Universidade Estadual de Campinas
Instituto de Matemática, Estatística e Computação Científica (IMECC)

# Modelos de regressão aplicados a dados referentes à temporada regular e pós temporada da *National Basketball Association* (NBA)

**Bolsista:**
Rubens Cortelazzi Roncato 236292

**Orientador:**
Rafael Pimentel Maia

**Campinas - SP**
**Setembro de 2023 - Agosto de 2024**

# 1 Introdução

A estatística desempenha um papel crucial em diversas áreas, incluindo a ciência dos esportes, auxiliando na contratação de jogadores e no *scouting*. No basquete, por exemplo, Giovanini et al (2014) analisaram como a pressão do jogo afeta a seleção de arremessos na NBA.

A *National Basketball Association* (NBA), uma das principais ligas de basquete do mundo, é composta por 30 times dos EUA e do Canadá, que jogam 82 partidas por temporada, com exceções durante a pandemia e por questões contratuais. Após a temporada regular, é realizada a pós temporada ou *playoffs* que determinam o campeão, com séries de melhor de sete jogos, sendo jogados pelos melhores 8 times de cada conferência (Leste e Oeste).

Os *playoffs* são a parte principal de toda a temporada, Morgado (2022) discute se há vantagens de se jogar em casa, ou seja, na cidade do time mandante nos *playoffs* da NBA. Na sua análise foram tomados como base os anos de 1946 até 2021 e o trabalho encontrou evidências de que existe vantagem de jogar em casa nos *playoffs*.

Diversas características podem influenciar uma vitória em jogo de NBA. Ajustar um modelo de regressão para modelar as variáveis preditoras com relação à variável resposta, neste caso, a proporção de vitórias é de extrema importância para compreender as decisões tomadas dentro de quadra. Além disso, nos *playoffs* a intensidade aumenta, o que pode afetar o desempenho dos jogadores e, consequentemente, os resultados.

Maciel (2021) desenvolveu um trabalho relacionando a regressão linear múltipla para determinar quais estatísticas do jogo de basquete apresentaram associação significativa com a quantidade total de vitórias dos times da NBA. Além disso, foi analisada apenas a temporada regular da NBA, sem fazer uma análise com relação à pós-temporada. Dessa forma, não sendo possível fazer uma comparação entre as variáveis preditoras que mais influenciam nos dois períodos da competição.

Para o presente trabalho, será apresentada a técnica de modelagem de regressão aplicada às estatísticas de 15 temporadas regulares da NBA (2008-2023) e com foco na temporada regular e na pós temporada. A variável de interesse é a proporção de vitórias por temporada ou pós temporada.

# 2 Materiais e Métodos

Foram utilizados dois bancos de dados, um para temporada regular e outro para os *playoffs*. As bases de dados analisadas possuem 31 variáveis, em que elas estão definidas abaixo, juntamente com uma explicação.

. **Posição** - Posição em que o time terminou na classificação geral da temporada;
. **TEAM** - Times que disputaram a temporada;
. **GP** – Quantidade de jogos jogados durante a temporada;
. **W** – Quantidade de vitórias ao longo da temporada;
. **L** - Quantidade de derrotas ao longo da temporada;
. **WINP** - Porcentagem de vitórias ao decorrer da temporada;
. **MIN** - Número médio de minutos jogados por partida;
. **PTS** – Número médio de pontos por partida na temporada;

. **FGM** - Número médio de arremessos de quadra acertados durante a temporada;
. **FGA** - Número médio de arremessos de quadra tentados durante a temporada;
. **FGP** - Porcentagem de arremessos de quadra acertados durante a temporada;
. **3PM** - Número médio de arremessos de 3 pontos acertados durante a temporada;
. **3PA** - Número médio de arremessos de 3 pontos tentados durante a temporada;
. **3PP** - Porcentagem de arremessos de 3 pontos acertados durante a temporada;
. **FTM** - Número médio de arremessos de lances livres acertados durante a temporada;
. **FTA** - Número médio de arremessos de lances livres tentados durante a temporada;
. **FTP** - Porcentagem de arremessos de lances livres acertados durante a temporada;
. **OREB** - Número médio de rebotes ofensivos por partida durante a temporada;
. **DREB** - Número médio de rebotes defensivos por partida durante a temporada;
. **REB** - Número médio de rebotes por partida durante a temporada;
. **AST** - Quantidade média de assistências por partida durante a temporada;
. **TOV** - Quantidade média de *turnovers* por partida durante a temporada;
. **STL** - Quantidade média de roubadas de bola por partida durante a temporada;
. **BLK** - Quantidade média de tocos por partida durante a temporada;
. **BLKA** - Quantidade média de tocos tomados por partida durante a temporada;
. **PF** - Quantidade média de faltas pessoais feitas por partida durante a temporada;
. **PFD** - Quantidade média de faltas pessoais sofridas por partida durante a temporada;
. **PlusMinus** - Média de *Plus/Minus* por partida durante a temporada;
. **Temporada** - Temporada em que foi realizado o jogo, que vai de 2008 até 2023.
. **Conferência** - Conferência que o time pertence, podendo ser leste ou oeste, dependendo da localização geográfica do time.

Com as 31 variáveis mostradas anteriormente, usamos para modelar a regressão 25 dessas variáveis, sendo que Posição, GP, W, L, MIN e Conferência não utilizamos para fazer a modelagem dos dados. Utilizamos essas variáveis em algumas análises descritivas e análises complementares, que ajudaram a entender de uma melhor forma o banco de dados analisado.

## 2.1 Metodologia

Análise de regressão é uma metodologia estatística que utiliza a relação entre duas ou mais variáveis de maneira que uma variável pode ser predita pela(s) outra(s). É uma metodologia amplamente utilizada em diversas áreas, pois é de relativamente fácil interpretação. Assim, serão apresentadas as metodologias que serviram de base para ser feita a modelagem.

### 2.1.1 Regressão Linear Múltipla

Gareth, James et al (2013) discutem que a técnica de regressão linear múltipla é utilizada para modelar a relação linear entre variáveis preditoras e à variável resposta, controlando pelas demais variáveis no modelo. O modelo geral com mais que duas variáveis preditoras é dado por:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_{p-1} X_{i,p-1} + \epsilon_i$$

para $i$ = 1,2, ..., n e $p$ o número de regressores, sendo $\beta_0$, $\beta_1$, ..., $\beta_{p-1}$ os parâmetros, $X_{i1}$, ..., $X_{i,p-1}$ constantes conhecidas e $\epsilon_i \sim N(0, \sigma^2)$.

Alguns pressupostos devem ser respeitados para poder ocorrer de forma válida a modelagem dos dados segundo esse método de análise, sendo eles homoscedasticidade (variância constante), ausência de multicolinearidade entre as variáveis explicativas (a correlação entre elas não pode estar perto de uma correlação perfeita), independência de erros e independência das variáveis preditoras.

Além disso, uma medida de ajuste do modelo que será utilizada pelas outras metodologias é o coeficiente de determinação ($R^2$). Sendo ela a proporção da variância explicada, assumindo um valor entre 0 e 1, e independente da escala de Y. O $R^2$ é calculado da seguinte forma: $R^2 = \frac{TSS-RSS}{TSS}$ sendo $TSS = \sum(y_i - \bar{y})^2$ e $RSS = \sum(y_i - \hat{y}_i)^2$, com $\bar{y}$ sendo a média da variável resposta e $\hat{y}_i$ sendo a estimativa do $y_i$.

### 2.1.2 Regressão Beta

O modelo de regressão beta, é um método estatístico aplicável quando os valores da variável resposta estão no intervalo (0,1). Nessa abordagem, a variável dependente segue uma distribuição beta, e sua média é relacionada a um conjunto de regressores por meio de um preditor linear com coeficientes desconhecidos e uma função de ligação. O modelo também considera um parâmetro de precisão, que pode ser constante ou depender de um conjunto diferente de regressores através de outra função de ligação.

Além disso, a regressão beta é útil para lidar com características como heterocedasticidade e assimetria, frequentemente presentes em dados que variam dentro do intervalo unitário padrão, como taxas ou proporções. Este modelo foi introduzido por Ferrari e Cribari-Neto em 2004. Quando a variável resposta inclui os extremos 0 e 1, uma transformação prática é $(y \cdot (n-1) + 0.5)/n$, em que n representa o tamanho da amostra (Smithson e Verkuilen, 2006).

Ferrari e Cribari-Neto (2004) propuseram uma parametrização para a função densidade da distribuição beta definindo $\mu = p/(p+q)$ e $\phi = p + q$ com

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}$$

com $0 < y < 1$, $0 < \mu < 1$ e $\phi > 0$. Sendo que escrevemos $y \sim \mathrm{B}(\mu, \phi)$.

Na parametrização de Ferrari e Cribari-Neto (2004) $E(y) = \mu$ e $Var(y) = \frac{\mu(1-\mu)}{1+\phi}$. Assim, o parâmetro $\phi$ é conhecido como parâmetro de precisão, pois para $\mu$ fixo, quanto maior $\phi$ menor é a variância de y. Além disso, $\phi^{-1}$ é um parâmetro de dispersão.

