UNIVERSIDADE FEDERAL DE SERGIPE

CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA

DEPARTAMENTO DE ESTATÍSTICA E CIÊNCIAS ATUARIAIS

ALEX CELMO SANTOS SOUZA

MODELAGEM DE REGRESSÃO RESISTENTE A *OUTLIERS* VERTICAIS E HORIZONTAIS DAS CAPTURAS DE CAMARÕES NO ESTADO DO RIO DE JANEIRO

São Cristóvão – SE

2025

ALEX CELMO SANTOS SOUZA

MODELAGEM DE REGRESSÃO RESISTENTE A *OUTLIERS* VERTICAIS E HORIZONTAIS DAS CAPTURAS DE CAMARÕES NO ESTADO DO RIO DE JANEIRO

Trabalho de Conclusão de Curso apresentado ao Departamento de Estatística e Ciências Atuariais da Universidade Federal de Sergipe, como parte dos requisitos para obtenção do grau de Bacharel em Ciências Atuariais.

Orientador: Prof. Dr. Luiz Henrique Gama Dore de Araújo

São Cristóvão – SE

2025

ALEX CELMO SANTOS SOUZA

MODELAGEM DE REGRESSÃO RESISTENTE A *OUTLIERS* VERTICAIS E HORIZONTAIS DAS CAPTURAS DE CAMARÕES NO ESTADO DO RIO DE JANEIRO

Trabalho de Conclusão de Curso apresentado ao Departamento de Estatística e Ciências Atuariais da Universidade Federal de Sergipe, como um dos pré-requisitos para obtenção do grau de Bacharel em Ciências Atuariais.

Aprovado em 31/03/2025, Nota Final: 10,0

Banca Examinadora

Documento assinado digitalmente
gov.br
LUIZ HENRIQUE GAMA DORE DE ARAUJO
Data: 10/04/2025 15:21:04-0300
Verifique em https://validar.iti.gov.br

_____

Prof. Dr. Luiz Henrique Gama Dore de Araújo

Orientador

Documento assinado digitalmente
gov.br
CLEBER MARTINS XAVIER
Data: 10/04/2025 11:10:30-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Cleber Martins Xavier

1° Examinador

Documento assinado digitalmente
gov.br
JOSAFA JOSE DO CARMO REIS JUNIOR
Data: 10/04/2025 10:20:33-0300
Verifique em https://validar.iti.gov.br

Dr. Josafá José Do Carmo Reis Junior - UFRPE

2° Examinador

Este trabalho é dedicado à minha mãe Arlete
e aos meus familiares.

AGRADECIMENTOS

Gostaria de expressar minha profunda gratidão a todas as pessoas que, de alguma forma, contribuíram para a realização deste trabalho e para minha jornada ao longo da graduação. A minha mãe, Arlete, dedico este trabalho com todo o amor e reconhecimento pelo apoio incondicional, que foi meu alicerce em todos os momentos. A minha namorada, Kananda, meu agradecimento especial por estar ao meu lado, ao trazer força, paciência e carinho durante essa trajetória.

Um agradecimento especial ao meu orientador prof. dr. Luiz Henrique Gama Dore de Araújo, por sua dedicação, paciência e sabedoria ao me guiar ao longo deste processo. Sua orientação foi fundamental para que eu pudesse superar os desafios e alcançar este resultado. Agradeço também a todos os meus amigos discentes, incluindo Jorge, Rodrigo, Bruno, Lenyr, Jessy, Romário, Matheus e Tiago, companheiros de tantos momentos ao longo da graduação, além de dona Neide, por tornarem essa caminhada mais leve e repleta de aprendizado mútuo.

Não poderia deixar de agradecer a todos os docentes que ministraram aulas durante meu percurso acadêmico no Departamento de Estatística e Ciências Atuariais da Universidade Federal de Sergipe. Cada aula, cada ensinamento, contribuiu para minha formação e para a realização deste trabalho. Por fim, minha gratidão se estende a todos os meus familiares, cujo apoio e incentivo foram essenciais para que eu chegasse até aqui.

A todos vocês, meu mais sincero obrigado!

"Segue os teus sonhos, trabalhe muito que as oportunidades vão surgir"

(C.Ronaldo)

RESUMO

Este trabalho propõe a aplicação de técnicas de modelagem de regressão linear simples para três espécies de camarões no Estado do Rio de Janeiro (rosa, barba-ruça e sete-barbas). A pesquisa é desenvolvida através do método de mínimos desvios absolutos (LAD) no ajustamento de modelos de regressão linear, no qual tenta-se eliminar a influência de dados discrepantes fora dos eixos verticais e horizontais presentes no conjunto estudado. O estudo segue a abordagem delineada no artigo de Dodge (1997), pois é um modelo resistente à presença de *outliers.* Neste modelo, o método identificação e supressão de pontos extremos, são feitas sobre os resíduos estandardizados através do *software* R para análise estatística. Os resultados mostram que a exclusão de pontos extremos tem efeito de relevância sobre a veracidade das estimativas de coeficiente de regressão, o que reduz a variabilidade residual e melhora a robustez dos modelos. Porém, vale ressaltar que para a remoção dos *outliers* deverá existir um motivo relativamente esclarecedor, que justifique a remoção. Contudo, estabeleceu-se correlações entre as espécies, que mostraram fortes relações positivas, relações negativas moderadas em algum grau-variável de acordo com cada caso. Este estudo contribui para o entendimento de técnicas resistentes à presença de *outliers*, presentes no banco de dados das capturas de espécies de camarões, e provê resultados consideráveis sobre a técnica do artigo de Dodge (1997).

Palavras-Chaves: Modelagem de regressão linear, mínimos desvios absolutos, *outliers,* captura de camarão, Rio de Janeiro.

ABSTRACT

This work proposes the application of simple linear regression modeling techniques to three shrimp species in the State of Rio de Janeiro (seven-beard, pink, and whitebeard). The research is developed using the Least Absolute Deviations (LAD) method in fitting linear regression models, which aims to eliminate the influence of discrepant data outside the vertical and horizontal axes present in the studied dataset. The study follows the approach outlined in Dodge's (1997) article, as it is a model resistant to the presence of outliers. In this model, the identification and suppression of extreme points are performed on standardized residuals using the R software for statistical analysis. The results show that the exclusion of extreme points has a relevant effect on the accuracy of regression coefficient estimates, reducing residual variability and improving the robustness of the models. However, it is worth noting that the removal of outliers should be justified by a reasonably clear reason. Nevertheless, correlations were established between the species, which showed strong positive relationships and moderately negative relationships to varying degrees depending on each case. This study contributes to the understanding of techniques resistant to the presence of outliers in the database of shrimp species catches and provides significant results regarding the technique from Dodge's (1997) article.

Keywords: Linear regression modeling, least absolute deviations, outliers, shrimp catch, Rio de Janeiro.

LISTA DE ILUSTRAÇÃO

Figura 1: **Simulação de dados e uma regressão onde são apresentados diferentes tipos de *outliers*.** **18**

Figura 2: **Ilustração de um *outlier* vertical e horizontal.....** **21**

Figura 3: **Boxplot da variável preditora.....** **32**

Figura 4: **Boxplot das espécies de camarões rosa, barba-ruça e sete-barbas.....** **33**

Figura 5: **Gráfico de dispersão do logaritmo do camarão rosa e camarão ind.....** **34**

Figura 6: **Gráfico de dispersão do logaritmo do camarão barba-ruça e camarão ind.....** **35**

Figura 7: **Gráfico de dispersão do logaritmo camarão sete-barbas e camarão ind.....** **36**

Figura 8: **Gráfico de dispersão do camarão rosa ajustado ao modelo proposto no artigo.....** **37**

Figura 9: **Gráfico de dispersão do camarão barba-ruça ajustado ao modelo proposto no artigo.....** **39**

Figura 10: **Gráfico de dispersão do camarão sete-barbas ajustado ao modelo proposto no artigo.....** **41**

LISTA DE TABELAS

Tabela 1: Simulação dos resultados da figura 2..... 22

Tabela 2: Tabela descritiva com os valores de correlação..... 31

Tabela 3: Os parâmetros de MMQ, LAD e Dodge (1997) para a espécie de camarão rosa..... 38

Tabela 4: Os parâmetros de MMQ, LAD e Dodge (1997) para a espécie de camarão barba-ruça... 40

Tabela 5: Os parâmetros de MMQ, LAD e Dodge (1997) para a espécie de camarão sete-barbas..... 42

Tabela 6: Tabela de porcentagem comparativa dos modelos em relação as três espécies..... 42

SUMÁRIO

1. INTRODUÇÃO..... 12
2. OBJETIVOS..... 14
2.1 GERAL..... 14
2.2 ESPECÍFICOS..... 14
3. JUSTIFICATIVA..... 15
4. REVISÃO LITERÁRIA..... 16
4.1. REFERENCIAL TEÓRICO..... 16
4.1.1 REGRESSÃO LINEAR SIMPLES E O MÉTODO DOS MÍNIMOS QUADRADOS..... 16
4.2 *OUTLIERS* – PONTO FLUENTE, PONTO ABERRANTE, PONTO DE ALAVANCA E AS TERMINOLOGIAS DE *OUTLIERS* VERTICAIS E HORIZONTAIS..... 17
4.2.1 SIMULAÇÃO DE PONTO FLUENTE E PONTO ABERRANTES..... 18
4.2.2 O MÉTODO DOS MÍNIMOS QUADRADOS E SUA SENSIBILIDADE A VALORES ATÍPICOS..... 19
4.2.3 GRÁFICO DE DISPERSÃO – SIMULAÇÃO NA PRESENÇA DE *OUTLIER*..... 21
4.3 FORMAS DE CONTORNAR O PROBLEMA NA PRESENÇA DE *OUTLIERS*..... 22
5. METODOLOGIA..... 25
5.1 MÍNIMOS DESVIOS ABSOLUTO..... 25
5.1.1 MINIMIZAÇÃO DOS MÍNIMOS DESVIOS ABSOLUTOS..... 25
5.1.2 MEDIANA COMO UMA MEDIDA DE CENTRALIDADE ROBUSTA..... 26
5.1.3 RESÍDUOS DA REGRESSÃO E IDENTIFICAÇÃO DE *OUTLIERS* VERTICAIS..... 26
5.2. ETAPAS DO MÉTODO PROPOSTO NO ARTIGO DODGE (1997)..... 26
5.2.1 DESCRIÇÃO DO MÉTODO DODGE UTILIZANDO O ALGORITMO..... 28
5.3 CONJUNTO DE DADOS..... 29
5.4 PLATAFORMA COMPUTACIONAL..... 29
6. RESULTADOS..... 31
6.1 ANÁLISE EXPLORATÓRIA DE DADOS..... 31
6.1.1 BOXPLOTS DA VARIÁVEL PREDITORA E DAS TRÊS ESPÉCIES DE CAMARÕES..... 32

6.2 MODELAGEM DE REGRESSÃO..... 37
7. CONCLUSÕES..... 44
BIBLIOGRAFIA..... 46
ANEXO A..... 48

1 INTRODUÇÃO

Muitos ramos da ciência, tecnologia e do conhecimento utilizam a análise e interpretação de dados, a fim de auxiliar na tomada de decisões, assessorar gestão de recursos e desenvolver previsões a partir de padrões de informações. Esse conhecimento vem se expandindo, e a habilidade de extrair informações consideráveis sobre os dados tornou-se um dos pilares da ciência contemporânea (HAIR JR. et. al 2019).

