UNIVERSIDADE FEDERAL DO PARANÁ

Adi M. A. Junior
Leonardo H. B. Krüger
Kristiany J. Martini
Konstanz T. Winter

ESTUDO DO NÚMERO DE PONTOS EFETUADOS EM UMA TEMPORADA DE BASQUETEBOL

CURITIBA
27 de novembro de 2017

Resumo

Em jogos de basquetebol é sempre visível que alguns times possuem vantagens em relação aos seus adversários, sejam estas vantagens relacionadas com o perfil dos jogadores, região de ocorrência do jogo, ou então com outras variáveis sociais ou econômicas. Diante deste cenário, o seguinte estudo tem por finalidade avaliar a influência que as variáveis fisiológicas e sociais têm sobre o número de pontos efetuados por jogadores de basquetebol em uma temporada de jogos ocorrida em 2014 e 2015.

Neste estudo foi utilizado um modelo linear generalizado baseado na distribuição binomial negativa, pois foi a distribuição que melhor acoplou os dados relacionados à temporada de basquete, além de ser um teste que apresentou boa qualidade (conforme teste AIC e de verossimilhança), um bom ajuste de resíduos e uma interpretação fácil e didática sobre as relações e efeitos vistos durante o ajuste.

Foram ajustadas predições para alguns perfis de jogadores e encontrados alguns pontos influentes, como o jogador LeBron James (jogador conhecido, excelente pontuação e excelente desempenho), Lance Stephenson e Nene (jogadores com baixa pontuação, ótimo desempenho, sendo Lance um jogador aleatório e Nene um jogador brasileiro).

Por fim, foram avaliados os efeitos de algumas variáveis em relação à resposta, como idade (jogadores mais velhos apresentaram menos desempenho), tempo de jogo (jogadores que participam de mais jogos possuem pontuações melhores), experiência (maior experiência, maior pontuação) e a posição onde o jogador se concentra (jogadores em posições de ataque possuem maior pontuação, como jogadores em posição de defesa possuem menor pontuação). Também foi avaliado um efeito interativo entre posição e time do jogador.

Palavras-chave: *Distribuição Binomial Negativa, GLM, Modelo linear generalizado, LeBronJames, Basquete, Temporada, Jogos.*

Sumário

1 Introdução

As temporadas de basquete são muito assistidas em países como Estados Unidos, Canadá, México e países da Europa, assim como também existe uma grande quantidade de telespectadores e admiradores do esporte em outros locais do mundo. Ao redor destas temporadas existe uma estrutura financeira e social que de certa forma estimula a entrada de jogadores de basquete, os quais são conduzidos à treinos fortíssimos, geração de experiência de jogo e diversas outras variáveis que os auxiliam na obtenção da maior pontuação possível em um jogo e/ou uma série de jogos.

Um grupo de pesquisa nos Estados Unidos coletou informações sobre os jogadores de basquete de todas as ligas participantes na temporada de basquete de 2014/2015, com o objetivo de agrupar uma grande quantidade de características socioeconômicas que poderiam ou não influenciar no desempenho dos jogadores que participavam desta temporada.

Dado este contexto, o objetivo do estudo é avaliar a influência que estas variáveis possuem sobre o número de pontos que os jogadores obtiveram nos jogos desta temporada. A base de dados advém do site de ciências de dados *Kaggle*, sendo este conjunto de dados composto por 422 jogadores e 34 variáveis.

2 Material e Métodos

2.1 Material

A coleta de dados foi efetuada nas temporadas de 2014 e 2015, com base na avaliação de 422 jogadores e 34 características relacionadas à pontuação que eles obtiveram na temporada e características sociais e relacionadas aos avaliados. A coleta dos dados se deu por meio do histórico de cada jogador. Destas 34 variáveis, 14 foram analisadas neste estudo. A seguir, é dada a descrição de cada variável e sua morfologia (fator, inteiro, numérico, lista, etc.):