Assuma que $y_1, ..., y_n$ seja uma amostra aleatória, em que $y_i \sim \mathrm{B}(\mu_i, \phi)$, $i = 1, ..., n$. O modelo de regressão beta é definido como:

$$g(\mu_i) = x_i^T \beta = \eta_i$$

em que $\beta = (\beta_1, ..., \beta_k)^T$ é um vetor $k \times 1$ de parâmetros de regressão desconhecidos ($k < n$), $x_i = (x_{i1}, ..., x_{ik})^T$ é um vetor de k regressores (ou variáveis independentes ou covariáveis) e $\eta_i$ é um preditor linear (por exemplo, $\eta_i = \beta_1 x_{i1} + ... + \beta_k x_{ik}$, normalmente $x_{i1} = 1$ para todo i então o modelo tem intercepto).

Nesse caso, $g(.) : (0,1) \mapsto \mathbb{R}$ é uma função de ligação, que é estritamente crescente e duas vezes diferenciável. Assim, temos dois motivos para utilizar uma função de ligação na estrutura

da regressão, sendo que primeiro ambos os lados da equação de regressão assumem valores na reta real quando uma função de ligação é aplicada a $\mu_i$. O segundo motivo é que há a flexibilidade adicional, uma vez que o profissional pode escolher a função que produz o melhor ajuste.

Algumas funções de ligação úteis são: logito $g(\mu) = \log(\mu/(1-\mu))$; probito $g(\mu) = \Phi^{-1}(\mu)$, em que $\Phi(.)$ é uma função de distribuição normal padrão; complementarmente log-log $g(\mu) = \log\{-\log(1-\mu)\}$; log-log $g(\mu) = -\log\{-\log(\mu)\}$ e Cauchy $g(\mu) = tan\{\pi(\mu - 0.5)\}$.

### 2.1.3 Modelos Aditivos Generalizados para posição, escala e forma (Gamlss)

Os modelos aditivos generalizados para posição, escala e forma (GAMLSS) foram propostos por Rigby e Stasinopoulos (2005). Esses modelos expandem os modelos lineares generalizados (MLG), permitindo a análise de distribuições além da família exponencial, como a distribuição beta e a distribuição beta inflacionada em 0.

Rigby e Stasinopoulos (2005) descrevem que os modelos GAMLSS assumem que a variável resposta Y segue uma função de densidade paramétrica $D(\mu, \sigma, \nu, \tau)$ em que $\mu$ e $\sigma$ são os parâmetros de localização e escala, respectivamente, e $\nu$ e $\tau$ são os parâmetros de forma, associados ao viés e à curtose da distribuição.

Cada componente do vetor de parâmetros pode ser modelada de forma independente, utilizando funções lineares, não lineares ou de suavização das variáveis preditoras. Um modelo GAMLSS é representado como $Y \sim D(\mu, \sigma, \nu, \tau)$, com funções de ligação relacionando os parâmetros da distribuição às variáveis preditoras: $\eta_1 = g_1(\mu) = X_1\beta_1 + \sum_{j=1}^{\alpha_1} Z_{j1}\gamma_{j1}$, $\eta_2 = g_2(\sigma) = X_2\beta_2 + \sum_{j=1}^{\alpha_2} Z_{j2}\gamma_{j2}$, $\eta_3 = g_3(\nu) = X_3\beta_3 + \sum_{j=1}^{\alpha_3} Z_{j3}\gamma_{j3}$ e $\eta_4 = g_4(\tau) = X_4\beta_4 + \sum_{j=1}^{\alpha_4} Z_{j4}\gamma_{j4}$.

$g_k(.)$ (para k = 1,2,3,4) são funções de ligação monótonas, $\mu, \sigma, \nu, \tau$ e $\eta_k$ são os vetores n-dimensionais, $\beta_k^T$ é o vetor de parâmetros, $X_k$ é a matriz de desenho associada ao vetor de efeitos fixos, $Z_{jk}$ é uma matriz de desenho fixa e por fim $\gamma_{jk}$ é uma variável aleatória.

A distribuição beta e a beta inflacionada em 0 usadas nos modelos possuem os seguintes parâmetros: $\mu, \sigma$ e $\mu, \sigma, \nu$, respectivamente. Vale destacar que Stasinopoulos e Rigby (2008) desenvolveram o pacote GAMLSS no hyperlinkRR (R Team 2024), que será utilizado. Comparando modelos GAMLSS para a distribuição normal com a regressão linear no R, encontramos resultados semelhantes, variando apenas nas estimativas dos parâmetros devido aos métodos de estimação empregados.

### 2.1.4 Modelos Lineares Mistos

ZUUR, Alain F. et al. (2009) oferecem uma extensa base sobre modelos mistos, destacando que os Modelos Lineares Mistos (LMM) são uma generalização dos modelos lineares que relaxam a premissa de independência entre as observações. Essa falta de independência pode surgir em experimentos por diversos motivos, sendo útil para controlar fatores de confusão.

No banco de dados da NBA, a dependência entre as observações ocorre porque o mesmo time é observado por 15 temporadas e há 30 observações por temporada. Assim, é esperado que um time tenha características mais similares ao longo das temporadas, compartilhando muitas condições associadas ao mesmo contexto.

Embora não estejamos interessados nas diferenças entre esses grupos, é necessário abordar a falta de independência. Os LMMs tratam dessa questão ao incorporar uma estrutura aleatória

nos modelos lineares, acomodando agrupamentos como variáveis categóricas preditoras aleatórias. Diferente das variáveis categóricas fixas, onde interpretamos as diferenças entre níveis, nas aleatórias queremos apenas estimar a variabilidade associada aos agrupamentos.

Os LMMs combinam variáveis fixas e aleatórias como preditoras. A estrutura aleatória pode influenciar o intercepto ou a inclinação do modelo. Quando influencia o intercepto, estimamos interceptos para as categorias a partir de uma distribuição normal definida por um intercepto médio e um desvio padrão.

A expressão geral para um modelo com uma preditora e uma variável aleatória afetando apenas o intercepto é: $y_{ij} = (\hat{\alpha} + \epsilon_j) + \beta x_{ij} + \epsilon_{ij}$, com $\epsilon_j = N(0, \sigma_{entre})$ e $\epsilon_{ij} = N(0, \sigma_{intra})$. O $\hat{\alpha}$ é o intercepto médio e $\sigma_{entre}$ é a estimativa do desvio padrão associado à distribuição de interceptos para a variável aleatória.

Para a modelagem dos dados no R (R Team 2024), utilizaremos dois pacotes: GAMLSS (Rigby et al., 2024) para a distribuição beta e lme4 (Bates et al., 2024) para a distribuição normal.

### 2.1.5 Métodos de escolha dos melhores modelos

Existe a possibilidade de conseguirmos selecionar as melhores variáveis para os modelos de forma automática que Gareth, James et al (2013) trazem que são: *Backward*, *Forward* e *Mixed selection*.

No método *Forward*, iniciamos com um modelo vazio, contendo apenas o intercepto, e adicionamos uma variável por vez. Selecionamos a variável que, quando adicionada, resulta no menor Critério de Informação de Akaike (AIC), que avalia a qualidade do modelo. Já no método *Backward*, começamos com um modelo completo, contendo todas as covariáveis, e removemos uma variável por vez. A variável removida é a que, ao ser excluída, resulta no menor AIC. Ambos os métodos serão utilizados para identificar os melhores modelos para análise.

Para os modelos de regressão beta, a seleção automática de variáveis não está implementada, sendo assim necessária outra forma de realizar a seleção de modelos. Utilizaremos o teste de razão de verossimilhança para verificar a significância da inclusão de cada variável no modelo. Esse teste também será aplicado para avaliar a relevância da inclusão de efeitos aleatórios.

### 2.1.6 Validação cruzada

Validação cruzada ou *Cross Validation* (CV) é um método utilizado para modelos de predição que incluem técnicas de aprendizado de máquina. O CV consiste em particionar os dados em conjuntos (partes), onde um conjunto é utilizado para treino e outro conjunto é utilizado para teste e avaliação do desempenho do modelo. A utilização do CV tem altas chances de detectar se o seu modelo está sobreajustado aos seus dados de treinamento, ou seja, sofrendo *overfitting*. Existem mais de um método de aplicação de CV, como métodos *K-fold*, *Leave one out*, entre outros.

O método de *Cross Validation*, que iremos utilizar, é o método *k-fold* que está descrito em Gareth, James et al (2013). O método que iremos analisar envolve dividir aleatoriamente o conjunto de observações em k grupos, ou *folds*, de tamanho aproximadamente igual. O primeiro *fold* é tratado como um conjunto de validação, e o método é ajustado nos k - 1 *folds* restantes.

O erro quadrático médio, MSE1, é então computado nas observações no *fold* mantido de fora. Este procedimento é repetido k vezes, cada vez um grupo diferente de observações é tratado como um conjunto de validação.

Esse processo resulta em k estimativas do erro de teste, MSE1, MSE2,..., MSEk. A estimativa de *k-fold* CV é calculada pela média desses valores, $CV_{(k)} = \frac{1}{k} \times \sum_{i=1}^{k} MSE_i$. Assim, o melhor modelo é aquele que apresenta menor RMSE e MAE, que são medidas de variabilidade dos modelos e maior $R^2$.

# 3 Resultados

Com as variáveis que utilizamos, explicadas anteriormente, o entendimento das análises ficará mais claro e caso de dúvida da variável que está sendo analisada, é possível reler o que foi dito na explicação. Dessa forma, ilustramos na Tabela 1 a análise descritiva da temporada regular, que contém o mínimo, 1°quartil, mediana, média, 3° quartil, máximo e variância, das 450 observações obtidas.