Por conta desse fato, foram desenvolvidas novas técnicas estatísticas que auxiliam na análise e interpretação de dados. Os modelos de regressão ocupam uma posição de destaque, permitindo por exemplo, estudar dados empíricos, relações entre variáveis e previsões.

Contudo, a regressão linear simples é uma área da estatística baseada em um procedimento de 'ajuste de curva', utilizado para desenvolver um modelo matemático de uma relação procurada. Sendo assim, a regressão linear modela a relação entre uma variável dependente e uma ou mais variáveis independentes (MONTGOMERY, 2012).

Entretanto, o modelo de regressão, atualmente é empregado pelos analistas através do método dos mínimos quadrados, como técnica comumente aplicada para se calcular parâmetros. Esse método procura minimizar a soma dos quadrados dos desvios dos valores observados em relação aos preditores. No entanto, o método de minimização dos quadrados é sensível à presença de valores anômalos (*outliers*), pois distorce as estimativas dos parâmetros o que fornece resultados viesados (MONTGOMERY et. al, 2021).

Isso acontece porque, os desvios extremos nos valores observados elevam o erro quadrático e, assim, diminuem a precisão do modelo. Uma maneira de reduzir a influência dos *outliers* sobre a regressão de parâmetro linear é usar os métodos robustos, que é um dos mais conhecidos métodos de mínimos quadrados absolutos.

Diferente dos mínimos quadrados, cujo objetivo é minimizar a soma dos resíduos ao quadrado, o método de mínimos desvios absolutos alcança estimativas mínimas a partir de valores que minimizam sua soma dos valores absolutos dos erros. No entanto, essa abordagem torna o estimador robusto e resistente à presença de *outliers*, pois os desvios extremos não são amplificados exponencialmente, como no caso dos quadrados (ROUSSEEUW et. al, 2003).

Como resultado, por vezes o estimador de mínimos desvios absolutos é referido como estimador de desvios absolutos, particularmente notável por sua capacidade de contornar à invalidez de dados contaminados com informações não confiáveis.

Os modelos de regressão que possuem *outliers*, podem ser verticais ou horizontais, dependendo da raiz destes no grupo de dados. O método dos mínimos desvios absolutos (*Least Absolute Deviations* - LAD), previne contra *outliers* verticais, que surgem quando o resultado observado ocupa ampla distância dos demais resultados previstos. Por conseguinte, o estimador LAD é robusto para valores discrepantes verticais, já que minimiza os desvios absolutos, no que resulta em menos importância dada às observações anormais na variável dependente (ROUSSEEUW et. al, 2003).

Porém, também pode haver *outliers* horizontais, ligados às variáveis independentes, que adulteram a estrutura do modelo ao inserir valores extremos em dimensões explicativas. De forma semelhante, valores discrepantes horizontais se fazem necessários às vezes de determinações e tratamentos, além de danos que podem comprometer a estabilidade do modelo (HAIR JR et. al, 2019).

O presente trabalho tem como objetivo aplicar o método proposto por Dodge (1997) para modelar as capturas anuais de três espécies de camarão, apenas as observadas entre 1950 a 2015 e não as capturas estimadas (reconstruídas) no estado do Rio de Janeiro: camarão rosa, barba-ruça e sete-barbas. Para mais detalhes sobre os dados analisados no presente estudo, consultar Freire et. al. (2021).

.

2 OBJETIVOS

O presente capítulo tem como foco apresentar o objetivo geral e específico. E assim, analisar de maneira detalhada o tempo do estudo, técnicas aplicadas, dentre outros passos empregados na formulação deste trabalho acadêmico.

2.1 GERAL

Utilizar o método de Mínimos Desvios Absolutos (LAD) e do artigo Dodge (1997), para construir modelos de regressão linear resistentes a *outliers* das três espécies do banco de dados da captura de camarões no estado do Rio de Janeiro.

2.2 ESPECÍFICOS

1. Analisar a influência de *outliers* verticais e horizontais nos dados de capturas anuais das espécies de camarão (camarão sete-barbas, camarão rosa e camarão barba-ruça) e o impacto nas estimativas dos modelos de regressão;
2. Estimar os coeficientes de regressão linear ao utilizar o método LAD, e comparar os resultados com e sem a presença de *outliers*, para avaliar a robustez e precisão das previsões;
3. Identificar variações nas tendências de captura das três espécies de camarões ao longo do período analisado;
4. Identificar os padrões de correlação entre as capturas das espécies de camarões e a variável preditora (camarão_ind), ao destacar relações positivas, negativas ou independentes;
5. Explorar a aplicabilidade dos modelos robustos de mínimos desvios absolutos, pelo método proposto no artigo de Dodge, no banco de dados da pesca no estado do Rio de Janeiro;

3 JUSTIFICATIVA

O método dos mínimos quadrados é o mais comumente utilizado para a estimação de parâmetros em modelos estatísticos, em especial os de regressão linear. No entanto, um dos seus principais desafios é a sensibilidade a *outliers*, devido à minimização da soma dos quadrados dos erros que resulta em uma minimização excessiva dos resíduos ligados a outliers. Como resultado, as estimativas são calculadas de modo viesado e com tendências que não se mostram fidedigna na realidade, e por consequência, aumenta a incerteza associada à cada termo – o que se mostra inaceitáveis estatisticamente – e assim, acabam por prejudicar as premissas e a confiabilidade do modelo.

Além dos já citados, os *outliers* verticais são os valores extremos diretamente ligados à variável dependente, sendo possível existir *outliers* horizontais em alguns casos. Isto refere-se a valores extremos da variável independente, e podem prejudicar de modo grave a qualidade do ajuste. Diante disto, é necessário utilizar outras técnicas estatísticas que sejam menos afetadas pela presença de valores discrepantes. Alguns métodos mais resistentes à presença de outliers são: a regressão por mínimos quadrados ponderados, a regressão robusta e abordagens não paramétricas.

Sendo assim, a presente pesquisa se justifica pela necessidade de investigar a linha de métodos alternativos que permitam a construção de modelos estatísticos mais fidedignos, precisos e aplicáveis a conjuntos de dados sujeitos a outliers. Dessa forma, uma das propostas desenvolvidas assegurará a robustez das inferências retiradas de modelos de regressão quando forem aplicadas ao algoritmo desenvolvido por Dodge (1997), e assim, contornará os problemas dos valores discrepantes que cria tendências inverossímeis e distantes dos valores reais.

4. REVISÃO LITERÁRIA

O capítulo apresenta trabalhos relevantes no meio acadêmico relacionados ao modelo de regressão linear. A começar pelo modelo de regressão linear e sua estimação por mínimos quadrados, ressaltando as hipóteses e limitações do modelo.

Em seguida, os estimadores por mínimos quadrados que são sensíveis a *outliers* em sua perda de eficiência e os impactos dos *outliers* nas estimações dos parâmetros. Juntamente com uma simulação de *outliers*, além de formas de contornar o problema na presença de *outliers* que é o principal objetivo do trabalho.

4.1 REFERENCIAL TEÓRICO

Tem-se por base do trabalho a revisão teórica da regressão linear e versões resistentes para manusear *outliers*. E, do mesmo modo, alterar a atenção para o modelo de regressão linear simples e métodos dos mínimos quadrados. Primeiro, todos aqueles que incluem seus pressupostos limitações, em seguida, examinar a sensibilidade desta abordagem aos valores aberrantes e demonstrar como contornar a questão.

4.1.1 REGRESSÃO LINEAR SIMPLES E O MÉTODO DOS MÍNIMOS QUADRADOS

A regressão linear simples é um dos modelos estatísticos mais comuns para descrever a relação entre uma variável resposta e a variável explicativa. A maioria dos procedimentos inferenciais sobre a regressão, presume que os erros seguem uma distribuição normal, o que facilita a construção de testes de hipóteses e intervalos de confiança (MONTGOMERY et. al, 2012).

A formulação básica do modelo de regressão linear simples começa com essas premissas de distribuição normal, hipóteses e construção de intervalos de confiança. Embora haja dois coeficientes, cada um tem um significado particular. A inclinação deste coeficiente reflete o aumento médio da variável de resposta para um incremento de uma unidade na variável

preditora. Desse modo, os coeficientes são essenciais para a interpretação prática do modelo, ao traduzir matematicamente a relação entre as variáveis (DRAPER et. al, 1998).

Dessa forma, a clareza interpretativa é importante para a aplicação do modelo em contextos reais. Neste sentido, o coeficiente de intercepto demonstra que o valor médio previsto para a variável preditora tende a aproximar-se de zero. Contudo, na regressão linear o método de estimativa dos mínimos quadrados é comumente utilizado para encontrar estimativas de parâmetros. Sendo assim, o método minimiza os resíduos quadrados a diferença entre os valores observados e os valores ajustados do modelo. Por outro lado, os resíduos exercem um papel de relevância na avaliação de qualidade do ajuste, pois permitem identificar se o modelo está capturando adequadamente os padrões dos dados, além de verificar a presença de heterocedasticidade ou não linearidade e avaliar se os pressupostos do método dos mínimos quadrados, como a normalidade e a independência dos erros estão sendo atendidos.

Apesar da sua eficácia, o método dos mínimos quadrados é sensível à presença de *outliers*, e os valores atípicos podem distorcer os resultados, o que exige cuidados na análise diagnóstica de verificação das premissas do modelo, podendo exercer um efeito considerável nas estimativas dos coeficientes (HAIR JR, et. al, 2019). Por conseguinte, a análise diagnóstica pode ser realizada para identificar tais valores para verificar se as premissas do modelo são atendidas como a independência dos erros, homoscedasticidade e a distribuição normal da resposta. Dependendo dos resultados da análise diagnóstica, o modelo pode ser corrigido e melhorado.

