UNIVERSIDADE FEDERAL DO PARANÁ
CURSO DE ESTATÍSTICA

MATEUS GEMELLI RAMOS
GUSTAVO S. SCHUTT

MODELO DE REGRESSÃO MULTIPLA: UM ESTUDO SOBRE O DESEMPENHO DOS JOGADORES DA NBA

CURITIBA
2018

MATEUS GEMELLI RAMOS

GUSTAVO S. SCHUTT

MODELO DE REGRESSÃO MULTIPLA: UM ESTUDO SOBRE O DESEMPENHO DOS JOGADORES DA NBA

Trabalho da disciplina de Análise de Regressão Linear (CE071) do curso de graduação em Estatística da Universidade Federal do Paraná.

Professor: Cesar Augusto Taconeli

Curitiba

2018

SUMÁRIO

INTRODUÇÃO

A National Basketball Association (NBA) é a principal liga profissional de basquetebol dos Estados Unidos. Atrai milhares de observadores anualmente e gera uma receita de milhões de dólares para as franquias (times), emissoras de televisão e patrocinadores.

A análise estatística é amplamente utilizada em cada temporada da NBA. O basquetebol se caracteriza por ser um esporte coletivo estratégico em que as equipes técnicas de cada franquia estudam as tendências dos seus oponentes, o comportamento e o potencial de cada jogador.

O presente estudo apresentará uma análise a respeito do desempenho em quadra dos jogadores da NBA na temporada 2016-2017. Para isso, considerou-se a variável resposta “pontos / minuto jogado” em função de outras diversas covariáveis relativas ao desempenho dos jogadores que serão modeladas por meio de uma regressão linear múltipla.

1 MATERIAIS E METODOS

1.1 MATERIAIS

Os dados utilizados foram coletados do site *Kaggle.* Trata-se de uma plataforma fundada em 2010 para competições de modelagem preditiva e analítica.

À medida que a empresa foi crescendo, ela foi expandindo a sua plataforma ao incluir novas funcionalidades. Assim, se tornou não somente um ambiente de competição, mas principalmente de colaboração entre cientistas de dados.

1.2 CONJUNTO DE DADOS

O conjunto de dados estudado possui ao todo 34 variáveis e 490 registros. As variáveis compreendem característica dos jogadores em campo e cada linha é um jogador que atuou na temporada da NBA de 2016-2017.

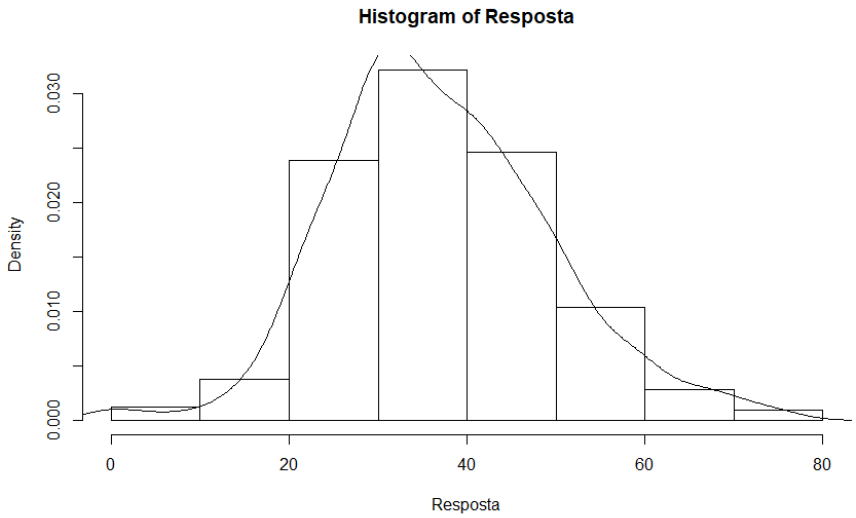
1.3 DEFINIÇÃO DA VARIAVEL RESPOSTA

Para a construção da variável dependente (resposta), utilizamos a combinação de duas variáveis: pontuação e tempo jogado em minutos. Sabe-se que a pontuação tem alta relação pelo tempo de participação nos jogos. Partindo disso, a pontuação dos jogadores foi combinada com o tempo de cada jogador em campo por minuto.

