Estimação do modelo de Regressão Segmentada utilizando uma abordagem robusta

João Victor Uliana Felix[†]

Universidade Federal do Espírito Santo - UFES
Departamento de Estatística[†]
Orientador: Prof. Dr. Alessandro José Queiroz Sarnaglia

Resumo

Este artigo propõe uma abordagem alternativa à metodologia usual da estimação do modelo de regressão segmentada com pontos de quebra desconhecidos. Especificamente, com base em Estimadores-M, é introduzida uma extensão da metodologia de estimação usual proposta por Muggeo [2003]. Devido as características bem conhecidas desse tipo de estimador, espera-se melhor desempenho da metodologia proposta num cenário de contaminação por outliers. Com o intuito de investigar o desempenho do método proposto, foi realizado um extensivo estudo de simulação de Monte Carlo sob vários cenários. O desempenho dos estimadores foi acessado via Viés e Raiz do Erro Quadrático Médio (REQM). A metodologia proposta se saiu melhor em ambas as medidas de desempenho. Para enfatizar a sua importância em situações práticas, a metodologia desenvolvida neste artigo foi ilustrada através de uma aplicação a dados de fisiologia do exercício. Os resultados mostraram que o erro padrão do método proposto por este artigo se saiu melhor do que quando comparado ao de Muggeo [2003].

1 Introdução

Compreender relações entre variáveis e realizar previsões mais acuradas é essencial em diversas áreas da ciência. O ramo da Estatística que estuda essas metodologias é denominado de Análise de Regressão [Hoffmann and Vieira, 1977] e, com o passar dos anos, sua teoria foi amplamente difundida e ramificada em busca da melhoria e adaptação dos procedimentos de estimação sob diversas particularidades inerentes aos dados.

Em uma análise de regressão linear, uma observação que é substancialmente diferente das outras pode causar uma grande diferença no resultado final de uma análise. Esses dados atípicos, ou outliers, como são comumente chamados, não são raros em situações aplicadas [Chatterjee and Hadi, 1986] e, muitas vezes, podem passar despercebidos na atualidade, onde frequentemente os dados são processados automaticamente.

A motivação deste estudo está na relação entre o consumo máximo de oxigênio (VO_2) e o aumento progressivo da velocidade de um corredor em uma esteira. Os dados considerados neste estudo fazem parte do conjunto de dados originalmente analisado por Abreu [2017], que estudou a incidência de platô [Taylor et al., 1955] em dados de VO_2. A relação entre essas duas variáveis, embora contínua, tende a apresentar uma ou mais mudanças bruscas em seu comportamento linear a partir de valores delimitadores de velocidade (desconhecidos *a priori*). A Figura 1 mostra a relação Velocidade $\times$ VO_2 para um determinado atleta.

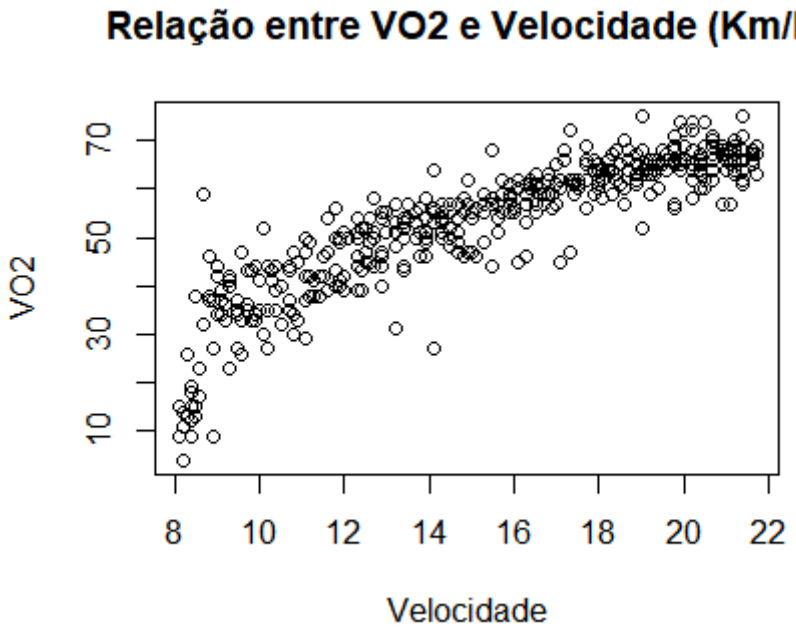

Figura 1: Velocidade × VO_2

Por meio da Figura 1, nota-se que a relação entre as variáveis muda significativamente de inclinação a partir de um ponto em torno de 11 km/h. Como destacado em Abreu [2017], esse tipo de comportamento pode ser convenientemente explicado através do modelo de Regressão Segmentada, originalmente introduzido por McZgee and Carleton [1970].

