Relatório Estatística 3

Gabriel Augusto L. Silva

Introdução

O objetivo deste relatório é analisar a correlação entre a quantidade de pontos dos dois melhores jogadores da NBB (National Basketball Brazil) de cada time ao longo das temporadas e sua respectiva eficiência em quadra, utilizando de **Regressão Linear Simples (MRLS)**, fixada a covariavel **PTS_TOTAL** (Pontos totais do jogador) e varivavel resposta **EF_TOTAL** (Eficiência total do jogador), estas informações se encontram no arquivo **Base_NBB.csv** disponivel no **Kaggle** para dominio publico por **Gabriel Pastorello** .

Contexto: A principal forma de interpretar o quão bom um jogador de basquete desempenha, é por meio de sua eficiência total, pois dentro dela engloba o quanto ele pontou, pegou rebotes, realizou assistências, defendeu quadra, realizou faltas, errou e todas as outras situações em quadra, a eficiencia leva em consideração toda a participação de um jogador em quadra, seja ela positiva ou não. Neste estudo buscaremos enxergar como a principal estatistica do esporte, a pontuação, é capaz de dizer sobre como um jogador desempenha.

Análise descritiva

Apresentamos inicialmente as medidas de resumo das duas variaveis observadas, sendo a variavel resposta a eficiencia do jogador **(EF_TOTAL)** e a covariavel sendo a quantidade de cestas dos jogadores **(PTS_TOTAL)**. Segue abaixo as informações na Table 1.

Table 1: Medidas de resumo

	Minimo	1° Quartil	Mediana	média	3° Quartil	Máximo
PTS_TOTAL	32.0	341.0	406.0	423.4	507.0	735.0
EF_TOTAL	21.0	380.0	450.0	480.5	573.0	789.0

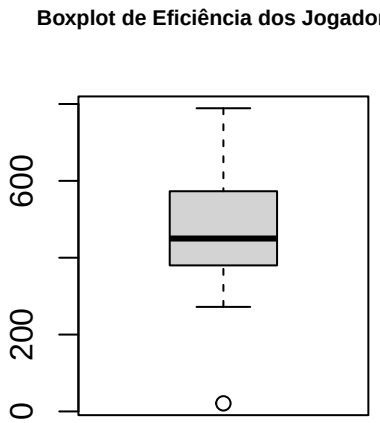


Figura 1

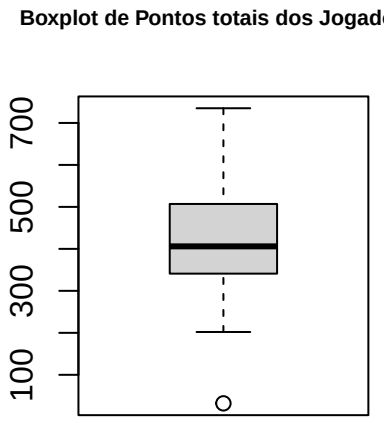
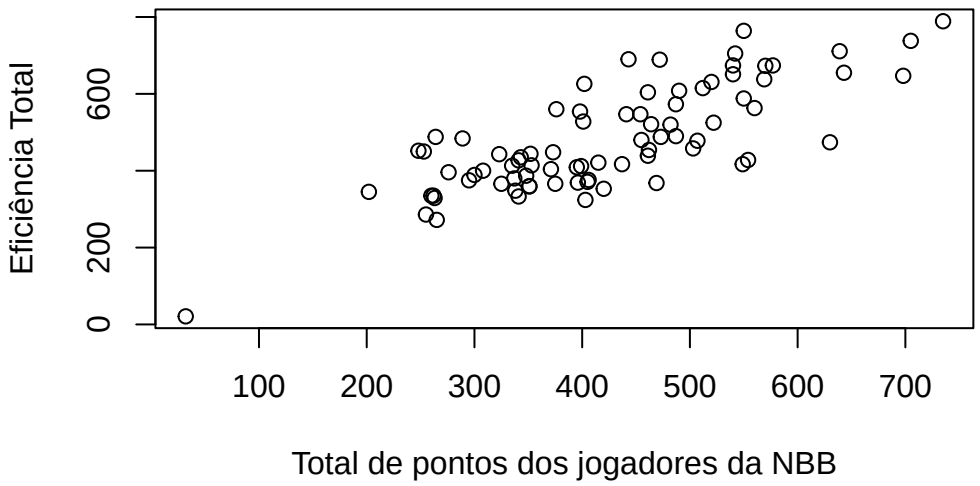


