Análise de Desempenho de Jogadores de Basquete

Alexandre Morales Diaz

Eduardo Pereira Lima

Vinicius Larangeiras

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

Curitiba

2017

Sumário

1. Resumo 3
2. Introdução 3
3. Material e métodos 4
3.1 Definição da Base de Dados 4
3.2 Ajuste de Modelos 4
3.3 Outros Modelos 8
4. Conclusão 9
5. REFERÊNCIAS BIBLIOGRÁFICAS 9

1. Resumo

O trabalho apresenta um estudo de desempenho de jogadores de basquete, com a aplicação de Modelos de Regressão Lineares, tentando relacionar estatísticas como peso, altura e aproveitamento de arremessos com a média de pontos de cada jogador individualmente. Para isso foram aplicados Modelos Lineares para descobrir as relações entre as variáveis.

Palavras-chave: Modelos Lineares, Modelos de Regressão Lineares, Modelo de Regressão Linear, Modelos Lineares Generalizados.

2. Introdução

Os Modelos de Regressão Lineares são utilizados para analisar a relação entre variáveis explicativas (peso, altura...) com a variável resposta (média de pontos por jogo), bem como analisar a significância de cada variável do modelo de regressão e a qualidade do ajuste.

A base de dados utilizada para este trabalho foi retirada de um repositório indicado pelo próprio professor da matéria, contendo informações de características físicas, aproveitamento de arremessos, e média de pontos por jogo, sendo esses dados disponíveis para cada jogador analisado na base de dados.

O objetivo deste trabalho é tentar ajustar um modelo de regressão linear satisfatório para a base de dados escolhida para o estudo. Para isso foram utilizados métodos e testes estatísticos que auxiliam no ajuste do modelo, como seleção de variáveis, calibração de parâmetros, etc.

3. Material e métodos

3.1 Definição da Base de Dados

Para definir a base de dados a ser utilizada foram consultados os repositórios disponibilizados pelo professor. Após analisar algumas bases de dados, levamos em conta número de variáveis e observações de cada base de dados, sendo que achamos satisfatórios esses critérios dos dados dos jogadores de basquete, bem como concluímos que essas variáveis têm uma alta relação entre si, podendo assim ajustar um modelo de regressão linear.

A base de dados contém informações de 54 jogadores de basquete, com as seguintes variáveis:

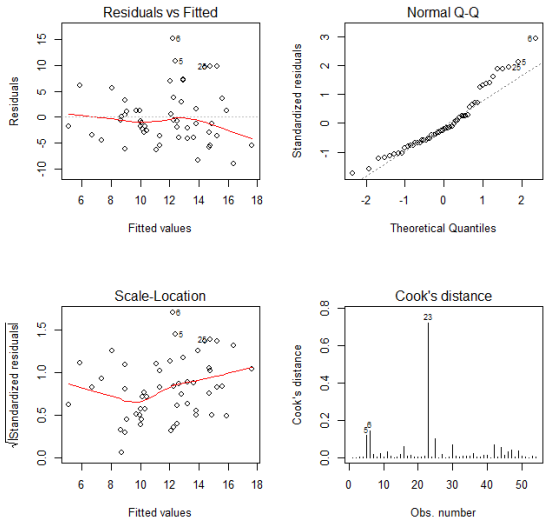
- Altura – altura do jogador em pés;
- Peso – peso do jogador em libras;
- Quadra – percentual de acerto de arremessos de quadra, que são arremessos realizados de qualquer lugar da quadra, sendo que a bola está em jogo;
- Livre – percentual de acerto de lances livres, que são arremessos realizados de um local predefinido quando o time adversário comete falta, sendo o jogador que recebeu a falta o responsável pelos arremessos, sendo na maioria das vezes dois arremessos, mas podendo também ser um ou três arremessos, dependendo das condições em que a falta foi sofrida;
- PPG – sigla em inglês para média de pontos por jogo (points per game).

