Análise de dados de contagem para jogadores de futebol americano

Alexandre Morales Diaz
Eduardo Pereira Lima
Pedro Guilherme Guimaraes
Vinicius Larangeiras

Trabalho de Modelos Lineares Generalizados (CE-225), Universidade Federal do Paraná, submetido ao professor Cesar Augusto Taconeli.

Curitiba
2017

Sumário

1. Resumo

O trabalho consiste em modelar uma distribuição de variável resposta do tipo contagem. Para isso foram analisados os dados de diversos quarterbacks, posição mais importante do futebol americano, e foi feito um estudo para analisar o desempenho desses jogadores e o quanto seu desempenho contribui para a pontuação final do time. Para isso foi utilizado a aplicação de modelos lineares generalizados para dados de contagem.

Palavras chave: Dados de contagem, Quarterback, Futebol Americano, Poisson, Binomial Negativa, Binomial Negativa Inflacionada de zeros, Modelos Lineares Generalizados, GLM, Deviance Residual, AIC

2. Introdução

Neste trabalho foram ajustados alguns modelos para dados de contagem com uma base de 13.188 observações de jogadores de futebol americano, sendo todos da posição de quarterback. Cada linha da base corresponde às estatísticas do jogador em uma partida, com isso, os jogadores terão mais de uma linha de estatísticas, já que essa base contém estatísticas de todos os jogos de 1996 até 2016.

Inicialmente apresentamos os materiais e métodos utilizados nesse trabalho. A base de dados que foi retirada do site www.kaggle.com, e fizemos uma breve análise descritiva destes dados para poder entender melhor como se comportam os dados. Depois apresentamos e citamos alguns modelos que foram ajustados para tentar chegar no melhor ajuste possível, considerando alguns métodos e estatísticas para se ter um bom modelo ajustado.

Para verificar a qualidade do ajuste do modelo levamos em consideração a análise dos resíduos e o quão bom foi o ajuste dos dados para com a distribuição escolhida em cada tentativa de ajuste, e outras estatísticas como Deviance Residual e Critério de Informação de Akaike.

Algumas das distribuições que foram utilizadas para essas análises foram a Poisson e a Binomial Negativa, porém esses dois ajustes não foram considerados satisfatórios, seja por problemas como superdispersão, ou então falta ajuste da distribuição aos dados, não atendendo os pressupostos dos resíduos, em especial à normalidade dos mesmos. A distribuição que melhor se ajustou aos dados foi a Binomial Negativa Inflacionada de Zeros, sendo esta a escolhida para ajustar o modelo em questão.

Com a definição do modelo a ser usado, iremos prosseguir com o objetivo do trabalho, que é relacionar o desempenho do quarterback com a pontuação final do time, ou seja, verificar quais estatísticas da posição influenciam mais na pontuação final do time, para então possivelmente poder assessorar jogadores da posição para que seu desempenho seja voltado à ajudar o máximo possível à vitória do time.

3. Material e métodos

3.1 Definição da Base de Dados

Após pesquisar diferentes bases de dados para este estudo, decidimos por utilizar a base de estatísticas de quarterbacks por atender aos requisitos de dados de contagem, e também por ser um esporte em que as estatísticas exercem enorme influência no resultado final da partida, diferentemente do futebol em que um lance pode definir a partida, no futebol americano o resultado é construído ao longo da partida.

3.2 Análises descritivas da base

Como citado anteriormente, a base estudada contém um total de 13.188 linhas, sendo cada linha correspondente às estatísticas de um quarterback num determinado jogo. Apresentamos na Tabela 1 as descritivas das variáveis contínuas.