$$Resposta = \frac{PontuaçãoDoJogador}{MinutosEmCampo}$$

Para melhor mensuração, multiplicamos a variável resposta por 100 apresentando-a na escala de percentual (%). Sabe-se que pela teoria de probabilidade que uma possível distribuição que pode se adequar bem a esse tipo de dados na escala percentual (entre zero e 1) é a Beta. Porém, essa metodologia foge do escopo do trabalho e, portanto vamos considerar que a variável resposta segue uma distribuição gaussiana. Pela Figura 1, nota-se que este pressuposto é plausível.

Figura 1: Densidade empírica estimada da variável resposta



2 MODELAGEM

Para a análise de dados foi utilizado o programa estatístico R. Primeiramente, foi realizada uma análise exploratória contendo a estatística descritiva dos dados e matriz de correlação linear. Para verificar a relação entre a pontuação dos jogadores e as demais covariáveis. Foi implementado o modelo de regressão linear múltiplo, de acordo com a expressão abaixo:

$$Y = \beta 0 + \beta 1xi1 + \beta 2xi2 + \ldots + \beta kxik + \varepsilon i$$

Após isso, foram retiradas variáveis categorias que não interessam para a análise, pois algumas variáveis têm características especificas do jogador que não agregam informação ao modelo. Além disso, foram retiradas variáveis com alto grau de correlação linear pela interpretação do coeficiente de correlação linear de Pearson. Para o ajuste do modelo final, utilizou-se métodos de seleção automática de variáveis e seleção manual por meio de verificação de significância do P-valor. Ao final compararam-se quatros possíveis modelos e foi adotado um modelo com covariáveis sem problemas de multicolinearidade.

2.1 **ANÁLISE DESCRITIVA**

Para iniciar o resumo e sumarização dos dados realizou-se uma análise descritiva. Foi observada uma série de variáveis categóricas com informações não relevantes como: Name Birth_Place Birthdate Collage Team.

2.2 MODELO DE REGRESSÃO MULTIPLA

Primeiramente é ajustado um modelo aditivo saturado, ou seja, incluem-se todas as covariáveis sem considerar as interações.

$$Y = \beta 0 + \beta 1xi1 + \beta 2xi2 + \ldots + \beta kxik + \varepsilon i$$

2.3 SELEÇÃO DE VARIAVEIS

Para a escolha das variáveis relevantes para o modelo, utilizaram-se os métodos *Backward, Forward e Stepwise,* assim como, uma quarta seleção manual observando os p-valores a cada novo modelo ajustado. Além disso, analisaram-se alguns critérios de bondade de ajuste para escolha do modelo final.

- Análise gráfica da matriz de correlação linear e interpretação do coeficiente de correlação linear de Pearson.
- P-valor – Significância estatística da covariável.
- VIF (fator de inflação da variância) – Para identificar possíveis problemas de multicolinearidade.
- Análise gráfica dos Resíduos – Para verificar pressupostos de independência, homocedasticidade e identificação de possíveis *out-liers.*

Para cada variável retirada é reajustado os parâmetros do modelo, a ordem de exclusão das variáveis e os motivos foram descritos na tabela abaixo.