Figura 2

Pela Table 1, observamos uma otima distribuição de valores inter-quartis, indicativo de que entre os times existe uma diferenças entre o desempenho dos jogadores, ao passo que, enquanto os 25% piores MVPs (os jogadores mais valiosos do time) de um time possuem até 380 scores de eficiência, outros times possuem 25% dos melhores MVPs com no minimo de 573 scores de eficiência, com uma média esperada de 480.5 scores de eficiência somado a grande variabilidade observados na Figura 1 [32.0, 735.0] e 2 [21.0, 789.0], podemos ver que existe uma diferença consideravel, até mesmo entre os melhores jogadores da liga. Também observamos a presença de 1 outlier nos boxplot de EF_TOTAL e PTS_TOTAL, ao olhar o banco de dados, notamos que ambas estatísticas pertencem ao mesmo jogador. Note que, o mesmo falado sobre a eficiência dos jogadores, também pode ser dito sobre a quantidade de pontos por jogador da liga.

Figura 3: Gráfico de dispersão das variáveis pontos e eficiência



Na Figura 3 observamos que a distribuição inter-quartis semelhantes na Figura 1 e 2 não é uma coincidência, notamos que a principio existe uma correlação linear positiva entre quantidade de pontos e eficiência total dos jogadores profissionais, ou seja, quanto maior a quantidade de pontos, maior o score de eficiência. A correlação é estimada em 0.786 onde segundo Hinkle, Wiersma & jurs (2003) é considerado uma forte correlação entre as variaveis. Baseado nisto, ajustamos um modelo normal linear do tipo:

$$eficiencia_i = \alpha + \beta pontos_i + \epsilon_i$$

Em que $\epsilon_i \sim Normal(0, \sigma^2)$ e $i = 1, ..., 81$ representam os melhores jogadores dos times da liga.

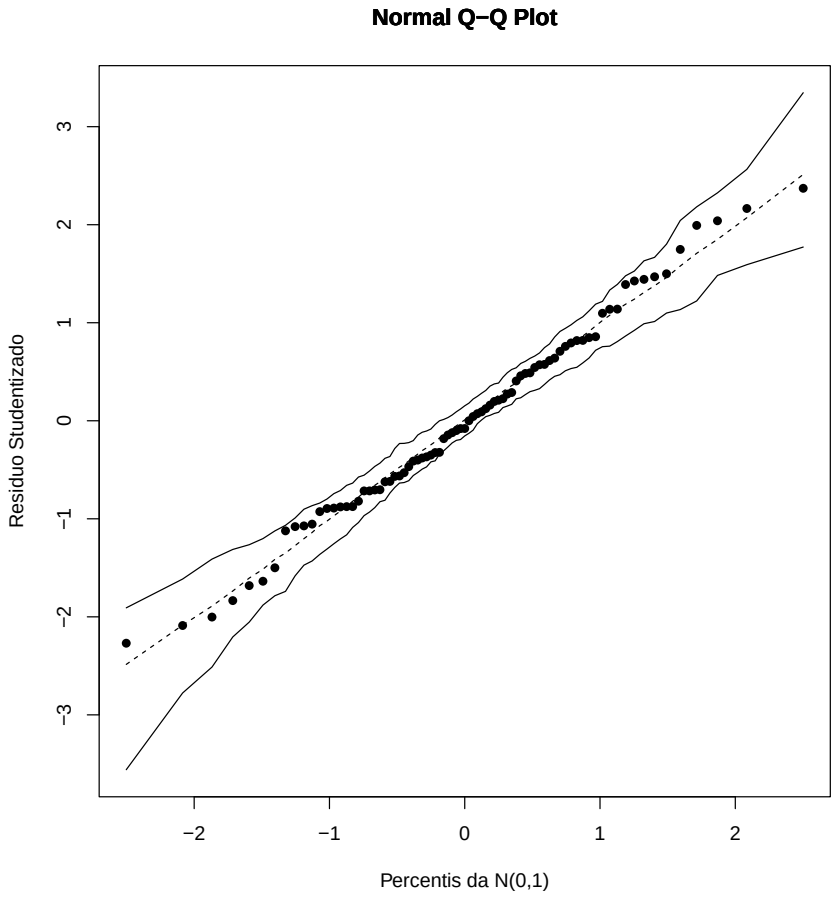
Rodando o ajuste obtemos os resultados demonstrados na Tabela 2.