Para melhor entendimento e esclarecimento nas análises foi realizada uma conversão nas unidades de medidas das variáveis Altura e Peso dos jogadores, para Metros e Quilogramas. Poderia ser escolhido como variável resposta Quadra, Livre ou PPG, sendo que num primeiro momento definimos como variável resposta PPG.

3.2 Ajuste de Modelos

O primeiro modelo apresentado incluindo todas as variáveis como explicativas (PPG ~ Altura + Quadra + Livre + Peso), se mostrou longe de ser satisfatório para análises dos dados, neste modelo somente a variável Quadra foi significativa e da variância total, apenas 22,2% foi explicada. Os Pressupostos também foram insatisfatórios, como mostrado na figura 1.

Figura 1 - Pressupostos obtidos a partir do primeiro modelo de análises



Como não foram atendidos os pressupostos necessários, foi recorrido ao método Stepwise, para ver quais variáveis deveriam ser mantidas no modelo. O modelo indicado com o método foi PPG ~ Altura + Quadra + Livre, ou seja, o modelo sugerido foi retirar apenas a variável Peso. Este novo modelo se mostrou ainda ineficaz para representar os dados, explicando somente 22,1% da variação total dos dados e persistindo com somente a variável Quadra sendo significativa. A perda foi considerada irrelevante, visto que a Variável Peso apenas contribuiu com 0,1% para ajudar a explicar os dados. Os pressupostos continuaram não sendo atendidos (Figura 2), o gráfico da distância de Cook mostrou que o individuo 23 destaca-se em relação aos demais, com isso foi decidido retira-lo da base.

Figura 2 - Pressupostos obtidos a partir do Segundo modelo de análises

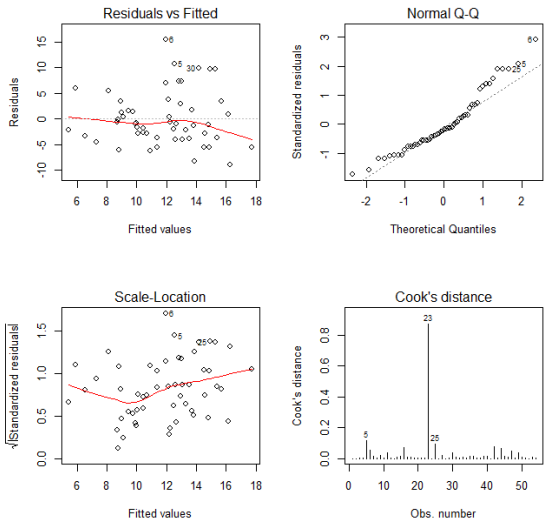
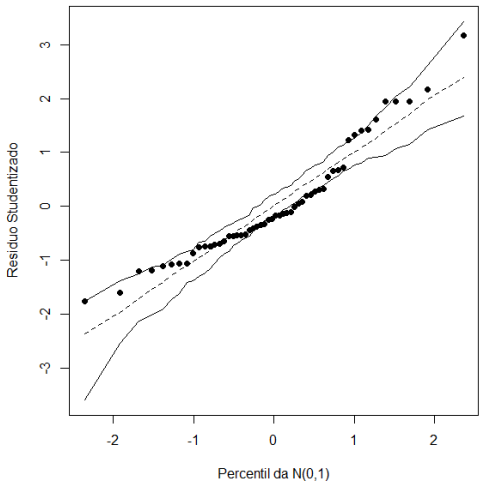


Figura 3 - Envelope Simulado para os Dados gerados a partir do Segundo Modelo



Para um melhor critério de avaliação foi realizado um teste de normalidade (Teste de Shapiro-Wilk) no segundo modelo, o p-valor apresentado foi de 0,0083, rejeitando a Hipótese Nula (Normalidade). Visualmente representado pela Figura 3.

O Terceiro modelo ajustado foi PPG ~ Altura + Quadra + Livre, sem a variável Peso conforme indicado pelo uso do Método Stepwise e retirando o individuo 23 da base de dados, pois ele apresentava muita discrepância em relação aos demais. A partir desta decisão as variáveis Quadra e Livre, tiveram efeito significativo para o modelo, mas a explicação para a variação permaneceu baixa com somente 25,9%, e ainda, os pressupostos continuaram a não ser atendidos, como mostrado nas figuras 4 e 5.

