Universidade Federal de Uberlândia
Faculdade de Matemática

Bacharelado em Estatística

REGRESSÃO LINEAR MÚLTIPLA NA MODELAGEM DE RESULTADOS NA LIGA NACIONAL DE BASQUETE (LNB)

Daniel Licnerski Borges

Uberlândia-MG
2022

Daniel Licnerski Borges

REGRESSÃO LINEAR MÚLTIPLA NA MODELAGEM DE RESULTADOS NA LIGA NACIONAL DE BASQUETE (LNB)

Trabalho de conclusão de curso apresentado à Coordenação do Curso de Bacharelado em Estatística como requisito parcial para obtenção do grau de Bacharel em Estatística.

Orientador: Profa. Dra. Maria Imaculada de Sousa Silva

Uberlândia-MG

2022

Universidade Federal de Uberlândia
Faculdade de Matemática

Coordenação do Curso de Bacharelado em Estatística

A banca examinadora, conforme abaixo assinado, certifica a adequação deste trabalho de conclusão de curso para obtenção do grau de Bacharel em Estatística.

Uberlândia, ____ de _____ de 20____

BANCA EXAMINADORA

Profa. Dra. Maria Imaculada de Sousa Silva

Prof. Dr. José Waldemar da Silva

Profa. Dra. Nádia Giaretta Biase

Uberlândia-MG

2022

Agradecimentos

Primeiramente gostaria de agradecer aos meus pais Thalma Licnerski e Marcílio Borges por sempre me apoiarem nas minhas decisões e sempre acreditarem no meu sonho de ter uma formação em uma Universidade Federal.

Agradeço aos meus professores, minha orientadora que tiveram paciência e atenção com a minha formação, agradeço aos meus amigos e colegas que compartilharam conhecimento durante a formação.

Não existem palavras para agradecer todas essas pessoas que ajudaram na minha formação. Foram dias difíceis que me fizeram pensar em desistir várias vezes mas minha família, amigos e professores foram essenciais para continuar nessa jornada e por isso o sentimento é de gratidão.

Resumo

Atualmente o basquetebol é um dos jogos olímpicos mais populares no mundo. O uso de dados estatísticos no basquete não é novidade. A utilização de análise dos números por treinadores e comissões técnicas permite condições melhores para aprimorar os resultados. Este trabalho busca, por meio de regressão linear múltipla, investigar quais variáveis são significativas para obter o maior número de vitórias dos times na Liga Nacional de Basquete (LNB). Na realização das análises, os dados foram coletados das estatísticas dos jogos ao final de 5 temporadas (2016-2020), obtendo os resultados com a utilização do software R. O modelo foi bem ajustado, conseguindo explicar 85% da variabilidade encontrada no número total de vitórias por temporada. Das 21 variáveis análisadas, 9 apresentaram significância estatística, entre essas, 7 demonstraram impactos positivos com a variável resposta (cestas de 3 pontos, cestas de 2 pontos, rebotes, índice de assistência por erro, índice de bolas recuperadas por erros, enterradas e violações).

Palavras-chave: Análise dos resíduos, Estatísticas de jogo, Esporte, Vitórias.

Abstract

Basketball is currently one of the most popular Olympic games in the world. Statistical data applied in basketball is nothing new. The use of data analysis, by coaches and technical teams allow better conditions to improve results. The main goal of this work is evaluate which variables are significant to obtain the highest number of wins of the teams in the National Basketball League through multiple linear regression. The data set were collected from game statistics at the end of five seasons (2016-2020), and the analysis were carrying out using the R software. The fitted model managing to explain 85% of the variability found in the total number of wins per season. Considering 21 variables analyzed, only 9 showed statistical significance, in which seven demonstrate positive impacts with the response (3-point baskets, 2-point baskets, rebounds, error assist rate, error recovery rate, dunks and violations).

Keywords: Residual Analysis, Game stats, Sports, Victories.

Sumário

Lista de Figuras

Lista de Tabelas

1. Introdução

Nas modalidades esportivas coletivas em geral, além da preparação do atleta e do grupo com relação às características físicas e psicológicas para um bom desempenho, torna-se necessário o estudo dos indicadores dos jogos, como também a forma como as componentes ou variáveis já observadas podem ser usadas para analisar e influenciar os jogos no futuro. Essas componentes, chamadas estatístícas dos jogos, podem e devem ser utilizadas pelas equipes para o planejamento técnico e até mesmo financeiro das equipes, como forma de maximizar os resultados positivos visando aprimorar o desempenho exequível, tanto individual quanto coletivo.

Entre os esportes coletivos, destaca-se o basquete como o segundo esporte mais popular do mundo, sendo, além disso, um dos esportes que mais cresce, perdendo somente para o futebol [13]. O esporte foi criado em 1891 pelo professor de Educação Física canadense James Naismit, na Associação Cristã de Rapazes de Springfield, Massachusetts, Estados Unidos [3]. Logo após cinco anos, o Brasil começou com a prática do esporte e foi ganhando cada vez mais espaço. A Liga Nacional de Basquete (LNB) foi lançada em dezembro de 2008, reunindo as principais lideranças e os mais representativos clubes do basquetebol brasileiro [10]. Atualmente a LNB possui 21 clubes associados de sete Estados mais o Distrito federal. Entre os 21 associados, 16 clubes participam do NBB (Novo Basquete Brasil), campeonato nacional adulto de basquetebol.

As competições são um negócio muito lucrativo para os donos das equipes e as cidades das equipes, trazendo lucratividade também para as emissoras que transmitem os jogos, além de empregar permanentemente muitos terceiros nos processos de planejamento, infraestrutura e organização dos jogos. Não se deve esquecer também de mencionar que o esporte é uma boa fonte de entretenimento para todos.

Hoje em dia o basquetebol vem adquirindo bastante popularidade e com isso grandes investimentos estão sendo feitos no esporte. Como exemplo, tem-se a parceria da Liga Nacional de Basquete (LNB) com a famosa marca de marterias esportivos, a Nike. A parceria prevê identificação de novos talentos em um intercâmbio global e acesso direto aos maiores especialistas de basquete do mundo [15]. Investimentos como o da Nike e de outros patrocinadores fizeram com que o esporte fosse levado a outro patamar, permitindo contratações até mesmo em profissões ainda relativamente novas e inovadoras, como exemplo os cientistas de dados. Diante da transformação digital em um ritmo cada vez mais acelerado, a demanda por cientistas de dados tem aumentado no mercado, apesar de ser uma profissão nova, e os times de basquetebol têm buscado esses profissionais para auxíliar no crescimento da equipe. No que se refere a aspectos técnicos e táticos, a análise de dados estatísticos de jogos de basquetebol tem sido um

dos instrumentos mais utilizados por estudiosos para definir perfis de atletas e de equipes.

Devido ao seu grande potencial, a análise de dados estatísticos de jogos tem sido muito utilizada em vários estudos com objetivo de definir perfis de atletas ou equipes, traçar estratégias de ataque ou defesa no jogo, ou simplesmente buscar a otimização de custos, desempenho e aproveitamento. Exemplos de estudos com esse foco são os trabalhos de Rose Junior, Tavares e Gitti [5] e Carneiro, Souza e Costa [2].

De acordo com Junior, Tavares e Gitti [5], a análise de dados quantitativos de um jogo de basquetebol é um processo importante na explicação de fatores que influenciam no desempenho individual e coletivo dos atletas.

O estudo de Carneiro, Souza e Costa [2] buscou investigar a importância da análise estatística no jogo do basquete e o seu poder no auxílio ao plano de jogo. Foram feitas investigações por meio de questionários aplicados a treinadores e análises descritivas das técnicas e das estatísticas mais utilizadas por eles na formação e na técnica utilizada para definir as equipes.

Baseando nestas premissas, a estatística é uma ferramenta de suma importância para otimização destes resultados, como também evidencia a relevância da estatística para o aperfeiçoamento do esporte. Nesse contexto, a modelagem por meio da técnica do modelo de regressão linear múltipla (MRLM) constitui-se um método eficiente para que possamos avaliar as covariáveis que efetivamente são significativas para explicar a variável resposta.

A regressão linear múltipla é uma das técnicas que podem ser utilizadas para estimar um caminho nas decisões de seus técnicos e jogadores por meio da escolha das variáveis que podem influenciar no número de vitórias de uma equipe. Em vista disso, este trabalho visa realizar um estudo acerca dos resultados da Liga Nacional de Basquete, utilizando um MRLM para identificar quais variáveis ou estatísticas de jogo têm maior significância nos resultados positivos da equipe, no intuito de melhorar suas estratégias de treinos e práticas em quadra.

No intuito de alcançar este objetivo, este trabalho foi dividido em quatro partes. Além desta introdução e das considerações finais, na segunda seção discute-se os fundamentos da metodologia aplicada. Na terceira seção, faz-se uma breve análise das variáveis e a base de dados empregada. Por fim, na quarta seção são apresentados os resultados alcançados.

2. FUNDAMENTAÇÃO TEÓRICA

A principal metodologia adotada neste trabalho é a regressão linear múltipla, mas para fundamentar tal método são apresentados nesta seção os testes de significância do modelo, uma breve análise sobre os métodos de regressão, os testes necessários para a estimação e o método *Stepwise*.

2.1 INTRODUÇÃO AOS MODELOS DE REGRESSÃO

Conforme explica Maroco [12]:

> "O termo 'regressão' foi proposto pela primeira vez por Sir Francis Galton em 1885 num estudo onde demonstrou que a altura dos filhos não tende a refletir a altura dos pais, mas tende sim a regredir para a média da população. Atualmente, o termo "Análise de Regressão" define um conjunto vasto de técnicas estatísticas usadas para modelar relações entre variáveis e predizer o valor de uma ou mais variáveis dependentes (ou de resposta) a partir de um conjunto de variáveis independentes (ou preditoras)" [12].