Table 2: Ajuste do modelo normal linear

	Estimativas	Erro Padrão	t-valor	p-valor
α	120.78910	33.12555	3.646	0.000475
β	0.84944	0.07504	11.319	$<$ 2e-16

Com os resultados da Tabela 2, podemos observar no modelo 1 que a variável **pontos** é significativa ao nível de 1%, confirmando o que observamos anteriormente quando comentamos

sobre o grau de correlação entre variável **eficiência** e variável **pontos**, olhando α observamos que no intercepto, ou seja, no ponto em que a zero pontos, é esperado 120.78 score de eficiência, e olhando para β , a cada cesta é esperado um aumento de 0.8494 de eficiência no score do jogador. A raiz do quadrado médio foi calculada em 84.23 em 79 graus de liberdade.

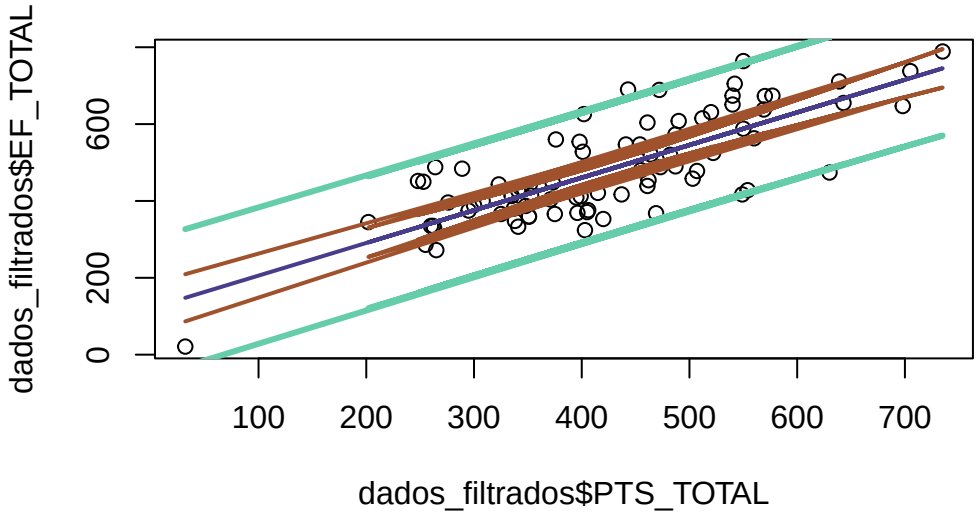


Pelo gráfico da Figura 4, podemos afirmar que é valido supor um modelo de regressão normal linear entre as variáveis **eficiência** e **pontos**, note que todos os pontos do gráfico se encontram dentro da região delimitada pelas linhas de corte. É possivel observar como ocorre o efeito de afunilamento no meio e logo depois a abertura nos extremos, isto se dá por conta do aumento de variabilidade nas pontas, já que os extremos possuem menos valores para se comparar em relação aos valores do centro. O coeficiente de determinação calculado ao se supor normalidade se deu por $R^2 = 0.6185866$, informando que em torno de 61% da variável **eficiência** pode ser informada por meio da variavel **pontos**, sendo os 39% restantes determinadas por outras estatistícas como rebote, assistência, roubo de bola, etc que não foram levadas em consideração.

Segue na Figura 5 e Table 3, informações sobre o **intervalo de confiança** da reta estimada.