Tabela 1: Medidas descritivas de cada variável da temporada regular

| Variável | Mínimo | 1° quartil | Mediana | Média | 3° quartil | Máximo | Desvio Padrão |
|---|---|---|---|---|---|---|---|
| WINP | 0.11 | 0.39 | 0.51 | 0.50 | 0.61 | 0.89 | 0.15 |
| PTS | 87.00 | 98.92 | 104.05 | 104.68 | 110.62 | 120.70 | 6.99 |
| FGM | 33.20 | 37.20 | 38.70 | 38.90 | 40.60 | 44.70 | 2.29 |
| FGA | 75.80 | 82.00 | 85.10 | 84.85 | 87.58 | 94.40 | 3.75 |
| FGP | 40.80 | 44.70 | 45.80 | 45.84 | 46.90 | 50.40 | 1.58 |
| 3PM | 3.80 | 6.80 | 9.00 | 9.23 | 11.38 | 16.70 | 2.82 |
| 3PA | 11.30 | 19.30 | 25.20 | 25.79 | 31.80 | 45.40 | 7.59 |
| 3PP | 29.50 | 34.52 | 35.60 | 35.68 | 36.90 | 41.60 | 1.80 |
| FTM | 12.20 | 16.40 | 17.55 | 17.65 | 18.70 | 24.10 | 1.78 |
| FTA | 16.60 | 21.40 | 23.00 | 23.08 | 24.50 | 31.10 | 2.30 |
| FTP | 66.00 | 74.70 | 76.75 | 76.53 | 78.50 | 83.90 | 2.99 |
| OREB | 7.60 | 9.70 | 10.50 | 10.57 | 11.40 | 14.60 | 1.22 |
| REB | 36.90 | 41.70 | 43.10 | 43.18 | 44.50 | 51.70 | 2.15 |
| DREB | 27.20 | 31.10 | 32.70 | 32.61 | 34.10 | 42.20 | 2.16 |
| AST | 17.40 | 21.10 | 22.70 | 22.85 | 24.30 | 30.40 | 2.30 |
| TOV | 11.10 | 13.50 | 14.20 | 14.24 | 15.00 | 17.70 | 1.09 |
| STL | 5.50 | 7.00 | 7.50 | 7.58 | 8.20 | 10.00 | 0.83 |
| BLK | 2.40 | 4.30 | 4.80 | 4.86 | 5.30 | 8.20 | 0.73 |
| BLKA | 3.00 | 4.40 | 4.80 | 4.86 | 5.30 | 6.90 | 0.70 |
| PF | 16.60 | 19.20 | 20.30 | 20.24 | 21.20 | 24.80 | 1.41 |
| PFD | 16.20 | 19.40 | 20.20 | 20.23 | 21.00 | 24.30 | 1.31 |
| *Plus/Minus* | -13.90 | -3.20 | 0.30 | -0.01 | 3.30 | 11.60 | 4.65 |

Por outro lado, a Tabela 2 trata das mesmas informações que foram analisadas na temporada regular, porém agora para a pós temporada ou os *playoffs*, que possui 240 observações.

Tabela 2: Medidas descritivas de cada variável para os *playoffs*

| Variável | Mínimo | 1° quartil | Mediana | Média | 3° quartil | Máximo | Desvio Padrão |
|---|---|---|---|---|---|---|---|
| WINP | 0.00 | 0.30 | 0.43 | 0.40 | 0.55 | 0.94 | 0.2 |

| Variável | Mínimo | 1° quartil | Mediana | Média | 3° quartil | Máximo | Desvio Padrão |
|---|---|---|---|---|---|---|---|
| PTS | 78.00 | 93.80 | 101.10 | 100.60 | 107.70 | 119.50 | 8.83 |
| FGM | 27.80 | 34.30 | 36.70 | 36.63 | 38.83 | 44.20 | 3.13 |
| FGA | 69.00 | 78.47 | 82.85 | 82.55 | 86.03 | 96.80 | 5.36 |
| FGP | 38.20 | 42.70 | 44.50 | 44.38 | 46.20 | 50.30 | 2.62 |
| 3PM | 2.30 | 6.60 | 8.80 | 9.07 | 11.30 | 18.00 | 3.07 |
| 3PA | 10.80 | 19.30 | 25.55 | 26.17 | 32.73 | 46.80 | 7.89 |
| 3PP | 20.00 | 32.08 | 34.40 | 34.39 | 37.30 | 46.50 | 3.82 |
| FTM | 11.30 | 16.38 | 18.30 | 18.29 | 20.02 | 28.20 | 2.96 |
| FTA | 14.40 | 21.30 | 23.70 | 23.95 | 26.23 | 36.80 | 3.81 |
| FTP | 60.80 | 73.08 | 76.70 | 76.47 | 79.92 | 88.60 | 4.90 |
| OREB | 5.20 | 8.90 | 10.00 | 10.19 | 11.60 | 16.70 | 1.96 |
| REB | 33.50 | 39.77 | 42.05 | 42.01 | 44.30 | 51.90 | 3.44 |
| DREB | 24.20 | 30.00 | 31.70 | 31.83 | 33.80 | 41.80 | 2.90 |
| AST | 12.60 | 18.30 | 20.45 | 20.56 | 22.52 | 28.40 | 2.98 |
| TOV | 8.40 | 12.38 | 13.25 | 13.45 | 14.50 | 19.00 | 1.74 |
| STL | 3.40 | 6.20 | 6.90 | 6.98 | 7.80 | 10.80 | 1.16 |
| BLK | 2.10 | 3.80 | 4.60 | 4.68 | 5.50 | 8.10 | 1.26 |
| BLKA | 2.30 | 4.08 | 4.80 | 5.08 | 6.00 | 10.30 | 1.41 |
| PF | 15.50 | 20.38 | 21.55 | 21.71 | 22.90 | 30.80 | 2.22 |
| PFD | 14.80 | 20.00 | 21.30 | 21.43 | 23.00 | 27.80 | 2.19 |
| *Plus/Minus* | -24.20 | -6.80 | -1.65 | -2.74 | 1.80 | 13.50 | 6.72 |

Com as informações passadas nas Tabelas, podemos notar na variável resposta (WINP) que o mínimo e o máximo da temporada regular foram atingidos pelo extinto Charlotte Bobcats na temporada 2011/2012 e pelo Golden State Warriors em 2015/2016, respectivamente. Interessante citar que na temporada 2015/2016, o Golden State Warriors ganhou 73 jogos, recorde de vitórias em uma única temporada, porém ao final não foi o time campeão da liga, perdendo para o Cleveland Cavaliers.

Podemos observar na Figura 1 a distribuição da variável resposta, porcentagem de vitórias na temporada durante a temporada regular. Assim, o primeiro quartil dos dados de temporada regular está entre 11 e 39% de vitórias na temporada. Por outro lado, o último quartil está entre 61 e 89% de vitórias na temporada, podendo se perceber uma simetria na distribuição. Já na Figura 1, olhando agora para os boxplots, podemos notar que existem apenas três *outliers* nas temporadas, sendo que a porcentagem mínima de vitórias, que aconteceu na temporada 2011/2012, não foi considerada *outlier*.

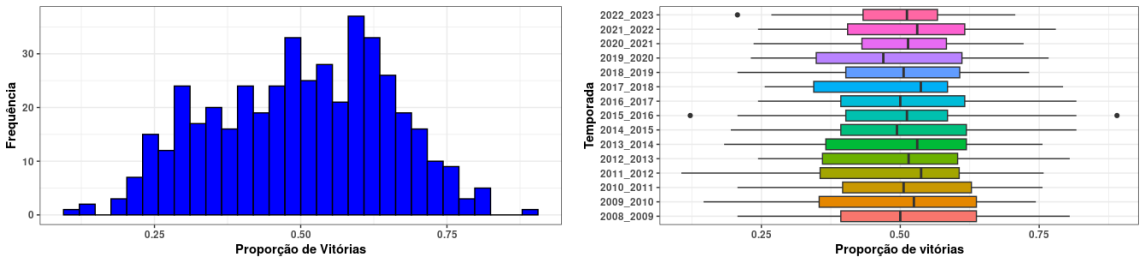


Figura 1: Histograma da proporção de vitórias na temporada regular e *Boxplot* da proporção de vitórias por temporada.

A Figura 2 mostra a distribuição da porcentagem de vitórias nos playoffs, podemos notar,

no histograma, que existem valores nas extremidades, ou seja, houve times que perderam todas as partidas que jogaram, no basquete isto é chamado de "varrida". Além disso, podemos notar que 75% dos dados presentes apresentam porcentagem de vitórias inferior a 55% de vitórias. Também, interessante citar que não houve nenhum time que não perdeu nenhum jogo, sendo que na temporada 2016/2017 houve o máximo de vitória que foi de 94% dos jogos vencidos.

