Revista da Estatística UFOP, Vol III(3), 2014, ISSN 2237-8111
Edição especial: 59ª Reunião Anual da Regional Brasileira da Sociedade Internacional de Biometria - RBRAS
Trabalhos apresentados na versão poster.

Uso de Métodos de Seleção Stepwise para Dados de Baseball

Andreza Jardelino da Silva [1]
Abrãao de Paula Taveira [2]
Tiago Almeida de Oliveira [3]

1 Introdução

O baseball é um jogo em que há duas equipes com nove jogadores cada, que alternadamente ocupam posições de ataque e defesa. O campeonato é composto por 30 times (29 americanos e 1 canadense) divididos em duas sub-ligas: *American League* (Liga Americana) e *National League* (Liga Nacional), cada uma com 15 times. A grande diferença entra as ligas é que enquanto na *National League* os *pitchers* (o arremessador ou lançador) também vão para o bastão em busca da rebatida, na *American League* existe o rebatedor designado que não participa do jogo nos turnos de defesa de seu time, ele apenas "troca" de lugar com o *pitcher* de sua equipe no ataque, sem que seja necessária ocorrer uma substituição.

O jogo consiste em rebater a bola com um bastão e em seguida correr pelas quatro bases, onde, a equipe que está atacando poderá parar em uma das bases e depois avançar com a ajuda da rebatida de um outro companheiro. Os times trocam de posição assim que três rebatedores são eliminados, ganha o jogo quem obtiver mais corridas no final.

Entretanto, a disparidade em um jogo de baseball é imensa, se o time tem condições financeiras ele pode comprar os melhores jogadores e assim garantir sua vitória no final do campeonato, por outro lado, se o time tem poucos recursos financeiros, ele só pode comprar jogadores subvalorizados, considerados impróprios para as grandes ligas. Lewis (2004) defende a ideia de uma forma estatística de ver o jogo, o qual sua abordagem está na relação existente entre o valor desse atleta na folha de pagamento e seu desempenho em campo.

A análise de Regressão Múltipla é uma metodologia estatística que permite realizar predição de valores de uma ou mais variáveis resposta por meio de um conjunto de variáveis explicativas.

Este trabalho teve como objetivo verificar a aplicabilidade do modelo de regressão linear múltipla, em conjunto com os critérios de seleção de modelos, para investigar quais das variáveis preditoras tem relação direta com o número de vitórias.

[1]Graduanda do Curso de Bacharelado em Estatística - UEPB. e-mail: *adajardel@hotmail.com*
[2]Graduando do Curso de Bacharelado em Estatística - UEPB.
[3]Professor Doutor da Universidade Estadual da Paraíba - UEPB.

2 Material e métodos

Os dados utilizados para a realização deste trabalho foram provenientes do site ESPN [1]. Para a realização do trabalho foram coletadas um total de 30 observações. As onze variáveis obtidas para análise foram: vitórias (W), média de rebatidas (AVG), erros (E), batidas (H), alcance da segunda base (2B), alcance da terceira base (3B), bola fora das dimensões do campo - *home-runs* (HR), corridas rebatidas para dentro (RBI), arremessador acerta a bola no batedor (BB), arremesso bem sucedido - *strikeouts* (SO) e por fim, o tipo de grupo - nacional ou americano (LIGA). Foi utilizado um modelo de regressão linear múltipla em que a variável dependente y (variável reposta) foi o número de vitórias (W) em função das demais variáveis coletadas. Matricialmente temos:

$$\mathbf{Y} = X\beta + \varepsilon, \tag{1}$$

em que X é a matriz de incidência do modelo;β é o vetor que contém: β_0 é o intercepto do modelo; β_1 coeficiente de regressão que relaciona x_1 a variável resposta; x_1 é covariável AVG; β_2 coeficiente de regressão que relaciona x_2 a variável resposta; x_2 é a covariável E; β_3 coeficiente de regressão que relaciona x_3 a variável resposta; x_3 é a covariável H; β_4 coeficiente de regressão que relaciona x_4 a variável resposta; x_4 é covariável 2B; β_5 coeficiente de regressão que relaciona x_5 a variável resposta; x_5 é a covariável 3B; β_6 coeficiente de regressão que relaciona x_6 a variável resposta; x_6 a covariável HR; β_7 coeficiente de regressão que relaciona x_7 a variável resposta; x_7 é a covariável RBI; β_8 coeficiente de regressão que relaciona x_8 a variável resposta; x_8 é a covariável BB; β_9 coeficiente de regressão que relaciona x_9 a variável resposta; x_9 é a covariável SO; β_{10} coeficiente de regressão que relaciona x_{10} a variável resposta; x_{10} é a covariável LIGA.

Ao realizar a análise de regressão, é necessário avaliar a existência da relação entre a variável resposta e as variáveis explicativas. Para verificar esta relação foi realizado o teste t de *Student*, o qual testa a hipótese de associação linear entre as variáveis envolvidas, e procedeu-se também o teste F de Snedecor. Foi realizado após os testes de hipóteses sobre os parâmetros, a verificação de algumas das pressuposições do modelo de regressão como, normalidade dos resíduos, homogeneidade de variância dos resíduos, independência dos resíduos. Os resíduos são obtidos pela seguinte equação: $\varepsilon = Y - X\beta$. A normalidade dos resíduos foi testada por meio da estatística de *Shapiro-Wilk*, sob hipótese H_0 de normalidade (SHAPIRO e WILK, 1965). As demais pressuposições foram verificadas por meio de análise gráfica dos resíduos. Realizou-se os métodos de Stepwise em conjunto com o Critério de Seleção de Akaike (AIC) e calculou-se o Cp de Mallows, R^2 ajustado para a seleção do modelo que melhor explique os dados. Todos os procedimentos das análises e gráficos se deu por meio do software R (R DEVELOPMENT CORE TEAM, 2013).

[1] *(http://espn.go.com/mlb/stats/team/-/stat/batting)*

3 Resultados e discussão

Na Figura 1 têm-se uma visualização geral dos dados, por meio dela sugere-se a existência de relação entre a variável resposta W e as variáveis explicativas. Especificamente, ao se verificar o gráfico obtida pela variável LIGA em relação a W, percebe-se a formação de dois grupos distintos. Uma outra observação está com relação aa variáveis HR, RBI, BB E X2B, a qual, sugere-se uma relação quadrática com W.