Table 3: Intervalo de confiança

IC	2.5%	97.5%
Intercepto	54.8543558	186.7238531
Pontos	0.7000681	0.9988113



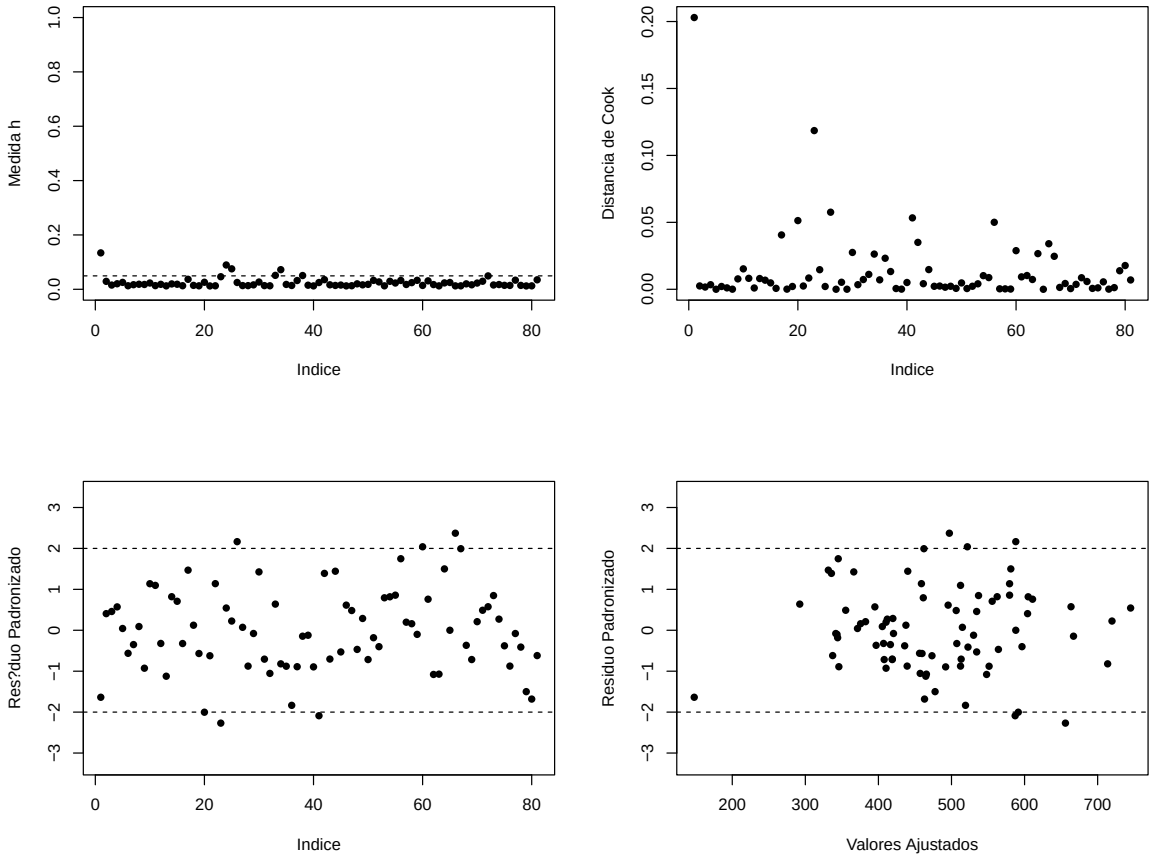
Comentando sobre a Table 3, podemos ver através do intercepto, os pontos critícos do intervalo de confiança, ou seja, com eles podemos dizer com 95% de confiança que os verdadeiros valores da variável **eficiência** estão no intervalo [54.8543558;186.7238531] e que o β real vai estar no intervalo [0.7000681;0.9988113].

A Figura 5 é uma representação gráfica da reta estimada (em roxo), das retas do intervalo de confiança (em laranja) e das retas de predição (em verde). Perceba que a reta de predição possui um "range" de valores muito maior do que a reta de intervalo de confiança, isto é obvio, pois a função das retas de predição é supor possiveis novas variaveis que poderiam aparecer, traduzindo para nosso estudo, ela nos permite prever a região que estaria um novo MVP de um time da liga baseado na quantidade de pontos que ele teve. Segue exemplo na Table 4

Table 4: Exemplo

Quantidade de pontos	900	200	450
Reta estimada	885.2848	290.677	503.037
Intervalo de predição	[702.1989:1068.371]	[118.7268;462.6273]	[334.31;671.7639]
Intervalo de confiança	[811.7023;958.8674]	[252.4561;328.898]	[483.9913;522.0826]

No Diagnostico da normal, representada pelos gráficos abaixos nos permitem fazer algumas observações.



Note que no Gráfico de pontos de alavanca, ele nos indica no total 4 valores que se encontram como possiveis pontos de alavanca os pontos 1, 25, 34 e 42, ou seja, jogadores que não seguiram o mesmo padrão dos outros, obtendo uma eficiência muito fora do esperado, observando a Table

5, percebemos que alguns destes pontos possuem uma leve influencia nas estimativas, como no ponto 1 que altera o alfa em 20 pontos.