Embora seja capaz de acomodar satisfatoriamente o fenômeno em estudo, o modelo de regressão segmentada possui a desvantagem de a função de log-verossimilhança não ser derivável nos pontos de quebra, ainda que os segmentos de reta sejam contínuos, isto é, essa função é apenas diferenciável por partes. Isso se torna um impeditivo ao procedimento de estimação de máxima verossimilhança, tendo em vista que métodos numéricos usuais não podem ser empregados e teoria assintótica é consideravelmente mais complexa [Feder et al., 1975, por exemplo]. Nesse contexto, Muggeo [2003] desenvolveu um algoritmo que visa contornar esses problemas através de aproximação de Taylor de primeira ordem da equação de regressão. Desde então, na literatura, essa proposta fomentou diversos procedimentos de estimação do modelo de regressão segmentada sob diferentes contextos [Muggeo, 2003, 2008, 2016, Patrício and Sarnaglia, 2019, por exemplo].

Por outro lado, outra característica marcante nos dados ilustrados na Figura 1 é que algumas poucas observações parecem destoar do padrão do restante dos dados. Do ponto de vista do fenômeno em questão, essas potenciais observações atípicas são, em geral, provocadas pelo deslocamento da máscara que coleta os dados, acarretando falha temporária da vedação da mesma. Os efeitos adversos provocados por outliers, compreendem: incremento do erro-padrão de estimadores clássicos; introdução de vícios e falta de precisão de previsores; incoerência em testes de hipóteses; entre outros. No contexto de análise de regressão, em diversos casos, a posição exata do outlier é desconhecida, o que torna o seu tratamento passível de incertezas devido a variabilidade oriunda da estimação da posição de ocorrência.

Uma solução alternativa muito popular na literatura é a utilização de métodos robustos, como Regressão-M [Stuart, 2011, Serneels et al., 2005, Rousseeuw et al., 2004, por exemplo]. Em poucas palavras, a Regressão-M busca fornecer estimativas robustas através da atribuição de menos peso aos erros discrepantes do que o método de mínimos quadrados.