4.2 *OUTLIERS* – PONTO FLUENTE, PONTO ABERRANTE, PONTO DE ALAVANCA E AS TERMINOLOGIAS DE *OUTLIERS* VERTICAIS E HORIZONTAIS

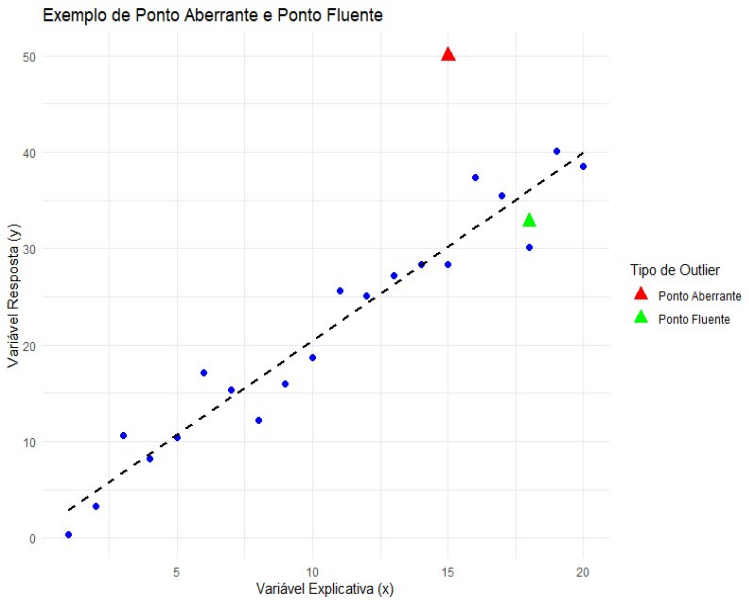
Os *outliers* são as observações que se afastam consideravelmente da tendência geral dos dados. Dependendo das características e do impacto na modelagem, é possível determinar diferentes tipos de *outliers*. Como exemplo, existem alguns tipos como os verticais, onde os dados têm amplitude em relação à variável de interesse, em outras palavras, são dados extremos, cujos valores estão fora da faixa desejada que segue a relação entre a variável dependente e independente. Nesse sentido, os *outliers* verticais são caracterizados por desvios anormais na variável resposta, enquanto os demais pontos seguem aproximadamente a tendência linear

esperada. Já os *outliers* horizontais são pontos de dados com valores que impactam a regressão linear como, por exemplo, os pontos de alta alavancagem (ROUSSEEUW et. al, 2003).

4.2.1 SIMULAÇÃO DE PONTO FLUENTE E PONTO ABERRANTE

Uma observação extrema na variável resposta que é desviante do padrão de dados é denominada ponto aberrante. O que pode influenciar na confiabilidade da estimativa e aumentar a variabilidade e a distorção das estatísticas da aritmética como médias, correlações ou apontar erros de medição, como demostrado na figura 1 a seguir.

Figura 1 – Simulação de dados e uma regressão onde são apresentados diferentes tipos de *outliers*



Fonte: Elaboração própria

Na figura 1, o ponto que não é fluente aos efeitos no fluxo do procedimento é denominado ponto fluente e extremo, mas conforme o padrão de dados não tem um impacto tão profundo no andamento da investigação, enquanto o ponto aberrante destorce a confiabilidade do modelo. Logo, o erro do ponto aberrante contribui para os enviesamentos dos coeficientes de regressão (ROUSSEEUW et. al, 2003).

Já os *outliers* horizontais são valores discrepantes em relação às variáveis explicativas, assim como, na variável resposta, o que deve ser retirados devido às suas influências no modelo. Os pontos de alavancagem são *outliers* horizontais, correspondentes a pontos que estão isolados na variável preditora. Desta forma, o isolamento destes pontos tem influência desproporcional sobre a reta de regressão. Pode acontecer de o ajuste aproximar-se da tendência do modelo, e apresentar um viés ou sair da tendência (ROUSSEEUW et. al, 2003).

4.2.2 O MÉTODO DOS MÍNIMOS QUADRADOS E SUA SENSIBILIDADE A VALORES ATÍPICOS

O método dos mínimos quadrados é uma técnica estatística que consiste em encontrar a melhor maneira de relacionar variáveis em um modelo de regressão. Várias soluções foram propostas no século XVIII- XIX, haja vista que, esse método foi inicialmente desenvolvido teoricamente, e com o passar do tempo a sua aplicação teórica foi designada em situações reais, o que na prática mostrou-se bem-sucedida. Um exemplo disso foi um livro publicado por Legendre (1805), *Nouvelles Méthodes pour la Détermination des Orbites des Comètes*, que uniu e generalizou várias técnicas conhecidas. Em todo caso, a descoberta dos mínimos quadrados é mais frequentemente atribuída a Gauss, que não somente foi o primeiro que demonstrou e aplicou a técnica em algum momento na década de 1790, mas também o documentou descrevendo detalhadamente o processo, publicado em 1809. Logo, o método dos mínimos quadrados permite determinar os elementos orbitais com base em observações, ao minimizar os erros de maneira sistemática (GAUSS, 1809).

A publicação, intitulada *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium*, consolidou sua contribuição, embora a prioridade tenha sido disputada com Legendre, que publicou uma versão anterior em 1805. Foi graças ao seu trabalho que se tornou claro que se usarmos mínimos quadrados para calcular meia – vida, então os erros cancelam-se mutuamente e obtém-se a melhor estimativa quanto possível. Dessa forma, o papel de Gauss nessa descoberta é importante porque a técnica tornou-se essencial para a prática científica da época. Deste modo, a aplicação do método dos mínimos quadrados por Gauss revolucionou a análise de dados astronômicos, e estabeleceu um padrão para a precisão científica no início do século XIX (ROBSON et. al, 2012).

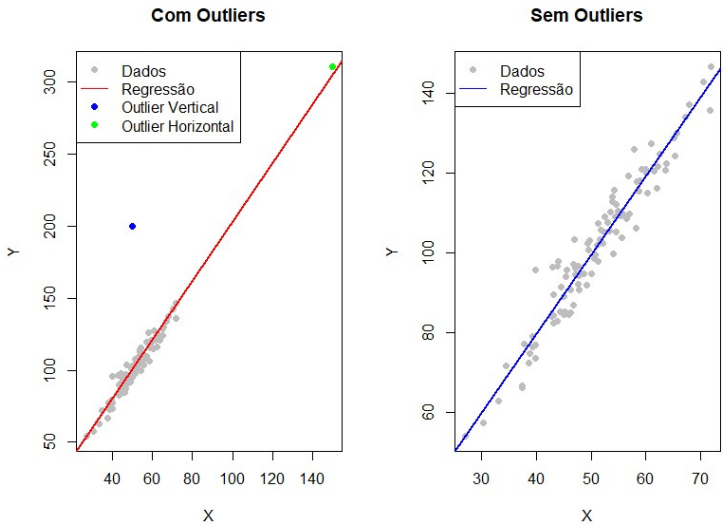
Em uma regressão linear, também utiliza o método para obter os coeficientes β_0, o qual determina onde a linha intercepta ao eixo vertical, e, β_1 a inclinação da linha. A chave deste método é traçar uma linha ou curva que caiba de maneira que a soma dos quadrados dos comprimentos das distâncias entre os verdadeiros valores e os valores que o modelo prediz, seja a menor possível. Assim, faz minimizar os erros entre o que observamos e que o modelo prediz, para obter uma melhor representação de como se relaciona as variáveis. O método dos mínimos quadrados fez parte de uma nova onda de aplicação de técnicas e estudos estatísticos às ciências sociais, com o mesmo grau de rigor que já vinha sendo aplicado em outras áreas científicas. Esse método foi influenciado pelo modelo científico como uma averiguação das hipóteses. Nesse sentido, o método dos mínimos quadrados, inicialmente aplicado a problemas astronômicos, tornou-se um marco na transição para uma abordagem quantitativa rigorosa que, no século XIX, começou a permear as ciências sociais (STIGLER, 1986).

O método científico foi criado por Galileu Galilei (1564-1642) e através destes princípios, foi desenvolvido o método dos mínimos quadrados. Assim, a técnica se mostrou fundamental para ajustar os modelos de dados experimentais, refletindo o compromisso do método científico com a precisão e a validação empírica (WOLBERG, 2006).

4.2.3 GRÁFICO DE DISPERSÃO – SIMULAÇÃO NA PRESENÇA DE *OUTLIER*

O gráfico de dispersão é apresentado com o objetivo de demonstrar a ausência e presença de *outliers* dos dados, com base na qual se identifica a presença de outliers verticais e horizontais. A seguir na figura 2, é ilustrado dois gráficos, sendo sucessivamente um com a presença de *outliers* e o outro sem a presença.

Figura 2: Ilustração de um *outlier* vertical e horizontal



Fonte: Elaboração própria

A comparação entre os pontos na reta, como aponta a dependência apresentada na figura 2, já que possuí um impacto considerável na dispersão dos pontos com as colorações verde e azul para *outliers* verticais e horizontais.

Depois de excluídos, a distribuição torna-se mais condensada no que pode ser interpretada como uma evidência de fato, que abafa o padrão visual de dependência entre as variáveis. Desta forma, o gráfico confirma a necessidade de identificar e tratar os *outliers*, sendo assim, a seguir na tabela 1 podemos verificar o quanto os dados são afetados.

Tabela 1: Simulação dos resultados da figura 2

Outliers	**Beta_0**	**Beta_1**
Com *Outliers*	-1.5383	2.0404
Sem *Outliers*	0.7977	1.9737

Fonte: Elaboração própria

Na tabela 1 que analisa os valores dos coeficientes beta 0 e beta 1 com e sem *outliers*, verifica-se que têm efeito significativo nos valores. Como consequência, os resultados apresentam uma flutuação considerável sem a presença dos *outliers*, porque os valores são afetados em comparativo com e sem *outliers*. Em conclusão, estes pontos distorcem a natureza dos padrões estatísticos. Portanto, sem a remoção dos *outliers* é difícil obter um nível normal de fixação dos valores em relação à reta no gráfico de dispersão.

4.3 FORMAS DE CONTORNAR O PROBLEMA NA PRESENÇA DE *OUTLIERS*

A maioria das técnicas de detecção de *outliers* foram desenvolvidas para contextos de aplicação específicos, e outras técnicas não são suficientemente genéricas para serem eficazes em qualquer tipo de cenário. Os métodos de detecção de *outliers* são frequentemente projetados para atender às características particulares de um domínio, como a distribuição dos dados ou o tipo de anomalia esperado (HAN et. al, 2011).

Para categorizar tais técnicas, diversos fatores foram levados em considerações, pois influenciam as escolhas e as aplicações delas. Segundo Chandola, Banerjee e Kumar (2009), a seleção de um método depende de aspectos como a definição do que constitui uma anomalia, o tipo de dados disponíveis e os recursos computacionais acessíveis. Dentre os fatores a serem considerados, podem ser citados: o tipo de *outlier* a ser identificado, a dimensionalidade dos dados, a natureza das informações, a disponibilidade de rótulos, e a necessidade ou não de qualquer tipo de parâmetros a serem definidos para o funcionamento do método.

Várias técnicas de detecção de anomalias foram desenvolvidas para serem aplicadas nos mais diversos campos, a fim de resolver os problemas relacionados às características e circunstâncias da coleta de dados. Ao contrário da classificação, em que os dados são

previamente caracterizados, as principais abordagens de detecção de anomalias são as baseadas em modelos estatísticos, onde parte-se do pressuposto de que os dados estão normalmente distribuídos, e assim, qualquer objeto que não se adeque a esse modelo é uma anomalia. A outra abordagem, são os métodos baseados em proximidade, ao classificar um objeto como anômalo, haverá um afastamento. Assim, deste modo, são subdivididos com base na distância e baseados em densidade. Há os métodos baseados em foco, em que um objeto é considerado normal se fizer parte de um grande grupo chamado *cluster*, caso contrário, é anômalo ou não possui classificação definida, como afirma Han et. al (2012).