A temática deste trabalho diz respeito à análise de regressão linear, no entanto realiza-se em seguida uma breve abordagem ao coeficiente de correlação e, consequentemente, ao coeficiente de determinação. A análise de correlação tem como objetivo a avaliação do grau de associação entre duas variáveis, X e Y ou seja, mede a "força" de relacionamento linear entre as variáveis X e Y. Para quantificar a relação entre duas variáveis quantitativas utiliza-se o coeficiente de correlação linear de Pearson.
O coeficiente de correlação linear de Pearson entre duas variáveis quantitativas, X e Y, é dado pela expressão (2.1) [8]:

$$R_{xy} = \frac{\sum_{i=1}^{n}(x_i - \widehat{x})(y_i - \widehat{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \widehat{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \widehat{y})^2}} \tag{2.1}$$

em que

$$\overline{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

e

$$\overline{y} = \sum_{i=1}^{n} \frac{y_i}{n}$$

A partir de R_{xy} podemos tirar conclusões sobre a direção e intensidade da relação existente entre as variáveis X e Y. Não existe uma "classificação" unânime de correlação, porém optou-se por seguir a considerada por Santos [17], apresentada na Tabela 2.1.

Tabela 2.1: Interpretação do coeficiente de correlação de Pearson

Coeficiente de Correlação	**Correlação**
$R_{xy} = 1$	Perfeita positiva
$0{,}8 \leq R_{xy} < 1$	Forte positiva
$0{,}5 \leq R_{xy} < 0,8$	Moderada positiva
$0{,}1 \leq R_{xy} < 0,5$	Fraca positiva
0	Nula
$-0{,}1 \leq R_{xy} < 0$	Infirma negativa
$-0{,}5 \leq R_{xy} < -0,1$	Fraca negativa
$-0{,}8 \leq R_{xy} < -0,5$	Moderada negativa
$-1 \leq R_{xy} < -0,8$	Forte negativa
$R_{xy} = -1$	Perfeita negativa

Para investigar a relação entre duas variáveis X e Y, podemos representar os valores das variáveis em um gráfico de dispersão. Afirma-se que existe uma relação linear entre as variáveis se os dados se aproximarem de uma linha reta. Com base na observação do diagrama de dispersão verificamos se a correlação entre as duas variáveis é mais ou menos forte, de acordo com a proximidade dos pontos em relação a uma reta.

Posteriromente, define-se o coeficiente de determinação, o qual faz-se igual ao quadrado do coeficiente de correlação de Pearson. Como vimos, o coeficiente de correlação linear de Pearson entre duas variáveis serve para medir a intensidade da relação linear entre elas. O coeficiente de determinação é mais indicado para medir a explicação da reta de regressão. Assim, quanto mais próximo de 1 estiver o valor do coeficiente de determinação, maior a porcentagem da variação de Y explicada pela reta estimada e, por conseguinte, maior a qualidade do ajustamento.

O coeficiente de determinação é dado pela expressão (2.2) [8]:

$$R_{xy}^2 = \frac{(\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}))^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{2.2}$$

O R_{xy}^2 toma valores entre zero e um. A qualidade do ajuste será tanto maior quanto mais

R^2_{xy} se aproximar de 1.

No caso da Regressão linear múltipla, o coeficiente de determinação também é usado para medir a qualidade do ajuste do modelo.

Em resumo, a presença ou ausência de relação linear pode ser averiguada a partir de dois pontos distintos:

a) quantificando a força dessa relação, para isso utiliza-se a análise de correlação;

b) ou explicando a forma dessa relação, fazendo uso da análise de regressão

Ambas as técnicas, apesar de intimamente ligadas, se diferem, pois na correlação todas as variáveis são aleatórias e desempenham o mesmo papel, não havendo nenhuma dependência, enquanto na regressão isso não acontece.

2.2 Introdução ao método de Regressão Linear Múltipla

A análise de regressão múltipla é uma técnica estatística, que pode ser usada para analisar a relação de causa e efeito entre uma única variável dependente e diversas variáveis independentes [7].

A análise de regressão múltipla tem por objetivo estimar o impacto do incremento de cada variável independente – que se traduz como peso de cada variável independente – sobre a respectiva variação da variável dependente. Os pesos denotam a contribuição relativa das variáveis independentes para a previsão geral e facilitam a interpretação sobre a influência de cada variável explicativa em fazer a previsão. [6][7].

Segundo Charnet [4], o modelo de regressão múltipla é dada pela expressão (2.3):

$$Y_i = \beta_o + \beta_{1i}x_{1i} + ... + \beta_k x_{ki} + u_i \tag{2.3}$$

em que: Y_i é o fenômeno em estudo (variável dependente); $\beta_1, \beta_2, ..., \beta_k$ são os coeficientes associados a cada variável independente (coeficientes angulares); x_{ki} são as variáveis explicativas (independentes) com i = 1,2,...,n; por fim, u_i é o termo do erro aleatório e não observável.

O erro ui é uma variável aleatória não observável, e representa possíveis variáveis que não

foram inseridas no modelo, mas que também contribuíram para a explicação de Y_i, em que u_i é conforme a expressão (2.4):

$$u_i = N(0, \sigma^2) \tag{2.4}$$

Ao estabelecer um modelo de regressão, é necessário segui algumas pressuposições, que segundo Hoffmann [8] são:

- a variável dependente Y_j é função linear das variáveis explanatórias $(X_{ij}, i = 1, ..., k)$;
- os valores das variáveis explanatórias são fixos;
- $E(u_i) = 0$, ou seja, $E(u) = 0$, onde 0 representa um vetor de zeros;
- os erros são homocedásticos, isto é, $E(u_j^2) = \sigma^2$;
- os erros são não-correlacionados entre si, isto é, $E(u_j u_h) = 0$ para $j \neq h$;
- os erros têm distribuição normal.

O modelo é estimado por meio do Método de Mínimos Quadrados. A seguir será apresentada a descrição dos Métodos Mínimos Quadrados Ordinários.

2.2.1 Método dos Mínimos Quadrados Ordinários

Recomendado pela sua precisão, o Método do Mínimos Quadrados Ordinários (MQO) consiste em determinar os estimadores que minimizam a soma de quadrados dos resíduos (HOFFMANN; VIEIRA, 1998). Considerando o modelo de regressão linear múltipla com k variáveis independentes na forma matricial tem-se a expressão (2.5):

$$y = X\beta + u \tag{2.5}$$

em que

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{m1} \\ 1 & x_{12} & x_{22} & \cdots & x_{m2} \\ 1 & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{mn} \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix}$$

$$u = \begin{bmatrix} u_o \\ u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

Sejam $\widehat{\beta}$ e $\widehat{u}$ os vetores dos estimadores dos parâmetros e dos erros (resíduos), respectivamente, isto é:

$$\widehat{\beta} = \begin{bmatrix} \widehat{\beta_0} \\ \widehat{\beta_1} \\ \widehat{\beta_2} \\ \vdots \\ \widehat{\beta_k} \end{bmatrix}$$

e

$$\widehat{u} = \begin{bmatrix} \widehat{u_0} \\ \widehat{u_1} \\ \widehat{u_2} \\ \vdots \\ \widehat{u_k} \end{bmatrix}$$

tem-se

$$\widehat{y} = X\widehat{\beta}$$

e os resíduos

$$\widehat{u} = y - X\widehat{\beta} = y - \widehat{y}$$

A soma dos quadrados dos desvios é dada por:

$$Z = e'e = (y' - b'X')(y - Xb) = y'y - y'Xb - b'X'y + b'X'Xb$$

A soma dos quadrados dos resíduos matricialmente é dada pela expressão (2.6):

$$Z = \widehat{\beta}'X'y + \widehat{\beta}'X'X\widehat{\beta} \tag{2.6}$$

A função Z apresenta ponto de mínimo para os valores de β que tornem a diferencial identicamente nula, isto é:

$$\frac{dz}{d\widehat{\beta}} = -2(d\widehat{\beta}')X'y + (d\widehat{\beta}')X'X\widehat{\beta} + \widehat{\beta}'X'X(d\widehat{\beta}) = 0 \tag{2.7}$$

Como X'X é uma matriz quadrada de ordem p e simétrica, então, pela propriedade reflexiva da transposta tem-se:

$$(d\widehat{\beta}')X'X\widehat{\beta} = \widehat{\beta}'X'X(d\widehat{\beta}) \tag{2.8}$$

Sendo assim, a expressão (2.7) pode ser escrita como:

$$-2(d\widehat{\beta}')X'y + 2(d\widehat{\beta}')X'X\widehat{\beta} \Leftrightarrow (d\widehat{\beta}')(X'X\widehat{\beta} - X'y) = 0 \tag{2.9}$$

Portanto, a diferencial de Z será identicamente nula para:

$$X'X\widehat{\beta} = X'y \tag{2.10}$$

que representa o sistema de equações normais (SEN).

Se X'X é não-regular, existe a matriz inversa $(X'X)^{-1}$. Pré-multiplicando os dois membros da expressão (2.10) por $(X'X)^{-1}$, obtém-se o estimador de β:

$$\widehat{\beta} = (X'X)^{-1}X'y \tag{2.11}$$

2.3 TESTE DE SIGNIFICÂNCIA DO MODELO

O teste de hipótese tem como base uma estatística de distribuição F, com k e (n – k – 1) graus de liberdade, sob H_0. Os k graus de liberdade se devem ao fato de termos $p = k + 1$ parâmetros das variáveis regressoras. As quantidades para calcular o valor observado da estatística são dispostas na Tabela 2.3 de Análise de Variância (ANAVA). São decorrentes da soma de quadrados total de Y (SQT), dívidas em 2 componentes: soma de quadrados da regressão ($SQReg$) e a soma de quadrados dos resíduos (SQE) [4].

$$SQT = SQReg + SQE \tag{2.12}$$

De acordo com a expressão (2.12), tem-se:

$$\sum_{i=1}^{n}(y_i - \widehat{y})^2 = \sum_{i=1}^{n}(\widehat{y} - \overline{y})^2 + \sum_{i=1}^{n}(y_i - \widehat{y})^2 \tag{2.13}$$

Por definição, os quadrados médios são obtidos dividindo as somas de quadrados pelos respectivos graus de liberdade.