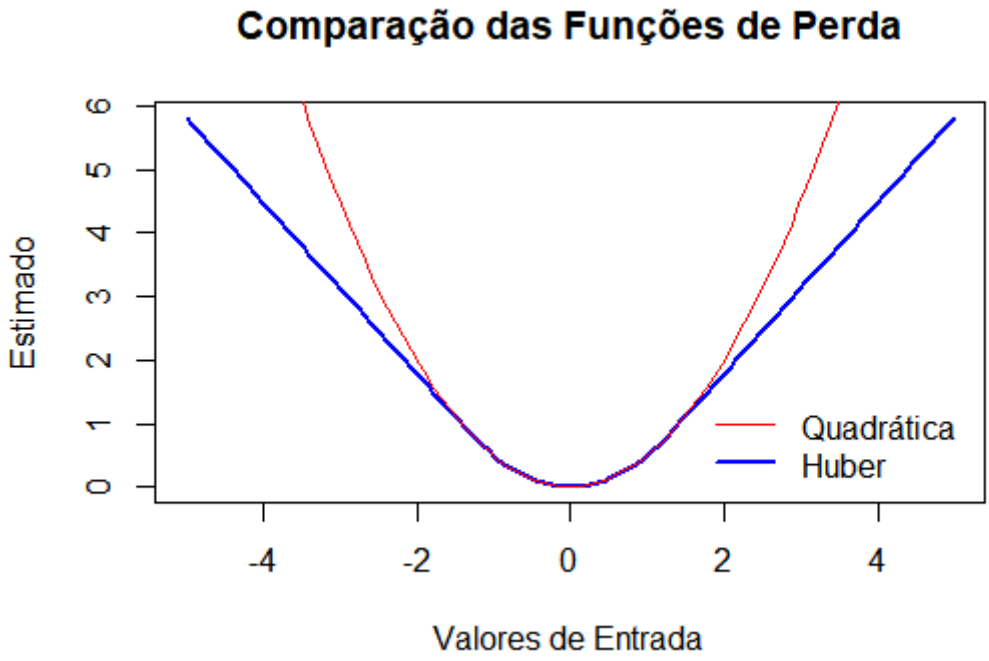

Figura 2: Comparação entre funções de perda

Como exemplo, a Figura 2 apresenta a perda quadrática (que origina o método de mínimos quadrados) e a perda de Huber (um caso particular da Regressão-M). Na Figura 2, fica claro que, se comparada a perda quadrática, a função de Huber atribui menos peso à valores muito distantes de zero, que no contexto de análise de regressão representariam resíduos oriundos de observações discrepantes.

Motivado por essa problemática, este estudo introduz método robusto para estimação do modelo de Regressão Segmentada que estende aquele proposto por Muggeo [2003] com auxílio de Regressão-M.

O desempenho da metodologia proposta foi investigado através de um estudo extensivo de simulação de Monte Carlo sob diversos cenários. Para efeitos de comparação, considerou-se o método usual [Muggeo, 2003]. Os resultados apontam que, em cenários não contaminados, ambos estimadores apresentam desempenho competitivo, enquanto que, sob a contaminação de outliers, como esperado, o método proposto é muito mais eficiente que o usual. Evidenciando suas características robustas.

Com intuito de ilustrar a importância do novo método proposto por este estudo, os dados de Velocidade $\times$ VO_2 (Figura 1) são revisitados. A aplicação mostra o ganho de eficiência ao se utilizar a metodologia robusta, em consonância com os resultados obtidos no estudo de simulação.

O restante desse trabalho é estruturado como se segue: na Seção 2, o modelo de Regressão Segmentada é definido, o método de estimação usual é brevemente apresentado e o método proposto é discutido; na Seção 3, os resultados do estudo de simulação e a discussão relacionada são apresentados; a Seção 4 discute a aplicação aos dados de Fisiologia do Exercício; e a Seção 5 apresenta conclusões e perspectivas futuras.

2 Metodologia

2.1 O Modelo de Regressão Segmentada

Na metodologia de regressão linear usual, o modelo é construído com intuito de explicar a tendência linear de variáveis explicativas representadas por um vetor $\underset{\sim}{X}' = (X_1, X_2, ..., X_p)$ em uma variável resposta Y, sob o efeito de um intercepto ζ_0 e um vetor de coeficientes $\zeta' = (\zeta_1, ..., \zeta_p)$ que mostra o impacto do incremento de uma unidade da covariável $\underset{\sim}{X}$ na resposta Y.

Esse modelo é expressado genericamente por:

$$Y_i = \zeta_0 + \underset{\sim}{X}_i'\zeta + \epsilon_i, i = 1, \ldots, n, \tag{1}$$

em que ϵ_i é um erro aleatório.

Embora a regressão linear (e suas extensões) seja de extrema importância, tendo merecido destaque na literatura nas últimas décadas, em alguns problemas práticos, a variável resposta não resulta de efeito linear das covariáveis, de modo que o modelo apresentado na Equação 1 não se adequa satisfatoriamente aos dados.

Em especial, na regressão segmentada, é possível ajustar vários modelos de regressão linear, porém com mudança de inclinação no que chamamos de pontos de quebra, que é o ponto de mudança responsável pela não-linearidade do modelo.

Sem perda de generalidade, iremos assumir apenas uma covariável (denotada por Z_i) com relacionamento não linear com respeito à resposta Y_i e que essa relação é governada pelo parâmetro escalar ψ. Nesse caso, o modelo pode ser expresso genericamente como

$$Y_i = \zeta_0 + \underset{\sim}{X}_i'\zeta + \alpha Z_i + \beta h(Z_i, \psi) + \epsilon_i, \quad i = 1, ..., n, \tag{2}$$

em que h é uma função derivável e β descreve o efeito não linear de Z_i em Y_i.