A eliminação do *outlier*, baseada na própria exclusão ou em transformações matemáticas, tem como o objetivo de diminuir o impacto desse tipo de observação. Um dos principais impactos conhecidos são os de logaritmos e raiz quadrada, por substituição do *outlier* por valores considerados mais adequados como média e mediana. Pode-se considerar a utilização de técnicas menos sensíveis a valores discrepantes como a regressão robusta, que minimiza o impacto no modelo. Desta forma, ajustam-se aos dados de forma a reduzir a influência de valores extremos sem eliminá-los completamente (HAMPEL et. al, 1986).

A remoção de *outliers* deve ser realizada com moderação, que visa garantir que não ocorra a perda de informações importantes. Os valores discrepantes não devem ser ignorados sem discricionariedade, e métodos alternativos como transformações matemáticas e imputação de dados mais eficazes. Além disso, métodos alternativos como regressão robusta ou algoritmos menos suscetíveis a *outliers*, podem ser empregados para melhor apresentação de informação. De qualquer maneira, alcançar a exclusão dos dados é um processo arriscado e a abordagem deve ser determinada pela natureza específica dos dados e da pesquisa. No entanto, a decisão de remover *outliers* deve ser fundamentada em uma análise cuidadosa do contexto, pois podem conter informações críticas sobre o fenômeno estudado, conforme Aggarwal (2017).

É indubitável que para contornar o problema na presença de *outliers* verticais e horizontais, Rousseeuw (1984), propõe uma metodologia baseada no estimador de Mínimos Quadrados Truncados (*Least Trimmed Squares - LTS*) que permite identificar e eliminar tanto *outliers* verticais quanto horizontais, o que ajusta o modelo com menos impacto da influência desses valores extremos.

Sendo assim, *outliers* devem ser removidos após uma análise cuidadosa de valores errados que não podem ser corrigidos ou que pertence a uma população diferente, sendo investigada em comparação com o tipo de estudo, a distribuição dos dados e a redução do nível

de realismo do impacto do *outlier*. Acima de tudo, a validade dos resultados que serão adquiridos deve ser garantida de forma a impedir que o método comprometa a qualidade da análise, o que geralmente é causado por erro humano, bem como por erros durante a coleta da amostra, gravação ou mesmo entrada. Sendo o mais indicado o tratamento a fim de que não haja interferência negativa em relação à acurácia.

5. METODOLOGIA

O presente capítulo descreve a metodologia que o estudo engajou, com foco nos procedimentos robustos para eliminar possíveis *outlier*, e descreve também o conjunto de dados utilizado.

5.1 MÍNIMOS DESVIOS ABSOLUTOS

O método dos mínimos desvios absolutos, é uma técnica de regressão que minimiza a soma dos valores absolutos das diferenças entre os valores observados e os previstos (BOSCOVICH, 1757), sendo:

$$S = \sum_{i=1}^{n} |y_i - (\beta_0 + \beta_1 x_i)|$$

onde busca minimizar a soma dos desvios absolutos, em que y_i, é o valor observado e $|y_i - (\beta_0 + \beta_1 x_i)|$ é o desvio absoluto entre o valor observado e o valor previsto. Isso faz com que o modelo seja mais robusto a *outliers* (BOSCOVICH, 1757).

5.1.1 MINIMIZAÇÃO DOS MÍNIMOS DESVIOS ABSOLUTOS

A soma dos desvios absolutos entre os valores observados y_i e os valores ajustados pelo modelo $\beta_0 - \beta_1 x_i$, é considerado um critério de ajuste que pode ser utilizado para minimizar o erro quadrático, sendo a SDA o erro absoluto, o que a torna mais robusta a *outliers*.

$$SDA(\beta_0, \beta_1) = \sum_{i=1}^{n} |y_i - \beta_0 - \beta_1 x_i|$$

5.1.2 MEDIANA COMO UMA MEDIDA DE CENTRALIDADE ROBUSTA

$$MED_x minimiza \sum_{i=1}^{n} |x_i - M|$$

A mediana minimiza a soma dos desvios absolutos dos dados x_i em relação a M, o que se mostra mais robusta em relação a *outliers*.

5.1.3 RESÍDUOS DA REGRESSÃO E IDENTIFICAÇÃO DE *OUTLIERS* VERTICAIS

A formula 1 está codificada ao erro da regressão, isto é a diferença entre o valor real observado y_i e aquele que foi predito com o modelo $\widehat{\beta}_0 + \widehat{\beta}_1 \, x_i$ (DODGE, 1997), o que sugere à seguinte conclusão:

$$r_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \quad (1)$$

$$Se|\frac{r_i}{\hat{\sigma}}| \geq 2{,}5; \text{Então é um } outlier \text{ vertical}$$

Se o resíduo padronizado ultrapassa 2.5 desvios padrão, esse ponto pode ser considerado um *outlier* (DODGE, 1997).

5.2 ETAPAS DO MÉTODO PROPOSTO NO ARTIGO DODGE (1997)

Na primeira etapa Dodge (1997), instrui a aplicar o método LAD para ajustar o modelo $y_i = \hat{\theta}_0 + \hat{\theta}_1 x_i$ usando todas as n observações disponíveis. Os parâmetros $\hat{\theta}_0 \, e \, \hat{\theta}_1$ são estimadores que minimizam:

$$\sum_{i=1}^{n} |y_i - \hat{\theta}_0 - \hat{\theta}_1 x_i|$$

Além disso, calcula-se uma estimativa robusta da escala dos resíduos, definida como $\hat{\sigma} = 1.4826 \text{ x MAD}$, onde MAD (*Median Absolute Deviation*) é a mediana dos valores absolutos dos resíduos não nulos de $|r_i| = |y_i - \hat{y}_i|$. O fator 1.4826 x MAD garante consistência com a variância de uma distribuição normal, caso os erros sejam gaussianos (DODGE, 1997).

Para a segunda etapa, identifica-se e remove *outliers* verticais, então para cada observação i, calcula-se o resíduo padronizado $\frac{r_i}{\hat{\sigma}}$, onde $r_i = y_i - \hat{y}_i$, é o resíduo bruto. Observações com $|\frac{r_i}{\hat{\sigma}}| \geq 2{,}5$, são classificadas como *outliers* verticais e removidas do conjunto de dados.

Na terceira etapa, aplica-se regressão inversa de x sobre y com as observações restantes, com o conjunto de dados reduzido após a remoção dos *outliers* verticais. Inverte-se os papéis das variáveis, nesse modelo, x é tratado como a variável de resposta e y como a variável explicativa. Ajusta-se o modelo $x_i = \hat{n}_0 + \hat{n}_1 y_i$, pelo LAD para minimizar:

$$\sum_{i=1}^{n} |x_i = \hat{n}_0 + \hat{n}_1 y_i|$$

Calcula-se novamente $\hat{\sigma}$ como na primeira etapa, mas com os novos resíduos da regressão inversa.

Na quarta etapa, identifica e remove os pontos de alavancagem, para as observações restantes, computam-se os resíduos padronizados $\frac{r_i}{\hat{\sigma}}$ da regressão inversa, onde $r_i = \hat{x}_i$. As observações com $|\frac{r_i}{\hat{\sigma}}| \geq 2{,}5$, então são consideradas pontos de alavancagem no modelo original, pois na regressão inversa esses valores extremos em x_i tornam-se *outliers* verticais e são removidas (DODGE, 1997).

Na última etapa, com o conjunto de dados limpo, livre de *outliers* verticais e pontos de alavancagem, aplica-se o LAD novamente para ajustar o modelo final $y_i = \hat{\theta}_0 + \hat{\theta}_1 x_i$ usando as observações remanescentes, em que fornece a equação definitiva do modelo robusto.

5.2.1 DESCRIÇÃO DO MÉTODO DODGE (1997) E A UTILIZAÇÃO DO ALGORITMO

O artigo propõe um algoritmo para detectar *outliers* ao longo da variável resposta e das variáveis explicativas por meio da regressão LAD. Mais especificamente, o fato de o LAD ser mais robusto a *outliers* na variável resposta, mas ainda pode ser sensível nas variáveis explicativas, especialmente nos pontos de alavancagem (DODGE, 1997). Dessa forma, a proposta da metodologia é alterar o papel das variáveis respostas e explicativas. Isto é, o algoritmo é proposto em cinco etapas principais.

1. Na primeira etapa, utiliza-se um modelo de regressão LAD por meio da variável resposta original y e as variáveis explicativas x.
2. Em seguida, na segunda etapa, calcula os resíduos padronizados para cortar o ponto na variável resposta, isto é, define o ponto como um *outlier* se o resíduo padronizado é igual ou superior a 2.5 pontos em valores absolutos (DODGE, 1997).
3. Terceiro passo, remover os pontos classificados como *outliers* na variável resposta.
4. Quarto passo, aplicar uma regressão LAD com os papeis trocados, isto é, tratar cada variável explicativa x_i como a variável resposta e das outras variáveis, incluindo y, que são preditores. Após esse processo, calcula-se para cada variável x_i, que são resíduos padronizados para definir os *outliers* para as variáveis explicativas.
5. Quinto passo a ser seguido para a utilização do algoritmo é remover os pontos classificados como *outliers* nas variáveis x_i. Por conseguinte, aplicar um modelo de regressão LAD final, aplica-se apenas os pontos que não for identificado como *outlier* em nenhuma das etapas anteriores e com a resposta original, variável preditora original e com a resposta e a variável preditora de papeis trocados.

5.3 CONJUNTO DE DADOS

O banco de dados a ser analisado compreende a captura de camarão de diferentes espécies no estado do Rio de Janeiro durante diversos anos. Nesta base, estão contidos dados acerca do camarão, tais como o camarão rosa, camarão barba-ruça e o camarão sete-barbas, quanto à produção anual do sistema pesqueiro do Rio de Janeiro.

A Base de dados utilizada para este trabalho contém os valores observados das capturas anuais de três espécies de camarões no estado do Rio de Janeiro, cobrindo os anos de 1950 a 2015 para o setor industrial. As coletas são para camarão rosa, camarão barba-ruça e camarão sete-barbas. Houve ainda uma variável preditiva geral chamada camarão_ind, que representa um índice de captura para camarão em geral. Para mais detalhes sobre os dados analisados no presente estudo, consultar Freire et. al. (2021).

Os dados foram organizados num formato tabular com colunas contendo captura anual para cada espécie, onde expressou unidades e toneladas que poderiam ser submetidas a análise estatística e armazenadas em um arquivo CSV (camaroes_ind_RJ.csv), como indicado em scripts R divulgado no anexo A (página 46). A base possui registros cobrindo décadas de atividade pesqueira, permitindo elucidar tendências transitórias bem como aplicar modelos preditivos. Durante o processamento, observações com valores ausentes foram tratadas por exclusão (usando o método na.exclude no R) para assegurar a integridade de análises posteriores.

Escolheu-se essa base de dados porque as espécies analisadas são economicamente relevantes, além de ser importante entender os padrões de captura a fim de informar políticas de gestão sustentável. A presença dos potenciais *outliers*, devido a variações sazonais, erros de medição ou condições ambientais extremas foi um fator considerado no pré-processamento e modelagem, cumprindo os objetivos de robustez estabelecidos no estudo.