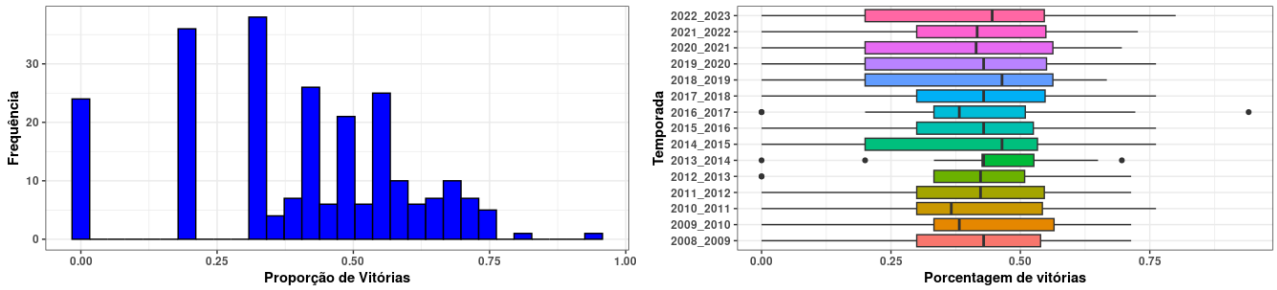


Figura 2: Histograma da proporção de vitórias nos *playoffs* e *Boxplot* da proporção de vitórias por temporada

Também, a quantidade mínima de pontos nos 15 anos analisados ocorreu nos *playoffs* e não na temporada regular, diferentemente do que seria pensado, pois na pós-temporada a competitividade é maior e existe um maior equilíbrio entre os times. Além disso, sobre pontuação, nas temporadas, podemos perceber pela Figura 3, que a média de pontos da temporada regular é maior do que a dos *playoffs* em aproximadamente 5 pontos ao longo dos 15 anos.

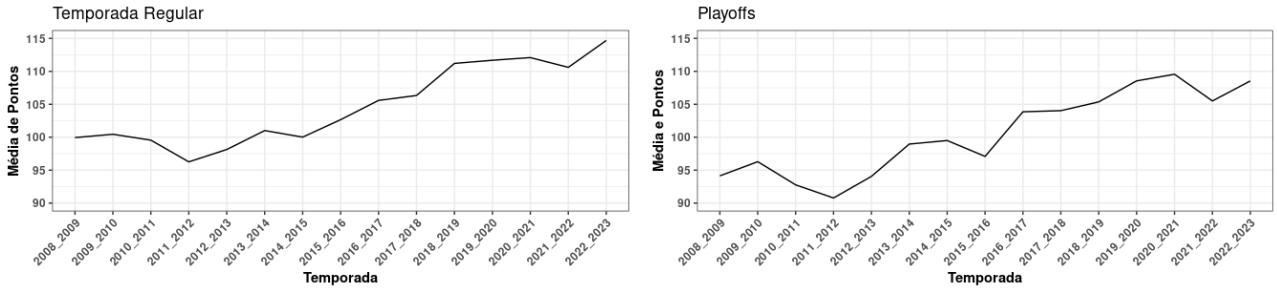


Figura 3: Média de Pontos por temporada da Temporada Regular e *playoffs*, respectivamente

Outra questão que, na atualidade do basquete da NBA, vem ganhando bastante notoriedade são as bolas de 3 pontos, que vêm revolucionando o basquete de antigamente. Sendo que no passado da liga os jogadores tinham maior tendência a arremessar bolas de 2 pontos, porém na atualidade a mentalidade vem sendo mudada. Um jogador que ajudou a alterar o modo de ver o jogo é Stephen Curry, que se tornou a pessoa com mais cestas de 3 pontos na NBA, já ultrapassando a marca de 3.000 bolas de 3 pontos acertadas.

Para mostrar em números o que foi falado anteriormente, temos que, dos arremessos tentados em quadra, 30.16% são de bolas de 3 pontos, enquanto de lances livres é de 27.29%. Por outro lado, dos arremessos certos, 23.51% deles são de bolas de 3 pontos. Assim, vamos ilustrar com a Figura 4 as bolas de 3 pontos durante as temporadas regulares.

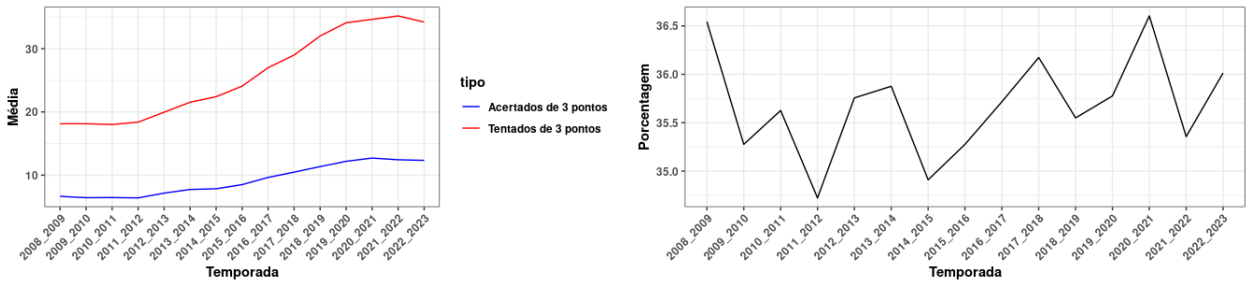


Figura 4: Gráficos de arremessos de 3 pontos na temporada regular

Já na Figura 5, iremos observar os arremessos de 3 pontos referente a pós-temporada da NBA. Assim, em ambos os gráficos notamos que os arremessos tentados e acertados têm uma tendência crescente, enquanto na porcentagem de 3 pontos não possui tendência, sendo que a quantidade de arremessos e acertos estão aumentando, porém a média de acertos não sofre uma diferença significativa. Ou seja, ao mesmo tempo que os jogadores estão arremessando mais, eles também estão acertando mais. Sendo que observaremos mais para frente a correlação entre essas duas variáveis.

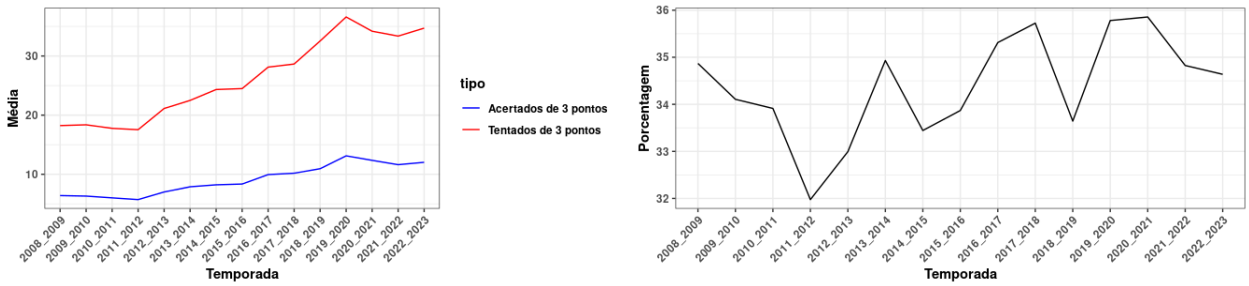


Figura 5: Gráficos de arremessos de 3 pontos nos *playoffs*

Com a parte de bolas de 3 pontos analisada, foi feito o desenvolvimento a respeito de uma análise de correlação de Pearson entre as variáveis. Na temporada regular a variável resposta (WINP) obteve correlação forte com *Plus/Minus* sendo de 0.97, enquanto nos *playoffs* as mesmas variáveis obtiveram uma correlação de 0.86. Além disso, na Tabela 3 observamos as correlações fortes, com valores maiores que 0.80.

Tabela 3: Correlações maiores que 0.80 entre as variáveis para os *playoffs* e temporada regular

| Variáveis | PFD/FTA | FTA/FTM | PTS/FGM | FGM/FGA | PTS/3PM | 3PA/3PM | PTS/3PA |
|---|---|---|---|---|---|---|---|
| Regular | 0.848 | 0.921 | 0.931 | 0.813 | 0.836 | 0.987 | 0.815 |
| *playoffs* | 0.866 | 0.920 | 0.907 | 0.718 | 0.745 | 0.951 | 0.683 |

Além disso, para os *playoffs* notamos que as correlações entre a variável resposta e as covariáveis se comportam de uma maneira estranha nos valores inferiores. Observando a Figura 6, notamos como está distribuído a relação entre a variável resposta (WINP_transformado, pois usamos a transformação da variável resposta citada em 2.1.2) e as variáveis preditoras, com

exceção de TEAM que é uma variável categórica e também notamos as correlações entre as variáveis. Dessa forma, percebemos que a relação entre a variável resposta e REB e *Plus/Minus*, nas pequenas porcentagens são retas verticais, ou seja, existem quando um time perde todas as partidas nos *playoffs*, ele joga apenas 4 jogos e fica com porcentagem de vitórias igual a 0, não tendo uma predição muito bem estabelecida nessa parte já que os valores estão muito dispersos.

Essa relação anterior vale também para quando o time ganha apenas um jogo, então necessariamente ele jogou 5 jogos, ou seja, tendo porcentagem de vitória de 20%. Isso vale para duas vitórias apenas, com 6 jogos, sendo 33,3% de vitórias. Assim por diante também vale a mesma relação dita anteriormente, levando a conclusão que é mais difícil fazer a predição dos *playoffs* do que a predição da temporada regular.