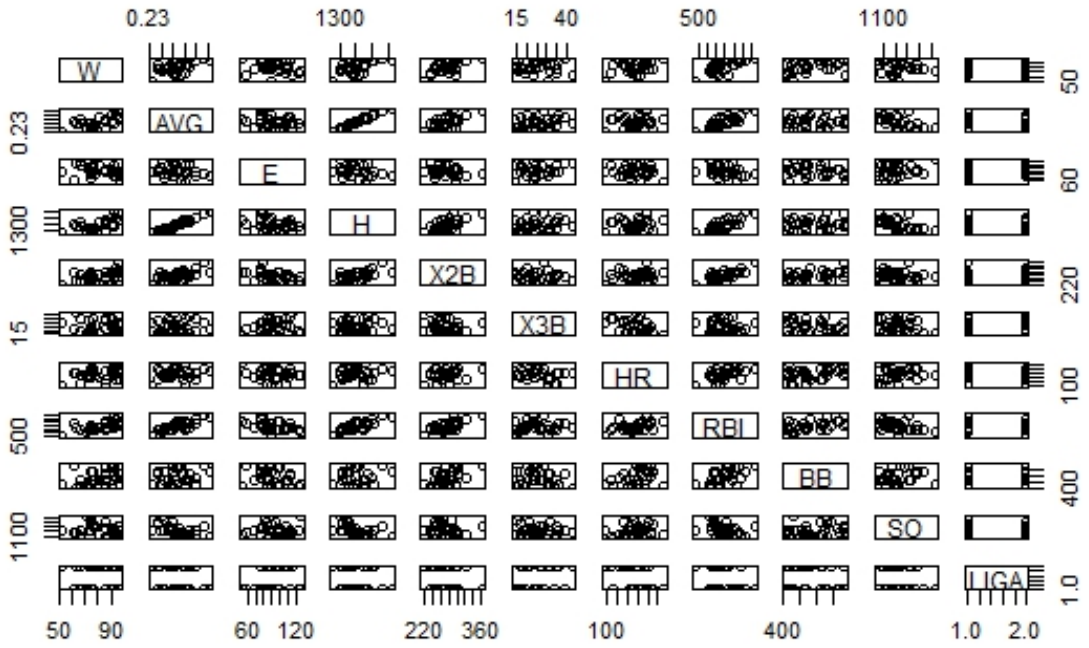

Figura 1: Scatter Plot para as variáveis W, AVG, E, H, 2B, 3B, HR, RBI, BB, SO e LIGA

Diante do que foi observado na Figura 1, procedeu-se o ajuste de um modelo de regressão múltipla, em que W é a variável resposta (y) e as variáveis explicativas com acréscimos de variáveis quadráticas, foi considerado como o modelo completo, foram: AVG, E, H, X2B, $X2B^2$, X3B, HR, HR^2, RBI, RBI^2, BB, BB^2, SO e LIGA , respectivamente estimados pelos seguintes coeficientes: $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5, \hat{\beta}_6, \hat{\beta}_7, \hat{\beta}_8, \hat{\beta}_9, \hat{\beta}_{10}, \hat{\beta}_{11}, \hat{\beta}_{12}, \hat{\beta}_{13}, \hat{\beta}_{14}$. Dessa maneira, os valores estimados estão na Tabela 1. Pôde-se observar que os preditores não obtiveram efeito signicativo, isto pode ter ocorrido devido ao fato que o teste t no modelo de regressão multipla é condicional, isto pode levar a má escolha do modelo ou no presente caso a afirmar que não existe um modelo de regressão adequado para o número de vitórias, daí procedeu-se o método de stepwise para a busca de um modelo que explique a variável resposta de interesse como se observa na tabela 2.

Tabela 1: Estimativas para o modelo completo referente aos parâmetros com respectivos erros padrões e estatística *t* para as variáveis explicativas.

Variáveis	Estimativas	E.P	valor t	Pr(>|t|)
(Intercept)	-285,7609	200,2844	-1,43	0,1741
AVG	-1759,1409	1056,4545	-1,67	0,1166
E	0,0997	0,1297	0,77	0,4542
H	0,2128	0,1443	1,47	0,1610
X2B	0,3714	1,2804	0,29	0,7758
X3B	-0,1071	0,2919	-0,37	0,7187
HR	-0,2159	0,7674	-0,28	0,7823
HR^2	0,0004	0,0025	0,17	0,8677
RBI	0,7744	0,6507	1,19	0,2525
RBI^2	-0,0004	0,0005	-0,82	0,4274
BB	0,6301	0,6636	0,95	0,3574
BB^2	-0,0007	0,0007	-1,03	0,3197
$X2B^2$	-0,0006	0,0023	-0,28	0,7828
SO	0,0052	0,0226	0,23	0,8192
LIGAn	4,5713	4,3762	1,04	0,3128

Para o modelo completo, o R^2 ajustado foi de $49,13\%$, não apresentando um bom ajuste do modelo de regressão (TABELA 1). Aplicando o critério de seleção o modelo selecionado conteve como variáveis explicativas RBI, RBI^2 e apresentou um ajuste R^2 ajustado melhor do que o do modelo completo de $55,56\%$.