- ***Games Played*:** número de jogos que o jogador participou na temporada **(inteiro)**
- ***MIN*:** Minutos jogados na temporada **(inteiro)**
- ***AST*:** Número de *Assists. Assist* (Basquetebol) é o nome dado ao momento quando um jogador passa a bola para outro do mesmo time. **(inteiro)**
- ***STL*:** Número de *Steals. Steal* (Basquetebol) é o nome dado ao momento quando um jogador retira a bola do adversário para continuar o jogo. **(inteiro)**
- ***BLK*:** Número de *Blocks. Block* (Basquetebol) é o nome dado ao momento quando um jogador bloqueia o seu adversário. **(inteiro)**
- ***Age:*** Idade do jogador **(inteiro)**
- ***Birth_Place*:** Local de nascimento **(fator)**

- ***Collage:*** é a universidade onde o jogador estudou **(fator)**
- ***Experience*:** Tempo de jogo em anos **(inteiro)**
- ***Height*:** Altura do jogador **(numérico)**
- ***Pos*:** Posição em que o jogador atua no basquetebol **(fator)**
- ***Team*:** Time do jogador. Ex: *ATL = Atlanta Team*. **(fator)**
- ***Weight*:** Peso do jogador **(numérico)**
- ***BMI*:** IMC - índice de massa corporal. **(inteiro)**

A variável resposta escolhida para a análise é o ***PTS*** (pontos efetuados no jogo), sendo esta uma variável contida no conjunto dos inteiros.

O *software* R foi utilizado para a análise do estudo. Os pacotes utilizados foram: *car*, *MASS*, *faraway*, *effects*, *statmod*, *corrplot e hnp*.

2.2 Métodos

Por ser uma variável resposta de contagem, o modelo inicial proposto foi o Poisson. Entretanto, foram encontradas características de falta de ajuste que prejudicavam os resultados obtidos pelo modelo. Portanto, foi utilizado o modelo baseado na Binomial Negativa, devido ao fato deste modelo apresentar melhor ajuste em relação à Poisson. Foi feita uma análise de resíduos e de efeitos, além de uma análise de predição.

3 Resultados e Discussão

3.1 Diagnóstico

Primeiramente, foi feita uma análise de correlação, por onde verificou-se uma correlação alta entre a variável resposta *(PTS)* e *MIN, STL, AST* e *Games Played*. Além disso, foi constatada correlação alta entre *Experience* e *Age*, *Games Played* e *MIN, BMI* e *Weight, STL* e *MIN* e *AST* e *MIN*. Após este gráfico de correlação, foi retirada a variável *Games Played* da análise, por esta ter uma correlação alta com *MIN* e a variável *MIN* possuir uma alta correlação com *PTS*.

