578 observações. No entanto, três jogadores foram identificados nas bases de estatísticas e que não estão presentes nas bases de salários. As observações foram acrescentadas à base após uma pesquisa na Basketball-Reference (2021) e estão listadas a seguir:

Tabela 5: Jogadores fora da base de salários original

Jogador	**Salário**
Devin Cannady	\$61.528
Frank Mason III	\$61.528
Robert Franks	\$198.040

Desta forma, a base final de salários contém 581 jogadores. Porém, nem todos estes jogadores estão presentes nas bases de estatísticas. Mais especificamente, 41 deles não estão. Assim, ao juntar as bases de salários e estatísticas, são contabilizadas 626 observações de 540 jogadores distintos. Com isto, é possível obter algumas estatísticas descritivas para melhor entender como os dados estão distribuídos:

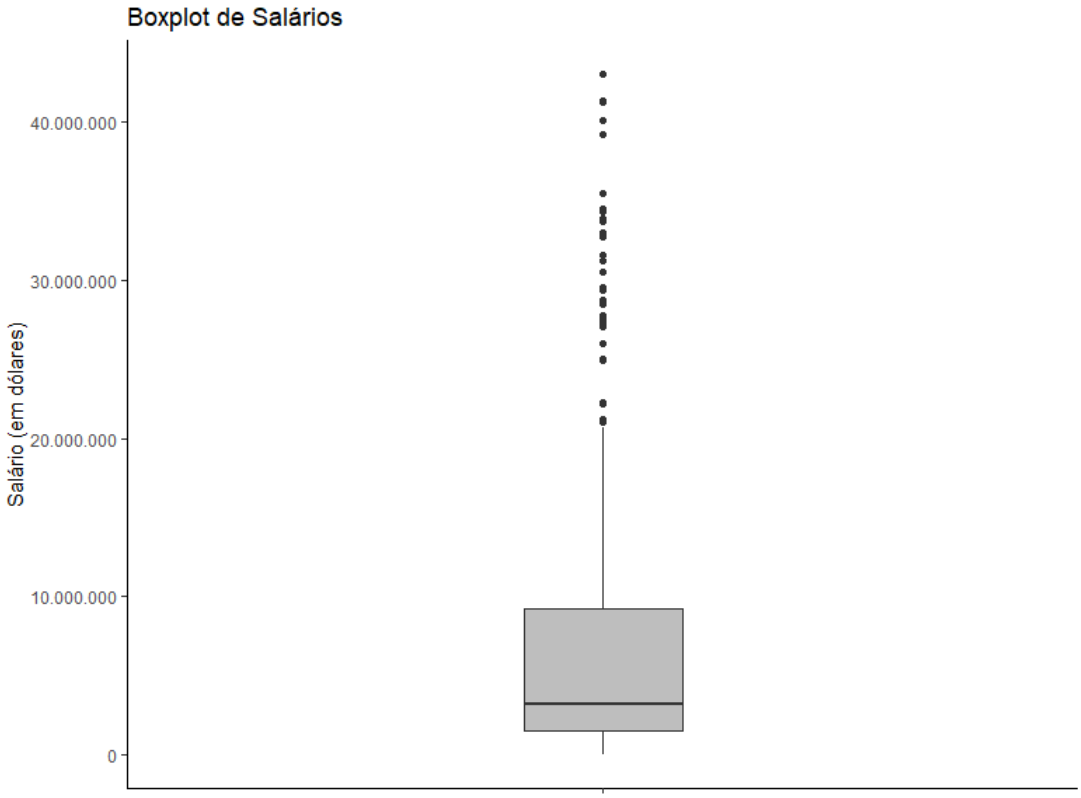
Tabela 6: Estatísticas descritivas da variável salários

Mín.	**1º Qu.**	**Mediana**	**Média**	**3º Qu.**	**Máx.**
24.611	1.517.981	3.160.540	6.998.108	9.256.000	43.006.362

Então, em média, o jogador da NBA recebeu na temporada 2020-2021 um salário de 6.998.108 dólares. O jogador mais bem remunerado foi Stephen Curry, do Golden State Warriors, com um salário de $43.006.362. Em contrapartida, o jogador com menor salário foi Elijah Bryant, membro do Milwaukee Bucks, campeão da temporada.

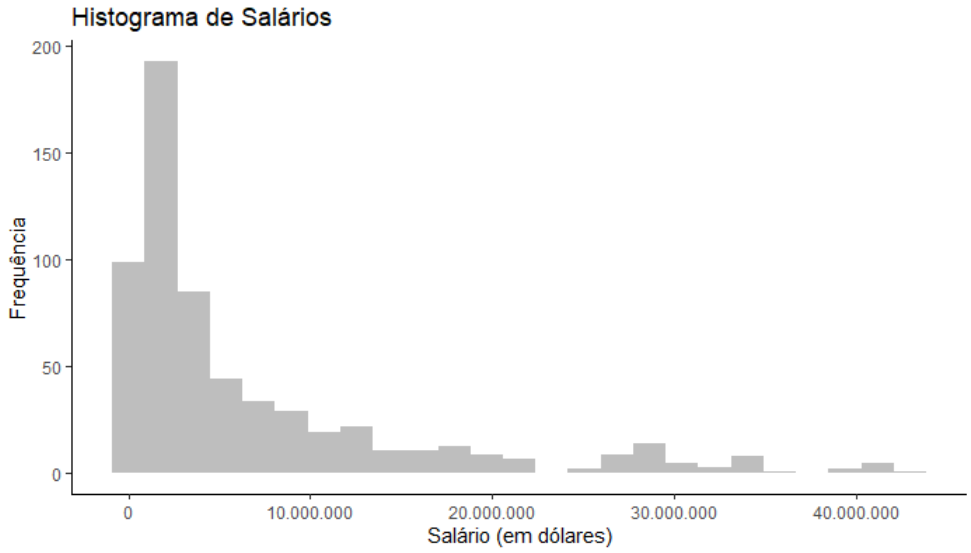
A seguir é apresentado um gráfico *boxplot* da variável em questão ajudando na representação visual das medidas vistas na Tabela 6.

Figura 2: *Boxplot* da variável salários



Como pode ser visto, jogadores com altos salários influenciam muito na distribuição dos salários. A mediana em torno de $3 milhões indica que há uma concentração grande de jogadores com salários baixos, se comparado aos dos melhores jogadores da liga. Ou seja, a distribuição é assimétrica, como pode ser vista pelo histograma apresentado a seguir:

Figura 3: Histograma da variável salários



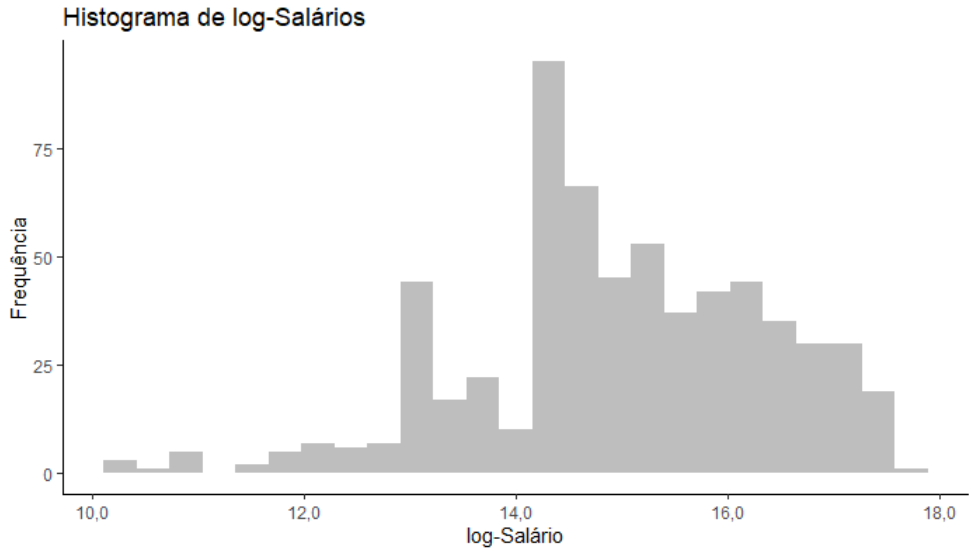
O desvio padrão da variável é de $8.911.043. Como citado em seções anteriores, os valores serão transformados pelo logaritmo natural. Esta transformação se deve para tentar lidar com a assimetria vista na Figura 3. As estatísticas descritivas da variável transformada podem ser vistas na tabela abaixo.