Tabela 2: Estimativas para o modelo reduzido referente aos parâmetros com respectivos erros padrões e estatística *t* para as variáveis explicativas.

Variáveis	Estimativa	E.P.	valor t	Pr(> t)
(Intercepto)	$-1,666 \times 10^2$	$9,485 \times 10^1$	$-1,756$	$0,0904$
RBI	$6,388 \times 10^{-1}$	$2,900 \times 10^{-1}$	$2,203$	$0,0363$
RBI^2	$-3,897 \times 10^{-4}$	$2,203 \times 10^{-4}$	$-1,769$	$0,0881$

Para o modelo selecionado conclui-se que para cada unidade de RBI tem-se um acréscimo de $0,6384$ no número de vitórias.

Pelos gráficos dos resíduos do modelo da tabela 2, observa-se que não houve desvios de normalidade que corrobora o teste de shapiro-wilk aplicado aos resíduos com valor P $>$ 0,05. Observa-se também que há presença de pontos no gráfico da distância de cook que merecem maiores estudos quanto a possibilidade de serem pontos de alavanca.

O critério de Cp foi aplicado porém não houve um modelo adequado que tivesse um valor próximo ao número de variáveis, desta maneira o critério foi desconsiderado para a análise dos dados de baseball.

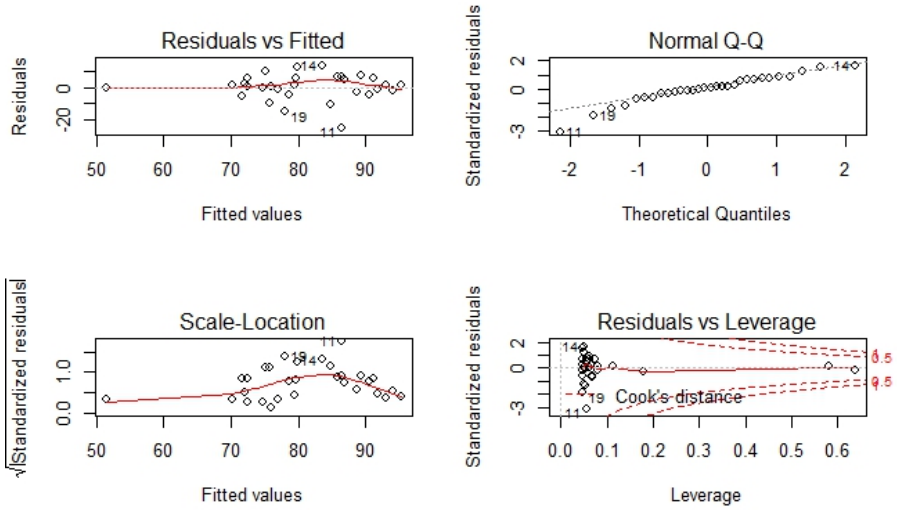

Figura 2: Gráfico de resíduos para o modelo ajustado com RBI e RBI^2

4 Conclusão

O melhor modelo escolhido a partir do método stepwise foi o que conteve como preditores as variáveis RBI e RBI^2, desta forma, mesmo com a existência de dois grupos distintos, o número de vitórias não está sendo influenciado devido as ligas, e sim, em relação a quantidade de rebatidas para dentro no jogo.

Referências

[1] HOFFMANN,A. ANÁLISE DE REGRESSÃO - Uma Introdução À Econometria. 4º ed. SÃO Paulo, SP: Hucitec, 2006. 378 p.

[2] LEWIS, M. MONEYBALL: The Art of Winning an Unfair Game. 1st edition. W. W. Norton & Company. 2004. 320 p.

[3] R Development Core Team (2012). R: A language and enviroment for statistical computing. Disponível em: <http://www.R-project.org>. Acesso em: 10 set. 2012.

[4] SHAPIRO, S. S.; WILK. M. B., Biometrika, Vol. 52, No. 3/4. Dec., 1965, pp. 591-611.