Tabela 1: Limpeza de variáveis

Variável	Motivo de Retirada
EFF	Não estimou os Parâmetros
REB	Não estimou os Parâmetros
BMI	P-valor próximo de 1
TOV	P-valor próximo de 1
FGM	Alta correlação com FGA e P-valor alto
STL	Alta correlação com AST e P-valor alto
OREB	Alta correlação
BLK	Alta correlação
Weight	P-valor alto
Age	P-valor alto
Height	P-valor alto
Experience	P-valor alto
AST.TOV	P-valor alto
FTM	Alta correlação com FTA
PF	Alta correlação com FTA
Games.Played	Sentido da variável
FGA	Alta correlação

2.4 ANÁLISE DE MULTICOLINEARIDADE

Deseja-se construir um modelo de regressão que melhor represente a explicação da variabilidade total da variável resposta em função de um conjunto de variáveis. Desta forma, para avaliar o diagnóstico de ajuste, é calculado o fator de inflação da variância – *VIF,* com o intuito de eliminar do modelo problemas de multicolinearidade. É preciso fazer esta análise para que não ocorram problemas de estimação dos parâmetros do modelo causando viés nas estimativas tais como baixa precisão (elevado erro).

Para diagnóstico de multicolinearidade utilizamos a métrica VIF:

$$VIFj = 1/(1 - R2j)$$

Utilizamos como regra para indicação de multicolinearidade de qualquer VIF superior a 5 para adoção do modelo final segundo seleção manual de variáveis, pois este se apresentou melhor ajustado aos dados comparados aos outros modelos testados por seleção automática.

Tabela 2: Análise de VIF para o modelo final

Variável	VIF
X3PA	63
X3PM	59.03
FGA	7.439
DREB	3.59
AST	2.348
X3P.	1.716
FG.	1.7
FT.	1.525
STL.TOV	1.078

Duas variáveis na análise apresentaram um elevado VIF, optamos em realizar a retirada dessas variáveis do modelo, estas são: X3PM, X3PA.

Feita a retirada das variáveis o modelo proposto se apresenta da seguinte maneira:

Tabela 3: Modelo Final

Coeficientes		Estimativa	Erro Padrão	Valor *t*	Pr(>|t|)
(Intercepto)	Intercepto	22.441	1.211	18.53	0.000
FG.	Media de aproveitamento de lances livres convertidos em relação aos cobrados	0.054	0.008	7.08	0.000
FGA	Tentativas de arremesso de 2 ou 3 pontos	0.034	0.002	14.06	0.000
FT.	Media de lances livre que tentou e realizou	0.025	0.006	4.2	0.000
X3P.	Media de tentativas e acertos de cestas de 3 pontos	0.023	0.007	3.16	0.002
AST	O número de assistências - passes que levam diretamente a uma cesta feita	-0.021	0.004	-4.87	0.000
DREB	Porcentagem de rebotes acumulados quando dada uma chance de rebote na defesa	-0.029	0.005	-6.43	0.000
STL.TOV	Passes que leva a cesta feita, comparado ao número de perda de bola para a defesa.	-0.07	0.012	-5.78	0.000

2.5 **COMPARAÇÃO DE MODELOS**

Para avaliarmos a qualidade dos modelos propostos, foram levados em consideração alguns critérios: coeficiente de determinação ajustado, *VIF* e análise de resíduos, abaixo na Figura 2.