Tabela 7: Estatísticas descritivas da variável log-salários

Mín.	**1º Qu.**	**Mediana**	**Média**	**3º Qu.**	**Máx.**
10,11	14,23	14,97	14,97	16,04	17,58

Agora, o desvio padrão da nova variável é 1,396. A distribuição se torna mais simétrica do que a variável original, apesar de ainda ter uma grande concentração de observações para valores de log-salário mais altos como pode ser visto no histograma a seguir:

Figura 4: Histograma da variável log-salários



3.2 Escolha do Modelo

Como discutido na Subseção 2.4, após remover as variáveis com correlações maiores que $|0,9|$ e o dados faltantes, a base que antes tinha 106 variáveis, agora tem 75 com 560 observações e nenhum valor faltante. O número de variáveis, apesar de ainda alto, é mais palpável para tentar ajustar um modelo de regressão. Entre as 75 variáveis, quatro não são de interesse para o ajuste: nome do jogador, time, posição e salário do atleta na escala original.

Através de algoritmos de seleção automática de modelos auxiliado pela função *stepAIC* da biblioteca *MASS* do *software* R, dois modelos foram encontrados baseando-se no Critério de Akaike (AIC) conforme variáveis iam entrando ou saindo do modelo. O primeiro, dito *backward*, parte do modelo cheio, isto é, com todas as variáveis possíveis como preditoras e vai avaliando se o AIC diminui ou não conforme cada variável sai do modelo:

$$\begin{aligned} logSalary = & Age.x + G.x + GS + FTA.x + BLK.x + PF.x + FG.1 + X3P.1 + \\ & FT.1 + ORB.1 + DRB.1 + TRB.1 + STL.1 + BLK.1 + PF.1 + PTS.1 + \\ & PER + DRB. + BLK. + TOV. + OWS + OBPM + MP.y + FGA.y + \\ & X3P.y + X2P.y + X2PA.y + X2P..y + TRB.y + STL.y + TOV.y + \\ & PTS.y + Age.y + DRtg + Pace. \end{aligned}$$

A descrição de cada variável pode ser vista no Apêndice A. O algoritmo de seleção automática acabou selecionando 35 variáveis como preditoras e registrando um $AIC = -287,97$. Apesar de conseguir diminuir as variáveis preditoras pela metade, o número ainda é muito grande para realizar um ajuste. Pelo algoritmo *forward* de seleção, que parte do modelo vazio (sem nenhuma variável preditora) e avalia o AIC conforme variáveis vão entrando no modelo, tem-se que:

$$\begin{aligned} logSalary =& MP.1 + Age.x + FG.1 + DRB. + ORtg + \\ & TOV. + Age.y + DRtg + OWS + X2PA.y. \end{aligned}$$

Este algoritmo registrou um $AIC = -267,26$ e encontrou menos variáveis preditoras do que o procedimento *backward* de seleção. Porém, antes de adotar este modelo para o restante do trabalho, é preciso analisar com cuidado. O modelo contém menos variáveis, mas tem um Critério de Informação de Akaike maior do que o visto na seleção anterior. Assim, por AIC, o modelo a ser considerado deveria ser o do algoritmo *backward*.

Além disto, ao analisar o significado das variáveis do algoritmo *forward*, visto no Apêndice A, nota-se que seis das dez variáveis registradas correspondem a estatísticas de time. O intuito de incluir estatísticas de time na base de dados é controlar possíveis efeitos da equipe na remuneração dos atletas. Ao ter mais variáveis de time do que individuais, o objetivo do trabalho de identificar fatores determinantes na remuneração dos atletas pode se tornar difícil pela limitação de variáveis particulares dos jogadores.

Desta forma, não é vantajoso escolher nenhum dos dois modelos para a continuidade do trabalho. Assim, se faz necessário outros métodos de seleção de modelos. Como discutido na Subseção 2.6, os métodos de regularização, em especial a regressão Lasso, podem ser muito úteis para a seleção de modelos. Esta opção é explorada a seguir.

3.3 Regressão Lasso

Os modelos resultantes da escolha por seleção automática não se provaram úteis para os objetivos do trabalho. A fim de explorar outras possibilidades de escolha de um modelo com menos variáveis, pode-se fazer a opção do ajuste pela regressão Lasso, como visto na Subseção 2.6. Através da troca de precisão por inclusão de viés, um modelo menos complexo é obtido.

O ajuste pode ser feito pelo *software* R através do pacote *glmnet*, que ajusta

um modelo linear generalizado por máxima verossimilhança penalizada. Desta forma é possível ajustar os modelos baseando nos métodos de regularização vistos anteriormente. O ajuste do modelo de regressão Lasso é feito ao especificar $\alpha = 1$ na função *glmnet*.

O ajuste é feito baseado nas 75 variáveis descritas no final da Subsubseção 2.4.2 ao utilizar log-salários dos jogadores como variável resposta e o restante, excluindo nome do jogador, salários na escala original, time e posição, como variáveis preditoras. O λ do modelo não é especificado. Isto porque, ao atribuir o ajuste a um objeto e utilizar a função *print* neste objeto, obtém-se 100 parâmetros de *tuning* com o respectivo número de parâmetros diferentes de zero e a porcentagem da variância explicada para análise.

Alguns dos resultados podem ser visto na tabela abaixo, onde G.L. representa o número de coeficientes diferentes de zero, que são os graus de liberdade para o Lasso. A tabela também apresenta a porcentagem da variância explicada para o respectivo lambda indicado.

Tabela 8: Resultados da regressão Lasso para diferentes parâmetros de *tuning*

G.L.	**%Variância**	**Lambda**
0	0%	0,884
1	14,88%	0,7339
2	29,62%	0,5552
3	46,02%	0,3487
6	56,49%	0,1818
8	59,97%	0,09479
9	60,26%	0,08637
11	60,57%	0,07869
17	61,82%	0,04503
27	63,98%	0,01776
37	65,09%	0,009261
47	67,25%	0,001441
64	67,87%	0,00027
66	68,11%	0,0001065
69	68,14%	0,0000884

É útil destacar que o próprio pacote oferece a função *cv.glmnet* que faz validação cruzada para descobrir o melhor λ. Mas, para este trabalho, o interesse estará em um parâmetro funcional e não necessariamente ótimo. Isto porque ao explorar a validação cruzada, o λ mínimo encontrado foi de 0,00368 e resultava em 44 parâmetros não nulos.

Como o intuito de se utilizar a regressão Lasso é encontrar um modelo com um

menor número de variáveis e o foco não é a capacidade preditiva do modelo em si, o lambda ótimo encontrado pela validação cruzada não foi explorado. Com isto, os resultados encontrados na função *glmnet* foram considerados satisfatórios para a continuidade do trabalho.