Então, para o caso de uma regressão linear múltipla, tem-se:

$$QMReg = \frac{SQReg}{k}$$

e

$$QMRes = \frac{SQE}{n - k - 1}$$

No modelo de regressão, o teste F é aplicado para validar se as variáveis regressoras, contribuem de maneira significativa para explicação da variável resposta. As hipóteses a serem testadas são: $H_0 : \beta_1 = 0, ..., \beta_k = 0$ contra H_1 pelo menos um dos betas difere de zero. Pode-se dizer que a hipótese específica que o modelo não é significativo, e em o modelo é significativo.

A estatística do teste é dada pela expressão (2.14):

$$F_0 = \frac{QMReg}{QMRes} \sim F_{(k,n-k-1)} \tag{2.14}$$

em que $F_0 = F_{calculado}$ e $F_{(k,n-k-1)} = F_{tabelado}$ é o valor da tabela F com k e $n-k-1$ graus de liberdade e um nível α de significância.

A hipótese $H_0 : \beta_1 = 0, ..., \beta_k = 0$ é rejeitada quando $F_{calculado} > F_{tabelado}$ ao ao n ível α de significância, com k e (n – k – 1) graus de liberdade. Para $F_{calculado} < F_{tabelado}$ aceita-se a hipótese H_0 ao nível α de significância, atestando que há indícios de ausência de relação linear entre as variáveis.

Tabela 2.2: ANAVA para o modelo regressão linear múltipla (MRLM)

FV (Fonte de Variação	**GL** (Graus de Liberdade)	**SQ** (Soma de Quadrados)	**QM** (Quadrado Médio)	F_0
Regressão	k	SQR	$\frac{SQR_{eg}}{k}$	$\frac{QMR_{eg}}{QMR_{es}}$
Erro	n-k-1	SQE	$\frac{SQE}{k-n-1}$	
Total	n-1	SQT		

Em um MRLM é interessante testar significância de cada um dos coeficientes de regressão, ou seja, existe interesse em verificar se a contribuição de uma variável regressora é significativa ou não [4].

Para testar individualmente as variáveis, $X_k, K = 1, 2, 3, ..., p$. Tem-se as seguintes hipóteses $H_0 : \beta_k = 0$ contra $H_1 : \beta_k \neq 0$.

A estatística do teste é dada pela expressão (2.15):

$$T_{\beta k} = \frac{\widehat{\beta_k} - \beta_k}{S(\widehat{\beta_k})} \sim t_{n-p} \tag{2.15}$$

em que $T_{\beta k} = T_{calculado}$ e $t_{(n-p)} = t_{tabelado}$ é o valor da tabela t de Student com n-p graus de liberdade em um nível α de significância. A hipótese H_0 é rejeitada ao nível de significância α se $t_{calculado} > t_{tabelado}$, atestando que a variável foi significativa para explicação do modelo. Para $t_{calculado} < t_{tabelado}$ ao nível de significância α a hipótese H_0 é aceita, conclui-se que a variável não foi significativa para o modelo linear.

2.4 Análise de Resíduos

De acordo com [11],os resíduos de um modelo de regressão linear têm uma relação muito forte com a qualidade do ajuste, bem como com a confiabilidade dos testes estatísticos sobre

os parâmetros do modelo. Nesse sentido, a análise de resíduos tem uma importância fundamental na verificação da qualidade do ajuste de modelo. Basicamente, essa análise fornece evidências sobre possíveis violações nas suposições do modelo, tais como a de normalidade, homocedasticidade, e quando for o caso ainda fornece indícios de falta de ajuste do modelo proposto.

O vetor de resíduos é definido por [4]:

$$\varepsilon = Y - X\beta$$

Assim, a esperança dos resíduos definidas respectivamente por:

$$E[\varepsilon] = E[Y - X\beta] = 0$$

e

$$Var[\varepsilon] = \sigma^2[I - X(X'X)^{-1}X']$$

em que I é a matriz identidade.

Caso o modelo de regressão esteja correto o conjunto de resíduos terão a mesma variância e serão adequados para a verificação de normalidade e homocedasticidade (variância constante) dos erros. As observações que possuírem os valores absolutos dos resíduos padronizados maiores que dois poderão ser consideradas pontos discrepantes. [21].

A análise dos resíduos do modelo pode ser realizada por meio de técnicas formais (testes de hipóteses) ou por meio de técnicas informais (verificação gráfica), em que existe a possibilidade de rejeitar com facilidade as pressuposições dos modelos.

Os gráficos mais utilizados nas técnicas informais são:

- QQ-plot dos resíduos: avaliar se os resíduos seguem distribuição normal de probabilidade;
- Resíduos versus medida *Leverage*: avaliar se há alguma observação influente (outliers);
- Gráfico de dispersão dos resíduos em função dos valores estimados;
- Gráfico de dispersão dos resíduos padronizados pelos valores ajustados.

2.5 MULTICOLINEARIDADE

Ademais, outro aspecto importante no ajuste de modelos de regressão linear múltipla se encontra na multicolinearidade. Tem-se como objetivo investigar se há multicolinearidade entre as variáveis regressoras, visto que a forte correlação entre elas pode resultar em efeitos problemas no ajuste do modelo. A multicolinearidade é um problema comum em regressão lienar,

indicando que existe uma relação de linearidade entre as variáveis independentes, prejudicando a estimação dos coeficientes da regressão [21].

Assim sendo, uma das formas de identificar a presença de multicolinearidade é avaliar o Fator de Inflação da Variância (VIF – *Variance Inflation Factor*). Esse fator mede a associação entre as variáveis regressoras de acordo com o coeficiente de determinação do modelo de regressão, apenas com as variáveis independentes. De acordo com Berk (1977) Fator de Inflação da Variância é definido de acordo com a expressão (2.16) como:

$$VIF_i = \frac{1}{1 - R_i^2} \tag{2.16}$$

onde R_i^2 é o coeficiente de determinação da regressão da variável explicativa X_i sobre as outras variáveis explicativas com i=1,2,...,k, sendo k quantidade de variáveis explicativas no modelo [21]. Geralmente, considera-se VIF $>$ 10 como um indicativo de problema de multicolinearidade.

2.6 TESTE DE SHAPIRO-WILK

No modelo de regressão o teste de normalidade é utilizado para determinar se a distribuição dos resíduos é adequada à uma distribuição normal, utiliza-se o teste de Shapiro-Wilk para realizar essa verificação. As hipóteses dos testes são: H_0 os resíduos seguem normalidade versus H_1 : os resíduos não seguem normalidade.

De acordo com Shapiro [18], a estatística do teste é dada pela equação (2.17):

$$W = \frac{\left[\sum_{i=1}^{\frac{n}{2}} a_i y_i\right]^2}{\sum_{i=1}^{n}(y_i - \widehat{y})^2} \tag{2.17}$$

desde que

$$(\sum_{i=1}^{\frac{n}{2}} a_i y_i) \leq \sum_{i=1}^{\frac{n}{2}} a_i^2 \sum_{i=1}^{\frac{n}{2}} y_i^2 = \sum_{i=1}^{\frac{n}{2}} y_i^2$$

em que, a_i é o melhor estimador não-viciado do valor esperado das estatísticas de ordem de uma amostra de tamanho n com distribuição normal. Realiza-se o teste de normalidade, rejeita-se

H_0 a um nível de significância α se p-valor $< \alpha$. A tabela W indica a porcentagem empírica aproximada dos pontos.

2.7 TESTE DE DUBIN-WATSON

Um dos testes mais conhecidos para verificação de autocorrelação temporal é a estatística d, de Durbin-Watson, que envolve o cálculo de um teste estatístico baseado nos resíduos do método de regressão de mínimos quadrados.

Sejam $\widehat{e} = (t = 1, 2, 3, ..., n)$, os resíduos da regressão ajustada por Mínimos Quadrados, então se tem a razão entre a soma das diferenças, elevadas ao quadrado, entre sucessivos resíduos e a soma do quadrado dos resíduos (SQR):

$$d = \frac{\sum_{t=2}^{n}(\widehat{e}_t - \widehat{e}_{t-1})}{\sum_{t-1}^{n}\widehat{e}_t^2} \tag{2.18}$$

Durbin e Watson tabularam os limites inferiores d_L, e os limites superiores, d_s , para vários valores e n (número de dados) e k (número de variáveis explicativas), tais que, se o d calculado estiver fora desses valores críticos, é possível a verificação da autocorrelação.

Para efetuar um teste de autocorrelação positiva, calcula-se d conforme a equação (2.18):

- Quando valores sucessivos de $\widehat{e}_t$ estão próximos um dos outros, a Estatística d será baixa, indicando a presença de correlação serial positiva, ou seja, se d$< d_L$, rejeita-se a hipótese de erros aleatórios e aceita a autocorrelação positiva;
- Se d$>d_u$, não se rejeita a hipótese e não existe autocorrelação (p = 0);
- Se $d_L < d < d_u$, o teste é inconclusivo e idealmente seriam necessárias observações adicionais.

A interpretação exata da estatística de Durbin-Watson é difícil porque a sequência dos termos de erro depende não só das sequências dos ê, mas também da sequência de todos os valores de X. [14].