No caso particular da regressão segmentada, $h(Z_i, \psi) = (Z_i - \psi)\mathbb{1}(Z_i > \psi)$ em (1), o que resulta no modelo

$$Y_i = \zeta_0 + \underset{\sim}{X}_i'\zeta + \alpha Z_i + \beta(Z_i - \psi)\mathbb{1}(Z_i > \psi) + \epsilon_i, \quad i = 1, ..., n, \tag{3}$$

em que $\mathbb{1}(.)$ denota a função indicadora. Nesse caso, o parâmetro ψ representa o ponto de mudança (ou de quebra).

2.2 Metodologia Usual

Como dito anteriormente, quando os pontos de mudança são desconhecidos, o procedimento de estimação de máxima verossimilhança dos parâmetros do modelo de Regressão Segmentada (Equação 3) é particularmente desafiador. Nesse sentido, o artigo seminal de Muggeo [2003] propõe uma estratégia para contornar a dificuldade numérica, aproximando o problema de otimização não linear por um procedimento interativo de mínimos quadrados. Mais especificamente, para estimação do modelo descrito na Equação 1, recorre-se à expansão de Taylor de h em torno de $\psi^{(0)}$, fornecendo

$$Y_i = \zeta_0 + \underset{\sim}{X}_i'\zeta + \alpha Z_i + \beta h(Z_i, \psi^{(0)}) + \gamma' \left[\frac{\partial h}{\partial \psi}(Z_i, \psi^{(0)})\right] + \epsilon_i, \quad i = 1, ..., n, \tag{4}$$

em que $\gamma' = \beta(\psi - \psi^{(0)})$. No caso particular da regressão segmentada (Equação 3), temos que $\frac{\partial h}{\partial \psi}(Z_i, \psi^{(0)}) = -\mathbb{1}(Z_i > \psi^{(0)})$.

O modelo de Regressão Segmentada (Equação 3) é então estimado via regressão linear usual do modelo aproximado em (4) (fixado $\psi^{(0)}$) e, então, o valor de ψ é atualizado através

de $\psi = \psi^{(0)} + \frac{\gamma}{\beta}$. O processo é repetido até atingir um critério de parada pré-especificado. No Algoritmo 1 abaixo, apresentamos um pseudo-código para o estimador descrito.

Algoritmo 1: Estimação do modelo de regressão segmentada

Entrada: $\hat{\psi}^{(1)}, Y_i, \underset{\sim}{X}_i, Z_i, i = 1, ..., n$

Saída: Ponto de quebra estimado

início

faça s = 0

Repita

Faça s = s + 1.

Defina as covariáveis

$$U_{ki}^{(s)} = (Z_i - \hat{\psi}_k^{(s)})_+ \text{ e } V_{ki}^{(s)} = -\mathbb{1}(Z_i > \hat{\psi}_k^{(s)}), \quad i = 1, ..., n.$$

para k = 1,...,q.

Ajuste o seguinte modelo de regressão linear múltipla:

$$Y_i = \zeta_0 + \underset{\sim}{X}_i'\zeta + \alpha Z_i + \sum_{k=1}^{q} \beta_k U_{ki}^{(s)} + \sum_{k=1}^{q} \gamma_k V_{ki}^{(s)} + \epsilon_i, \quad i = 1, ..., n,$$

para obter as estimativas de $\hat{\zeta}_0^{(s)}$, $\hat{\zeta}^{(s)}$, $\hat{\alpha}^{(s)}$, $\hat{\beta}^{(s)}$ e $\hat{\gamma}^{(s)}$.

Atualize as estimativas dos pontos de quebra por:

$$\hat{\psi}_k^{(s+1)} = \frac{\hat{\gamma}_k^{(s)}}{\hat{\beta}_k^{(s)}} + \hat{\psi}_k^{(s)}, \quad k = 1, ..., q.$$

Retorne ao passo 3.

fim

Até convergir

2.3 Estimador-M

Como citado anteriormente, observações discrepantes podem ter consequências severas no processo de inferência. Por esse motivo, métodos robustos são de suma importância. Neste artigo,os Estimadores-M serão utilizados como alternativa robusta à métodos clássicos. Os Estimadores-M são conhecidos por serem os mais flexíveis e costumam resolver problemas sem muita complexidade [Stuart, 2011].