Como a Tabela 8 indica, o ganho na variância explicada é muito pouco a partir de certo lambda. O ganho entre a escolha de um lambda que retorna um modelo com nove variáveis para um modelo com 11 variáveis é de 0,3%. Assim, o modelo escolhido foi o de nove variáveis, com $\lambda = 0,08637$. Através da função *coef*, é possível retornar o modelo com as variáveis associadas aos coeficientes não nulos encontrados:

$$logSalary = Age.x + GS + MP.1 + FG.1 + DRB.1 + STL.1 + TOV.1 + PTS.1 + ORtg$$

Tabela 9: Descrição das variáveis do modelo Lasso

Variável	**Descrição**
logSalary	Salários transformados em escala logarítmica
Age.x	Idade do jogador
GS	Partidas como titular
MP.1	Minutos jogados por jogo
FG.1	Arremessos convertidos por jogo
DRB.1	Rebotes defensivos por jogo
STL.1	Roubos de bola por jogo
TOV.1	Perdas de posse por jogo
PTS.1	Pontos marcados por jogo
ORtg	Estimativa de pontos feitos por 100 posses de bola.

É importante notar que o modelo encontrado possui um AIC de -242,64, mais alto do que os encontrados na Subseção 3.2. Mas, devido às discussões anteriores sobre o número de variáveis e seu significado no modelo, foi feita a escolha da equação acima como o modelo final a ser trabalhado.

Com o modelo definido, é possível analisar o conjunto de variáveis preditoras mais afundo. A seguir, para cada variável explicativa, foi feito um gráfico com os logsalários no eixo Y. Já adiantando o objetivo do trabalho centrado na regressão quantílica, foram ajustados modelos lineares simples de regressão com apenas a variável em questão como variável preditora. E, somado aos ajustes, foram traçadas linhas de regressão para a média e para cinco quantis. Para tanto, a seguinte legenda se faz útil:

Figura 5: Legenda utilizada na Figura 6

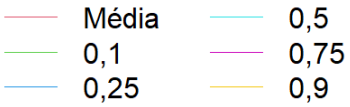
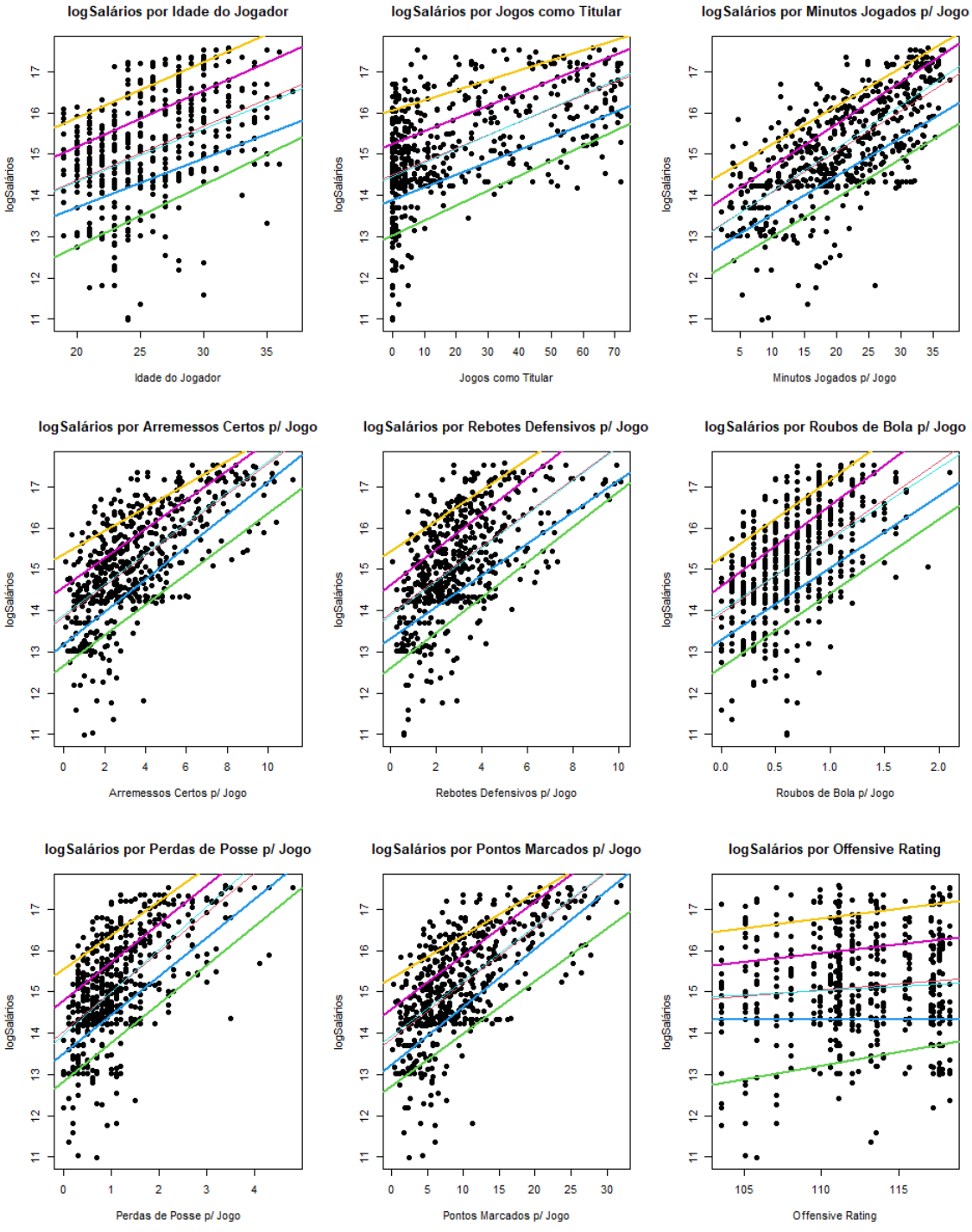


Figura 6: Gráfico da variável logsalários por cada uma das variáveis do modelo



Ao analisar a Figura 6, observa-se uma relação positiva entre a variável de logsalários e todas as variáveis explicativas, com exceção da variável de ritmo ofensivo (*Offensive Rating*). Para a variável de ritmo, o ajuste para o quantil 0,25 apresenta uma inclinação levemente negativa e para os ajustes de média e mediana é levemente positiva. Com relação aos quantis restantes, a relação positiva pode ser identificada para esta variável.

É interessante notar que ritmo ofensivo é a única variável referente a estatística de time. No gráfico, é possível ver que há 30 valores da variável para ritmo ofensivo, correspondente aos 30 times da liga. Neste caso, o ajuste não é tão bom quanto para as variáveis individuais. Isto porque se algum jogador, mesmo com remuneração alta e bons números, faz parte de uma equipe que foi mal na temporada como um todo, o *Offensive Rating* será baixo.

Quanto às outras variáveis, a relação positiva com logsalários é esperada ao considerar que jogadores mais bem remunerados devem ter melhores números. A única variável com conotação negativa é a de perdas de posse por jogo. Mas até esta relação pode ser explicada pelo volume de jogo imposto pelos atletas mais bem pagos.

Uma alta remuneração ter relação positiva com o número de atuações como titular e minutos por jogo indica que jogadores com maiores salários acabam atuando muito. E, por consequência, há mais oportunidades de cometerem erros, como explicitado pelo número de perdas de posse de bola.

Por fim, vale destacar a relação entre o ajuste pela média e mediana. Em sua maioria são ajustes que bem parecidos com uma linha sobrepondo a outra em algumas faixas. Para as variáveis de roubos de bola e perdas de posse, porém, a distância entre os ajustes é mais perceptível e suas diferenças serão exploradas na próxima subseção.

3.4 Resultados da Estimação

Utilizando o modelo encontrado ao fim da Subseção 3.3 e com o auxílio do *software* R, foram ajustados diferentes modelos de regressão para análise. O primeiro, de regressão por estimação de Mínimos Quadrados Ordinários, é apresentado na primeira coluna da Tabela 10. Nas colunas subsequentes, estão os resultados da regressão quantílica para os quantis 0,1, 0,25, 0,5 (mediana), 0,75 e 0,9.