Figura 4 - Pressupostos obtidos a partir do Terceiro Modelo de Análises

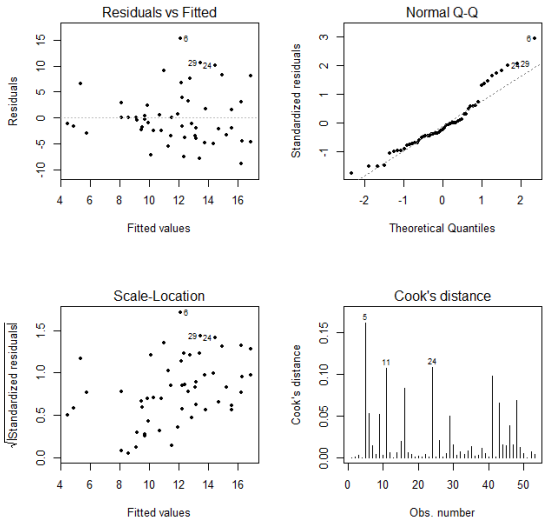
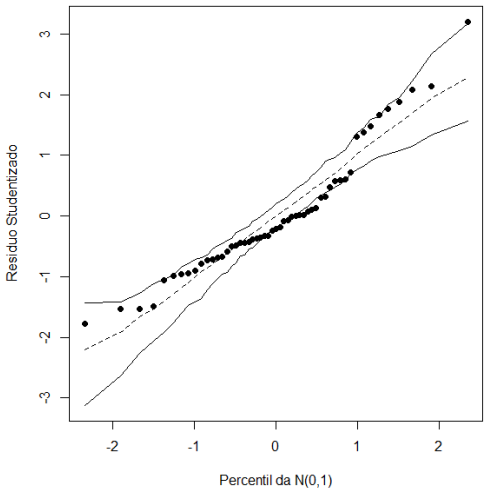
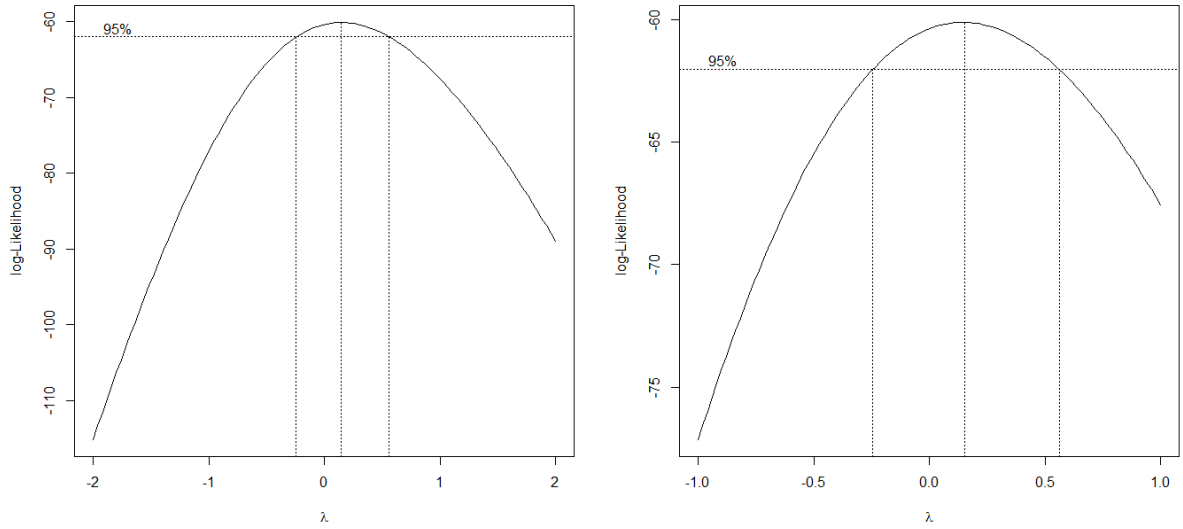