2.8 TESTE DE BREUSCH-PAGAN

Uma das maneiras de verificar a suposição de homogeneidade de variâncias, ou seja, se a variância dos erros permanece constante para todas as observações $Var(e_i) = \sigma_e^2$, com i=1,. . . ,n é a aplicação do teste de Breusch-Pagan [1]. A dinâmica do teste é descrita a seguir, conforme Neto [16].

Para obtenção da estatística do teste, ajusta-se o modelo de regressão, encontram-se os resíduos (e_i) e o valores ajustados ($\widehat{y}$) para as n observações e, ainda, consideram-se os resíduos (μ_i) padronizados dados por:

$$\mu_i = \frac{e_i^2}{\frac{SQE}{n}} \tag{2.19}$$

em que

$$SQE = \sum_{i=1}^{n} e_i^2$$

.

Posteriormente, ajusta-se um modelo de regressão auxiliar tendo como variável resposta os resíduos padronizados $\mu = (\mu_1, ..., \mu_n)$ e como variável explicativa os valores ajustados $\widehat{y} = (\widehat{y}_1, ..., \widehat{y}_n)$. Em seguida obtêm-se a estatística X_{BP}^2, também conhecida por estatística LM dada por $LM = n * R_e^2$, em que R_e^2 é o coeficiente de determinação do modelo de regressão auxiliar. Sob a hipótese nula de homocedasticidade, a estatística de teste segue uma distribuição qui-quadrado com k graus de liberdade, sendo k o número de variáveis explicativas do modelo.

2.9 MÉTODO STEPWISE

Existem três formas de se realizar a seleção de variáveis em uma regressão: (1) *forward* - quando a equação começa apenas com o intercepto e cada preditor entra, um por um, na equação; (2) *backward* – quando todos os preditores são incluídos de uma só vez na equação, e depois são retiradas, um a um, até que se identifiquem os melhores preditores; (3) *blockwise* ou *stepwise* – assemelha-se à regressão *stepwise forward*, mas ao invés de os preditores serem incluídos individualmente, o processo adiciona sistematicamente a variável mais significativa ou remove a variável menos significativa durante cada etapa. [9] [19].

Geralmente, a estratégia escolhida para estudos exploratórios é a regressão stepwise. Quando é utilizando este tipo de regressão, o pesquisador, desprovido de uma teoria consistente sobre os fenônemos estudados, está interessado em apenas em descrever relacionamentos poucos conhecidos entre variáveis, e não em os explicar de forma estrita. Neste tipo de regressão, a seleção da sequência de entrada dos preditores na equação é feita estatisticamente, sem um modelo teórico consistente a ser seguido. Em estudos exploratórios, os pesquisadores elaboram um modelo teórico de investigação que inclui hipóteses sobre relacionamentos entre variáveis, mas que ainda impossibilita afirmações consistentes sobre a magnitude ou direção desses relacionamentos. Além disso, este tipo de estudo ainda não encontra apoio empírico às hipóteses a serem testadas.

Segundo Alves (2013) no modelo de regressão pode possuir variáveis que discriminam pouco a resposta. O método *Stepwise* é utilizado para variáveis que mais discriminam a resposta, diminuindo o número de variáveis no modelo final. Considerando que esse método é realizado de forma interativa, adicionando variáveis (*forward*) e removendo variáveis (*backward*), neste trabalho utilizaremos o metodo stepwise para selecionar as melhores variáveis que correspondem positivamente a variável resposta, partir de um critério de seleção com base no teste F, correlação linear múltipla e erro quadrático total.

3. METODOLOGIA

3.1 MATERIAL E MÉTODOS

Para o atual trabalho, os dados foram coletados no site oficial da *Liga Nacional de Basquete* – LNB [10]. Todas as informações foram obtidas das temporadas regulares de 2016 a 2020, sendo um total de 5 temporadas. Cada temporada é composta por 16 equipes, sendo que cada equipe participou de 28 jogos por temporada. O conjunto de dados tem um total de 80 observações, das quais cada linha de informações contém os valores das variáveis preditores ou explicativas do trabalho. No banco de dados considerado, foram coletados 29 variáveis para cada equipe por temporada de acordo com a Tabela 3.1, porém 6 dessas variáveis foram excluídas. Variáveis como 2PC, 3PC e LLC correspondiam aos mesmo valores que X2P, X3P e LL respectivamente. Como as variáveis JO e MIN são iguais para todos as equipes também não foram utilizadas nas análises e a variáveis PLL e EF é uma estatística calculada apartir de outras variáveis. Portanto sobraram 21 covariáveis e mais variável resposta (VITORIA) para o ajuste inicial do modelo.

Neste trabalho, utilizou-se uma estatística descritiva para analisar as características dos dados.A estatística descritiva fornece resumos simples sobre a amostra e sobre as observações que foram feitas. Tal resumo pode ser quantitativo ou visual. Esses resumos podem formar a base da descrição inicial dos dados, tanto como parte de uma análise estatística mais extensa ou como valores suficientes por si mesmos.

Utilizou-se ainda uma técnica de ajuste de modelos de regressão linear múltipla em que a escolha das variáveis preditivas é realizada automaticamente por meio do método de *stepwise* com auxílio do software R [20] .

Tabela 3.1: Variáveis Utilizadas para as Análises

Estatística	Descrição
JO	Número de partidas
MIN	Minutos em quadra
PTS	Pontos
X3P	Cestas de 3 pontos
X2P	Cestas de 2 pontos
LL	Lances Livres
RT	Rebotes
RD	Rebotes defensivos
RO	Rebotes ofensivos
IA	Índice de assistências por erros
2PC	Arremessos de 2 pontos certos
X2PT	Arremessos de 2 pontos tentados
3PC	Arremessos de 3 pontos certos
X3PT	Arremessos de 3 pontos tentados
PLL	Porcentagem de lances livres convertidos
LLC	Lances livrres certos
LLT	Lances livres tentados
EM	Enterradas
BE	Índice de bolas recuperadas por erros
BR	Bolas recuperadas
ER	Erros
FC	Faltas cometidas
TO	Tocos (ação de defesa)
ET	Total de erros (soma de erros e violações)
EF	Eficiência
TFC	Índice de tocos or faltas cometidas
VI	Violações
AS	Assistências
VITORIA	Quantidade de vitórias

4. RESULTADOS

A análise empírica foi feita visando investigar quais variáveis são significativas para obter o maior número de vitórias dos times na *Liga Nacional de Basquete* (LNB). Nessa seção foi feito uma análise da estatística descritiva, estimação do modelo, analise do gráfico de correlação das variáveis, analise de resíduos para avaliar a qualidade do modelo e também estimação do modelo indicando quais variáveis foram significativas.

4.1 ESTATÍSTICA DESCRITIVA

A partir dos dados coletados das cinco temporadas da LNB, na Tabela 4.1, apresentam-se as estatísticas descritivas das 21 variáveis explicativas e também da variável resposta VITORIA (quantidade de vitórias) que analisamos durante o trabalho.

Tabela 4.1: Estatística Descritiva da Variável Resposta e Variáveis Regressoras

Variáveis	**Mín**	**Média**	**Máx**	**Mediana**	**Var(S^2)**	**Desv. Padrão (S)**
PTS	66,89	78,95	89,18	78,31	26,47	5,14
X3P	17,25	26,79	39,95	26,11	17,56	4,19
X2P	31,14	37,85	45,78	37,84	9,05	3,01
LL	10,50	14,34	19,0	14,20	3,21	1,79
RT	32,18	36,87	42,79	36,53	4,84	2,20
RO	8,19	10,40	13,05	10,39	1,36	1,17
RD	22,36	26,47	30,77	26,27	2,88	1,70
AS	12,46	16,92	20,74	17,02	3,84	1,96
ER	10,87	13,65	16,25	13,67	1,43	1,20
IA	0,88	1,25	1,82	1,20	0,05	0,22
X3PT	20,04	26,33	37,13	26,16	9,85	3,14
X2PT	30,97	37,01	42,41	36,93	6,64	2,58
LLT	15,10	19,60	25,16	19,54	4,24	2,06
EM	0,50	1,44	3,56	1,30	0,32	0,57
BR	5,19	6,97	8,71	7,11	0,60	0,78
BE	0,39	0,51	0,66	0,51	0,01	0,07
TO	1,08	2,02	3,10	1,92	0,22	0,47
FC	15,47	20,54	23,75	20,39	3,31	1,82
TFC	0,05	0,099	0,19	0,095	0,00	0,03
VI	1,43	2,44	5,64	2,29	0,51	0,71
EF	65,39	87,60	108,20	86,77	97,01	9,85
VITORIA	3,00	13,62	28	13,50	58,01	7,62

Nesta tabela 4.1 encontram-se os resultados, os valores mínimos, média, valores máximos, mediana, variância e desvio padrão, respectivamente.

O boxplot pode ajudar a visualizar o centro, a dispersão e a assimetria de um conjunto de dados. Além disso, ele é excelente para identificar e controlar valores extremos (outliers). Com ele, é possível identificar facilmente, qual é o valor que representa a mediana, que em alguns casos é o valor que representa melhor os dados.

Para visualizar melhor esses gráficos, o conjunto de covariáveis foi dividido em 3 grupos, e os box plots dessas covariáveis, mais a variável resposta foram representados nas Figuras 4.1, 4.2 e 4.3, como forma de avaliar a distribuição e a dispersão das variáveis de forma individual, não havendo interesse em comparara a variabilidade entre elas.

Na Figura 4.1 apresenta-se o Box plot para 8 covariáveis do conjunto de dados, observando-se uma aparente simetria na distribuição de cada variável, com a linha que representa a mediana presente no centro dos retângulos. Para as covariáveis X3P, RT e RD destacaram-se alguns valores discrepantes, porém, estas não significam nenhuma irregularidade, sendo apenas característica intrínseca do tipo de dados sob estudo.