Para cada estimativa, em parênteses é apresentado o p-valor correspondente do teste t de significância. A não rejeição da hipótese nula do teste, $H_0 : \hat{\beta}_i(\theta) = 0$, indica

Tabela 10: Resultados do ajuste para diferentes quantis
***: 0,001; **: 0,01; *: 0,05

Variável	**MQO**	**Q10**	**Q25**	**Q50**	**Q75**	**Q90**
Intercepto	7,540(0,000)*** (5,573, 9,507)	6,305(0,002)** (2,466, 9,499)	6,603(0,000)*** (3,698, 7,979)	8,263(0,000)*** (5,279, 10,340)	10,257(0,000)*** (7,066, 12,494)	9,541(0,000)*** (6,392, 12,777)
Age.x	0,092(0,000)*** (0,075, 0,109)	0,091(0,000)*** (0,046, 0,14)	0,118(0,000)*** (0,082, 0,137)	0,099(0,000)*** (0,079, 0,120)	0,075(0,000)*** (0,063, 0,092)	0,082(0,000)*** (0,051, 0,109)
GS	0,003(0,259) (-0,002, 0,008)	0,004(0,405) (-0,004, 0,016)	0,004(0,252) (-0,003, 0,01)	0,004(0,127) (-0,002, 0,010)	0,001(0,801) (-0,004, 0,005)	-0,005(0,168) (-0,010, 0,003)
MP.1	0,032(0,004)** (0,011, 0,053)	0,027(0,279) (-0,015, 0,078)	0,024(0,094) (0,009, 0,061)	0,031(0,026)* (0,013, 0,058)	0,037(0,018)* (0,011, 0,070)	0,056(0,007)** (-0,005, 0,075)
FG.1	0,100(0,397) (-0,131, 0,331)	0,259(0,237) (-0,145, 0,94)	0,193(0,132) (-0,13, 0,515)	0,107(0,286) (-0,082, 0,266)	0,038(0,706) (-0,164, 0,151)	-0,064(0,704) (-0,401, 0,348)
DRB.1	0,085(0,005)** (0,026, 0,144)	0,110(0,023)* (-0,034, 0,19)	0,075(0,049)* (-0,012, 0,131)	0,053(0,074) (0,000, 0,100)	0,068(0,022)* (0,024, 0,105)	0,096(0,048)* (0,014, 0,168)
STL.1	0,123(0,389) (-0,157, 0,403)	0,110(0,698) (-0,437, 0,536)	0,169(0,464) (-0,422, 0,491)	0,304(0,034)* (-0,174, 0,367)	0,112(0,400) (-0,088, 0,393)	0,018(0,930) (-0,302, 0,703)
TOV.1	0,197(0,018)* (0,034, 0,36)	0,292(0,061) (0,057, 0,598)	0,210(0,100) (-0,023, 0,472)	0,117(0,205) (-0,055, 0,287)	0,125(0,174) (0,000, 0,296)	0,207(0,062) (-0,118, 0,384)
PTS.1	0,002(0,958) (-0,083, 0,087)	-0,053(0,504) (-0,276, 0,097)	-0,017(0,750) (-0,157, 0,069)	0,005(0,906) (-0,047, 0,067)	0,020(0,585) (-0,021, 0,107)	0,032(0,627) (-0,108, 0,194)
ORtg	0,032(0,000)*** (0,015, 0,049)	0,034(0,046)* (0,008, 0,067)	0,030(0,002)** (0,019, 0,057)	0,024(0,030)* (0,006, 0,048)	0,018(0,177) (-0,003, 0,049)	0,025(0,030)* (-0,002, 0,051)

que o coeficiente estimado não contribui para explicar o modelo. A estatística do teste é calculada dividindo o valor do parâmetro estimado pelo erro padrão associado à estimação. Porém, para a execução do teste, algumas ressalvas precisam ser feitas.

O pacote utilizado no R para a estimação dos resultados foi o *quantreg*, que tem autoria de Roger Koenker, um dos coautores do artigo que introduziu a técnica de regressão quantílica em 1978. Nele, diferentes possibilidades são oferecidas ao especificar na função *rq*, utilizada para o ajuste do modelo, o comando *se*, referente ao erro padrão (*standard error*). Algumas dessas opções foram citadas na Subseção 2.3 e mais informações podem ser vistas na documentação do pacote ou no Apêndice A de Koenker (2005).

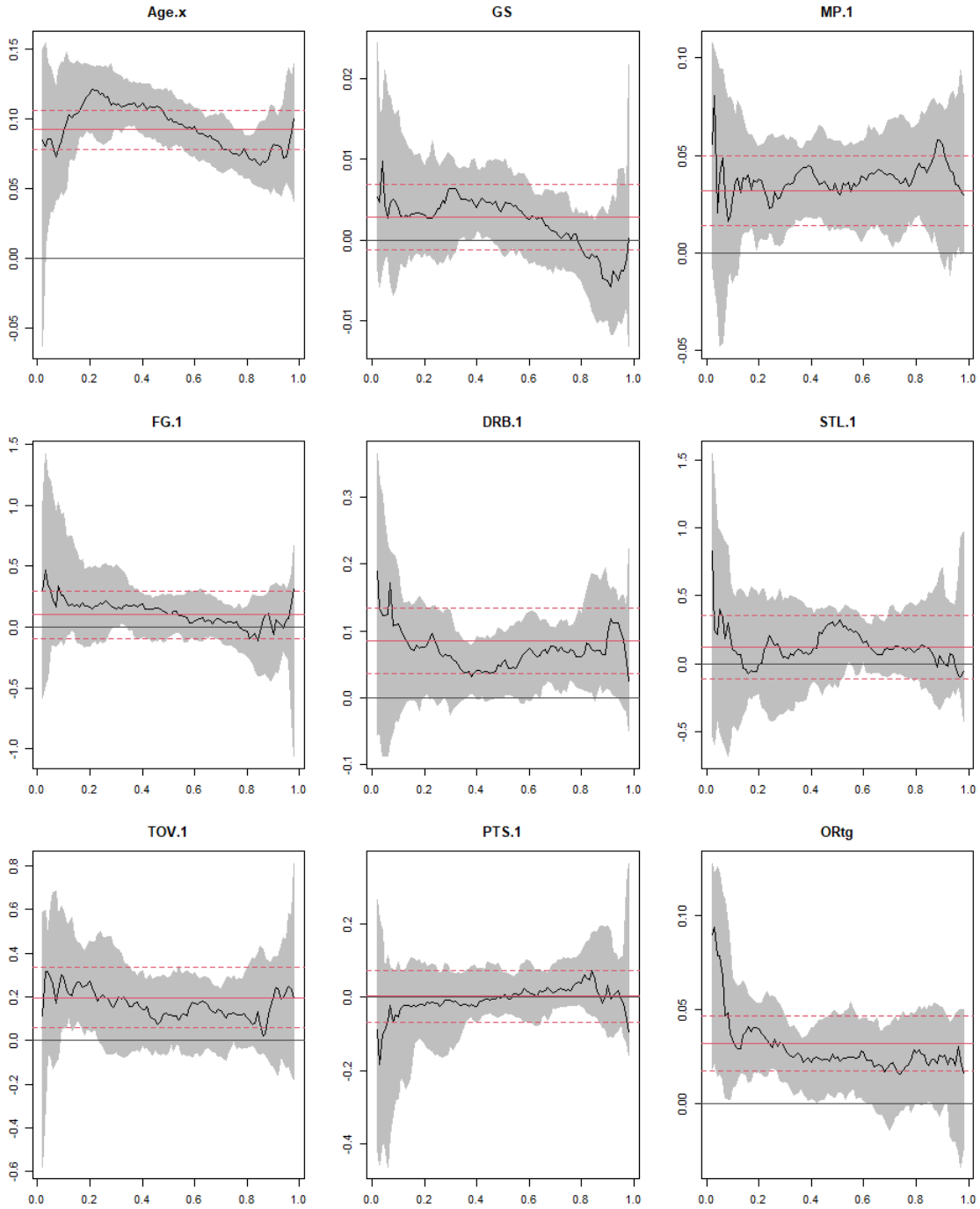
Dentre as opções que podem ser especificadas, estão: *nid*, que considera erros não independentes e identicamente distribuídos; *ker*, que usa a estimativa da matriz de covariância kernel proposta em Powell (1991); e *boot* que obtém as estimativas por método de *bootstrap*. Ao se utilizar um método a favor do outro, a estimativa do erro padrão é diferente e, consequentemente, a estatística do teste muda para cada método, rendendo diferentes resultados para o teste t de significância e a sua interpretação.