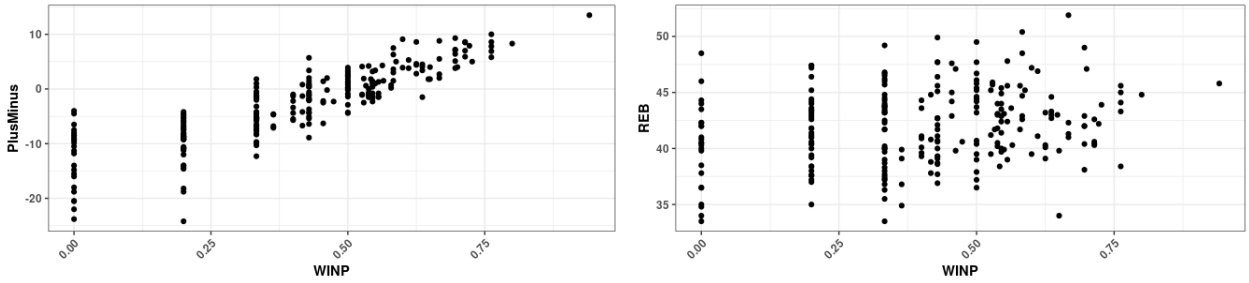


Figura 6: Relação entre WINP e as variáveis *Plus/Minus* e REB, respectivamente nos *playoffs*

Para finalizar, iremos fazer uma breve análise sobre aparições em *playoffs*, sendo que a conferência leste possui os times com mais aparições nos *playoffs* nesses últimos 15 anos, sendo o Boston Celtics, Atlanta Hawks e Miami Heat, com 14, 12 e 12 aparições, respectivamente. Por outro lado, a conferência oeste possui o Portland Trail Blazers, San Antonio Spurs e Dallas Mavericks com 11, 11 e 10 aparições em *playoffs*, respectivamente.

Os times que tiveram menos aparições em *playoffs* foram o Sacramento Kings pela conferência oeste com apenas uma classificação, sendo que somente conseguiram se classificar na temporada 2022/2023, enquanto na conferência leste o Charlotte Hornets obteve apenas uma aparição na pós-temporada. Além disso, importante citar que dos 15 últimos campeões da NBA, dez deles eram da conferência oeste e cinco da conferência leste, com Golden State Warriors (Oeste) sendo o maior campeão do período com 4 títulos, seguido pelo Los Angeles Lakers (Oeste) com 3 títulos e já no Leste o Miami Heat obteve dois triunfos nesses últimos 15 anos.

## 3.1 Temporada Regular

Foram desenvolvidos modelos de regressão lineares múltiplos, regressão beta, utilizando modelos generalizados e também utilizando efeitos aleatórios. Na regressão linear foram testados os modelos completos (com todas as variáveis do banco de dados), modelos em que as variáveis foram significantes com 5% de significância e utilizando dois métodos *stepwise* (*backward selection* e *forward selection*), sendo aplicados os mesmos modelos para modelos generalizados (gamlss) em que foi testado para densidade beta, já que com a densidade normal seria encontrado o mesmo modelo da regressão linear múltipla.

No modelo de efeitos aleatórios foram testados modelos com a densidade beta e normal, utilizando os times e o número da temporada como efeito aleatório. Por fim, realizamos a regressão beta utilizando o pacote no R chamado betareg, em que testamos para 5 funções de ligação diferentes (logito, loglog, probito, cloglog, cauchito).

Dessa forma, foram desenvolvidos diversos modelos para serem testados, sendo que após a realização dos métodos automáticos de escolhas de variáveis também foram realizados testes de razão de verossimilhança e análise Anova para verificar se a adição de uma variável específica seria relevante para o modelo ou não, dependendo do p-valor do teste.

### 3.1.1 Validação Cruzada dos melhores modelos

Como foi falado na 2.1.6 sobre a importância do *Cross Validation*, será feita esta análise para os 11 melhores modelos encontrados. Assim, quanto maior o $R^2$ melhor é o modelo, além de quanto menor for RMSE e MAE melhor é a performance do modelo.

Com as características de uma boa avaliação de *Cross validation* descritas anteriormente, observando a Tabela 4 visualmente podemos observar alguns modelos que não se adequaram muito bem como os modelos mistos para a densidade beta, tanto com efeito aleatório de Team quanto da Temporada, em que RMSE e MAE para os dois modelos foram bem maior do que os encontrados nos demais modelos. Além disso, outro modelo que pode ser retirado dos melhores modelos é Gamlss Beta *Forward*, que é o de modelos generalizados realizado no pacote gamlss no R, pois RMSE e MAE encontrados foram muito grandes.

Tabela 4: Resultado do *Cross Validation* para os 11 melhores modelos

| Modelo | $R^2$ | RMSE | MAE |
|---|---|---|---|
| Regressão Linear | 0.943 | 0.036 | 0.029 |
| Betareg Logito | 0.946 | 0.035 | 0.028 |
| Betareg loglog | 0.940 | 0.037 | 0.029 |
| Betareg probito | 0.945 | 0.035 | 0.028 |
| Betareg cloglog | 0.943 | 0.036 | 0.029 |
| Betareg Cauchit | 0.940 | 0.037 | 0.030 |
| Gamlss Beta | 0.943 | 0.702 | 0.569 |
| Misto Normal TEAM | 0.941 | 0.037 | 0.029 |
| Misto Normal TEMPORADA | 0.943 | 0.036 | 0.029 |
| Misto Beta TEAM | 0.939 | 0.701 | 0.569 |
| Misto Beta TEMPORADA | 0.936 | 0.702 | 0.569 |

Dessa forma, sobraram oito modelos que apresentam de forma adequada para o prosseguimento das análises, que seria a análise de resíduos e a interpretação do modelo escolhido. Assim, desenvolvemos a análise de resíduos em sequência para descobrirmos qual o modelo a ser escolhido.

### 3.1.2 Análise de resíduos

Antes de iniciarmos a análise de resíduos, será feita uma análise de variância (Anova) dos modelos para verificar se pelo menos uma das variáveis independentes não tem um efeito significativo na variável resposta, levando em consideração o efeito das outras variáveis independentes

no modelo. Assim, para a regressão beta obtivemos que as funções de ligação logito (PF e 3PP), loglog (OREB e FGP) e probito (PF) obtiveram pelo menos uma variável que não foi significante.

Além disso, tivemos que o efeito aleatório de Temporada para o modelo misto normal não foi significante, quando testamos incluir o efeito aleatório em um modelo vazio. Assim, os modelos restantes são regressão linear, regressão beta cloglog e cauchito e modelos misto normal com efeito aleatório TEAM.

Assim, realizando as testagens das pressuposições dos modelos, observamos na Tabela 5 que os quatro modelos selecionados seguiram todas as pressuposições do modelo. Além disso, se observarmos o *Cross Validation* realizado, as medidas são próximas umas das outras desses quatro modelos. Dessa forma, os modelos são bem ajustados e se verificarmos o modelo misto normal TEAM e testarmos se é significativo inserir o efeito aleatório de TEAM no modelo vazio, encontramos que a inserção do efeito aleatório é significante. Então, o modelo misto com distribuição normal e efeito aleatório TEAM será escolhido como melhor modelo e partiremos para a análise de resíduos.

Tabela 5: Análise das pressuposições dos modelos identificados como melhores

| Modelo | Normalidade | Independência | Homoscedasticidade |
|---|---|---|---|
| Regressão Linear | 0.370 | 0.273 | 0.137 |
| Betareg cloglog | 0.372 | 0.140 | 0.277 |
| Betareg Cauchit | 0.181 | 0.185 | 0.001 |
| Misto Normal TEAM | 0.250 | | |

Após a análise de resíduos, poderemos verificar se o modelo escolhido (efeito aleatório com TEAM com distribuição normal) satisfaz as pressuposições do modelo e se os resíduos estão bem-comportados. Dessa forma, nas Figura 7 podemos observar o histograma dos resíduos na esquerda e ao lado o *boxplot* dos resíduos. Podemos notar, pelo *boxplot*, que existem alguns *outliers* presentes nos resíduos e já no histograma aparenta seguir uma distribuição normal pela forma que o histograma tomou, se ajustando bem.

Além disso, se observamos o gráfico mais à direita dos três, podemos observar o gráfico quantil-quantil dos resíduos que tem como utilidade verificar se os resíduos se ajustam à distribuição normal. Podemos perceber que grande parte dos pontos estão em cima da linha, indicando que visualmente os resíduos seguem a distribuição normal.

Pelo histograma, *boxplot* e gráfico quantil-quantil dos resíduos aparentam seguir uma distribuição normal e realizando o teste de normalidade de Shapiro-Wilk observamos a estatística do teste de 0.996 e o p-valor de 0.250. Dessa forma, chegamos à conclusão de que não temos evidências para rejeitar a hipótese se nula e assim podemos assumir que os resíduos seguem uma distribuição normal.