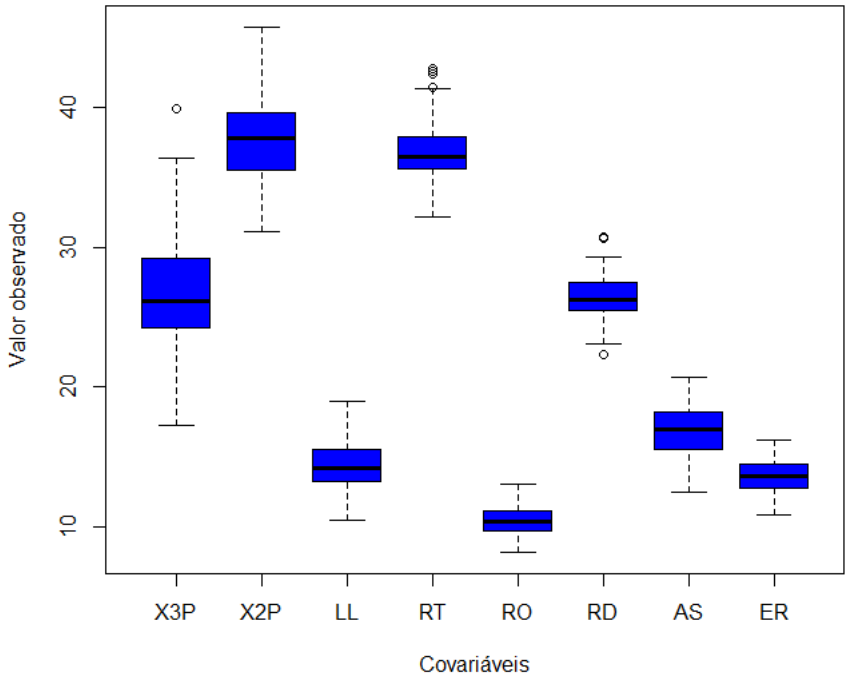

Figura 4.1: Box plot para covariáveis 1

Observa-se na Figura 4.2 que as covariáveis BE e TFC tiveram as menores variações diante das 8 covariáveis apresentadas. Conclui-se que as covariáveis EM, TFC e VI apresentaram alguns valores discrepantes, porém estas não significam que há irregularidade.

Com base na representação das 5 covariáveis na Figura 4.3, nota-se que todas as covariáveis apresentam uma simetria na sua distribuição. Já a variável resposta (VITORIA) demonstrou

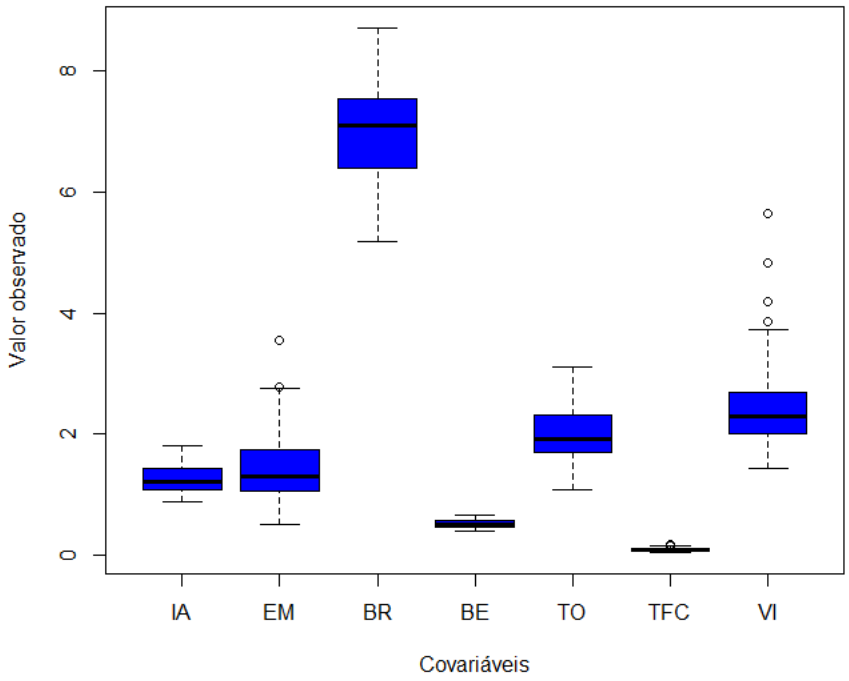

Figura 4.2: Box plot para covariáveis 2

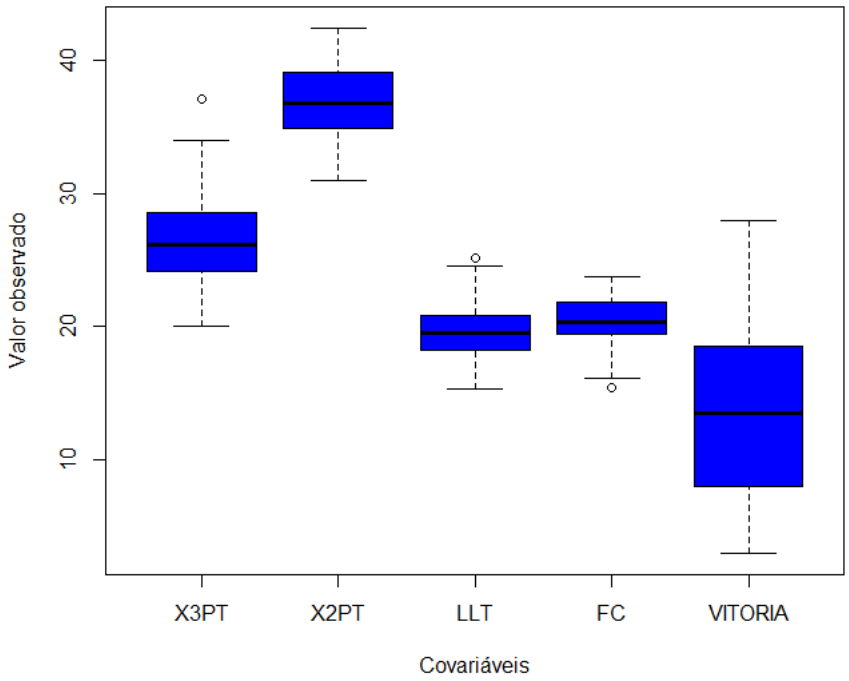

Figura 4.3: Box plot para covariáveis 3

uma maior variação em relação às outras covariáveis. Há alguns valores discrepantes nas covariáveis X3PT, LLT e FC, mas não resultando nenhuma irregularidade.

Uma ferramenta muito utilizada para análises descritivas é o histograma. Na Figura 4.4 apresenta-se um histograma da variável resposta (quantidade de vitória), possibilitando verificar características que se aproximam de uma distribuição normal de probabilidade, o que é desejável para a variável resposta. Observamos nos valores do histograma que a maioria dos times obtiveram quantidades de vitórias entre 10 e 15, e também que menos de 5 times tiveram quantidades de vitórias entre 25 e 30.

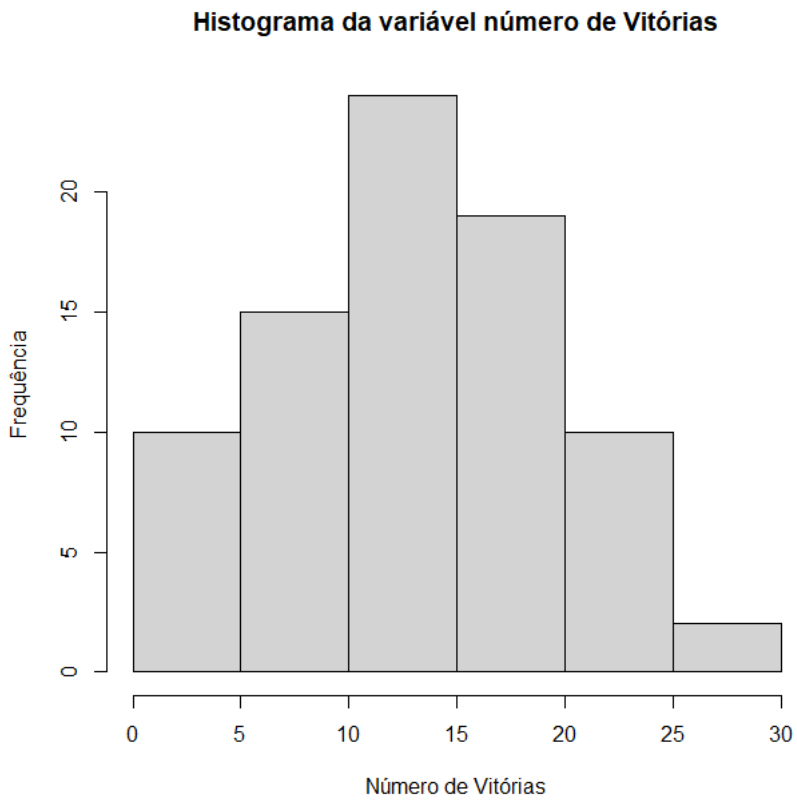

Figura 4.4: Histograma da Variável Resposta (número de vitórias)

A dispersão entre as variáveis PTS e VITORIA apresentada Figura 4.5, demonstra uma tendência linear crescente, ou seja , conforme o número de pontos aumenta o número de vitórias também está aumentando. Pelo fato de que uma variável é consequência da outra, a covariável Pts foi retirada da análise.

Verifica-se a Figura 4.6 que há uma relação evidente entre a variável EF e a quantidade de vitórias. De acordo com o aumento de EF, o número de vitórias também está aumentando. Portanto graficamente a variável EF tem uma relação positiva muito forte com a variável número de vitórias. Além disso, a variável EF é calculada com base nas demais covariáveis, sendo desnecessário considerá-la no modelo.