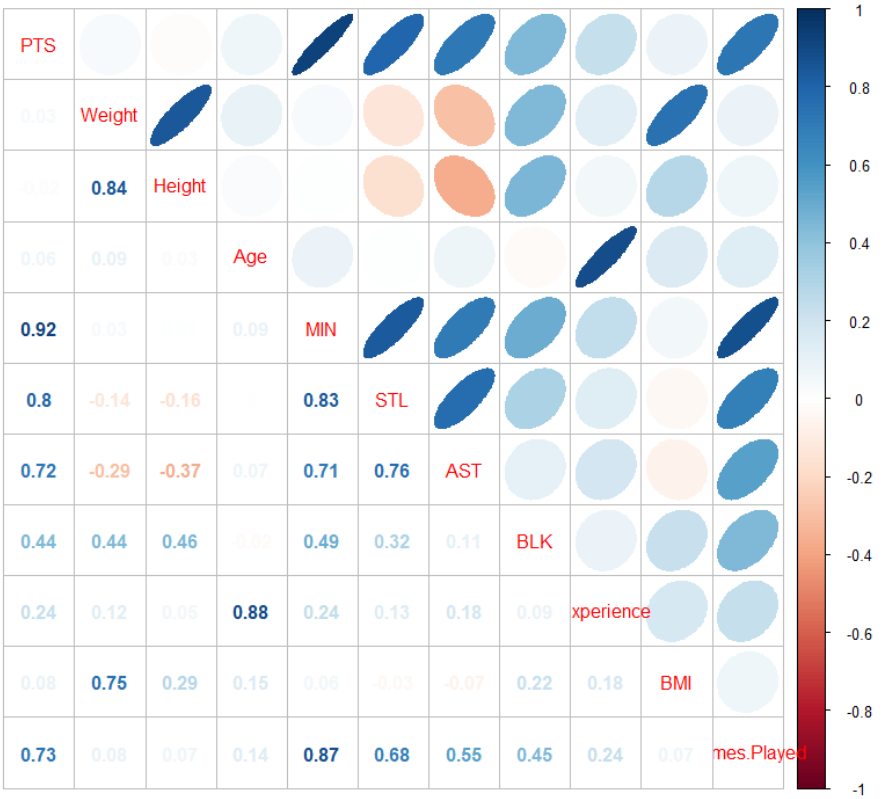


Gráfico 1: Gráfico de Correlação entre Variáveis

Também foi feito um gráfico de dispersão, que comprovou as correlações acima e nos mostrou indicativo de relações como *BLK* com *Weight* e *Height*, *Experience* com *AST, BLK* e *STL*, além de IMC com altura (mas não com peso, o que é uma característica comum em corpos atléticos).

O primeiro ajuste efetuado foi pelo modelo de Poisson. Entretanto, houve dificuldades relacionadas a qualidade do teste, como mau ajuste dos resíduos e efeitos, por exemplo. Portanto, foi ajustado um modelo Binomial Negativo, o qual apresentou um *AIC* 5x (cinco vezes) menor que o resultante da Poisson, além da verossimilhança 6x (seis vezes) maior que o do ajuste pela Poisson.

Na análise de resíduos por envelopes simulados, foi constatada a total falta de ajuste por parte do modelo de Poisson (99% dos dados para fora do envelope) e um ajuste com apenas 15% dos dados fora do envelope, no caso da Binomial Negativa. Após análise de pontos influentes e tratativa da base de dados para remoção dos *outliers*, foi criado o modelo final (chamado de *M3*), o qual será usado como base para este estudo. Resultados sobre o *M3* são apresentados a seguir:

Ajuste	AIC	Verossim.
M2	4.887	-2.402
M3	4.957	-2.323
M2 - M3	**70**	**79**

Tabela 1: Comparação de Ajustes

Variables	Estimate	Std. Error	Z-Value	P-Value	Significance
(Intercept)	4.477	251	17.864	< 2e-16	<0,1%
Age	(1)	0	(3.642)	0.00027	<0,1%
MIN	1	0	41.656	< 2e-16	<0,1%
Experience	58	15	4.004	6.23e-05	<0,1%
PosSG:TeamMIN	715	258	2.774	0.00554	0,1%
PosC:TeamSAS	856	322	2.656	0.00791	0,1%
PosPF:TeamATL	681	288	2.365	0.01803	1%
PosC:TeamCHI	(691)	331	(2.088)	0.03676	1%
PosPG:TeamIND	579	269	2.156	0.03110	1%
PosC:TeamLAC	(877)	405	(2.168)	0.03014	1%
PosSF:TeamNYK	692	321	2.158	0.03093	1%
PosSG:TeamNYK	525	258	2.039	0.04141	1%
PosSF:TeamOKC	614	287	2.138	0.03251	1%
PosPG:TeamATL	591	321	1.842	0.06546	5%
PosC:TeamBOS	667	403	1.658	0.09735	5%
PosPF:TeamCHI	529	321	1.650	0.09887	5%
PosPG:TeamCHI	616	319	1.931	0.05351	5%
PosPG:TeamCLE	(579)	330	(1.755)	0.07928	5%
PosPG:TeamDET	534	289	1.851	0.06418	5%
PosPF:TeamIND	508	289	1.754	0.07946	5%
PosSF:TeamLAC	(574)	296	(1.938)	0.05263	5%
PosSF:TeamLAL	(583)	328	(1.779)	0.07517	5%
PosC:TeamNYK	554	289	1.918	0.05508	5%
Others[1]		**Non Significative**			-
Null Deviance:	**2927,08**	DF:	**382**		
Residual Deviance:	**404,08**	DF:	**232**		
Dispersion Parameter:	**1**				