5.4 PLATAFORMA COMPUTACIONAL

Para aplicação deste algoritmo, foi utilizada a plataforma computacional *RStudio version* 2023.12.1+402 *(Ocean Storm)*, um ambiente integrado de desenvolvimento (IDE) para a linguagem R, amplamente reconhecido pela capacidade de análise estatística e

disponibilidade de pacotes robustos (POSIT TEAM, 2024). O *RStudio* foi escolhido para executar essas etapas devido à sua facilidade de uso e a existência de funções específicas para LAD, como as disponíveis no pacote *L1pack* (OSORIO; WOLODZKO, 2024). No *script* desenvolvido, o pacote *L1pack* foi empregado para realizar os ajustes LAD, o que permite a aplicação prática do algoritmo em dados reais. Especificamente, o *script* processou o conjunto de dados camaroes_ind_RJ.csv, contendo variáveis relacionadas às diferentes espécies de camarões, como cam_sete_barbas_ind, cam_rosa_ind e cam_barba_ruca_ind, em relação à variável preditora camarao_ind. Através do *RStudio*, foi possível carregar os dados, calcular correlações, ajustar modelos LAD, identificar *outlier*s verticais e pontos de alavancagens, e visualizar os resultados graficamente, replicando as etapas do algoritmo fornecido pelo artigo de Dodge(1997).

6. RESULTADOS

Neste capítulo é apresentado as figuras e tabelas no qual os *outliers* são visualizados com base em gráficos de dispersão e coeficientes de correlação. Além de descrever graficamente o comportamento dos *outliers*, será contabilizada a seguir pela análise de regressão, na qual a estimativa do coeficiente é comparada com duas estimativas que também são derivadas e feitas a partir de diferentes métodos, e finalmente, a análise descritiva adicional que aborda os resultados, baseado na técnica aplicada pelo algoritmo abordado pela pesquisa.

6.1 ANÁLISE EXPLORATÓRIA DE DADOS

Uma análise das correlações entre diferentes espécies de camarões tem como objetivo de explorar as interações e possíveis influências entre elas. Serão examinadas as correlações positivas ou negativas, entre espécies como camarão_ind, camarão barba-ruça, camarão sete-barbas e camarão rosa. Assim é ilustrado na tabela 2 a seguir.

Tabela 2: Tabela descritiva com os valores de correlação

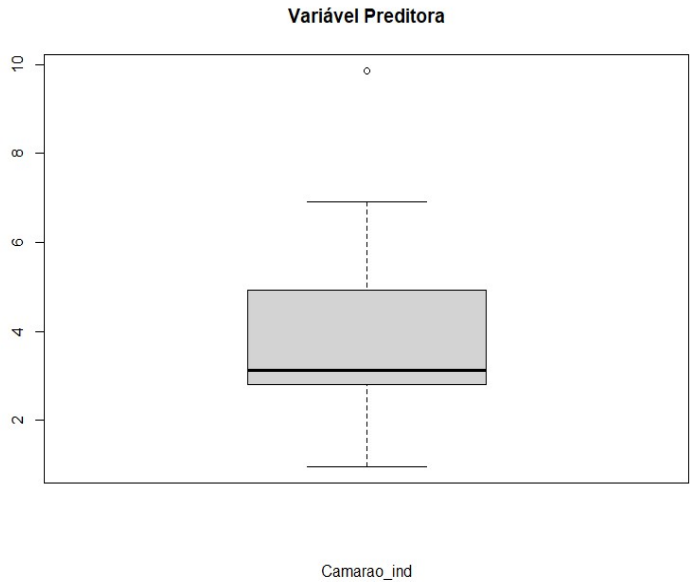
Espécie preditora: Camarão_ind > Comparativo	**Correlação**	**Interpretação**
Cam_barba_ruca_ind	0.7430	Correlação positiva forte
Cam_sete_barbas_ind	0.7164	Correlação positiva forte
Cam_rosa_ind	-0.6104	Correlação negativa moderada

Fonte: Elaboração própria

A forte correlação positiva na tabela 2 está acima de 0.7 para camarão_ind na espécie de camarão barba-ruça, sendo a correlação de 0.7430 entre camarão_ind e camarão sete-barbas. Em oposição, há uma moderada correlação negativa -0.6104 entre camarão rosa e camarão_ind, indicando que o aumento de uma das espécies de camarões, podendo influenciar na diminuição de outra espécie de camarão.

6.1.1 BOXPLOTS DA VARIÁVEL PREDITORA E DAS 3 ESPÉCIES DE CAMARÕES

Figura 3: Boxplot da variável preditora

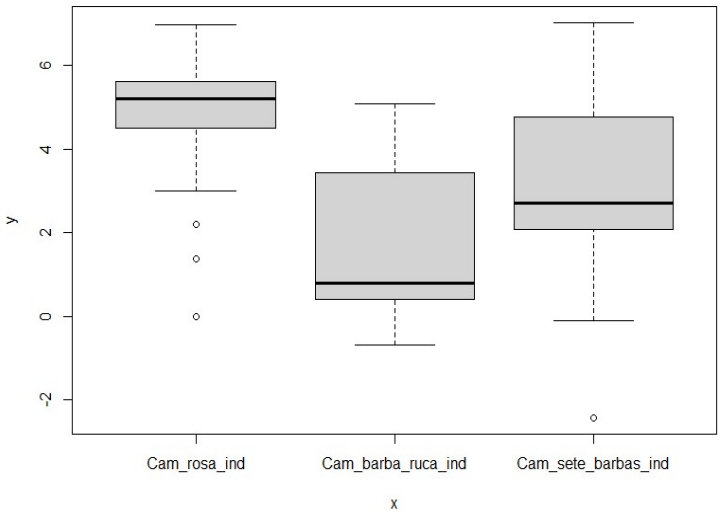


Fonte: Elaboração própria

Na figura 3 a variável preditora o total da espécie camarão_ind, observa-se que a mediana está próxima ao quartil inferior, pois, um outlier é perceptivo pelo ponto acima do limite superior.

Assim, sugere um ponto de dados atípico, não de acordo com o que é esperado do restante da distribuição. Isso pode ser interpretado como um alto valor de camarão, percebido em uma única observação, possivelmente dada a condições ambientais, sazonais ou erro na coleta de dados.

Figura 4: Boxplot das espécies de camarões rosa, barba-ruça e sete-barbas

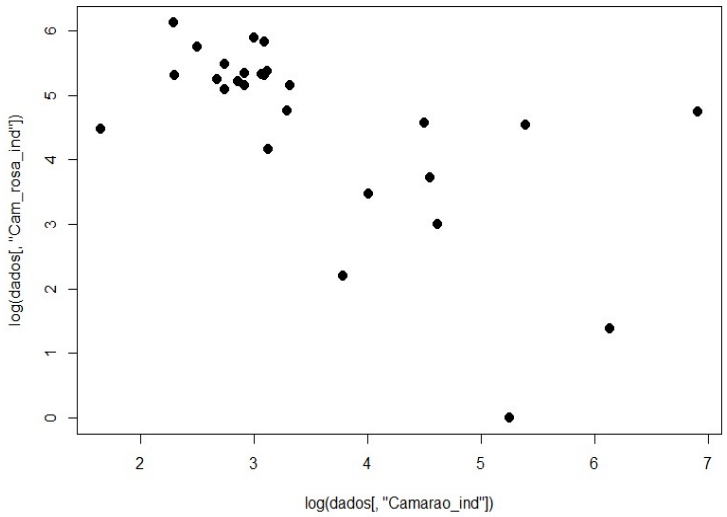


Fonte: Elaboração própria

Na figura 4 os logs estão distribuídos no total observado das diferentes espécies de camarão, pode-se observar que o camarão rosa possui maior mediana, o que pressupõe que a maioria dos valores desse registro será significativo. Por outro lado, o camarão sete-barbas tem desvio-padrão mais próximo de zero, o que significa que os valores oscilam entorno da mediana.

Os pontos mais dispersos são camarão sete-barbas e camarão barba-ruça, o que implica uma alta propagação dos dados, sendo que os três *outliers* para camarão rosa, podem influenciar no comportamento do próprio conjunto de dados.

Figura 5: Gráfico de dispersão do logaritmo do camarão rosa e camarão ind

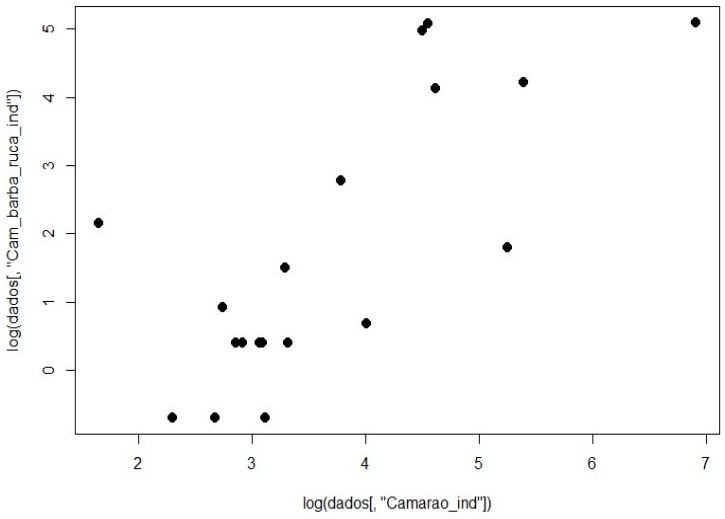


Fonte: Elaboração própria

Na figura 6 o padrão exibido pelos pontos é o declínio da dispersão, uma queda que ocorre devido à variabilidade constante, em razão dos pontos expandidos.

Contudo, a expansão da maioria dos pontos demonstra a tendência com concentração com pontos dentro e fora da tendência linear. Enquanto, alguns pontos estão dispersos, o que revela uma tendência decrescente, sendo os *outliers* horizontais e verticais.

Figura 6: Gráfico de dispersão do logaritmo do camarão barba-ruça e camarão ind

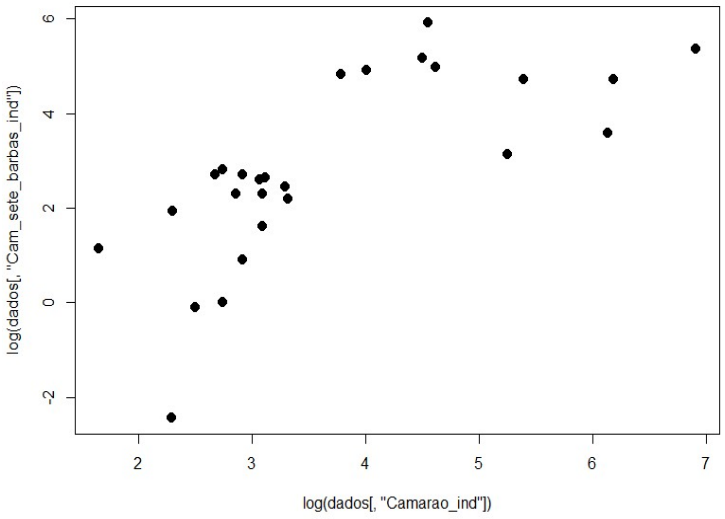


Fonte: Elaboração própria

Na figura 5 do camarão barba-ruça, ocorre uma alta dispersão dos dados, devido à variabilidade constante, observa-se uma tendência de crescimento de alguns pontos, outros com razão linear e os valores de barba-ruça, o que se concentra mais entre 2 e 4 nas escalas em toneladas y e x logarítmica.