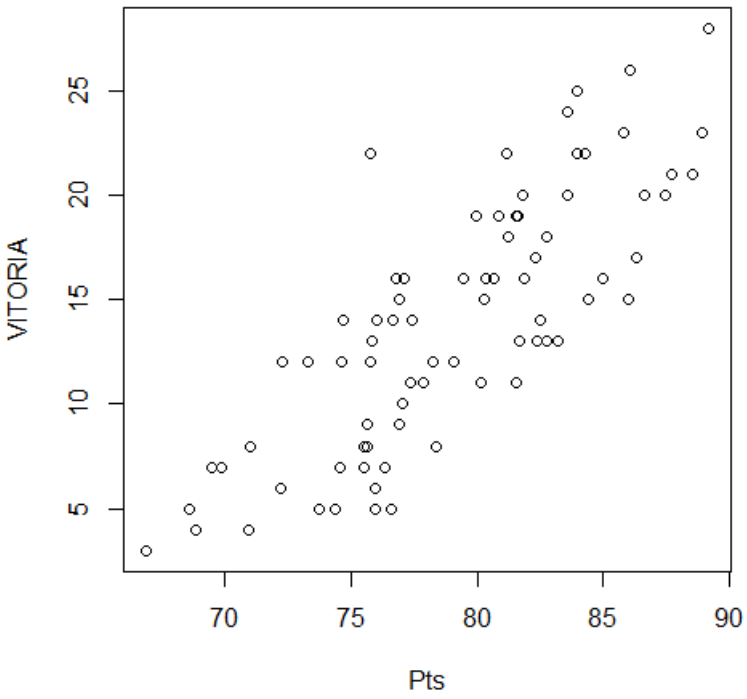

Figura 4.5: Dispersão entre PTS e VITORIA

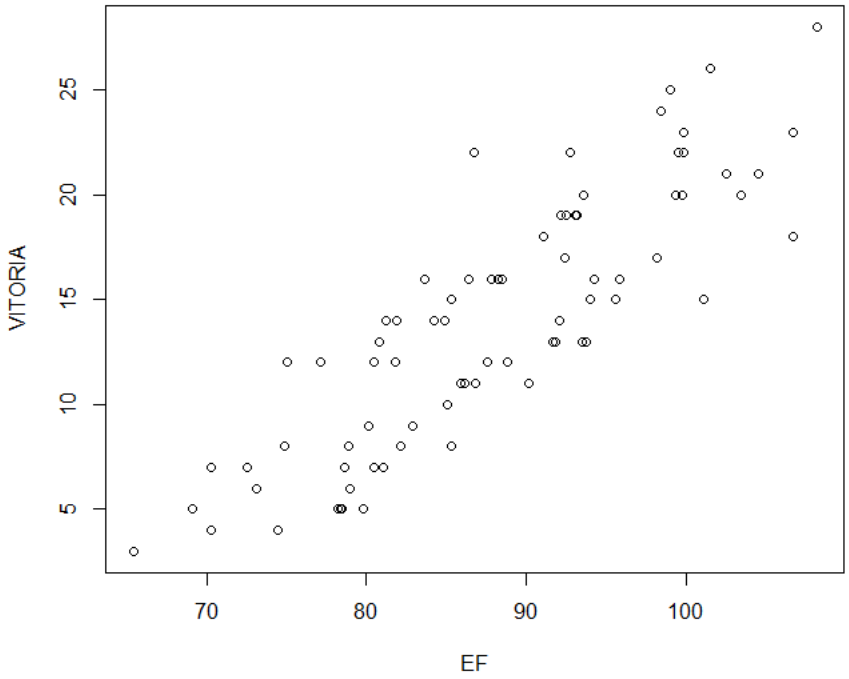

Figura 4.6: Gráfico de Dispersão entre EF e VITORIA

Para avaliar a correlação entre as variáveis explicativas, utilizou-se a representação gráfica apresentada no Figura 4.7.

De acordo com a Figura 4.7, as variáveis mais correlacionadas de forma positiva com a variável resposta (VITORIA) são: PTS, X3P, X2PT, RT, RD, AS, IA, BR, BE, TO e TFC. Entre as variáveis dependentes que estão mais associadas, a correlação de Pts com X3P e X2P é alta (isso é esperado devido a variável Pts ser a soma da X3P e X2P), porém ela também tem uma correlação alta com AS e IA.

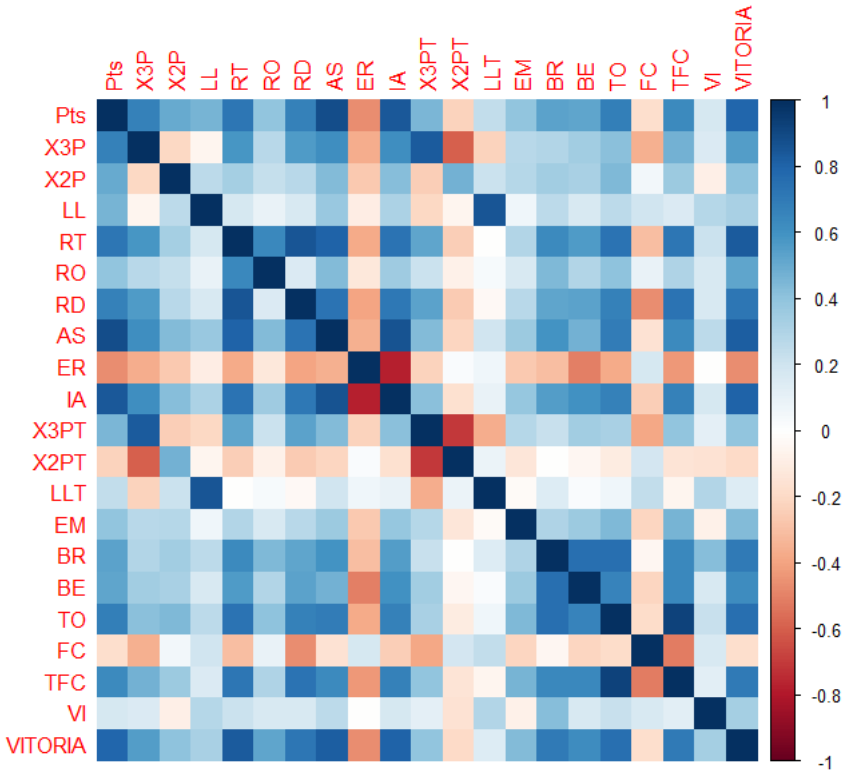

Figura 4.7: Gráfico de Correlação das Variáveis

Nota-se que as variáveis explicativas que estão correlacionadas com a variável resposta, também estão muito correlacionadas entre elas, podendo gerar um problema de multicolinearidade no modelo. No entanto, quando se analisa os valores dessas correlações, não se encontra muitos valores maiores que 0,8, sendo assim, espera-se que o próprio método de seleção de variáveis já elimine aquelas variáveis mais correlacionadas com outras, isentando o modelo final dessa multicolinearidade. Além disso, variável resposta já é uma consequência direta da variável número de pontos (PTS), sendo assim, ela não foi incluída no modelo inicial.

4.2 Estimação do Modelo de Regressão

Partindo do princípio do modelo completo, com todas as 21 variáveis regressoras, foi utilizado o método stepwise de seleção de variáveis que resultou em um modelo de regressão linear

múltipla com 9 variáveis regressoras, com significância dos parâmetros, a um nível de significância de 10%, conforme a Tabela 4.2. Isso pode ser observado nesta Tabela, já que os valores de pvalor apresentados são todos menores que esse nível de significância.

As variáveis regressoras do modelo final foram: X3P (cestas de 3 pontos) ; X2P (cestas de 2 pontos); RT (rebotes); IA (índice de assistências por erros); X3PT (arremessos de 3 pontos tentados); X2PT (arremessos de 2 pontos tentados); EM (enterrradas); BE (índices de bolas recuperadas por erros); VI (violações).

O coeficiente de determinação deste modelo foi igual a 85,6% afirmando que boa parte da variabilidade da variável resposta pode ser explicada pelo modelo de regressão.

Tabela 4.2: Estimativa para o Modelo Final de Regressão

Coeficientes	**Estimativa**	**Erro**	**Estatística t**	**p-valor**
Intercepto	-47,21	9,551	-4,943	<6e-06
X3P	0,414	0,162	2,555	0,0127
X2P	0,278	0,156	1,782	0,079
RT	1,376	0,216	6,378	<2e-08
IA	4,515	2,683	1,683	0,0969
X3PT	-0,655	0,201	-3,254	0,0017
X2PT	-0,319	0,175	-1,824	0,0723
EM	1,556	0,569	2,735	0,0079
BE	11,768	5,361	2,195	0,0314
VI	1,472	0,417	3,525	0,0007

Para averiguar se existe multicolinearidade entre as variáveis explicativas, calculou-se o fator de inflação da variância (VIF – *Variance Inflation Factor*), como apresentado na Tabela 4.3. Considerando que os valores foram todos menores que 10, pode-se concluir que não há multicolinearidade no modelo.

Tabela 4.3: Valores do Fator de Inflação da Variância para as Covariáveis

Variáveis	X3P	X2P	RT	IA	X3PT	X2PT	EM	BE	VI
VIF	6,00	2,86	2,93	4,40	5,19	2,65	1,37	1,88	1,15

4.3 Análise dos Resíduos do modelo

Para avaliar a qualidade do ajuste do modelo de Regressão Linear Múltipla, utilizou-se a verificação gráfica e também a verificação formal, por meio de testes que avaliam se as pressuposições do modelo estão sendo atendidas.