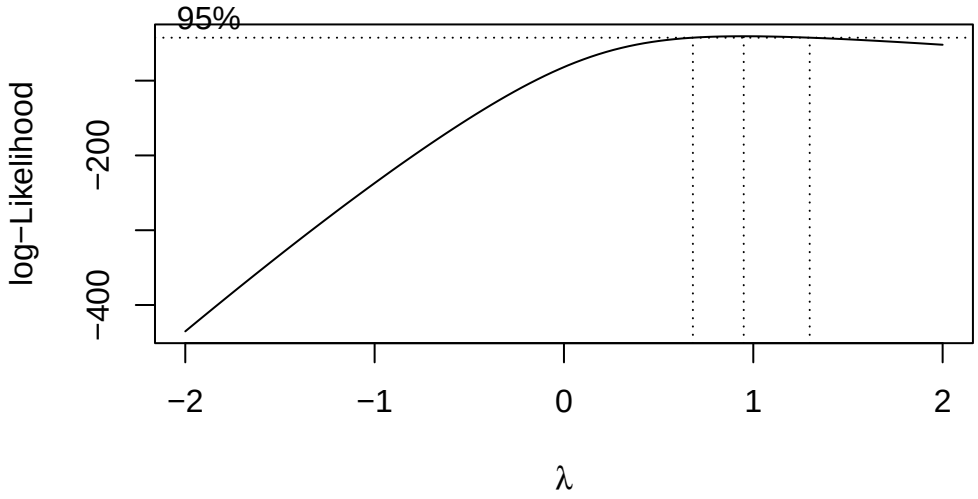
Table 5: Outliers observados

		Estimativa	Erro Padrão	t-valor	p-valor
Completo	α	120.78910	33.12555	3.646	0.000475
	β	0.84944	0.07504	11.319	<2e-16
Sem 1	α	141.8901	35.2221	4.028	0.000129
	β	0.8039	0.0793	10.137	6.89e-16
Sem 25	α	122.42279	34.10633	3.589	0.000577
	β	0.84500	0.07803	10.830	< 2e-16
Sem 34	α	115.04186	33.92544	3.391	0.0011
	β	0.86511	0.07759	11.150	<2e-16
Sem 42	α	112.53753	33.46276	3.363	0.0012
	β	0.86547	0.07549	11.464	<2e-16

Ao observar o Gráfico de pontos influentes, notamos de cara um valor extremamente distante em comparação ao resto (o ponto 1), entretanto ao calcular o distanciamento de cook, chegamos no valor de 0.3496, como nenhum dos pontos se encontrava na zona de rejeição, concluimos que não existiam pontos influentes, com o maior valor encontrado sendo 0.20.

Notamos um possivel indicio de heterocedasticidade no modelo no Gráfico de Homocedasticidade, a partir disso, utilizaremos o método de boxcox, para identificar se este modelo é valido ou não.

Figura 7: Gráfico de Boxcox



Após observarmos que o intervalo de confiança possui valores no intervalo [0,1] podemos concluir que não é necessário uma transformação para esse modelo, logo, é correto afirmar que o modelo 1 é valido para observar a relação linear entre a variável **eficiência** e a variável **pontos**.

Conclusão

O estudo sobre a relação de eficiência dos melhores jogadores por time da NBB e suas respectivas quantidade de pontos buscou tentar ver o quanto a quantidade de pontos dizia sobre a qualidade de um jogador, tendo em vista que existem diferentes estatísticas no basquete que podem ser levadas em consideração quando se fala sobre um jogador profissional. Observamos que a eficiência e a quantidade de pontos possuem uma correlação de 0.78 o que mostra a forte relaçao entre elas. O R^2 de 0,61 nos diz que 61% da variável pode ser explicada por meio da covariavel **pontos**, enquanto o resto pode ser explicada por outras estatistícas, como rebote, defesa e etc. Com o envelope foi confirmado a normalidade das variáveis o que nos permitiu fazer as afirmações sobre a regressão ser linear e valida, juntamente a isto, realizamos o teste de boxcox para que pudessemos ver se o modelo era valido para o teste, o que foi confirmado como valido durante a análise. O gráfico de diagnosticos nos forneceu alguns outliers que nos entregam indicios de heterocedasticidade, e alguns pontos de alavanca (pontos 1, 25, 34 e 42), que após a retirada deles, apresentaram algumas leves mudanças, mas não foram fortes o suficientes, para alterar os testes de boxcox e envelope. Portanto, podemos super que existe uma relação relevante entre a eficiência total dos melhores jogadores da liga NBB e sua quantidade total de pontos ao longo dos jogos.