Tabela 2: Sumário do Modelo 3 *(M3)*

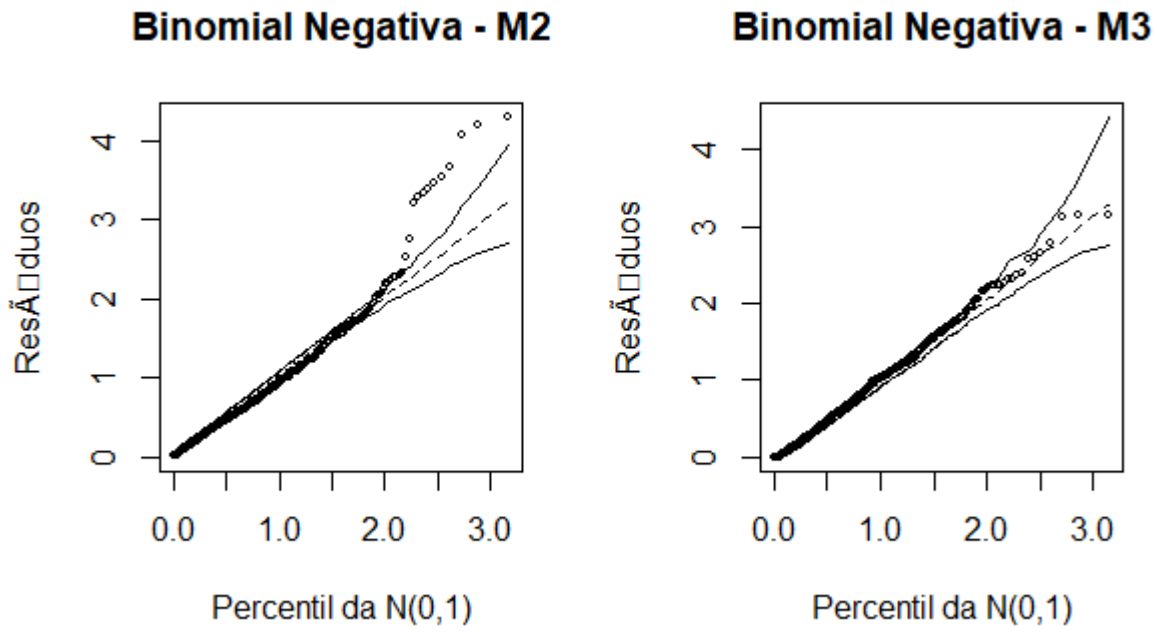


Gráfico 2: Envelopes Simulados - Modelo 2 *(M2)* x Modelo 3 *(M3)*

Fica evidente a melhor qualidade do ajuste pelo modelo *M3*. Apesar do ganho no *AIC*, tal medida penaliza o número de parâmetros do ajuste e, portanto, a análise fica melhor baseada na medida de verossimilhança. O gráfico de envelope simulado demonstra que houve uma melhoria no ajuste do modelo *M3*, com todos os dados acoplados, enquanto que o modelo *M2* apresentou uma sequência de dados fora do envelope.

No modelo *M3*, além dos ajustes de pontos influentes, foi verificada a interação entre *Pos* (posição do jogador) e *Team* (time do jogador). Tal interação também é responsável pelo melhor resultado deste ajuste.

Foi concluída com 90% de confiança, pelo teste da anova, a presença de melhorias no modelo *M3* por conta da interação entre *Pos* e *Team*. Outras interações não foram significativas e, portanto, não acrescentadas na análise.

Portanto, o modelo utilizado será o da Binomial Negativa. A seguir é apresentada a expressão do modelo:

$$y_i|x_i \; BN(\mu_i, \phi)$$

$$g(\mu_i) = x^t\beta$$

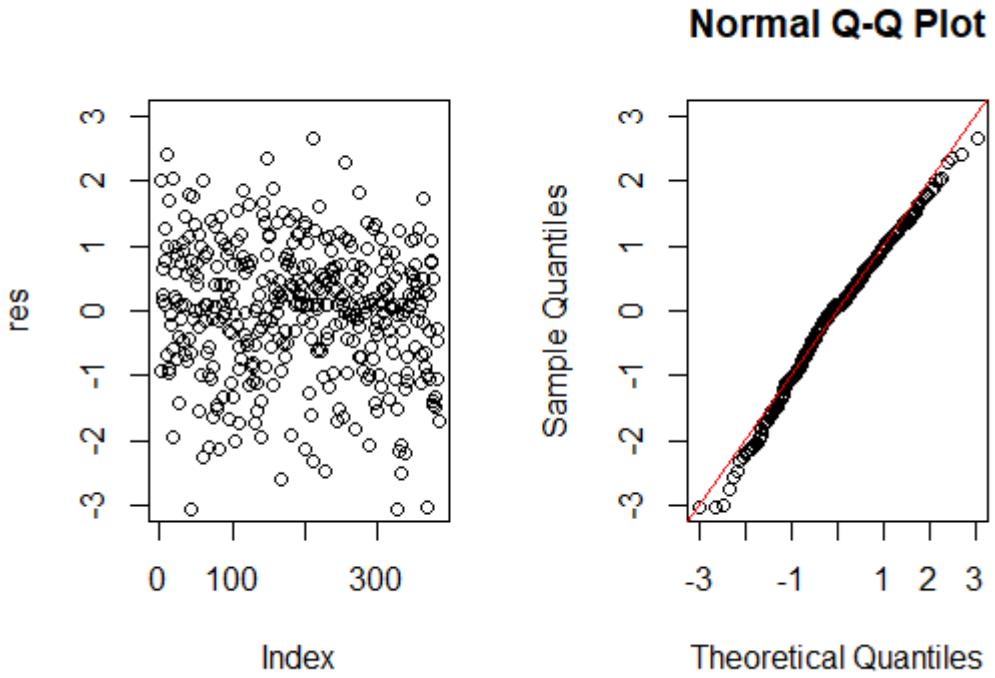
$$g(I) = \beta_0 + \beta_1 * Age + \beta_2 * Min + \beta_3 Experience + \beta_i * (POS * Team)$$

O modelo binomial negativo foi feito com a função de ligação identidade, pois foi a que obteve melhor resultado no ajuste do modelo.

3.2 Resíduos

Nos gráficos de diagnóstico padrões há uma dificuldade de interpretação do comportamento dos resíduos, por estarem baseados nos componentes da *deviance*. Entretanto, já foi possível identificar por eles os indícios de bom ajuste. A seguir é apresentada uma análise de resíduos baseada no modelo *M3*, o que demonstra com mais confiança o comportamento dos resíduos em relação ao ajuste.

No gráfico *Normal Q-Q Plot*, por exemplo, é visível que as escalas estão alinhadas entre -3 e 3, tanto para o eixo y quanto para o x. Para o gráfico *Index x Res*, é verificado que os pontos têm uma distribuição aleatória dentro do perímetro, o que também é indicativo de bom ajuste. Outros gráficos de resíduos também não mostraram algum comportamento que indicasse falta de ajuste.



Gráficos 3 e 4: Análise de resíduos do modelo 3 *(M3)*

3.3 Efeitos e Predições

A seguir é apresentado um gráfico dos principais efeitos em relação à resposta. Percebe-se um efeito crescente para *Experience* e *MIN*, assim como comportamento decrescente em *Age*. Vale atentar quanto ao fato de que o comportamento de uma variável é representado com a premissa de que todas as outras variáveis estão centradas na média.