No trabalho, o método *bootstrap* foi o utilizado por não assumir distribuições para os erros. Foi utilizado o *bootstrap* padrão do pacote, por par "xy" e o número de 200 replicações para ajuste. Diferentes *bootstraps* podem resultar em diferentes estimativas. Para tanto, a função nativa do R *set.seed* foi utilizada para que os resultados possam eventualmente ser replicados.

Além disto, a Tabela 10 ainda apresenta abaixo das estimativas e dos p-valores dos testes, um intervalo de 95% de confiança para os parâmetros estimados. Este intervalo é feito pelo próprio pacote *quantreg* utilizando o método de inversão de rank, descrito na seção 3.5.5 de Koenker (2005) ao não se especificar o método de cálculo do erro padrão como mencionado antes. Este método não consegue estimar o erro padrão, sendo necessário utilizar um dos métodos anteriores para realizar o teste de significância. Assim, dois métodos foram utilizados no cálculo e os intervalos do método de inversão de rank comparados com os resultados dos testes utilizando o método *bootstrap*.

Para auxiliar na visualização, os parâmetros estimados para cada quantil, além dos apresentados na Tabela 10, bem como seus intervalos de confiança, podem ser vistos na Figura 7. As linhas horizontais correspondem às estimativas feitas utilizando Mínimos Quadrados Ordinários e seus intervalos de confiança.

Figura 7: Parâmetros estimados para os diferentes quantis



Analisando os resultados, para o modelo de regressão linear clássica todos os parâmetros estimados foram positivos. Para o teste t de significância, o efeito da idade do jogador e do ritmo ofensivo do time foram significativos ao nível de significância de 0,1%. As variáveis de minutos jogados e rebotes defensivos por jogo foram significativos ao nível de 1%. Já a variável de posses de bola perdidas por jogo foi significativa ao nível de 5%.

A regressão quantílica permite expandir esta análise de forma mais minuciosa. A variável de idade do jogador é a única que é significativa para todos os quantis com significância ao nível de 0,1% e com todos os parâmetros estimados positivos. Ou seja, quanto mais velho o jogador, maior é sua remuneração.

Este resultado já era esperado, pois jogadores que recém ingressaram na liga têm contratos menos lucrativos do que jogadores mais experientes. Isto porque nos esportes americanos, os chamados contratos de calouro são pré-estabelecidos, não permitindo aos jogadores mais jovens negociarem valores em seus primeiros anos. Esta relação pode ser vista facilmente na Figura 6 que demonstra log-salários pela idade do jogador, onde é possível identificar um salto na remuneração dos atletas por volta dos 24, 25 anos.

Para o décimo quantil, as variáveis significativas no modelo, além de idade, foram as de rebotes defensivos e ritmo ofensivo do time. Porém, o p-valor do teste t de significância da variável de ritmo está no limite do nível de 5% de significância. A variável de perdas de posse por jogo também está bem próxima do nível de significância, com p-valor de 6,1%.

Além disto, o intervalo de 95% de confiança para a variável de rebotes contém o zero, o que não é o esperado ao rejeitar o teste de significância. Desta forma, não é possível atribuir claramente um fator determinante na remuneração ao analisar os jogadores com os menores salários na liga.

Para o 25º quantil, além da idade, a variável de rebotes defensivos por jogo é significativa ao nível de 5%. Porém, além de estar em cima do limite, seu intervalo de confiança contém o zero, não podendo ser considerado um fator determinante. A outra variável significativa encontrada foi a de ritmo ofensivo do time sendo significativa com p-valor de 0,002, significativa ao nível de 1%, quase 0,1%. Este resultado é concordante com a análise por MQO, com valores muito próximos inclusive. É possível então levantar a hipótese de que jogadores neste quantil são remunerados não por características individuais, mas sim pela produção do time. Esta relação não é possível ser identificada na Figura 6.

Olhando para o 50º quantil, correspondente ao efeito na mediana da distribuição de log-salários, três variáveis, além de idade, são significativas para o modelo ao nível de 5% de significância: minutos jogados, roubos de bola e ritmo ofensivo. Apesar de seu intervalo de confiança conter o zero, é interessante notar que roubos de bola foi considerado significativo apenas para este quantil, não sendo relevante nem para a análise por MQO. É a única variável do modelo que denota uma característica defensiva.

Já para o 75º quantil, idade, minutos jogados e rebotes defensivos foram considerados significativos, estes dois últimos ao nível de 5% de significância. Este é o único quantil que não identificou ritmo ofensivo, uma estatística de time, como fator determinante de remuneração.

Por último, para o 90º quantil, as variáveis que foram significativas ao modelo juntas à idade foram minutos jogados por partida, significativa ao nível de 1%, rebotes defensivos e ritmo ofensivo, ambas significativas ao nível de 5%. Para os intervalos de 95% de confiança, tanto minutos jogados como ritmo ofensivo apresentaram intervalos contendo o zero. A variável de rebotes ainda apresentou um p-valor bem próximo ao limite do nível de 5% de significância. Assim, não é possível atribuir um fator determinante claro para este quantil.

Cabe destacar que esta seção de resultados é inteiramente baseada no modelo encontrado na subseção anterior, por meio da regressão Lasso. Ao escolher o modelo, foi feita a opção por menos variáveis explicativas e variáveis que fizessem sentido para a identificação de possíveis fatores determinantes de remuneração. É possível, para futuros trabalhos, a escolha de um outro modelo baseado em diferentes aspectos.

O enfoque do trabalho não abrangeu a capacidade preditiva do modelo e este é um caminho que poderia ser explorado. O próprio modelo final teve um Critério de Akaike pior do que os encontrados na Subseção 3.2. Também não foi explorada a possibilidade de transformação de alguma variável explicativa no modelo, como a sua presença estando elevada à segunda potência, o que agregaria ao modelo trabalhado.

Outro caminho que se prova interessante é o de seleção de variáveis por quantis. O pacote *rqPen* do R traz essa funcionalidade. Identificar quais modelos se ajustam melhor a depender do quantil pode trazer informações adicionais para a análise. O modelo ajustado na mediana traz, por exemplo, mais variáveis de time e engloba assistências, uma das principais estatísticas do basquete, e também bloqueios (tocos), uma outra variável defensiva além de roubos de bola e que poderia ser contemplada no modelo trabalhado.

4 Conclusão

O principal objetivo do trabalho era o de identificar fatores determinantes na remuneração dos atletas da NBA baseado em estatísticas da temporada de 2020-2021. Durante o desenvolvimento o principal problema encontrado foi o da escolha do modelo a ser trabalhado, devido ao excesso de variáveis na base de dados. A solução encontrada foi a de utilizar a técnica de regressão Lasso para a escolha de modelo. Após, as análises de regressão clássica e quantílica foram feitas e apresentadas.