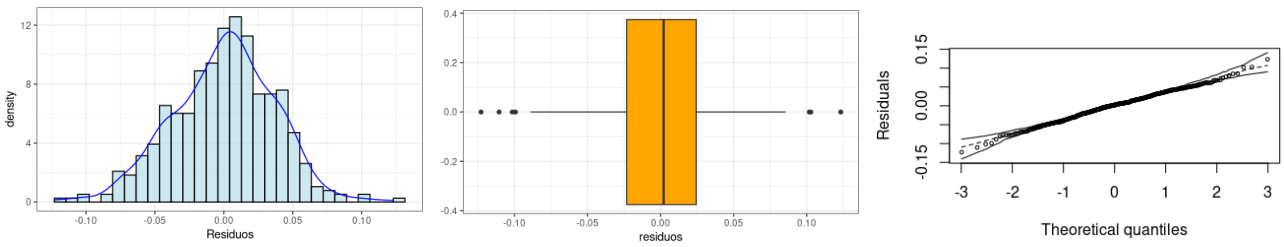


Figura 7: Histograma dos resíduos, *Boxplot* dos resíduos e gráfico quantil-quantil dos resíduos do modelo de efeito aleatório TEAM

Na Figura 8, podemos observar o restante da análise de resíduos, em que ambos os gráficos são para verificar a homoscedasticidade (variância constante) dos resíduos. No gráfico da esquerda, observamos os gráficos de valores ajustados x resíduos, em que se observarmos a linha azul não vemos nenhuma tendência aparente, dando mais indícios de homoscedasticidade. Já no gráfico da direita, observamos o diagrama de Dispersão de Valores Ajustados vs. Resíduos Padronizados, em que é possível observar que não tem tendência aparente e também identificar alguns *outliers* presentes no banco de dados.

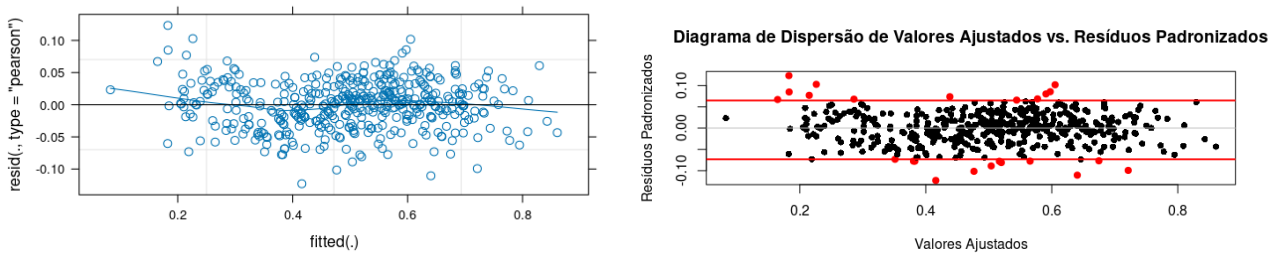


Figura 8: Gráfico de valores ajustados x resíduos (homoscedasticidade) e Diagrama de Dispersão de Valores Ajustados vs. Resíduos Padronizados

Assim, com as análises realizadas, podemos perceber que os resíduos seguem uma distribuição normal e apresentam homoscedasticidade. Dessa forma, o modelo escolhido apresenta boas características para ser implementado e podemos seguir para a interpretação do modelo.

### 3.1.3 Interpretação do modelo da temporada regular

O modelo escolhido foi o modelo misto com efeito aleatório TEAM e distribuição normal, que contém as variáveis escolhidas OREB, PF, 3PA e *Plus/Minus*. Na Tabela 6, notamos que o intercepto é 0.609, indicando a taxa de vitórias esperada quando as outras covariáveis são zero. O coeficiente para OREB é -0.004, indicando que um aumento de uma unidade em rebotes ofensivos está associado à uma diminuição de 0.004 na taxa de vitórias, mantendo todas as outras variáveis constantes. O coeficiente para *Plus/Minus* é 0.031, indicando que um aumento de uma unidade na diferença entre pontos marcados e pontos sofridos está associado a um aumento de 0.031 na taxa de vitórias, mantendo todas as outras variáveis constantes.

Já o coeficiente PF é -0.003, indicando que um aumento de uma unidade em rebotes ofensivos está associado a uma diminuição de 0.003 na taxa de vitórias, mantendo todas as outras variáveis constantes. Sendo a mesma lógica para 3PA, mas com coeficiente de -0.0005.

Além disso, na Tabela 6, observamos a análise de variância (Anova) para o modelo escolhido, que mostra a decomposição da variância explicada pelos preditores do modelo. O F-value indica a significância estatística de cada preditor, com os preditores, OREB, PF, 3PA e *Plus/Minus*, têm F-values significativos, indicando que são importantes para explicar a variação na taxa de vitórias.

Tabela 6: Sumário dos efeitos fixos e Análise de variância do modelo misto com efeito aleatório TEAM.

| Variável | Estimativa | Erro Padrão | Valor t | Soma dos Quadrados | Valor F |
|---|---|---|---|---|---|
| (Intercept) | 0.609 | 0.033 | 18.446 | | |
| *Plus/Minus* | 0.031 | 0.0004 | 75.512 | 8.5895 | 6074.7027 |
| OREB | -0.004 | 0.002 | -2.426 | 0.0072 | 5.1021 |
| PF | -0.003 | 0.001 | -1.994 | 0.0053 | 3.7344 |
| 3PA | -0.0005 | 0.0003 | -1.762 | 0.0044 | 3.1045 |

O intercepto no modelo representa a média geral da taxa de vitórias entre todas as equipes, porém, para entender as diferenças específicas entre as equipes, pode-se utilizar os efeitos aleatórios de time. Sendo que os efeitos aleatórios de time capturam as variações específicas de cada equipe que não são explicadas pelas variáveis independentes incluídas no modelo. Eles fornecem uma maneira de modelar as diferenças entre as equipes que podem surgir de características únicas de cada uma, como estratégias de jogo, habilidades dos jogadores, treinamento, entre outros fatores.

Assim, o efeito aleatório para cada time pode ser visto na Figura 9, em que o time que apresenta maior efeito aleatório é o Memphis Grizzlies, seguido por Portland Trail Blazers e Los Angeles Lakers. Já os menores efeitos aleatórios pertencem ao Minnesota Timberwolves e Detroit Pistons. Dessa forma, se fôssemos escolher os times que têm mais chances de ficar com mais porcentagens de vitórias ao longo da temporada, escolheríamos segundo o modelo misto com efeito aleatório TEAM Memphis Grizzlies e Portland Trail Blazers.

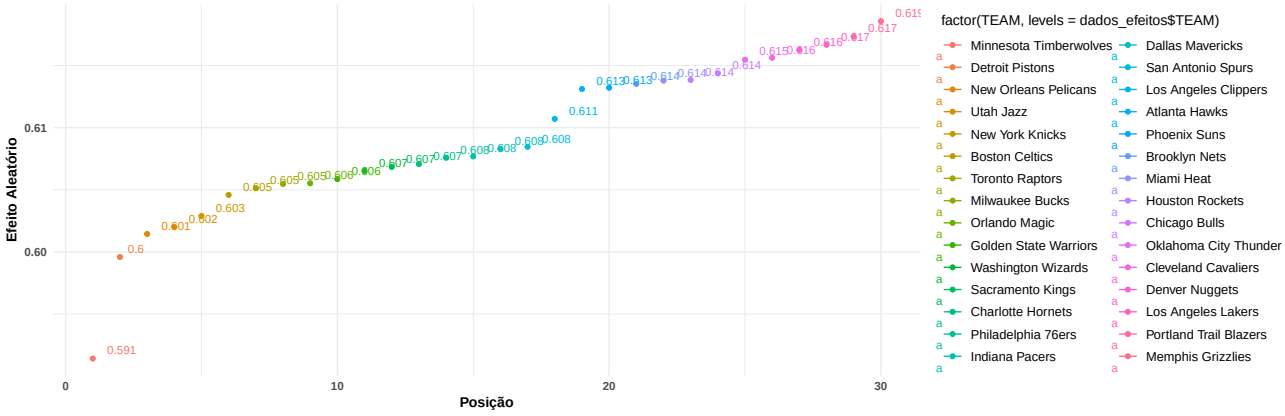


Figura 9: Efeitos aleatórios preditos do modelo misto com efeito aleatório TEAM.

## 3.2 *Playoffs*

Os métodos utilizados na temporada regular foram os mesmos aplicados na pós-temporada. Assim, foram desenvolvidos diversos modelos para serem testados, sendo que após a realização dos métodos automáticos de escolhas de variáveis também foram realizados testes de razão de verossimilhança para verificar se a adição de uma variável específica é relevante para o modelo ou não, dependendo do p-valor do teste.

### 3.2.1 Validação Cruzada dos melhores modelos

Com as características de uma boa avaliação de *cross validation* descritas anteriormente, observando a Tabela 7, visualmente podemos observar alguns modelos que não se adequaram muito bem como os modelos mistos para a densidade beta, tanto com efeito aleatório de TEAM quanto da Temporada, em que RMSE e MAE para os dois modelos foram bem maior do que os encontrados nos demais modelos. Além disso, outro modelo que pode ser retirado dos melhores modelos é Gamlss Beta *Forward*, que é o modelo generalizado realizado no pacote gamlss no R, pois RMSE e MAE encontrados foram muito grandes. Foram os mesmos modelos que não foram utilizados no *Cross validation* da Temporada regular.