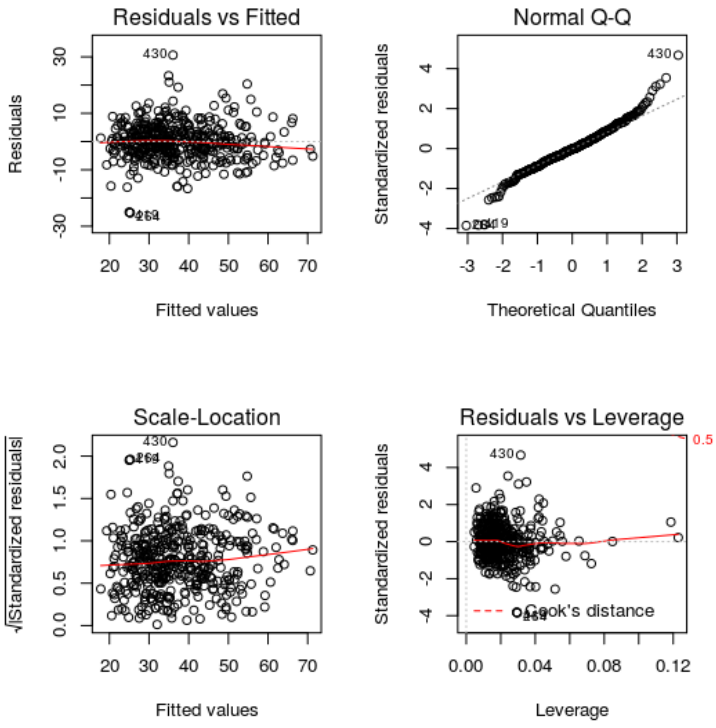
Tabela 4: Comparação de diagnósticos

Modelo	Coef_Det.Ajus	VIF Max
Seleção manual	0.6541180	5.411
Seleção Backward	0.7381523	68.51
Seleção Forward	0.7342847	191.4
Seleção Stepwise	0.7381523	68.51

O modelo selecionado manualmente mantém o mesmo nível de qualidade dos modelos com seleção automática, todavia os modelos comparativos apresentaram problemas de multicolinearidade em seu ajuste. Devido a isso, optamos por manter a seleção manual das variáveis.

3 ANÁLISE DE RESÍDUOS

Figura 2: distribuição dos resíduos



Pela análise gráfica dos resíduos apresentada pela Figura 2, pode-se dizer que os resíduos não violam os pressupostos exigidos para a validade de ajuste de uma regressão linear, ou seja, os resíduos parecem ser identicamente distribuídos (distribuição gaussiana), independentes e homocedásticos (variância comum). Há possíveis candidatos a *out-liers,* porém pela natureza das informações optou-se não retirá-los do conjunto de dados, ou seja, eles representam a realidade.

4 CONCLUSÃO

Pode-se concluir através da análise que o modelo se ajustou de forma mais satisfatória, através da seleção manual, com variáveis que melhor explicam a variável resposta e que fazem sentido para a o estudo proposto. As variáveis FG., FGA, FT., X3P, AST, DREB e STL.TOV foram as resultantes do modelo final de regressão que apresentou um coeficiente de determinação ajustado de 0,65 sendo um valor consideravelmente adequado. Esse valor foi um pouco abaixo dos coeficientes de determinação ajustados pelos métodos de seleção automáticos (backward, forward e stepwise). Entretanto, o modelo manual ficou mais bem ajustado sem problema de multicolinearidade e com uma quantidade menor de covariáveis sendo assim mais parcimonioso, ou seja, capaz de proporcionar um bom ajuste com menor quantidade de parâmetros.

REFERÊNCIAS

LOPES, Jose G. **Conheça o Kaggle, o portal para Ciência de Dados**. Disponível em:
< http://joseguilhermelopes.com.br/kaggle-o-portal-para-ciencia-de-dados/> Acesso em: 01 Julho 2018.

Base de dados: Social Power NBA
< https://www.kaggle.com/noahgift/social-power> Acesso em: 01 Julho 2018.

CHARNET, Reinaldo ET al. **Análise de modelos de regressão linear com aplicações**. Campinas, São Paulo, Unicamp, 356p, 1999.

CHARNET, Reinaldo et al. **Análise de modelos de regressão linear com aplicações**. Campinas, São Paulo, Unicamp, 356p, 1999.

CHARNET, Reinaldo et al. **Análise de modelos de regressão linear com aplicações**. Campinas, São Paulo, Unicamp, 356p, 1999.

MONTGOMERY, Douglas C.; PECK, Elizabeth A.; VINING, G. Geoffrey. **Introduction to linear regression analysis**. John Wiley & Sons, 2012.

DRAPER, Norman R.; SMITH, Harry. **Applied regression analysis**. John Wiley & Sons, 2014.

FARAWAY, Julian J. **Linear models with R**. CRC press, 2014.