Para os ajustes de regressão quantílica foi necessário fazer a opção por alguns métodos: o de estimação de erro padrão por método *bootstrap* e o de intervalo de confiança por método de inversão de rank. Desta forma, foi possível realizar os testes t de significância de coeficientes e calcular intervalos de 95% de confiança.

O intuito de comparar a análise de regressão por mínimos quadrados e quantílica trouxe resultados interessantes. A regressão clássica identificou a variável de perdas de posse por jogo como um fator determinante na remuneração média dos atletas. Mas, ao analisar cada quantil separadamente, esta variável não apareceu como significante em nenhum deles.

A idade apareceu como principal fator determinante de remuneração. Este resultado era esperado, pois jogadores mais jovens não podem negociar seus contratos em seus primeiros anos, possuindo contratos pré-estabelecidos. Ou seja, mesmo jogando bem e, consequentemente, registrando boas estatísticas, a remuneração continua baixa.

Nos quantis inferiores, entre 0,1 e 0,5, a variável ritmo ofensivo apareceu como determinante. Esta foi a única variável de time usada no modelo. O resultado também está alinhado com o que era esperado. Jogadores com menor remuneração são mais afetados por variáveis de time. Isto porque partindo do pressuposto de que jogadores mais bem remunerados são os jogadores mais habilidosos, espera-se que sua habilidade não seja traduzida em estatísticas do time, mas sim em estatísticas pessoais.

Olhando para os quantis superiores, a variável de minutos jogados por partida apareceu como significativa. No 90^{o} quantil o intervalo de 95% de confiança contém o zero, mas o p-valor do teste t é o mais baixo dos quantis que identificaram a relevância da variável. Este é um resultado também esperado. Partindo do mesmo pressuposto de que maior salário é sinônimo de maior habilidade, espera-se que os melhores jogadores estejam em quadra por mais tempo. Assim, minutos jogados possa ser considerado não um fator determinante, mas sim uma consequência de maior remuneração.

A última variável significante do modelo foi a de rebotes defensivos, significante para quase todos os quantis, exceto a mediana. É interessante notar que para a regressão clássica, avaliando o efeito médio, a variável foi significativa, mas ao considerar a mediana não. Mas até neste quantil, o p-valor do teste t de significância não foi alto, sendo significante no nível de 10%, que apesar de não ter sido adotado neste trabalho, é utilizado na literatura para avaliar se a variável é significante.

Para o 10º e 25º quantis, os intervalos de 95% de confiança para o parâmetro desta variável continham o zero. A variável de rebotes defensivos sendo significativa para os maiores quantis pode indicar a presença de um fator determinante de remuneração. Mas, como visto para a variável minutos jogados, é possível levantar a hipótese de que jogadores mais bem remunerados, por estarem mais em quadra, recuperam um maior número de rebotes. A correlação entre rebotes defensivos e minutos jogados é 0,715. Esta hipótese pode ser explorada em outro trabalho futuramente.

Todas as variáveis que foram significativas para algum quantil em específico foram significativas também para a regressão clássica. Com o auxílio da regressão quantílica, o trabalho pôde destrinchar para quais faixas de salário estas variáveis se destacam mais. Ao considerar o teto salarial, os resultados são úteis ao nortear quais características times podem procurar ao montar seus elencos.

Para atletas com menores salários, buscar jogadores com um perfil adequado ao estilo de jogo do time, explicitado no modelo pela variável ritmo ofensivo. Para salários medianos e acima, procurar jogadores que vão jogar bastante e que agarram um grande número de rebotes. E, em geral, levar em conta a experiência do jogador na sua remuneração. Um maior salário está atrelado a uma maior idade.

Referências

BADENHAUSEN, K.; OZANIAN, M. *NBA Team Values 2021: Knicks Keep Top Spot At $5 Billion, While Warriors Seize No. 2 From Lakers*. 2021. Disponível em: ⟨https://www.forbes.com/sites/kurtbadenhausen/2021/02/10/nba-team-values-2021-knicks-keep-top-spot-at-5-billion-warriors-bump-lakers-for-second-place/?sh=5bf0fae9645b⟩. Acesso em: 23/08/2021.

BASKETBALL-REFERENCE. *2020-21 NBA Season Summary*. 2021. Disponível em: ⟨https://www.basketball-reference.com/leagues/NBA_2021.html⟩. Acesso em: 22/08/2021.

CARVALHO, C. M.; POLSON, N. G.; SCOTT, J. G. The horseshoe estimator for sparse signals. *Biometrika*, [Oxford University Press, Biometrika Trust], v. 97, n. 2, p. 465–480, 2010. ISSN 00063444, 14643510.

DAVINO, C.; FURNO, M.; VISTOCCO, D. *Quantile Regression: Theory and Applications*. [S.l.]: Wiley, 2013. (Wiley Series in Probability and Statistics). ISBN 9781119975281.

DEWAN, J.; ZMINDA, D.; STATS, I. S. *STATS Basketball Scoreboard, 1993-94*. New York: Harpercollins Publishers, 1993. ISBN 9780062730350.

GILCHRIST, W. *Statistical Modelling with Quantile Functions*. New York: Chapman and Hall/CRC, 2000. ISBN 9780429119200.

GLEESON, S. *Brooklyn Nets sale nearly complete, will be worth about $2.35 billion*. 2019. Disponível em: ⟨https://www.usatoday.com/story/sports/nba/nets/2019/08/16/brooklyn-nets-sale-2-35-billion-highest-price-tag-us-sports-team/2028663001/⟩. Acesso em: 23/08/2021.

HAO, L.; NAIMAN, D. Q. *Quantile Regression*. Thousand Oaks, California: SAGE Publications, Inc., 2007. ISBN 9781412926287.

HELIN, K. *Report: NBA to keep most of relaxed two-way player rules for next season*. 2021. Disponível em: ⟨https://nba.nbcsports.com/2021/07/18/report-nba-to-keep-most-of-relaxed-two-way-player-rules-for-next-season/⟩. Acesso em: 25/08/2021.

HOERL, A. E.; KENNARD, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, [Taylor & Francis, Ltd., American Statistical Association, American Society for Quality], v. 12, n. 1, p. 55–67, 1970. ISSN 00401706.

HOOPSHYPE. *2020/21 NBA Player Salaries*. 2020. Disponível em: ⟨https://hoopshype.com/salaries/players/2020-2021/⟩. Acesso em: 22/08/2021.

HUBER, P. J. Behavior of maximum likelihood estimates under nonstandard conditions. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley, CA, USA: University of California Press, 1967. p. 221.

KASABIAN, P. *NBA Announces Salary Cap, Luxury Tax Information for 2021-22 at Start of Free Agency*. 2021. Disponível em: ⟨https://bleacherreport.com/articles/10009477-nba-announces-salary-cap-luxury-tax-information-for-2021-22-at-start-of-free-agency⟩. Acesso em: 25/08/2021.

KEEFER, Q. A. W. Compensation discrimination for defensive players: Applying quantile regression to the national football league market for linebackers. *Journal of Sports Economics*, v. 14, n. 1, p. 23–44, 2013. Disponível em: ⟨https://doi.org/10.1177/1527002511413288⟩.

KOENKER, R. *Quantile Regression*. [S.l.]: Cambridge University Press, 2005. (Econometric Society monographs 38). ISBN 0521845734,9780521845731,9780511130342,9780521608275,0521608279.

KOENKER, R.; BASSETT, G. Regression quantiles. *Econometrica*, [Wiley, Econometric Society], v. 46, n. 1, p. 33–50, 1978. ISSN 00129682, 14680262. Disponível em: ⟨http://www.jstor.org/stable/1913643⟩.