Tabela 7: Resultado do *Cross Validation* para os 11 melhores modelos

| Modelo | $R^2$ | RMSE | MAE |
|---|---|---|---|
| Regressão Linear | 0.808 | 0.090 | 0.069 |
| Betareg Logito | 0.790 | 0.097 | 0.078 |
| Betareg loglog | 0.788 | 0.096 | 0.075 |
| Betareg probito | 0.790 | 0.097 | 0.077 |
| Betareg cloglog | 0.778 | 0.100 | 0.081 |
| Betareg Cauchit | 0.775 | 0.099 | 0.080 |
| Gamlss Beta | 0.782 | 0.930 | 0.778 |
| Misto Normal TEAM | 0.779 | 0.096 | 0.074 |
| Misto Normal TEMPORADA | 0.808 | 0.090 | 0.069 |
| Misto Beta TEAM | 0.760 | 0.936 | 0.777 |
| Misto Beta TEMPORADA | 0.782 | 0.930 | 0.778 |

Dessa forma, sobraram oito modelos que se apresentam de forma adequada para o prosseguimento das análises, que seria a análise de resíduos e a interpretação do modelo escolhido. Assim, iremos desenvolver a análise de resíduos em sequência para descobrirmos qual o modelo que escolheremos.

### 3.2.2 Análise de resíduos

Antes de iniciarmos a análise de resíduos, será feita uma análise de variância (Anova) dos modelos para verificar se pelo menos uma das variáveis independentes não tem um efeito significativo na variável resposta, levando em consideração o efeito das outras variáveis independentes no modelo. Assim, para a regressão beta obtivemos que as funções de ligação logito (FTP e REB), probito (FTP e REB), cloglog (FTP e REB) e cauchito (TEAM e FTP) obtiveram pelo menos uma variável que não foi significante.

Além disso, tivemos que o efeito aleatório de Temporada para o modelo misto normal não foi significante, quando testamos incluir o efeito aleatório em um modelo vazio. Assim, os modelos restantes são regressão linear, regressão beta loglog e modelos misto normal com efeito aleatório TEAM.

Assim, realizando as testagens das pressuposições dos modelos, observamos na Tabela 8 que a regressão linear não seguiu nem a normalidade e nem a homoscedasticidade, enquanto o modelo misto não seguiu a normalidade também. Enquanto a regressão beta loglog seguiu a normalidade e independência. Com essas informações, será escolhida a regressão beta loglog, pois cumpriu duas pressuposições do modelo e também apresentou todas as variáveis relevantes e apresentou boas medidas no *Cross Validation*. Também, importante ressaltar que o modelo misto normal TEAM apresentou o efeito aleatório TEAM relevante, porém não cumpriu com a normalidade dos resíduos.

Tabela 8: Análise das pressuposições dos modelos identificados como melhores para os *Playoffs*

| Modelo | Normalidade | Independência | Homoscedasticidade |
|---|---|---|---|
| Regressão Linear | 0.006 | 0.067 | 0.004 |
| Betareg loglog | 0.419 | 0.062 | 0.007 |
| Misto Normal TEAM | 0.0003 | | |

Após a análise de resíduos, poderemos verificar se o modelo escolhido (Betareg loglog) satisfaz as pressuposições do modelo e se os resíduos estão bem-comportados. Dessa forma, na Figura 10, podemos observar o histograma dos resíduos na esquerda e ao lado o *boxplot* dos resíduos. Podemos notar pelo *boxplot* que existem alguns *outliers* presentes nos resíduos e já no histograma aparenta seguir uma distribuição normal pela forma que o histograma tomou, sendo importante notar que na cauda da esquerda os resíduos estão mais dispersos.

Além disso, se observarmos o gráfico mais à direita, podemos observar o gráfico quantil-quantil dos resíduos que tem como utilidade verificar se os resíduos se ajustam à distribuição normal. Podemos perceber que grande parte dos pontos estão em cima da linha, porém a cauda esquerda da distribuição não se encontra muito bem ajustada. Realizando o teste de normalidade de Shapiro-Wilk, observamos o p-valor de 0.419. Dessa forma, chegamos à conclusão de que não temos evidências para rejeitar a hipótese se nula e assim podemos assumir que os resíduos seguem uma distribuição normal.

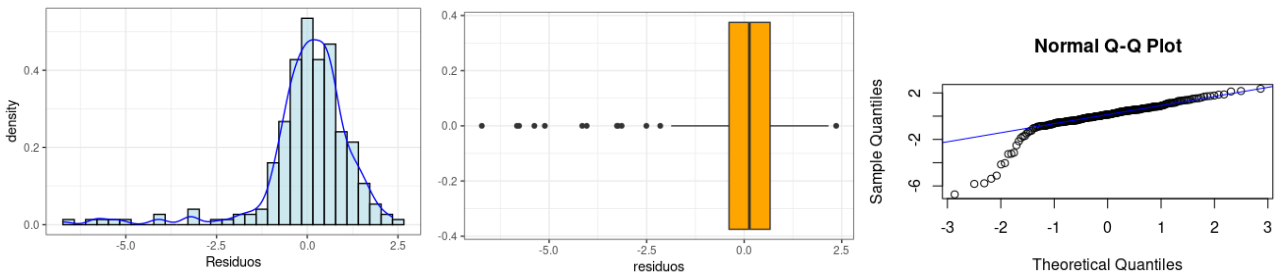


Figura 10: Histograma dos resíduos, *Boxplot* dos resíduos e gráfico quantil-quantil dos resíduos do modelo Betareg loglog.

Na Figura 11, podemos observar o restante da análise de resíduos, em que o gráfico da esquerda é para verificar a suposição de independência e o da direita a suposição de homoscedasticidade. No gráfico da esquerda notamos que os resíduos estão bem-comportados, pois não aparentam ter nenhuma tendência e nem estar muito dispersos. Já no gráfico da direita, notamos que os resíduos apresentam três linhas retas na esquerda, sendo que isso não é usual nos resíduos, sendo esse o motivo da heterocedasticidade.

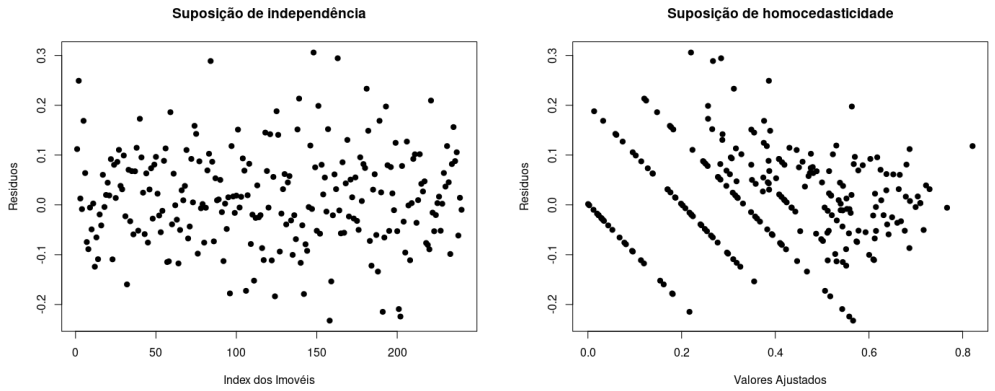


Figura 11: Gráfico da suposição de independência e Diagrama de Dispersão de Valores Ajustados vs. Resíduos (homoscedasticidade)

Assim, como visto nos gráficos de resíduo anteriores, notamos que os resíduos nos valores inferiores possuem algo de estranho neles. Sendo explicado pela Figura 6 que mostra a correlação entre a variável resposta e suas covariáveis.

Assim, com as análises realizadas, podemos perceber que os resíduos seguem uma distribuição normal e apresentam independência, porém tem heterocedasticidade, devido à relação entre a variável resposta e suas covariáveis. Dessa forma, o modelo escolhido apresenta boas características para ser implementado e podemos seguir para a interpretação do modelo.

### 3.2.3 Interpretação do modelo dos *playoffs*

O modelo escolhido foi o de regressão beta com função de ligação loglog, que contém as variáveis TEAM, REB e *Plus/Minus*. Na Tabela 9, observamos a análise de variância (Anova) para o modelo escolhido, que mostra a decomposição da variância explicada pelos preditores do modelo. O Pr(>Chisq) indica a significância estatística de cada preditor, com os três preditores, TEAM, REB e *Plus/Minus*, têm p-valor significativo com valores menores que 0.1, indicando que são importantes para explicar a variação na taxa de vitórias. Este modelo fornece uma estrutura para entender como os times, rebotes e a diferença de pontos influenciam a taxa de vitórias das equipes.

Tabela 9: Análise de variância do modelo Betareg loglog

| Variável | Df | Chisq | Pr(>Chisq) |
|---|---|---|---|
| TEAM | 31 | 116.5069 | 7.841e-12 |
| REB | 1 | 3.6914 | 0.05469 |
| *Plus/Minus* | 1 | 476.8766 | 2.2e-16 |

Além disso, na Figura 12, observamos as estimativas dos coeficientes do modelo betareg loglog. Nele observamos as estimativas de REB, *Plus/Minus* e phi, com os valores de 0.014, 0.082 e 12.481 respectivamente e os valores das estimativas para cada time. Também, o intercepto do modelo é de -0.149, mas importante dizer que a interpretação do modelo de regressão beta com função de ligação loglog não é simples de se explicar e as metodologias que envolvem o modelo são complexas. Porém, com o modelo apresentado observamos que cada time apresenta um incremento diferente no intercepto.