Figura 5 - Envelope Simulado para os Dados gerados a partir do Terceiro Modelo



Os dados parecem ter uma variância não constante, e a normalidade não está sendo atendida. Uma solução para corrigir este problema, foi aplicar uma transformação na variável resposta utilizando o método de Box Cox (Figura 6). O método indicou a Transformação 0,152, mas como essa transformação se assemelha muito a uma Transformação Logarítmica, foi optado a utiliza lá por uma questão de praticidade.

Figura 6 - Transformação Sugerida para a Variável Resposta pelo Método de Box-Cox



Um quarto ajuste do modelo foi realizado com a Transformação da Variável Resposta, sem a Variável Peso e sem o individuo 23 da base de dados. Este ajuste apresentou uma melhora na explicação dos Dados, porém insuficiente para se considerar um ajuste adequado, explicando 34,6% da variabilidade total, porém enfim os Pressupostos foram atendidos (Figuras 7 e 8), indicando finalmente que o modelo teve alguma adequação aos dados.

Figura 7 - Pressupostos obtidos a partir do Quarto Modelo de Análises

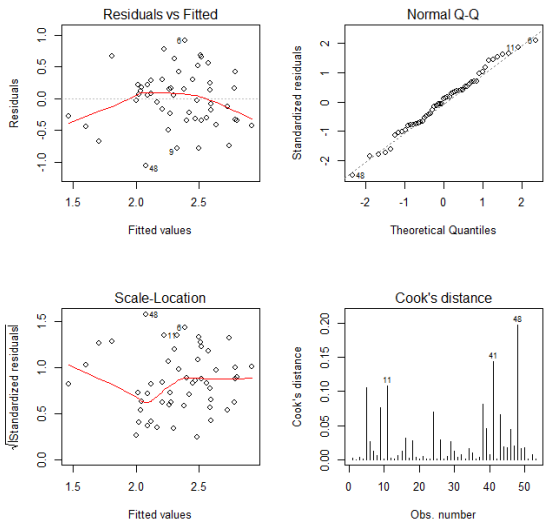
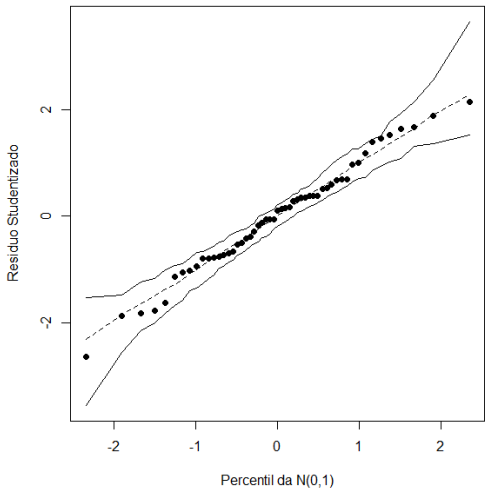


Figura 8 - Envelope Simulado para os Dados Gerados a partir do Quarto Modelo



O teste de Shapiro Wilk também confirmou a Normalidade, com um p-valor apresentado de 0,8864.

A fim de melhorar a explicação da variação do modelo, foram procurado outras possíveis causas que estivessem comprometendo o modelo final e analisando os Gráficos de Pressupostos alguns indivíduos se mostravam como possíveis pontos discrepantes, foram eles os indivíduos 9 e 48 de nossa base de dados, e como a Variável Explicativa Altura não

apresentou significância em nenhum modelo proposto até então, foi retirada do Modelo Final que ficou descrito como PPG ~ Quadra + Livre, com uma Transformação na Variável Resposta, retirando as variáveis Altura e Peso como Explicativas e sem 3 indivíduos de nossa base de dados, essa perda de informação foi estatisticamente aceitável, pois teve relevância na explicação da variação total dos Dados. Este Modelo, entretanto, apesar de todas as Recorrências necessárias não se mostrou satisfatório para a adequação desejada dos Dados, contudo sua melhora foi há 38,4%, quase 40% da variação total. Os Pressupostos e Normalidade (Figuras 9 e 10, respectivamente) tiveram melhoras e correspondências possíveis dentre os modelos apresentados.

