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

Palavras-Chave: Regressão beta, Modelos Mistos, NBA

Autores(as):

Rubens Cortelazzi Roncato, IMECC - UNICAMP

Prof. Dr. Rafael Pimentel Maia, IMECC - UNICAMP

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.

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 a 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 apresentação 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 a 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 proporção de vitórias por temporada ou pós temporada.

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 e 450 linhas para temporada regular e 240 linhas para os *playoffs*.

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 <* μ *< 1 e* $\phi > 0$*. Sendo que escrevemos* $y \sim B(\mu,\ \phi)$*.*

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

Assuma que $y_1, \dots, y_n$ *seja uma amostra aleatória, em que* $y_i \sim B(\mu_i, \phi),\ i = 1, \dots, n$*. O modelo de regressão beta é definido como:* $g(\mu_i) = x_i^T\beta = \eta_i$*, em que* $\beta = (\beta_1, \dots, \beta_k)^T$ *é um vetor* $k \times 1$ *de parâmetros de regressão desconhecidos (k < n),* $x_i = (x_{i1}, \dots, x_{ik})^T$ *é um vetor de k regressores (ou variáveis independentes ou covariáveis) e* η_i *é um preditor linear(por exemplo,* $\eta_i = \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$*, 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* μ_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 $g(\mu) = log\big(\mu/(1-\mu)\big)$*; 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* $g(\mu) = \tan\{\pi(\mu - 0.5)\}$*.*

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: $y_{ij} = \big(\hat{\alpha} + \epsilon_j\big) + \beta x_{ij} +$

ϵ_{ij}, com $\epsilon_j = N(0, \sigma_{entre})$ e $\epsilon_{ij} = N(0, \sigma_{intra})$. O $\hat{\alpha}$ é o intercepto médio e σ_{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.

VALIDAÇÃO CRUZADA (CV):

A validação cruzada (CV) é um método usado em modelos de predição com técnicas de aprendizado de máquina, onde os dados são particionados em conjuntos para treino e teste, ajudando a detectar o sobreajuste do modelo aos dados de treinamento.

Utilizamos o método *k-fold*, conforme descrito em Gareth, James et al (2013). Este método envolve dividir o conjunto de dados em k grupos (ou *folds*) aproximadamente iguais. Em cada iteração, um *fold* é usado como conjunto de validação, e o modelo é treinado nos *folds* restantes. O erro quadrático médio (MSE) é calculado para o *fold* de validação. Este processo é repetido k vezes, cada vez com um *fold* diferente como validação.

RESULTADOS:

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*.

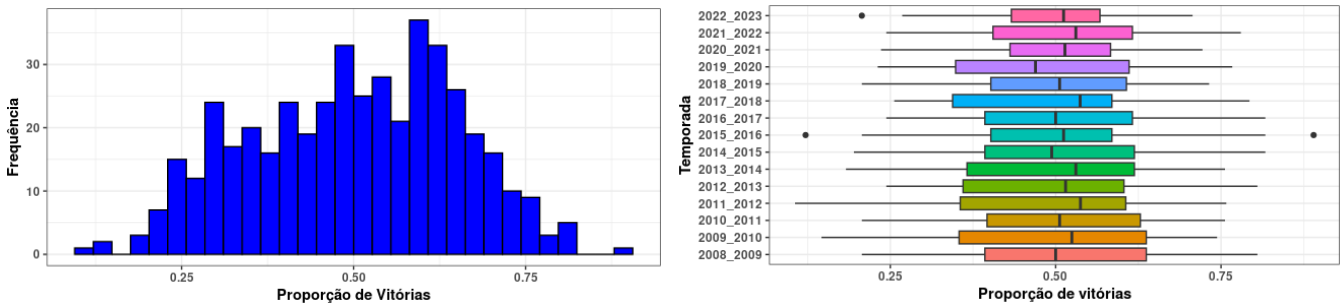

Figura 1: Boxplot da proporção de vitórias por temporada e Histograma da proporção de vitórias na temporada regular.

A Figura 2 mostras 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.