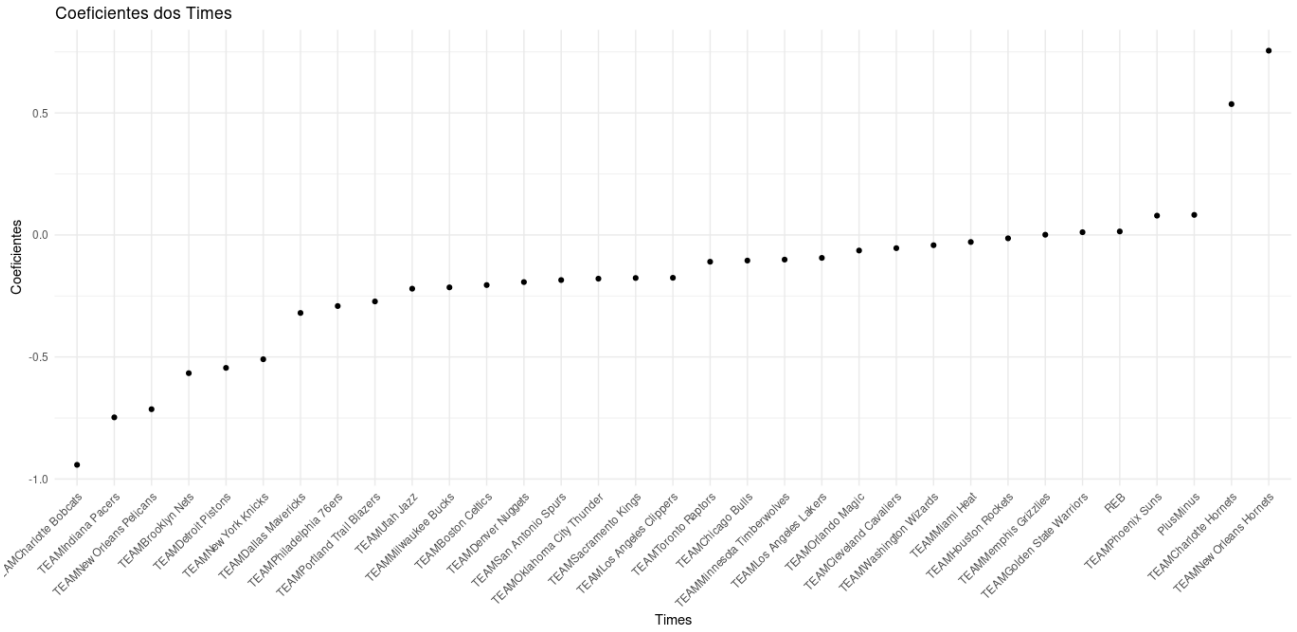


Figura 12: Coeficientes estimados do modelo de regressão beta loglog.

# 4 Discussão e Conclusões

Como mencionado foram testadas quatro diferentes formas de modelar os dados. Sendo que para a temporada regular o modelo escolhido foi da metodologia de modelos mistos e já na pós-temporada foi a regressão beta com função de ligação loglog.

Dessa forma, foi identificado para a temporada regular o seguinte modelo $WINP = 0.609 + 0.031 \times Plus/Minus - 0.004 \times OREB - 0.003 \times PF - 0.0005 \times 3PA$ com todas as variáveis do modelo significativas e também possui o efeito aleatório de time. Já para os *playoffs*, obtivemos que o modelo de regressão beta com função de ligação loglog foi o que melhor se adequou e que contém as variáveis TEAM, REB e *Plus/Minus*, com TEAM sendo uma variável categórica e que os diferentes times têm um incremento diferente no intercepto.

Assim, percebemos que *Plus/Minus* é significante para ambas as partes da temporada, trazendo de reflexão que o mais importante é se ter um equilíbrio entre ataque e defesa, pois a quantidade de pontos não foi significante para o modelo, ou seja, não adianta marcar muitos pontos, se sua defesa não consegue controlar o outro time.

Também, nos *playoffs*, obtivemos que REB é significativo e quanto mais rebotes, maior a chance de obter a vitória. Assim, é preciso na pós temporada ter uma consistência entre ataque

e defesa, sendo importante conquistar os rebotes nos dois lados da quadra para que as chances de vitórias sejam maiores.

Já na temporada regular é preciso ter o equilíbrio nos dois lados da quadra, além de ter que prestar atenção na quantidade de bolas de 3 pontos tentadas, pois como as bolas de 3 pontos são mais difíceis do que uma bandeja ou enterrada, é preciso achar o momento correto para arremessar a bola de 3 pontos. Outro ponto importante a ser levado em consideração são as faltas feitas sobre o adversário, pois quanto mais faltas você cometer mais chances seu adversário terá em um arremesso sem marcação (lances livres). Importante ressaltar que normalmente quando um time está perdendo por poucos pontos e falta pouco tempo no cronômetro para acabar o jogo esse time faz faltas para parar o relógio e ter mais oportunidades de ataque, em contrapartida oferece lances livres ao time adversário, porém como foi visto, lances livres não são uma das variáveis significativas nos modelos, portanto pode ser interessante usar esta tática nos finais dos jogos, caso esteja perdendo.

Outro ponto a ser destacado no modelo de temporada regular é que rebotes ofensivos (OREB) têm impacto negativo na porcentagem de vitórias na temporada regular. Ou seja, quanto mais rebotes ofensivos o time pegar, menor a porcentagem de vitórias ao longo da temporada, o que nos traz que é preciso ter o arremesso acertado e não depender dos rebotes ofensivos para ganhar os jogos. Assim, podendo notar as diferenças nas duas partes da temporada.

# 5 Matéria encaminhada para publicação

Nada a declarar

# 6 Bibliografia


BATES, D.; MÄCHLER, M.; BOLKER, B.; WALKER, S. **Fitting Linear Mixed-Effects Models Using lme4**. 2024. Disponível em: <https://cran.r-project.org/web/packages/lme4/index.html>. Acesso em: 29 julho de 2024.

FERRARI, Silvia; CRIBARI-NETO, Francisco. Beta regression for modelling rates and proportions. **Journal of applied statistics**, v. 31, n. 7, p. 799-815, 2004. Disponível em: <https://doi.org/10.1080/0266476042000214501>. Acesso em: 21 de maio de 2024.

GARETH, James et al. **An introduction to statistical learning: with applications in R.** Spinger, 2013. Disponível em: <https://www.statlearning.com/>. Acesso em: 21 de maio de 2024.

GIOVANINI, Bruno et al. Does game pressure affect hand selection of NBA basketball players?. **Psychology of Sport and Exercise**, v. 51, p. 101785, 2020. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S1469029220302764>. Acesso em: 21 de maio de 2024.

MACIEL, Luiz Felipe Vieira. Regressão linear múltipla na modelagem de resultados na National Basketball Association (NBA). 2019. Disponível em: <https://repositorio.ufu.br/bitstream/123456789/28341/3/Regress%C3%A3oLinearM%C3%BAltipla.pdf>. Acesso em: 21 de maio de 2024.

MORGADO, Gabriel Ferreira de Melo. Vantagens de jogar em casa nos playoffs da NBA (1946-2021). 2022. Disponível em:<https://repositorio.unesp.br/server/api/core/bitstreams/cc331e04-0f45-4ad8-ac21-d877271782a0/content>. Acesso em: 21 de maio de 2024.

R Core Team . R: **A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. 2023. Disponível em: <https://www.r-project.org/>. Acesso em: 29 julho de 2024.

RIGBY, Robert A.; STASINOPOULOS, D. Mikis. Generalized additive models for location, scale and shape. **Journal of the Royal Statistical Society Series C: Applied Statistics**, v. 54, n. 3, p. 507-554, 2005. Disponível em: <https://doi.org/10.1111/j.1467-9876.2005.00510.x>. Acesso em: 21 de maio de 2024.

SMITHSON, Michael; VERKUILEN, Jay. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. **Psychological methods**, v. 11, n. 1, p. 54, 2006. Disponível em: <https://psycnet.apa.org/record/2006-03820-004>. Acesso em: 21 de maio de 2024.

STASINOPOULOS, D. Mikis; RIGBY, Robert A. Generalized additive models for location scale and shape (GAMLSS) in R. **Journal of Statistical Software**, v. 23, p. 1-46, 2008. Disponível em: <https://www.jstatsoft.org/article/view/v023i07>. Acesso em: 21 de maio de 2024.

ZUUR, Alain F. et al. **Mixed effects models and extensions in ecology with R.** New York: springer, 2009. Disponível em: <https://link.springer.com/book/10.1007/978-0-387-87458-6>. Acesso em: 21 de maio de 2024.

# 7 Perspectivas de continuidade ou desdobramento do trabalho

Pode ser feita uma análise para diferentes ligas de basquete como a WNBA (basquete norte americano feminino), NBB (Liga de Basquete Masculino do Brasil) e LBF (Liga de Basquete Feminino do Brasil), para identificar certos padrões e diferenças entre as ligas.

# 8 Outras atividades de interesse universitário

Nada a declarar.

# 9 Apoio

Programa Institucional de Bolsas de Iniciação Científica e Tecnológica (PIBIC)

# 10 Agradecimentos

Gostaria de agradecer ao meu professor e orientador Rafael Pimentel Maia por ter aceitado ser meu orientador e ter aceitado o tema. Além disso, gostaria de agradecer meus pais, Juliane e

Wandmar, por sempre me incentivarem e sempre estarem do meu lado para tudo que eu preciso. Também, queria agradecer a minha irmã, Natalia, que me apoia em todas as decisões que eu tomo e sempre está presente quando preciso. Assim, gostaria de agradecer a toda minha família porque sem eles eu não conseguiria realizar nada.