No primeiro gráfico da Figura 4.8, tem-se os resíduos em função dos valores estimados pelo modelo. Neste gráfico é possível observar a independência e a homocedasticidade, considerando que os pontos se distribuem de maneira razoavelmente aleatória e com amplitudes similares em

torno do zero. Verifica-se ainda que os pontos não estão seguindo uma tendência específica, indicando que o modelo proposto apresenta um bom ajuste. Pode-se observar também alguns valores discrepantes, no entanto, para o tipo de variável que se está considerando, é natural que apareçam times com número muito discrepante de vitórias dos demais, tanto para mais como para menos. Dessa forma, não se faz necessário excluir esses pontos da análise.

O gráfico Normal Q-Q plot indica a normalidade dos resíduos, visto que praticamente todos os pontos muito próximos da reta de referência. No terceiro gráfico da Figura 4.8, dos resíduos padronizados pelos valores ajustados, pode-se fazer as mesmas observações em relação ao primeiro gráfico. A aleatoriedadade dos pontos e a ausência de um padrão nas observações indica a independência dos resíduos e a homogeneidade de variâncias, levando também à conclusão de um bom ajuste.

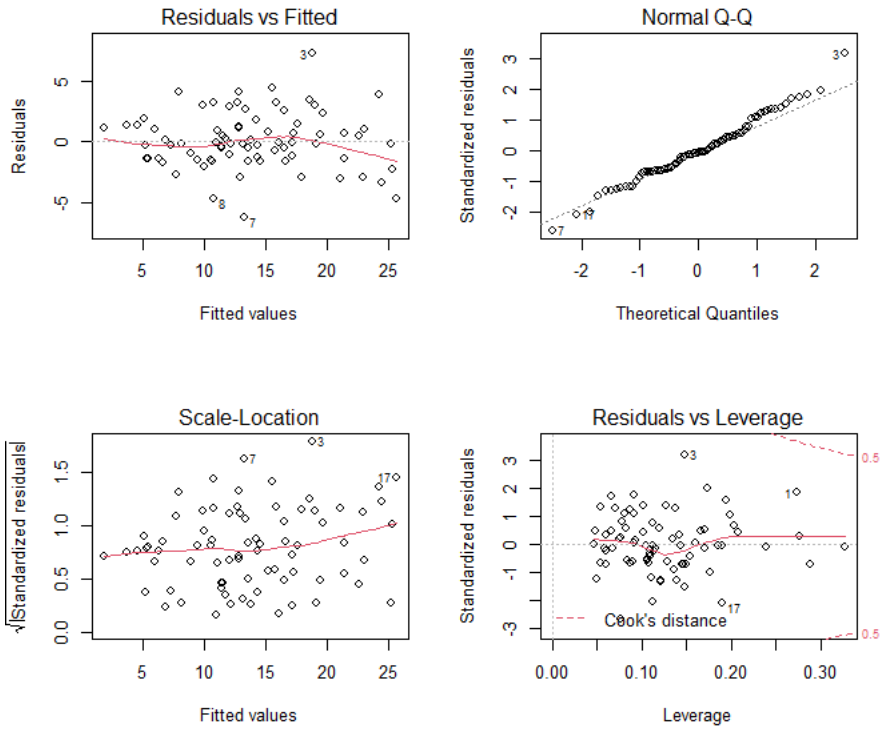

Figura 4.8: Gráficos para análise dos resíduos do modelo ajustado

Já no Gráfico 4 da Figura 4.8, referente aos resíduos padronizados contra a leverage, a distância de Cook maior que 1 pode indicar presença de outliers. Neste gráfico, verifica-se comportamento aleatório semelhante, como também se observam alguns pontos discrepantes dos demais, podendo ser possíveis pontos de alavancagem. Porém, não devem ser retirados da análise, pois estes possuem grande relevância para indicar as variáveis mais importantes na predição do número de vitórias.

O teste de Shapiro Wilk para normalidade apresentou uma estatística de teste de 0,98 com um p-valor de 0,517, podendo-se assim não rejeitar a hipótese nula de normalidade dos resíduos.

Desse modo, o teste permite concluir, como na verificação gráfica, que os resíduos do modelo ajustado são normalmente distribuídos.

A verificação da suposição de variância constante foi feita por meio do teste de Breusch-Pagan, cuja hipótese nula se encontra na afirmação de que os resíduos possuem variância constante. Isto é, o modelo é homocedástico, contra a hipótese alternativa de que os resíduos não possuem variância constante, ou seja, que o modelo é heterocedástico. O resultado foi uma estatística de teste de 11,09, com um pvalor de 0,1967, não rejeitando a hipótese nula de variância constante, o que permitiu concluir que o modelo é homocedástico.

O teste de Durbin Watson para verificação da suposição de independência dos resíduos apresentou uma estatística de teste de 1,4288, com um pvalor de 0,00227. Neste caso, a hipótese nula de independência dos resíduos foi rejeitada, levando a uma indicação de autocorrelação significativa. No entanto, a verificação gráfica já realizada não apresentou indícios de ausência de independência. Outros gráficos também utilizados para essa verificação são o diagrama dos resíduos contra a ordem das observações e o diagrama dos resíduos contra as variáveis indicativas, como pode ser visto nas Figuras 4.9 e 4.10 .

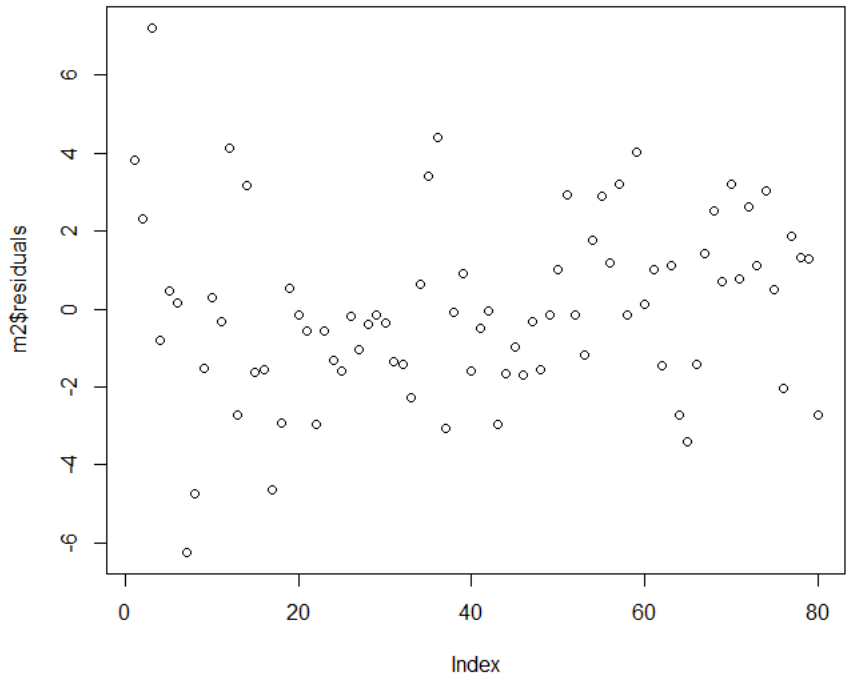

Figura 4.9: Diagrama dos resíduos em relação a ordem das observações

Na Figura 4.9 observa-se uma distribuição aleatoria dos resíduos, não se observando nenhum padrão que indique algum tipo de dependência no tempo.

A suposição de independência dos resíduos também foi analisada por meio dos diagramas de dispersão das variáveis preditoras contra os resíduos. Na figura 3, apresentou-se como exemplo essa análise contra as variáveis preditoras VI, X3PT, IA e RT. Nessas representações gráfícas não se verificou nenhum padrão que indique ausência de independência, sendo verificado também situações similares na análise gráfica com as demais covariáveis, cujos gráficos não foram apresentados aqui.

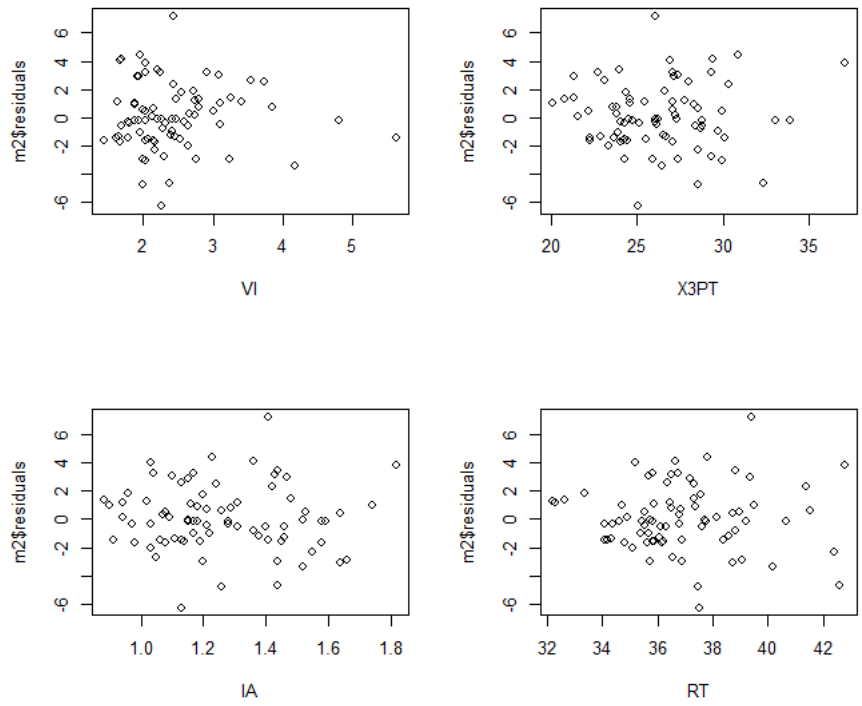

Figura 4.10: Diagramas de dispersão dos resíduos contra as variáveis preditoras VI, X3PT, IA e RT.