Dessa forma, fatores que apresentam *outliers*, seja em padrões horizontais ou verticais, possuem valores fora do conjunto de pontos concentrados.

Figura 7: Gráfico de dispersão do logaritmo camarão sete-barbas e camarão ind



Fonte: Elaboração própria

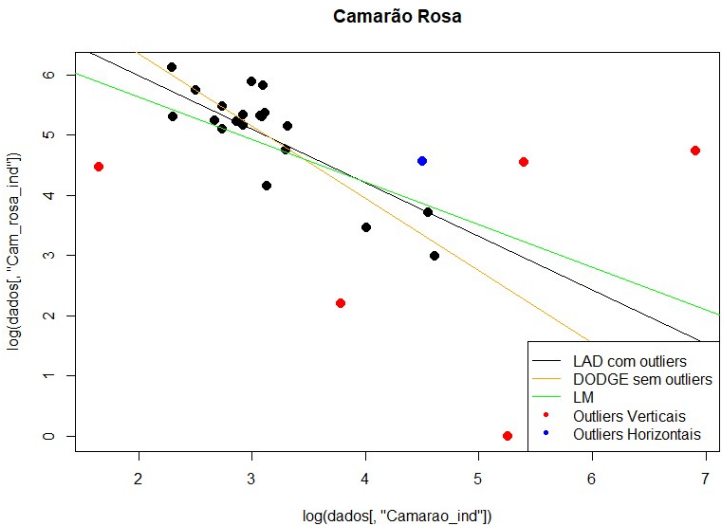
Na figura 7, cabe observar que no gráfico de dispersão do camarão sete-barbas, a dispersão dos outliers verticais e horizontais não é tão discrepante quanto na situação representada pelo gráfico 5 do camarão barba-ruça.

Entretanto, aproxima-se do cenário apresentado pelo gráfico de dispersão do camarão rosa. Porém, o comportamento do gráfico demonstra uma relação positiva.

6.2 MODELAGEM DE REGRESSÃO

As próximas figuras ilustram a dispersão do camarão rosa, barba-ruça e sete-barbas em relação ao modelo de regressão proposto no artigo DODGE (1997) para ajustar os dados. Os gráficos buscam representar visualmente a relação entre as variáveis analisadas, considerando diferentes categorias de dados e suas respectivas distribuições nas retas, como demostra na figura 8 a seguir.

Figura 8: Gráfico de dispersão do camarão rosa ajustado ao modelo proposto no artigo



Fonte: Elaboração própria

Na figura 8 apresenta a dispersão do camarão rosa no eixo y em escala logarítmica em função da variável preditora no eixo x, também em escala logarítmica com pontos azuis os horizontais e vermelhos os verticais. São exibidas três retas de regressão e são ajustadas nas cores preta com dados completos o que inclui os outliers na cor verde por mínimos quadrados, e, laranja no modelo proposto sem outliers.

As retas indicam uma tendência negativa sugerindo que o total de camarão rosa diminui ao longo do tempo. A reta laranja é resistente a *outliers,* mostra uma inclinação menos acentuada comparada às demais, o que indica a exclusão de *outliers* e suaviza a relação inversa, e sugere menor influência sobre a variável preditora. O destaque é a reta laranja – proposto pelo

método de Dodge (1997) – que segue um percurso oposto as demais retas dos modelos, devido à presença de *outliers.* A seguir segue a tabela 3 com os parâmetros.

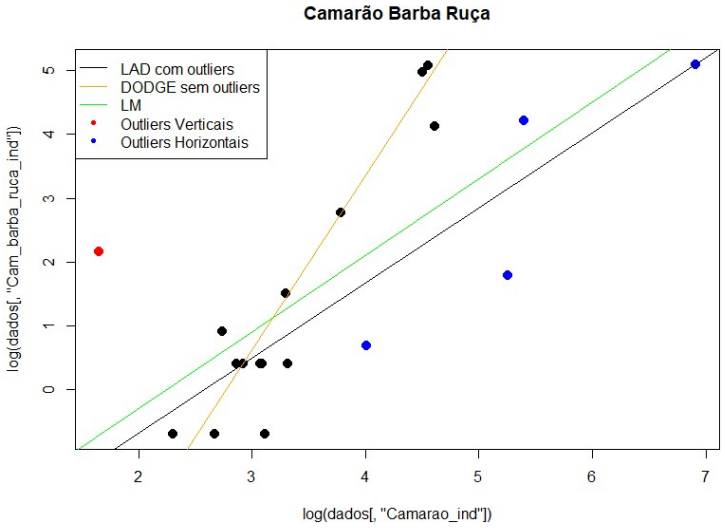
Tabela 3: Os parâmetros de MMQ, LAD e Dodge (1997) para a espécie de camarão rosa

COEFICIENTES	MMQ (Mínimos quadrados)	LAD (Mínimos desvios absolutos)	DODGE (1997)
Beta_0	7.0484	7.7624	8.7531
Beta_1	-0.7067	-0.8891	-1.1998

Fonte: Elaboração própria

A tabela 3 apresenta os parâmetros dos modelos ajustados para o camarão rosa com valores de Beta 0 e Beta 1 para mínimos quadrados sendo igual a 7.0484 e -0.7067 para LAD demonstra o resultado 7.7624 e -0.8891. Já para o método de Dodge as informações de Beta 1 e Beta 0 respectivamente são 8.7531 e -1.1998. Ao comparar os modelos, o Dodge exibe o maior Beta 0 de 8.7531, contra 7.0484 do MMQ e LAD é 7.7624, o que indica um intercepto mais elevado. Já o Beta 1 e Doge obtém um resultado igual a -1.1998, o que possui valor absoluto maior que o MMQ representa -0.7067 e LAD é de -0.8891, o que sugere uma relação mais intensa. Os resultados do modelo Dodge sem *outliers* amplifica de forma resistente aos valores discrepantes.

Figura 9: Gráfico de dispersão do camarão barba-ruça ajustado ao modelo proposto no artigo



Fonte: Elaboração própria

Na figura 9 é apresentado pontos azuis horizontais e vermelhos verticais calculados em escalas logarítmicas. Da mesma forma, é demonstrado a dispersão do camarão barba-ruça no eixo y em escala logarítmica em função da variável preditora no eixo x.

As três retas de regressão são exibidas: a reta de cor preta é ajustada com dados completos, no qual inclui *outliers*; já a reta de cor verde é demonstrada com o método dos mínimos quadrados; e por fim, a reta laranja é calculado pelo algoritmo Dodge (1997), o qual fornece um modelo resistentes à presença de *outliers*. A reta laranja sem *outliers*, exibe inclinação mais acentuada indicando maior sensibilidade à variação em relação a variável preditora.

Isso sugere que ao excluir *outliers*, a relação entre as variáveis se torna mais clara e pronunciada. Em suma, os valores dobram a tendência quando comparado com a linha na cor preta do ajuste com os dados completos – em consideração aos valores discrepantes – a reta dispara o que torna a angulação da reta diferente das demais. Isso fornece uma inclinação distinta, e resulta na demonstração de como os *outliers* impactam na tendência das retas. A seguir segue a tabela 4 com os parâmetros.

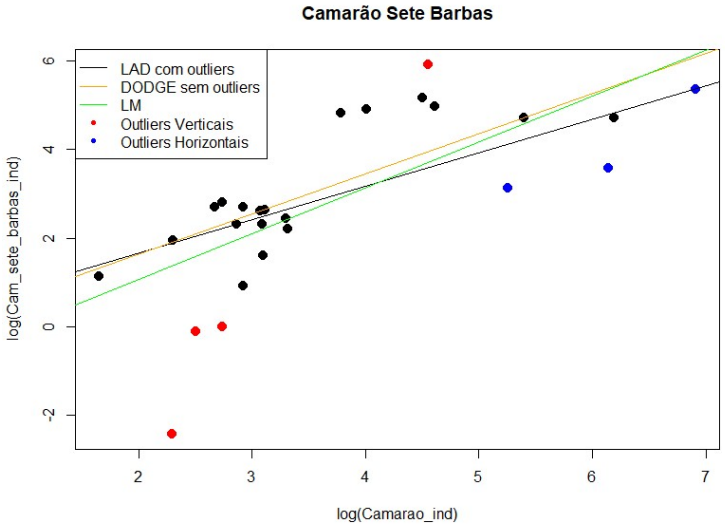
Tabela 4: Os parâmetros de MMQ, LAD e Dodge (1997) para a espécie de camarão barba-ruça

COEFICIENTES	MMQ (Mínimos quadrados)	LAD (Mínimos desvios absolutos)	DODGE (1997)
Beta_0	-2.6842	-3.0195	-7.5661
Beta_1	1.1973	1.1738	2.7321

Fonte: Elaboração própria

Em comparativo com tabela 3 do camarão rosa, o algoritmo Dodge (1997) é apresentado nessa tabela com o resultado significativo. O que revela que na tabela 4 houve um aumento considerável no valor do Beta 1 ao utilizar o modelo proposto pelo artigo. Os parâmetros dos modelos ajustados para o camarão barba-ruça com Beta 0 e Beta 1 para os MMQ é igual a -2.6842 e 1.973, já para o LAD é retornado o valor de -3.0195 e 1.1738. E, para o algoritmo Dodge (1997), desempenha com o resultado de -7.5661 e 2.7321 sucessivamente. Comparado com os modelos, a técnica do artigo apresenta o menor Beta 0 relativos ao valor de -7.5661, contra -2.6842 do MMQ, e, -3.0195 do LAD. Conforme visto, os resultados indicam um intercepto mais baixo, enquanto o Beta 1 do Dodge (1997) é 2.7321 é maior que 1.1973 do MMQ, com o valor de 1.738 do LAD.

Figura 10: Gráfico de dispersão do camarão sete-barbas ajustado ao modelo proposto no artigo



Fonte: Elaboração própria

Na figura 10 a dispersão dos pontos azuis horizontais e vermelhos verticais, como também das três retas de regressão do camarão sete-barbas pode ser vista no eixo y, onde foram calculados em escala logarítmica em função da variável preditora como base no eixo x.

As retas mostram uma tendência positiva, ao indicar que os pontos de camarão sete-barbas aumentam consideravelmente. Quando analisado especificamente, a reta laranja resistentes a *outliers*, apresenta uma inclinação acentuada sobre a reta dos mínimos quadrados, e sugere que a exclusão de *outliers* intensifica a relação entre as duas retas, o que se mostra discrepante sobre a reta preta traçada formalmente sobre o conjunto de dados geral, sem a exclusão.

Comparado ao gráfico 9 do camarão barba-ruça que também mostrou tendência positiva, mas discrepante, chegando a uma inclinação alta na reta laranja. Porém, como demonstrado no gráfico 8 do camarão rosa, possui tendências semelhantes, mas de maneira negativa com o declínio. A seguir segue a tabela 5 com os parâmetros.