Tabela 1: Descritivas das variáveis contínuas da base utilizada

	Tentativas	Completos	Jardas	Jds_Tentativa	Touchdown	Interceptado	Mais_longo	Sack	Jds_Perdidas	Nota_NFL	Pontos_time	Ano
Mínimo	0,00	-6,00	-11,00	-11,00	0,00	0,00	-11,00	0,00	0,00	0,00	0,00	1996,00
1º Quartil	20,00	11,00	115,00	5,20	0,00	0,00	22,00	0,00	0,00	58,60	13,00	2001,00
Mediana	29,00	17,00	197,00	6,70	1,00	1,00	31,00	2,00	9,00	80,90	21,00	2006,00
Média	26,87	16,12	168,10	6,88	1,11	0,80	33,00	1,86	11,97	80,24	21,38	2006,00
3º Quartil	36,00	22,00	260,00	8,10	2,00	1,00	44,00	3,00	12,00	102,00	28,00	2011,00
Máximo	69,00	58,00	527,00	81,00	7,00	7,00	99,00	12,00	91,00	158,30	62,00	2016,00

Como podemos ver, temos 15 variáveis na base, sendo elas:

- Nome: Identificação dos jogadores. Os nomes podem se repetir, já que um jogador pode ter realizado mais de uma partida.
- Tentativas: Tentativas de lançamentos realizadas do quarterback durante o jogo.
- Completos: Lançamentos completos realizados pelo jogador, ou seja, tentativas de lançamento em que o jogador alvo conseguiu completar a recepção.
- Jardas: Total de jardas dos passes completados.
- Jds_Tentativa: É uma média de jardas totais por tentativa de passe.
- Touchdown: Touchdown é o “gol” do futebol americano. Essa variável contempla apenas os touchdowns realizados a partir de um passe do quarterback.
- Interceptado: Interceptações sofridas pelo quarterback. Interceptação se define quando um jogador do time adversário recebe a bola durante uma tentativa de lançamento.
- Mais_Longo: Passe completo pelo quarterback mais longo do jogo.
- TD_Longo: Variável que indica se o passo mais longo do jogo teve como resultado um touchdown. Com n igual a 10301 e t igual a 2887, com n sendo não e t sendo touchdown.

- Sack: Esse é um evento que ocorre quando um jogador do time adversário consegue derrubar o quarterback antes que o mesmo realize o lançamento, resultando numa perda de jardas.
- Jds_Perdidas: Quantidade de jardas perdidas resultadas de sacks.
- Nota_NFL: Nota dada pela NFL (organizadora da competição) para o desempenho do quarterback.
- Pontos_Time: Quantidade de pontos marcados pelo time do quarterback.
- Casa_Fora: Indica se o time do quarterback era mandante ou não da partida. Tendo 6629 em casa e 6559 fora de casa, com alway sendo fora e home casa.
- Ano: Indica o ano em que foi realizada a partida.

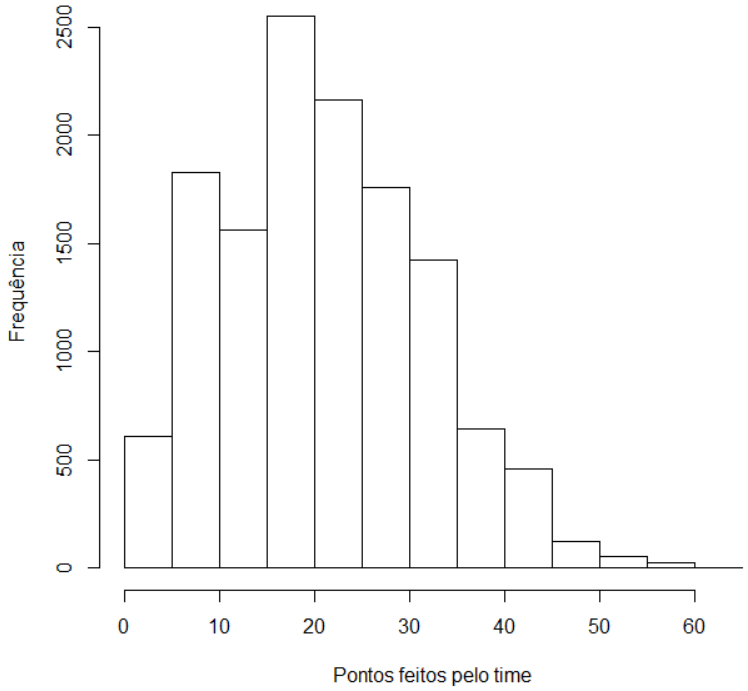
3.3 Modelos Ajustados

Para ajustar um modelo de GLM, removemos da base a variável “Nome”, pois o modelo de GLM pressupõe de que as linhas são independentes entre si. Para poder utilizar a identificação do jogador no modelo teríamos que aplicar técnicas de estatística multivariada, porém como não temos o arcabouço necessário para utilizar tais técnicas, iremos assumir que a linhas são independentes.