Sendo assim, apesar da suposta subjetividade da análise gráfica, optou-se por aceitar a hipótese de independência dos erros. Neste caso, acredita-se que o teste de Durbin Watson tenha rejeitado erroneamente a hipótese nula, dado que sua regra de decisão e interpretação envolve também alguma subjetividade.

4.4 Interpretação do Modelo

O modelo de regressão linear múltipla é apresentado pela expressão (4.1):

$$Y = -47.21 + 0.414X3P + 0.278X2P + 1.37RT + 4.51IA - 0.65X3PT$$

$$\text{-0.32X2PT+1.56EM+11.77BE+1.47VI} \quad (4.1)$$

A patir das estimativas dos parâmetros do modelo é possível verificar que a variável X3P, referente à quantidade média de cestas de 3 pontos, tem uma relação positiva com a variável resposta. Desse modo, quanto mais o time faz cestas de 3 pontos as chances de vitória do time aumentam, sendo o acréscimo de 0,414 no número de vitórias a cada aumento de uma cesta de 3 pontos.

Analisando a variável X2P, também observa-se uma relação positiva com a quantidade de vitórias, portanto o coeficiente de X2P apresenta cerca da metade em relação ao valor estimado para X3P. Desse modo, a quantidade de cestas de 2 pontos mesmo sendo positiva, apresenta menor impacto comparando a cestas de 3 pontos.

Observando a variável RT que corresponde à quantidade média de rebotes do time na temporada, verifica-se que esta também foi siginificativamente positiva, ou seja, quanto mais rebotes o time conseguir em uma temporada maior será a quantidade de vitórias conquistadas. O aumento no número de vitórias é de 1,37 a cada aumento de uma unidade no número de rebotes.

Considerando a variável IA, que significa índice de assistência por erro, quanto maior o índice, mais eficiente é o time, em questão de movimentação da bola antes de uma finalização. A variável tem uma relação positiva de 4.51 em relação à quantidade de vitórias, então aumentando em uma unidade esse índice a quantidade de vitórias conquistadas pelo time aumenta em 4,51.

Uma variável que apresentou significância estatística e coeficiente negativo no modelo é a X3PT. Isto significa que a quantidade média de cestas de 3 pontos tentados, ou seja, quanto mais o time tentar cestas de 3 pontos, menor é a quantidade de chance para conquistar maior número de vitórias. Pode-se afirmar o mesmo para a variável X2PT, que corresponde a quantidade média de cestas de 2 pontos tentados. Portanto, tanto quanto a quantidade de cestas tentadas de 3 pontos e de 2 pontos, resulta num resultado negativo para o modelo.

Um resultado interessante para o modelo foi verificado com relação à variável EM, a qual representa a quantidade média de enterradas feita por um time na temporada. Quanto mais o time realiza enterradas em uma temporada, maior a quantidade total de vitórias na temporada.

Avaliando a variável BE, representando o índice de bolas recuperadas por erros, verificou-se que esta tem maior peso no modelo. De acordo com o valor positivo bem grande da variável BE, pode-se dizer que à medida que o time recupera as bolas de erros cometidos na partida, tem-se um aumento bem siginificativo na quantidade de vitórias, mantendo as demais variáveis

fixas.

Outra relação positiva à variável resposta se encontra na variável VI, a qual significa violação em quadra, ou seja, o jogador comete uma das violações de regras do basquete. Este resultado não era esperado, pois supõe-se que a violação não seja um fator positivo para bons resultados no basquete e no modelo ajustado apresentou-se como significante e positivo.

Por fim, com o modelo de regressão ajustado, existem 9 variáveis que foram significativa para o modelo, podemos extrair informações importantes para ter um aumento significativo na quantidade de vitórias de um time de basquete na temporada. O treinador do time pode ter um foco maior na quantidade de cestas de 3 pontos e de 2 pontos por jogo, porém, sem cometer um número alto de tentativas fracassados.

Com relação ao rebote (RT), pode ser adotado um ajuste nos treinos, de maneira que conforme o time melhore esse recurso maior sucesso obterá em quadras, além de influenciar no IA, que também representa uma oportunidade para corrigir os erros do time. O trabalho demonstrou que os erros dos times adversários influenciam no aumento do sucesso do time em quadra de acordo com as variáveis RT, IA e BE. Desse modo, por outro lado, podemos afirmar que BE (bola recuperada pelos erros do próprio time), torna-se de suma importância para obter quantidades maiores de vitórias. Faz-se necessário ter ressalvas para o resultado da variável VI, pois a quantidade de violações apresentou valor positivo para a quantidade de vitórias, porém, isso de fato é um caso a se pensar, mostrando que em quadra não faz sentido ter violações para aumentar a quantidade de vitórias.

5. CONSIDERAÇÕES FINAIS

Este trabalho avaliou com êxito a relação entre as 22 variáveis iniciais e, utilizando o modelo de regressão linear múltipla, concluiu-se que somente 9 delas foram significativas para definir a quantidade de vitórias de um time de basquete da LNB. As variáveis que tiveram relação positiva com o modelo estudado foram cestas de 3 pontos (X3P), cestas de 2 pontos (X2P), rebotes (RT), índice de assistência por erro (IA), índice de bolas recuperadas por erros (BE), enterradas (EM) e violações (VI). Dentre estas, a que apresentou maior coeficiente foi a variável BE, portanto, a cada unidade acrescentada desta variável, tem-se maior impacto no número de vitórias na temporada. Já no resultado do teste, a variável que apresentou uma maior significância foi a variável RT.

Em contrapartida, duas variáveis apresentaram relação negativa com a quantidade de vitórias. Ao analisar a relação de arremessos de 3 pontos tentatados (X3PT) e arremessos de 2 pontos tentados (X2PT) com a variável resposta, concluiu-se que quanto mais arremessos de 3 e 2 pontos tentados durante o jogo, menor é a quantidade de vitória da equipe na temporada.

Determinadas variáveis como assistência (AS) e lances livres (LL), não foram significativas no resultado, ou seja, não influenciaram na quantidade de vitórias da equipe na temporada e portanto não entraram no modelo final. Levando em conta os métodos utilizados para a realização do diagnóstico, confirmou-se que o ajuste do modelo foi adequado, explicando 85% da variabilidade da variável número de vitórias na temporada.

Referências Bibliográficas

[1] Breusch, T. S. e Pagan, A. R.: *A simple test for heteroscedasticity and random coefficient variation*. Econometrica: Journal of the econometric society, pp. 1287–1294, 1979.

[2] Carneiro, F. F. B., Souza, D. d. e Costa, F. d. C.: *Contribuições do uso da estatística para a formação de equipes de basquetebol*. Coleção Pesquisa em Educação Física, 14(3):31–40, 2015.

[3] CBB: *Confederação Brasileira de basquete*, 2021. https://www.cbb.com.br/basquete, acessado em 10/12/2021.

[4] Charnet, R., Freire, C. d. L., Charnet, E. M., Bonvino, H. *et al.*: *Análise de modelos de regressão linear com aplicações*. Campinas: Unicamp, 1999.

[5] De Rose Junior, D., Tavares, A. C. e Gitti, V.: *Perfil técnico de jogadores brasileiros de basquetebol: relação entre os indicadores de jogo e posições específicas*. Rev. bras. educ. fís. esp, pp. 377–384, 2004.

[6] Fáveiro, L., Belfiore, P., Silva, F. e CHAM, B.: *Análise de dados: modelagem multivariada para tomada de decisão*. São Paulo: Campus, p. 235, 2009.

[7] HAIR, J. F., Anderson, R. e Tatham, R.: *BLACK WC Análise Multivariada de Dados*. Tradução Adonai Schlup SantAnna e Anselmo Chaves Neto, 5, 2005.

[8] Hoffmann, R. e Vieira, S.: *Análise de regressão: uma introdução à econometria*. São Paulo, 1998.

[9] Keppel, G.: *Design and analysis: A researcher's handbook*. Prentice-Hall, Inc, 1991.

[10] LNB: *Liga Nacional de Basquete*, 2021. https://lnb.com.br/nbb/estatisticas/, acessado em 10/06/2021.

[11] Maciel, L. F. V. *et al.*: *Regressão linear múltipla na modelagem de resultados na National Basketball Association (NBA)*. 2019.

[12] Maroco, J.: *Análise Estatística, com utilização do SPSS.(2003) Edições Sílabo*. Lisboa, Portugal.

[13] MELHORES, M. E.: *Quais os 14 tipos de esportes mais populares do mundo*. https://www.maioresemelhores.com/tipos-de-esportes-mais-populares-do-mundo, Acessado em 06/12/2021.

[14] Pindyck, R. S. e Rubinfeld, D. L.: *Econometria: modelos & previsões*. Elsevier, 2004.

[15] R7, E.: *Liga anuncia parceria com Nike, enquanto CBB sofre fuga de investimento no basquete brasileira*, 2021. https://esportes.r7.com/olimpiadas/liga-anuncia-parceria-com-nike-enquanto-cbb-sofre-com-fuga-de-investimento-no-basquete-brasileiro-23082021, Acessado em 13/12/2021.

[16] Sallum Neto, F.: *Comparação de ajustes do modelo de Gompertz a dados de crescimento*. 2013.

[17] Santos, C.: *Estatística descritiva*. Manual de auto-aprendizagem, 2, 2007.

[18] Shapiro, S. S. e Wilk, M. B.: *An analysis of variance test for normality (complete samples)*. Biometrika, 52(3/4):591–611, 1965.

[19] Tabachnick, B. G. e Fidell, L. S.: *Using multivariate statistics . Northridge*. Cal.: Harper Collins, 1996.

[20] Team, R. C.: *A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria,*. https://www.R-project.org/.

[21] Villa, T. E. d.: *Predição do custo de milho por meio de modelos de regressão linear múltipla*. 2016.