Tabela 5: Os parâmetros de MMQ, LAD e Dodge (1997) para a espécie de camarão sete-barbas

COEFICIENTES	MMQ (Mínimos quadrados)	LAD (Mínimos desvios absolutos)	DODGE (1997)
Beta_0	-1.0065	0.1407	-0.4140
Beta_1	1.0349	0.7553	0.9491

Fonte: Elaboração própria

Em comparativos com as outras duas espécies, o barba-ruça foi o que mais obteve um número elevado. Já os resultados do sete-barbas, se assemelha com a tabela 3, com tendências próximas nos três modelos, e, dessa forma, a tabela 5 detalha os parâmetros dos modelos ajustados com Beta 0 e Beta 1 para MMQ obtendo valores de -1.0065 e 1.0349, para o LAD é igual a 0.1407 e 0.7553 e para a técnica do artigo resulta em -0.4140 e 0.9491. Quando comparado com os modelos de Dodge (1997), o Beta 0 é igual a -0.4140 sendo maior que -1.0065 em comparação com MMQ, mas para o resultado do LAD, o modelo desempenha o valor de -0.1407, enquanto o Beta 1 do Dodge (1997) o valor é 0.9491, e menor que 1.0349 do MMQ, porém o MMQ é maior que 0.7553 em relação ao LAD, e indica uma inclinação intermediária. Logo a seguir, é exibido a tabela 6 de porcentagem comparativa dos modelos em relação as três espécies.

Tabela 6: Tabela de porcentagem comparativa dos modelos em relação as três espécies

Espécie	Coeficientes	Dodge x MMQ (Porcentagem)	Dodge x LAD (Porcentagem)
Camarão rosa	Beta_0	+24.18% (Aumento)	+12.78% (Aumento)
	Beta_1	-69.81% (Redução)	-34.95% (Redução)
Camarão barba-ruça	Beta_0	-181.87% (Redução)	-150.59% (Redução)
	Beta_1	+128.23% (Aumento)	+132.78% (Aumento)
Camarão sete-barbas	Beta_0	-58.87% (Redução)	-394.24% (Redução)
	Beta_1	-8.30% (Redução)	+25.68% (Aumento)

Fonte: Elaboração própria

Na tabela 6, os percentuais resultantes dos Dodge (1997) sobre às variações relativas aos mínimos quadrados (MMQ) e mínimos desvios absolutos (LAD), obteve um resultado percentual elevado no camarão barba-ruça, em relação às outras duas espécies. O camarão barba-ruça, o Beta 0 passa a ser menor com -181,87% para MMQ e -150,59% para LAD. Por outro lado, o Beta 1 atingiu valores maiores, e como resultado dobrou em +128,23% em comparativo ao mínimos quadrados (MMQ) e +132,78% para LAD.

7. CONCLUSÕES

Este trabalho buscou aplicar métodos resistentes a outliers ao modelar as capturas anuais de três espécies de camarão no estado do Rio de Janeiro entre 1950 e 2015, resistindo a observações extremas. Os resultados do método de desvios mínimos absolutos (LAD) e da abordagem de Dodge (1997) mostrou que remover dados discrepantes de modo justificado impacta significativamente as estimativas muito importantes para reconstrução pesqueira, reduzindo a variabilidade residual e aumentando a robustez dos modelos.

A análise preliminar revelou fortes correlações positivas entre o camarão sete-barbas e a barba-ruça com a variável preditora, e uma correlação negativa moderada para o camarão rosa. Essas tendências foram confirmadas pelos modelos, que comparou os mínimos quadrados ordinários, o LAD e o método de Dodge (1997). Para o camarão rosa, destacou uma tendência negativa mais pronunciada, enquanto para as outras espécies mostraram tendências positivas mais intensas, evidenciando a influência dos dados atípicos na suavização ou exacerbação das relações.

Os objetivos foram totalmente alcançados, analisou-se a influência de observações extremas, estimaram-se e compararam-se os coeficientes, identificaram-se as tendências de captura, destacaram-se os padrões de correlação e explorou-se o método de Dodge (1997) na aplicação ao banco de dados. A remoção de dados discrepantes baseada nos resíduos padronizados e no algoritmo de Dodge (1997) mostrou-se uma estratégia eficaz para contornar a sensibilidade na presença dos valores discrepantes.

Dessa forma, observou-se que tanto os valores absolutos quanto as variações no comportamento da espécie se mostrou distintos em relação ao modelo, que obteve os resultados de beta 0 e beta 1 ao empregar a técnica descrita no artigo, apresentando valores superiores aos outros comparativos, uma vez que o coeficiente resultou em um aumento considerável, o camarão barba-ruça apresentou um desempenho notável, com os coeficientes beta 0 e beta 1 dobrando seus percentuais em ambas as técnicas avaliadas, Dodge x MMQ e Dodge x LAD, alcançando, respectivamente, +128,23% e +132,78%.

Assim, este estudo contribui para avançar as técnicas de modelagem em dados imprecisos, oferecendo uma aplicação prática na continuação de pesquisa sobre o projeto já iniciado anteriormente. Os resultados reforçam a relevância do LAD e do algoritmo Dodge

(1997) como ferramentas valiosas para análises robustas, com potencial para subsidiar a gestão sustentável da pesca no Rio de Janeiro e em todo o Brasil.

BIBLIOGRAFIA

AGGARWAL, Charu C. *Outlier analysis. Ludhhavna: Sadbhavna Publications*, [s.d.]. Disponível em: <https://sadbhavnapublications.org/research-enrichment-material/2-Statistical-Books/Outlier-Analysis.pdf>. Acesso em: 2 fev. 2025.

BOSCOVICH, R. J. *De literaria expeditione per pontificiam ditionem et synopsis amplioris operis. Commentarii de Bononiensi Scientiarum et Artium Instituto Atque Academia*, v. 4, p. 353-396, 1757.

CHANDOLA, V.; BANERJEE, A.; KUMAR, V. *Anomaly detection: a survey. ACM Computing Surveys,* v. 41, n. 3, p. 1-58, 2009. Disponível em: <http://cucis.ece.northwestern.edu/projects/DMS/publications/AnomalyDetection.pdf>. Acesso em: 03 fev. 2025.

DRAPER, N. R.; SMITH, H. *Applied regression analysis*. 3. ed. New York: Wiley, 1998. Disponível em: <https://www.wiley.com/en-us/Applied+Regression+Analysis%2C+3rd+Edition-p-9780471170822>. Acesso em: 09 fev. 2025.

FREIRE K.M.F. et.al. 2021. *Reconstruction of marine comercial landings for the Brazilian industrial and artesanal fisheries from* 1950 *to* 2015. Front. Mar. Sci.

GAUSS, C. F. *Theoria motus corporum coelestium in sectionibus conicis solem ambientium. Hamburg*: F. Perthes e I. H. Besser, 1809. Disponível em: <https://archive.org/details/bub_gb_ORUOAAAAQAAJ>. Acesso em: 03 fev. 2025.

HAIR JR., J. F. et al. *Multivariate data analysis. 8. ed. Boston: Cengage Learning*, 2019. Disponível em: <https://www.amazon.com/Multivariate-Data-Analysis-Joseph-Hair/dp/9353501350>. Acesso em: 08 jan. 2025.

HAN, J.; PEI, J.; KAMBER, M. *Data mining: concepts and techniques*. 3. ed. Burlington: Morgan Kaufmann, 2012. Disponível em: <https://www.sciencedirect.com/book/9780123814791/data-mining-concepts-and-techniques>. Acesso em: 08 fev. 2025.

HAMPEL, F. R.; RONCHETTI, E. M.; ROUSSEEUW, P. J.; STAHEL, W. A. *Robust statistics: the approach based on influence functions.* New York: Wiley, 1986. Disponível em: <https://www.amazon.com/Robust-Statistics-Approach-InfluenceFunctions/dp/0471735779>. Acesso em: 24 jan. 2025.

LEGENDRE, A. M. *Nouvelles méthodes pour la détermination des orbites des comètes*. Paris: Firmin Didot, 1805. Disponível em: <https://archive.org/details/nouvellesmthode00legegoog>. Acesso em: 26 jan. 2025.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. *Introduction to linear regression analysis.* 6. ed. Hoboken: Wiley, 2021. Disponível em: <https://a.co/d/1gPdTtz>. Acesso em: 07 fev. 2025.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. *Introduction to linear regression analysis*. 5. ed. Hoboken: Wiley, 2012. Disponível em: <https://www.amazon.com/Introduction-Regression-Analysis-Douglas-Montgomery/dp/0470542810>. Acesso em: 09 jan. 2025.

OSORIO F.; WOLODZKO T. *Routines for L1 estimation. R package version* 0.52, 2024. Disponível em: <https://github.com/faosorios/L1pack>. Acesso em: 20 mar. 2025.

POSIT TEAM. RStudio*: Integrated Development Environment for* R. Boston, MA: Posit Software, PBC, 2024. Disponível em: <http://www.posit.co/>. Acesso em: 20 mar. 2025.

ROBSON, E.; STEDALL, J. (Eds.). *The Oxford handbook of the history of mathematics. Oxford: Oxford University Press*, 2012. Disponível em: <https://global.oup.com/academic/product/the-oxford-handbook-of-the-history-of-mathematics-9780199213122>. Acesso em: 12 fev. 2025.

ROUSSEEUW, P. J.; LEROY, A. M. *Robust regression and outlier detection*. New York: Wiley, 2003. Disponível em: <https://www.researchgate.net/publication/303193534_Robust_Regression_Outlier_Detection_John_Wiley_Sons>. Acesso em: 08 fev. 2025.

ROUSSEEUW, P. J. *Least median of squares regression. Journal of the American Statistical Association*. Disponível em: <https://www.jstor.org/stable/2288718>. Acesso em: 21 jan. 2025.

STIGLER, S. M. *The history of statistics: the measurement of uncertainty before* 1900. Cambridge, MA: Harvard University Press, 1986. Disponível em: <https://archive.org/details/historyofstatist0000stig>. Acesso em: 02 fev. 2025.

WOLBERG, J. Data *analysis using the method of least squares: extracting the most information from experiments*. Berlin: Springer, 2006. Disponível em: <https://www.amazon.com/Analysis-Using-Method-Least-Squares/dp/B01A0BTCGG>. Acesso em: 04 fev. 2025.

ANEXO A

Segue o Script em RStudio desenvolvido para a análises apresentadas para a pesquisa.