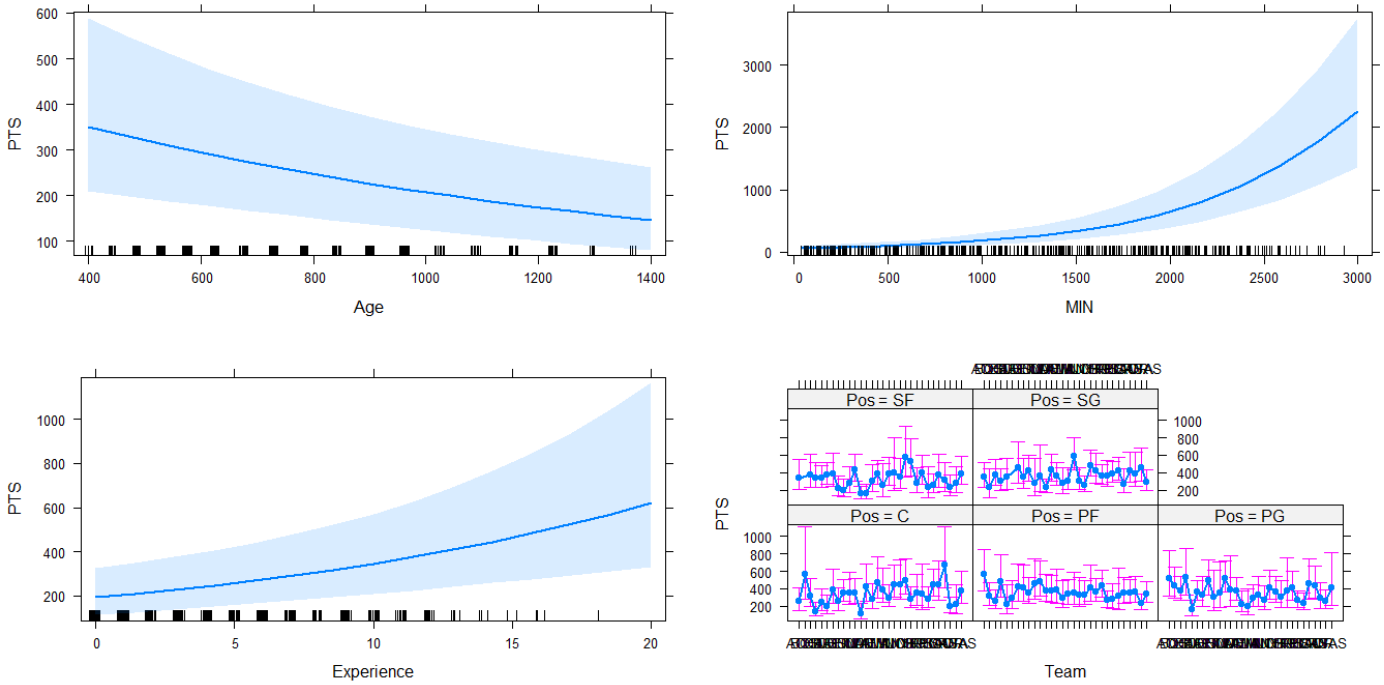


Gráfico 5: Análise de efeitos relacionados à variável resposta *PTS* (pontos efetuados)

Na análise de predição, foram considerados os perfis a seguir:

Informações	Jogador 1	Jogador 2	Jogador 3
Weight	112.5	103.5	112.5
Team	CLE	CHA	WAS
Age	961	625	1089
Min	2493	1.573	1.694
Experience	11	4	12
Pos	SF	SG	PF

Tabela 3: Perfis para predição

Foi predita a pontuação de 1743 para o perfil 1, 600 para o perfil 2 e 682 para o perfil 3. O valor real para o LeBron James foi de 1743. Para o Lance Stephenson foi de 501 e para o Nene, 682. Vale lembrar que o LeBron James é um ponto fora da curva, devido ao fato de ter uma das melhores pontuações no campeonato e, mesmo assim, conseguiu ser previsto pelo modelo.

4.0 Conclusão

O modelo identifica uma série de causas que influenciam no número de pontos obtidos por um jogador na temporada de basquetebol. Apesar de ser um estudo amostral, seus resultados podem ser utilizados para pesquisas mais técnicas sobre o desempenho de jogadores de basquetebol nas temporadas de jogos.

Foi obtida significância entre idade, tempo de jogo e experiência dos jogadores com número de pontos efetuados pelo jogador. Além disso, foi constatada uma relação de interação significativa entre posição e time com o número de pontos.

É perceptível que, neste estudo, a experiência e os minutos jogados na temporada afetam consideravelmente a quantidade de pontos, que jogadores mais novos possuem menor pontuação e que jogadores em posição de ataque têm um resultado melhor que jogadores em posição de defesa.

Dentre os perfis analisados, o LeBron James foi um dos principais outliers, por ter efetuado uma grande quantidade de pontos. Vale lembrar que o LeBron James já é um jogador extremamente conhecido no mundo do basquetebol por possuir habilidades inquestionavelmente boas.