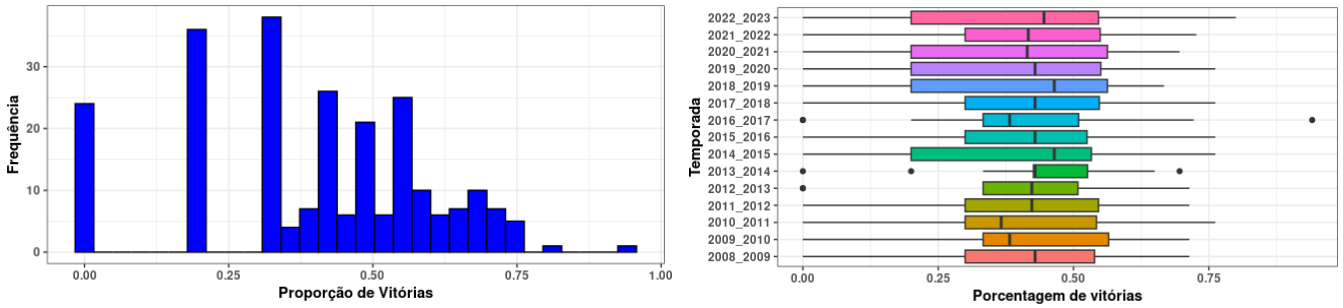

Figura 2: Boxplot da proporção de vitórias por temporada e Histograma da proporção de vitórias nos playoffs.

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 relevante para o modelo ou não, dependendo do p-valor do teste.

Após a realização da validação cruzada, 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, iremos desenvolver a análise de resíduos em sequência para descobrirmos qual o modelo que escolheremos.

Após a análise de resíduos verificamos que o modelo escolhido foi de modelos mistos com efeito aleatório TEAM com distribuição normal e verificamos se satisfaz as pressuposições do modelo e se os resíduos estão bem-comportados. Dessa forma, nas Figura 3 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.

Pelo histograma, boxplot e gráfico quantil-quantil dos resíduos aparenta 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 na conclusão de que não temos evidências para rejeitar a hipótese se nula e assim podemos assumir que os resíduos seguem uma distribuição normal.

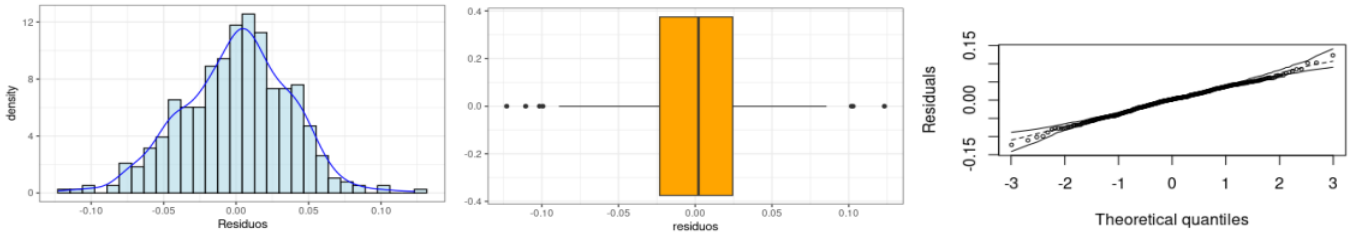

Figura 3:Histograma dos resíduos, Boxplot dos resíduos e QQplot

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.

PLAYOFFS:

Os métodos utilizados na temporada regular foram os mesmos aplicados na pós-temporada. Assim, foram desenvolvidas diversas 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 relevante para o modelo ou não, dependendo do p-valor do teste.

Foi realizada a Validação cruzada dos modelos, selecionando os melhores modelos para realização da análise de resíduos, em que foi testado a normalidade, homoscedasticidade e independência dos resíduos dos modelos. Foi encontrado que o melhor modelo é o modelo de regressão beta com função de ligação loglog com as variáveis *Plus/Minus*, TEAM e REB.

CONCLUSÃO:

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 é 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) tem 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.

BIBLIOGRAFIA:

BATES, D.; MÄCHLER, M.; BOLKER, B.; WALKER, S. Fitting Linear Mixed-Effects Models Using lme4. 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.

GARETH, James et al. **An introduction to statistical learning: with applications in R.** Spinger, 2013.

GIOVANINI, Bruno et al. Does game pressure affect hand selection of NBA basketball players?. **Psychology of Sport and Exercise**, v. 51, p. 101785, 2020.

MACIEL, Luiz Felipe Vieira. Regressão linear múltipla na modelagem de resultados na National Basketball Association (NBA). 2019.

MORGADO, Gabriel Ferreira de Melo. Vantagens de jogar em casa nos playoffs da NBA (1946-2021). 2022.

R Core Team . R: **A language and environment for statistical computing.** R Foundation for Statistical Computing, Vienna, Austria. 2023.

RIGBY, R. A.; STASINOPOULOS, D. M.; HELLER, G. Z.; DE BASTIANI, F. gamlss: Generalized Additive Models for Location Scale and Shape. 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.

ZUUR, Alain F. et al. **Mixed effects models and extensions in ecology with R.** New York: springer, 2009.