Figura 9 - Pressupostos obtidos a partir do Modelo Final de Análises

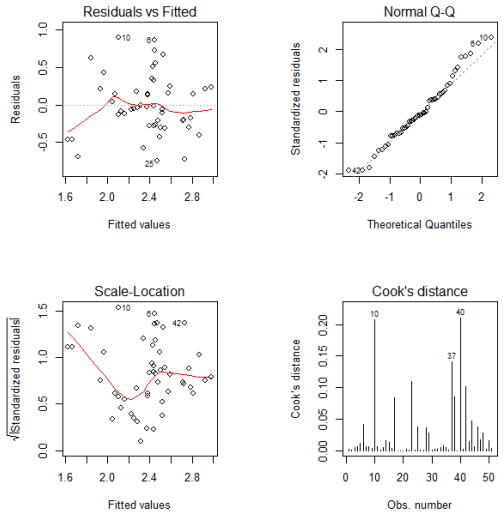
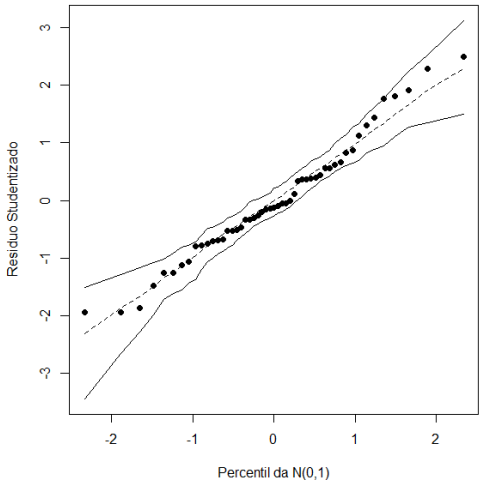


Figura 10 - Envelope Simulado dos Dados Gerados a partir do Modelo Final



3.3 Outros Modelos

Visto a insatisfação do Modelo final, trocamos a variável resposta em busca de algo mais ideal. Iteramos o mesmo procedimento aplicado para encontrar este Modelo Final trocando as Variáveis Respostas e Explicativas entre si, o resultado foi encontrar uma variável que poderia ser explicada há um nível mais razoável, porém, a interpretação da correlação se tornou espúria, ou seja, não havia sentido prática lógico para acreditar que a explicação era de fato verdadeira. Este Modelo foi representado por Quadra ~ Altura + PPG, o que conseguiu explicar, através de alguns recursos, 38,8% da variação Total dos Dados e atendendo os Pressupostos. Todos os modelos apresentando Livre como Variável Resposta se mostraram abaixo em relação aos modelos com PPG e Quadra como Variáveis Resposta, tendo seu máximo em 30,0% de explicação sobre a variação Total.

4. Conclusão

O modelo de regressão linear que concluímos ter o melhor ajuste foi o PPG ~ Quadra + Livre por atender bem os pressupostos e as variáveis explicativas terem boa significância ao modelo, apesar de que o coeficiente de determinação não foi considerado muito bom.

As variáveis Peso e Altura foram excluídas do modelo por não acusarem significância relevante ao modelo.

Por conta do modelo não apresentar um coeficiente de determinação satisfatório, acreditamos que se a base de dados tivesse mais variáveis, como tempo em quadra e tentativas de arremesso, seria possível ajustar melhor um modelo de regressão linear.

Cabe destacar que seria interessante implementar outras técnicas estatísticas, como Modelos Lineares Generalizados (GLM), que possivelmente se ajustariam melhor aos dados.

5. REFERÊNCIAS BIBLIOGRÁFICAS

GIOLO, Suely Ruiz. **Análise de Regressão Linear**. Curitiba: Universidade Federal do Paraná, 2017.