Utilizando o método de seleção Stepwise, em que são analisadas quais variáveis são significativas para o ajuste do modelo, foi definido de que as variáveis “Jds_Tentativa” e “TD_Longo” não são significativas para o ajuste do modelo final.

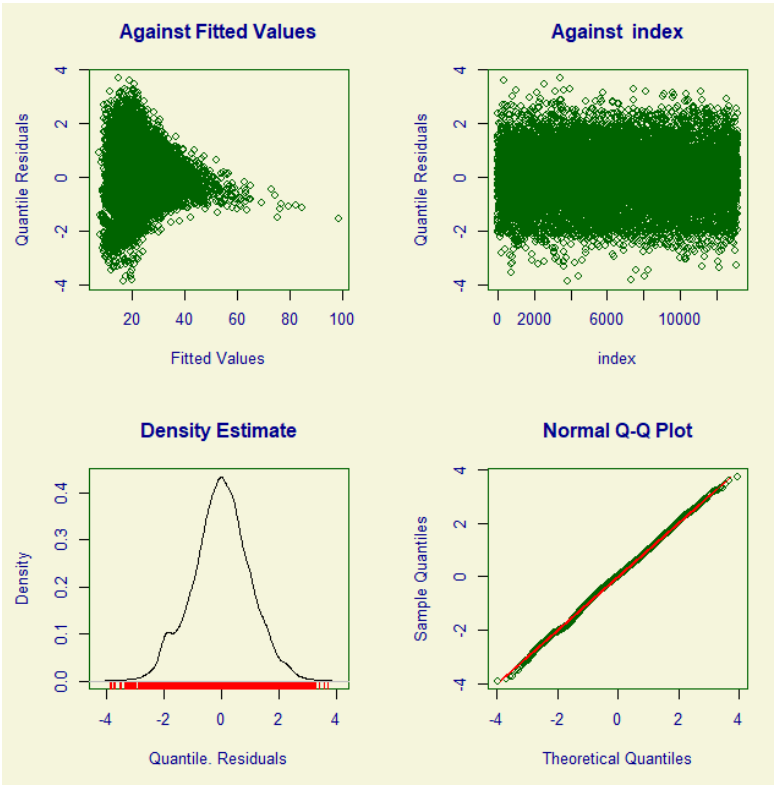
O modelo que melhor se adequou aos dados foi a Binomial Negativa Inflacionada de Zeros, que é um modelo de mistura de uma distribuição para contagens, no caso uma Binomial Negativa, com uma distribuição degenerada em zero (igual a zero com probabilidade um). Apesar de a variável resposta não conter zeros em sua maioria, foi o modelo que melhor se ajustou aos dados. Na Figura 1, podemos observar o histograma da variável “Pontos_Time”.

Figura 1 - Histograma dos pontos feitos por time



Na Figura 2, temos os gráficos para análise dos resíduos do modelo final.