Relatório: Cálculos para o Trabalho de Conclusão de Curso

```
library(L1pack)

dados0 <- read.csv("C:\\Users\\User\\Desktop\\meu
tcc\\tcc_alex_celmo\\dados_originais\\camaroes_ind_RJ.csv", dec=",", sep=";")

nDados0 <- dim(dados0)[1]

plot(log(dados0[,"Camarao_ind"]), log(dados0[,"Cam_rosa_ind"]), main = "Cam_rosa_ind vs
Camarao_ind")

plot(log(dados0[,"Camarao_ind"]), log(dados0[,"Cam_barba_ruca_ind"]), main =
"Cam_barba_ruca_ind vs Camarao_ind")

plot(log(dados0[,"Camarao_ind"]), log(dados0[,"Cam_sete_barbas_ind"]), main =
"Cam_sete_barbas_ind vs Camarao_ind")

cor(na.exclude(cbind(log(dados0[,"Camarao_ind"]), log(dados0[,"Cam_rosa_ind"]))))

cor(na.exclude(cbind(log(dados0[,"Camarao_ind"]), log(dados0[,"Cam_barba_ruca_ind"]))))

cor(na.exclude(cbind(log(dados0[,"Camarao_ind"]), log(dados0[,"Cam_sete_barbas_ind"]))))

y <- c(log(dados0[,"Cam_rosa_ind"]), log(dados0[,"Cam_barba_ruca_ind"]),
log(dados0[,"Cam_sete_barbas_ind"]))

x <- as.factor(c(rep(1, nDados0), rep(2, nDados0), rep(3, nDados0)))

levels(x) <- c("Cam_rosa_ind", "Cam_barba_ruca_ind", "Cam_sete_barbas_ind")

boxplot(y ~ x)
```

```
boxplot(log(dados0[,"Camarao_ind"]), main = "Variável Preditora", sub = "Camarao_ind")


#-------------------------------------------------------------------------------------------------------------
------------------

dados <- na.exclude(dados0[, c("Cam_rosa_ind", "Camarao_ind")])

nDados <- dim(dados)[1]

Index <- 1:nDados

dados <- cbind(Index, dados)

aj1 <- lad(log(Cam_rosa_ind) ~ log(Camarao_ind), data = dados)

r1 <- residuals(aj1)

sigma1 <- 1.4826 * median(abs(r1))

I1 <- abs(r1 / sigma1) >= 2.5

dados1 <- dados[!I1,]

out1 <- dados[I1, "Index"]

aj2 <- lad(log(Camarao_ind) ~ log(Cam_rosa_ind), data = dados1)

r2 <- residuals(aj2)

sigma2 <- 1.4826 * median(abs(r2))

I2 <- abs(r2 / sigma2) >= 2.5

dados2 <- dados1[!I2,]

out2 <- dados1[I2, "Index"]
```

```
ajComOut <- lad(log(Cam_rosa_ind) ~ log(Camarao_ind), data = dados)

ajSemOut <- lad(log(Cam_rosa_ind) ~ log(Camarao_ind), data = dados2)

plot(log(dados[,"Camarao_ind"]), log(dados[,"Cam_rosa_ind"]), pch = 20, cex = 2, main =
"Camarão Rosa")

points(log(dados[out1, "Camarao_ind"]), log(dados[out1, "Cam_rosa_ind"]), pch = 20, cex =
2, col = "red")

points(log(dados[out2, "Camarao_ind"]), log(dados[out2, "Cam_rosa_ind"]), pch = 20, cex =
2, col = "blue")

abline(ajComOut)

abline(ajSemOut, col = "orange")

ajLM <- lm(log(Cam_rosa_ind) ~ log(Camarao_ind), data = dados)

abline(ajLM, col = "green")

legend("bottomright", legend = c("LAD com outliers", "DODGE sem outliers", "LM",
"Outliers Verticais", "Outliers Horizontais"),

      col = c("black", "orange", "green", "red", "blue"), lty = c(1, 1, 1, NA, NA), pch = c(NA,
NA, NA, 20, 20))

summary(ajComOut)

summary(ajSemOut)

summary(ajLM)



#-----------------------------------------------------------------------------------------------------------
----------
```

```
dados <- na.exclude(dados0[, c("Cam_barba_ruca_ind", "Camarao_ind")])

nDados <- dim(dados)[1]

Index <- 1:nDados

dados <- cbind(Index, dados)

aj1 <- lad(log(Cam_barba_ruca_ind) ~ log(Camarao_ind), data = dados)

r1 <- residuals(aj1)

sigma1 <- 1.4826 * median(abs(r1))

I1 <- abs(r1 / sigma1) >= 2.5

dados1 <- dados[!I1,]

out1 <- dados[I1, "Index"]

aj2 <- lad(log(Camarao_ind) ~ log(Cam_barba_ruca_ind), data = dados1)

r2 <- residuals(aj2)

sigma2 <- 1.4826 * median(abs(r2))

I2 <- abs(r2 / sigma2) >= 2.5

dados2 <- dados1[!I2,]

out2 <- dados1[I2, "Index"]

ajComOut <- lad(log(Cam_barba_ruca_ind) ~ log(Camarao_ind), data = dados)

ajSemOut <- lad(log(Cam_barba_ruca_ind) ~ log(Camarao_ind), data = dados2)

plot(log(dados[,"Camarao_ind"]), log(dados[,"Cam_barba_ruca_ind"]), pch = 20, cex = 2,
main = "Camarão Barba Ruça")
```

```
points(log(dados[out1, "Camarao_ind"]), log(dados[out1, "Cam_barba_ruca_ind"]), pch = 20, cex = 2, col = "red")

points(log(dados[out2, "Camarao_ind"]), log(dados[out2, "Cam_barba_ruca_ind"]), pch = 20, cex = 2, col = "blue")

abline(ajComOut)

abline(ajSemOut, col = "orange")

ajLM <- lm(log(Cam_barba_ruca_ind) ~ log(Camarao_ind), data = dados)

abline(ajLM, col = "green")

legend("topleft", legend = c("LAD com outliers", "DODGE sem outliers", "LM", "Outliers Verticais", "Outliers Horizontais"),

      col = c("black", "orange", "green", "red", "blue"), lty = c(1, 1, 1, NA, NA), pch = c(NA, NA, NA, 20, 20))

summary(ajComOut)

summary(ajSemOut)

summary(ajLM)



#-------------------------------------------------------------------------------------------------------------------

dados <- na.exclude(dados0[, c("Cam_sete_barbas_ind", "Camarao_ind")])

nDados <- dim(dados)[1]

Index <- 1:nDados
```

```
dados <- cbind(Index, dados)

aj1 <- lad(log(Cam_sete_barbas_ind) ~ log(Camarao_ind), data = dados)

r1 <- residuals(aj1)

sigma1 <- 1.4826 * median(abs(r1))

I1 <- abs(r1 / sigma1) >= 2.5

dados1 <- dados[!I1,]

out1 <- dados[I1, "Index"]

aj2 <- lad(log(Camarao_ind) ~ log(Cam_sete_barbas_ind), data = dados1)

r2 <- residuals(aj2)

sigma2 <- 1.4826 * median(abs(r2))

I2 <- abs(r2 / sigma2) >= 2.5

dados2 <- dados1[!I2,]

out2 <- dados1[I2, "Index"]

ajComOut <- lad(log(Cam_sete_barbas_ind) ~ log(Camarao_ind), data = dados)

ajSemOut <- lad(log(Cam_sete_barbas_ind) ~ log(Camarao_ind), data = dados2)

plot(log(dados[,"Camarao_ind"]), log(dados[,"Cam_sete_barbas_ind"]), pch = 20, cex = 2,

    main = "Camarão Sete Barbas", xlab = "log(Camarao_ind)", ylab =
"log(Cam_sete_barbas_ind)")

points(log(dados[out1, "Camarao_ind"]), log(dados[out1, "Cam_sete_barbas_ind"]), pch =
20, cex = 2, col = "red")
```

```
points(log(dados[out2, "Camarao_ind"]), log(dados[out2, "Cam_sete_barbas_ind"]), pch = 20, cex = 2, col = "blue")

abline(ajComOut)

abline(ajSemOut, col = "orange")

ajLM <- lm(log(Cam_sete_barbas_ind) ~ log(Camarao_ind), data = dados)

abline(ajLM, col = "green")

legend("topleft", legend = c("LAD com outliers", "DODGE sem outliers", "LM", "Outliers Verticais", "Outliers Horizontais"),

      col = c("black", "orange", "green", "red", "blue"), lty = c(1, 1, 1, NA, NA), pch = c(NA, NA, NA, 20, 20))

summary(ajComOut)

summary(ajSemOut)

summary(ajLM)

#simulação:

############ simulação de ponto aberrante e ponto fluente ###############

library(ggplot2)

set.seed(123)

x <- 1:20

y <- 2 * x + rnorm(20, mean = 0, sd = 3)

x_aberrante <- 15

y_aberrante <- 50
```

```
x_fluente <- 18

y_fluente <- 2 * x_fluente + rnorm(1, mean = 0, sd = 3)

dados <- data.frame(x, y)

dados_extra <- data.frame(

  x = c(x_aberrante, x_fluente),

  y = c(y_aberrante, y_fluente),

  tipo = c("Ponto Aberrante", "Ponto Fluente")

)

ggplot(dados, aes(x, y)) +

  geom_point(color = "blue", size = 2) +  # Pontos originais

  geom_smooth(method = "lm", se = FALSE, color = "black", linetype = "dashed") +

  geom_point(data = dados_extra, aes(x, y, color = tipo), size = 4, shape = 17) +

  scale_color_manual(values = c("Ponto Aberrante" = "red", "Ponto Fluente" = "green")) +

  labs(title = "Exemplo de Ponto Aberrante e Ponto Fluente",

       x = "Variável Explicativa (x)",

       y = "Variável Resposta (y)",

       color = "Tipo de Outlier") +

  theme_minimal()

############ Simulação de outliers verticais e horizontais ###########
```

```
set.seed(123)

n <- 100

x <- rnorm(n, mean = 50, sd = 10)

y <- 2 * x + rnorm(n, mean = 0, sd = 5)

data <- data.frame(x, y)

outlier_vertical <- data.frame(x = 50, y = 200)

outlier_horizontal <- data.frame(x = 150, y = 2 * 150 + rnorm(1, 0, 5))

data_with_outliers <- rbind(data, outlier_vertical, outlier_horizontal)

modelo_com_outliers <- lm(y ~ x, data = data_with_outliers)

modelo_sem_outliers <- lm(y ~ x, data = data)

coef_com_outliers <- coef(modelo_com_outliers)

coef_sem_outliers <- coef(modelo_sem_outliers)


par(mfrow = c(1, 2))


##########################################################################
##########################################

##################################### Regressão com outliers
#############################################################

##########################################################################
##########################################
```

```
plot(data_with_outliers$x, data_with_outliers$y, pch = 19, col = "gray",
    main = "Com Outliers", xlab = "X", ylab = "Y")
abline(modelo_com_outliers, col = "red", lwd = 2)
points(outlier_vertical, col = "blue", pch = 19)
points(outlier_horizontal, col = "green", pch = 19)
legend("topleft", legend = c("Dados", "Regressão", "Outlier Vertical", "Outlier Horizontal"),
      col = c("gray", "red", "blue", "green"), pch = c(19, NA, 19, 19), lty = c(NA, 1, NA, NA))


##########################################################################################
########################################

##################################### Regressão sem outliers
#############################################################

##########################################################################################
########################################

plot(data$x, data$y, pch = 19, col = "gray",
    main = "Sem Outliers", xlab = "X", ylab = "Y")
abline(modelo_sem_outliers, col = "blue", lwd = 2)
legend("topleft", legend = c("Dados", "Regressão"),
      col = c("gray", "blue"), pch = c(19, NA), lty = c(NA, 1))
```

```
##############################################################################
##########################################

########################################## Comparação dos coeficientes
##################################################

##############################################################################
##########################################

cat("Coeficientes com outliers:", coef_com_outliers, "\n")

cat("Coeficientes sem outliers:", coef_sem_outliers, "\n")
```