KUTNER, M.; NACHTSHEIM, C.; NETER, J. *Applied Linear Regression Models*. New York: McGraw-Hill/Irwin, 2004. (Irwin/McGraw-Hill series in operations and decision sciences). ISBN 9780073014661.

LEG/UFPR, L. de Estatística e G. *Regularização*. 2018. Disponível em: ⟨http://cursos.leg.ufpr.br/ML4all/apoio/Regularizacao.html⟩. Acesso em: 18/02/2022.

MATANGE, Y. *NBA Salaries: All players to have signed contracts of $200 million or more*. 2021. Disponível em: ⟨https://ca.nba.com/news/nba-salaries-all-players-to-have-signed-contracts-of-200-million-stephen-curry/ubmfz2o9slruzp1pl883w9hq⟩. Acesso em: 24/08/2021.

PARZEN, M. I.; WEI, L.-J.; YING, Z. A resampling method based on pivotal estimating functions. *Biometrika*, Oxford University Press, v. 81, n. 2, p. 341–350, 1994.

POWELL, J. *Estimation of Monotonic Regression Models under Quantile Restrictions*. In: BARNETT, W.; POWELL, J.; TAUCHEN, G. (Ed.). *Nonparametric and Semiparametric Methods in Econometrics and Statistics: Proceedings of the Fifth International Symposium in Economic Theory and Econometrics*. New York: Cambridge University Press, 1991, (International Symposia in Economic Theory and Econometrics). ISBN 9780521424318.

RAO, C. R. Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. In: *Mathematical Proceedings of the Cambridge Philosophical Society*. United Kingdom: Cambridge University Press, 1948. v. 44, n. 1, p. 50–57.

ROA, A. D. *Salary comparison: NBA players vs. Premier League, NFL and MLB players*. 2021. Disponível em: ⟨https://hoopshype.com/lists/salary-comparison-nba-players-vs-premier-league-nfl-and-mlb-players/⟩. Acesso em: 24/08/2021.

ROSEN, C. *The First Tip-Off: The Incredible Story of the Birth of the NBA*. New York: McGraw-Hill Education, 2008. v. 1. 231-236 p.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, [Royal Statistical Society, Wiley], v. 58, n. 1, p. 267–288, 1996. ISSN 00359246.

VINCENT, C.; EASTMAN, B. Determinants of pay in the nhl: A quantile regression approach. *Journal of Sports Economics*, v. 10, n. 3, p. 256–277, 2009. Disponível em: ⟨https://doi.org/10.1177/1527002508327519⟩.

WATSON, P. *How Many NBA Players Are Playing in the Tokyo Olympics?* 2021. Disponível em: ⟨https://www.sportscasting.com/how-many-nba-players-are-playing-in-the-tokyo-olympics/⟩. Acesso em: 24/08/2021.

ZOU, H.; HASTIE, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, [Royal Statistical Society, Wiley], v. 67, n. 2, p. 301–320, 2005. ISSN 13697412, 14679868.

Apêndice

A Descrição de Variáveis

Variável	**Descrição**
Player	Nome do jogador
Salary	Salário na temporada 2020-2021
Tm	Time
Pos	Posição
Age.x	Idade
G.x	Jogos
GS	Jogos como titular
FG..x	Aproveitamento de arremessos
X3P.x	Arremessos de 3 pontos convertidos
X3P..x	Aproveitamento de arremessos de 3 pontos
X2P.x	Arremessos de 2 pontos convertidos
X2P..x	Aproveitamento de arremessos de 2 pontos
FT.x	Lances livres convertidos
FT..x	Aproveitamento de lances livres
ORB.x	Rebotes ofensivos
DRB.x	Rebotes defensivos
AST.x	Assistências
STL.x	Roubos de bola
BLK.x	Bloqueios (tocos)
PF.x	Faltas pessoais
MP.1	Minutos jogados p/ jogo
FG.1	Arremessos convertidos p/ jogo
X3P.1	Arremessos de 3 pontos convertidos p/ jogo
X2P.1	Arremessos de 2 pontos convertidos p/ jogo
FT.1	Lances livres convertidos p/ jogo
ORB.1	Rebotes ofensivos p/ jogo
DRB.1	Rebotes defensivos p/ jogo
TRB.1	Rebotes totais p/ jogo
AST.1	Assistências p/ jogo
STL.1	Roubos de bola p/ jogo
BLK.1	Bloqueios (tocos) p/ jogo

Variável	**Descrição**
TOV.1	Perdas de posse p/ jogo
PF.1	Faltas pessoais p/ jogo
PTS.1	Pontos marcados p/ jogo
PER	Uma medida de produção por minuto padronizada tal que a média da liga é 15.
TS..x	Uma medida de eficiência de arremesso que leva em conta arremessos de 2 e 3 pontos e lances livres.
X3PAr.x	Porcentagem de tentativas de arremessos de 3 pontos.
FTr.x	Número de lances livres por tentativa de arremesso.
ORB.	Uma estimativa da porcentagem dos rebotes ofensivos disponíveis agarrados pelo jogador enquanto ele estava em quadra.
DRB.	Uma estimativa da porcentagem dos rebotes defensivos disponíveis agarrados pelo jogador enquanto ele estava em quadra.
STL.	Uma estimativa da porcentagem das posses adversárias que terminaram com um roubo de bola do jogador enquanto ele estava em quadra.
BLK.	Uma estimativa da porcentagem dos arremessos de 2 pontos bloqueados pelo jogador enquanto ele estava em quadra.
TOV.	Uma estimativa do número de posses perdidas em 100 jogadas.
USG.	Uma estimativa da porcentagem de jogadas do time que utilizaram o jogador enquanto esteve em quadra.
OWS	Uma estimativa da contribuição do jogador no número de vitórias devido ao seu ataque.
DWS	Uma estimativa da contribuição do jogador no número de vitórias devido à sua defesa.
WS.48	Uma estimativa da contribuição do jogador por 48 minutos no número de vitórias (a média da liga é aproximadamente .100).
OBPM	Uma estimativa do saldo de pontos por 100 posses em que o jogador contribuiu ofensivamente em relação a um jogador médio da liga jogando em um time médio.
DBPM	Uma estimativa do saldo de pontos por 100 posses em que o jogador contribuiu defensivamente em relação a um jogador médio da liga jogando em um time médio.

Variável	**Descrição**
VORP	"Uma estimativa do saldo de pontos por 100 posses em que o jogador contribuiu em relação a um jogador médio da liga jogando em um time médio em uma temporada de 82 jogos. Multiplique por 2,7 para obter vitórias sobre um jogador médio"
MP.y	Minutos jogados p/ jogo pelo time
FGA.y	Arremessos tentados p/ jogo pelo time
FG..y	Aproveitamento de arremessos pelo time
X3P.y	Arremessos de 3 pontos convertidos p/ jogo pelo time
X3P..y	Aproveitamento de arremessos de 3 pontos pelo time
X2P.y	Arremessos de 2 pontos convertidos p/ jogo pelo time
X2PA.y	Arremessos de 2 pontos tentados p/ jogo pelo time
X2P..y	Aproveitamento de arremessos de 2 pontos pelo time
FT.y	Lances livres convertidos p/ jogo pelo time
FT..y	Aproveitamento de lances livres pelo time
ORB.y	Rebotes ofensivos p/ jogo pelo time
DRB.y	Rebotes defensivos p/ jogo pelo time
TRB.y	Rebotes totais p/ jogo pelo time
AST.y	Assistências p/ jogo pelo time
STL.y	Roubos de bola p/ jogo pelo time
BLK.y	Bloqueios (tocos) p/ jogo pelo time
TOV.y	Perdas de posse p/ jogo pelo time
PF.y	Faltas pessoais p/ jogo pelo time
PTS.y	Pontos marcados p/ jogo pelo time
Age.y	Média de idade do time em 1º de fevereiro
W	Vitórias
ORtg	Estimativa de pontos feitos por 100 posses de bola.
DRtg	Estimativa de pontos cedidos por 100 posses de bola.
Pace	Estimativa do número de posses de bola por 48 minutos.
logSalary	Log-salário na temporada 2020-2021.