Figura 2 – Gráficos das análises dos resíduos



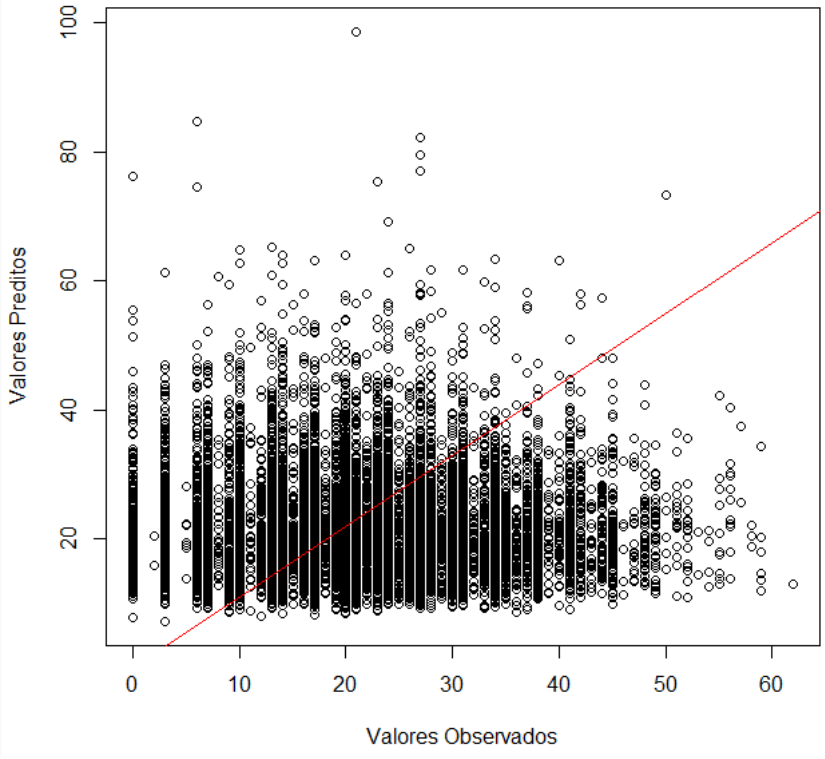
Através desses gráficos, podemos tirar as seguintes conclusões:

- O gráfico de quantis residuais versus valores ajustados mostra uma tendência dos dados em forma de cone, porém levando em conta os outros gráficos, este não se mostra um problema relevante ao ajuste do modelo.
- O gráfico de quantis residuais versus o índice de cada observação nos mostra que o dados estão homogeneamente distribuídos ao redor da média, e a dispersão dos desvios, eixo Y, está dentro da amplitude satisfatória, pois a grande maioria dos resíduos se encontram entre -3 desvios padrão e 3 desvios padrão.
- O gráfico de densidade versus os quantis residuais nos mostra que os resíduos aparentam seguir uma distribuição normal, que é nosso objetivo nesse estudo.
- E por último o gráfico de normalidade, conhecido por gráfico Quantil-Quantil, que corrobora o terceiro gráfico, indicando normalidade dos resíduos.

O AIC deste modelo foi de 92985,44. Como falamos anteriormente, tentamos ajustar outros modelos com outras distribuições, e como parte de nossa escolha pelo melhor modelo, levamos em conta essa estatística, que também indicou a Binomial Negativa Inflacionada de Zeros como o modelo que melhor se ajusta aos dados

Escolhido o modelo que melhor se ajusta aos dados, realizamos as predições para cada observação, e comparamos com os valores observados. Segue a comparação na Figura 3.

Figura 3 - Preditos vs Observados



Podemos observar que existe uma forte associação entre os valores preditos e os valores observados, trazendo mais evidências de que o modelo se ajustou adequadamente aos dados.

4. Conclusão

Após toda a análise realizada neste estudo concluímos que as variáveis que mais influenciam positivamente para a pontuação final do time são “Touchdown” e “Casa_Fora”, sendo que jogar em casa aumenta significativamente os pontos do time, e as que mais influenciam negativamente são “Sack” e “Interceptado”.

Dito isso, em uma possível assessoria a um quarterback, seria indicado treinar lançamentos que visam jogadores perto da linha de touchdown, e também algum treinamento psicológico para que quando o time jogue fora de casa tenha um desempenho melhor.

Para o caso dos sacks, teria que ser feito um trabalho com a linha ofensiva, que são os jogadores que protegem o quarterback, pois eles evitam que o quarterback seja sacado. Além disso, um treinamento com o quarterback para fazer lançamentos mais rápidos, evitando assim o sack.

Para os passes interceptados, seria indicado um treinamento de lançamento, pois a maior causa de interceptações é o lançamento errado.

5. REFERÊNCIAS BIBLIOGRÁFICAS

Paulo, Gilberto A. Modelos de Regressão com apoio computacional. Instituto de Matemática e Estatística Universidade de São Paulo.