B Códigos em R

```
### Pacotes utilizados:
lapply(c("tidyverse","dplyr", "scales", "MASS",
         "quantreg", "glmnet"),
       library, character.only = TRUE)

### Base consolidada (626x106)
basetotal <- read.csv("basetotal.txt")

### 2.4.1 Correlações
cor(basetotal[,-c(1:4)],
    use = 'complete.obs')

basefinal <- basetotal[-c(8,9,10,14,16,18,21,24,
                          28,30,33,35,37,39,55,
                          56,63,67,70,74,80,
                          93:98,101,103:105)]

### 2.4.2 Valores Missing (NA)
# Tabela 1
colnames(basetotal)[colSums(is.na(basetotal)) > 0]

# Tabela 2
table(basetotal[rowSums(is.na(
  basetotal)) > 0,]$G.x)

# Tabela 3
for (i in 1:10) {
  print(length(filter(basetotal,
                      G.x >= i)$Player))
}

# Reintroduzindo X3PA.x e FTA.x
basefinal <- basetotal[-c(8,9,10,14,16,18,21,24,
                          28,30,33,35,37,39,55,
                          56,63,67,70,74,80,
                          93:98,101,103:105)]
```

```
basefinal <- filter(basefinal, G.x >= 8)

### 2.5.1 Análise de Componentes Principais
pca <- prcomp(basefinal[, c(5:74)],
              center = T, scale = T)

# Tabela 4
summary(pca)

### 3.1 Análise Descritiva: Salários
# Tabela 6
summary(basetotal$Salary)

# Figura 2
ggplot(basetotal, aes(x="", y=Salary)) +
  geom_boxplot(fill="gray", width=0.2) +
  labs(title="Boxplot de Salários",
       x="", y="Salário (em dólares)") +
  theme_classic() +
  scale_y_continuous(
    labels = comma_format(big.mark = ".",
                          decimal.mark = ","))

# Figura 3
ggplot(basetotal, aes(x=Salary)) +
  geom_histogram(fill="gray",bins=25) +
  labs(title="Histograma de Salários",
       x="Salário (em dólares)",
       y="Frequência") +
  theme_classic() +
  scale_x_continuous(
    labels = comma_format(big.mark = ".",
                          decimal.mark = ","))

# Tabela 7
summary(basetotal$logSalary)

# Figura 4
```

```
ggplot(basetotal, aes(x=logSalary)) +
  geom_histogram(fill="gray",bins=25)+
  labs(title="Histograma de log-Salários",
       x="log-Salário", y="Frequência")+
  theme_classic() +
  scale_x_continuous(
    labels = comma_format(big.mark = ".",
                          decimal.mark = ","))

### 3.2 Escolha do Modelo
m1 <- lm(logSalary ~ 1,
         data = basefinal[-c(1:4)])

m2 <- lm(logSalary ~ .,
         data = basefinal[-c(1:4)])

stepAIC(m2, direction = "backward")

stepAIC(m1, direction = "forward",
        scope=list(upper=m2,lower=m1))

### 3.3 Regressão Lasso
X <- as.matrix(basefinal[c(5:74)])
Y <- as.matrix(basefinal[75])

l <- glmnet(X, Y, alpha = 1,
               standardize = T)

# Tabela 8
print(l)

# Tabela 9
coef(l, s = 0.08637)

# Figura 6
grafico <- function(variavel, nome_x) {
  plot(logSalary ~ variavel, data = basefinal,
       pch = 16,
       main = paste("logSalários por", nome_x),
```

```
        xlab = paste(nome_x),
        ylab = "logSalários")

  abline(lm(logSalary ~ variavel,
            data = basefinal), col = 2,
         lty = 1, lwd = 1)

  abline(rq(logSalary ~ variavel,
            data = basefinal, tau = 0.1),
         col = 3, lty = 1, lwd = 2)
  abline(rq(logSalary ~ variavel,
            data = basefinal, tau = 0.25),
         col = 4, lty = 1, lwd = 2)
  abline(rq(logSalary ~ variavel,
            data = basefinal, tau = 0.5),
         col = 5, lty = 1, lwd = 1)
  abline(rq(logSalary ~ variavel,
            data = basefinal, tau = 0.75),
         col = 6, lty = 1, lwd = 2)
  abline(rq(logSalary ~ variavel,
            data = basefinal, tau = 0.9),
         col = 7, lty = 1, lwd = 2)
}

grafico(basefinal$Age.x, "Idade␣do␣Jogador")
grafico(basefinal$GS, "Jogos␣como␣Titular")
grafico(basefinal$MP.1,
        "Minutos␣Jogados␣p/␣Jogo")
grafico(basefinal$FG.1,
        "Arremessos␣Certos␣p/␣Jogo")
grafico(basefinal$DRB.1,
        "Rebotes␣Defensivos␣p/␣Jogo")
grafico(basefinal$STL.1,
        "Roubos␣de␣Bola␣p/␣Jogo")
grafico(basefinal$TOV.1,
        "Perdas␣de␣Posse␣p/␣Jogo")
grafico(basefinal$PTS.1,
        "Pontos␣Marcados␣p/␣Jogo")
grafico(basefinal$ORtg, "Offensive␣Rating")
```

```
### 3.4 Resultados da Estimação (Tabela 10)
# MQO
# Estimativas e p-valores do teste
mqo <- lm(logSalary ~ Age.x + GS + MP.1 +
            FG.1 + DRB.1 + STL.1 + TOV.1 +
            PTS.1 + ORtg,
          data = basefinal)
summary(mqo)

# Intervalo de confiança
round(confint(mqo, level = .95), 3)

# Regressão Quantílica
# Estimativas e intervalos de confiança
quantile_ci <- function(quantile) {
  fit <- rq(logSalary ~ Age.x + GS + MP.1 +
              FG.1 + DRB.1 + STL.1 + TOV.1 +
              PTS.1 + ORtg,
            tau = quantile, data = basefinal)
  est <- round(summary(
    fit, se = 'boot')$coefficients[,1],3)
  li <- round(summary(
    fit, alpha = .05)$coefficients[,2],3)
  ls <- round(summary(
    fit, alpha = .05)$coefficients[,3],3)
  list(Estimativa = est, LI = li, LS = ls)
}

for (i in c(0.1, 0.25, 0.5, 0.75, 0.9)) {
  print(i)
  print(quantile_ci(i))
}

# p-valores do teste
quantile_pval <- function(quantile) {
  set.seed(160013224)
  fit <- rq(logSalary ~Age.x + GS + MP.1 +
              FG.1 + DRB.1 + STL.1 + TOV.1 +
```

```
                PTS.1 + ORtg,
              tau = quantile, data = basefinal)
  pval <- round(summary(
    fit, se = 'boot')$coefficients[,4],3)
  list("p-valor" = pval)
}

for (i in c(0.1, 0.25, 0.5, 0.75, 0.9)) {
  print(i)
  print(quantile_pval(i))
}

# Figura 7
fit1 <- summary(rq(logSalary ~Age.x + GS + MP.1 +
                     FG.1 + DRB.1 + STL.1 + TOV.1 +
                     PTS.1 + ORtg, tau = 2:98/100,
                   data = basefinal), alpha = .05)

plot(fit1,c(2:10), type = "l", mfrow = c(3,3))
```