Para introduzirmos o conceito de Regressão-M, relembre que o método de mínimos quadrados consiste em minimizar a perda quadrática

$$(Y_i - \mu(\underset{\sim}{X}_i; \theta))^2$$

em termos do vetor de parâmetros θ, que indexa $\mu(\underset{\sim}{x}; \theta)$, o efeito (possivelmente não linear) esperado quando as covariáveis $\underset{\sim}{X} = \underset{\sim}{x}$. No entanto, é notório que a perda quadrática atribui muito peso à eventuais erros discrepantes, o que pode comprometer demasiadamente as estimativas. Nesse sentido, a Regressão-M é obtida substituindo-se a perda quadrática, e^2, por alguma outra função perda, que denotaremos por $\rho(e)$. Portanto, as estimativas via Regressão-M são obtidas minimizando-se

$$\rho\left(\frac{Y_i - \mu(\underset{\sim}{X}_i; \theta)}{\sigma}\right), \tag{5}$$

onde a função ρ satisfaz $\rho \geq 0$; $\rho(0) = 0$; $\rho(-e) = \rho(e)$; e $\rho(e) \geq \rho(e^*)$, se $|e| > |e^*|$, e σ é o desvio-padrão do erro $Y_i - \mu(\underset{\sim}{X}_i; \theta)$. Em geral, o termo σ é desconhecido e sua estimação conjunta com o vetor θ pode demandar alto custo computacional. No caso em que $\mu(\underset{\sim}{x}; \theta)$ descreve uma relação linear, uma alternativa comum na literatura [Huber, 2011, por exemplo] consiste em estimar σ usando o Desvio Absoluto Mediano (MAD, em inglês), dos resíduos obtidos de alguma estimativa preliminar de θ. Para um valor fixo de θ, esse estimador de σ é dado por

$$\hat{\sigma} = k \cdot \text{MAD},$$

em que $\text{MAD} = \text{med}\{|Y_1 - \mu(\underset{\sim}{X}_1; \theta)|, \ldots, |Y_n - \mu(\underset{\sim}{X}_n; \theta)|\}$ e o fator de escala constante k é escolhido para garantir consistência para σ. No caso de normalidade dos dados $k = 1.4826$.

No caso de efeito linear, isto é, $\mu(\underset{\sim}{x}; \theta) = \underset{\sim}{x}'\theta$, a minimização na Equação 5 pode ser reformulada como o problema de solução do sistema de equações não lineares

$$\sum_{i=1}^{n} \underset{\sim}{X}_i \psi(Y_i - \underset{\sim}{X}_i'\beta) = \mathbf{0},$$

em que $\psi(e) = \frac{\partial \rho}{\partial e}$. Nesse caso, o procedimento de estimação pode ser reformulado como um método de mínimos quadrados ponderados iterativo. Para maiores detalhes, veja Huber [2011], por exemplo.

2.4 Metodologia Proposta

Neste trabalho, é proposta a utilização de Regressão-M em substituição de regressão usual para a etapa de estimação do modelo aproximado em (4), com intuito de obter estimativas robustas do modelo segmentado em (1). Essa metodologia promete melhor resultado e assertividade no

ajuste de modelos para dados com observações discrepantes.

Algoritmo 2: Estimação do modelo de regressão m-segmentada

Entrada: $\hat{\psi}^{(1)}, Y_i, \underset{\sim}{X}_i, Z_i, i = 1, ..., n$
Saída: Ponto de quebra estimado

1 **início**
2 faça $s = 0$
3 Repita
4 Faça $s = s + 1$.
5 Defina as covariáveis

$$U_{ki}^{(s)} = (Z_i - \hat{\psi}_k^{(s)})_+ \text{ e } V_{ki}^{(s)} = -\mathbb{1}(Z_i > \hat{\psi}_k^{(s)}), \quad i = 1, ..., n.$$

para $k = 1, ..., q$.
6 Ajuste o seguinte modelo de regressão linear múltipla via Regressão-M:

$$Y_i = \zeta_0 + \underset{\sim}{X}_i'\zeta + \alpha Z_i + \sum_{k=1}^{q} \beta_k U_{ki}^{(s)} + \sum_{k=1}^{q} \gamma_k V_{ki}^{(s)} + \epsilon_i, \quad i = 1, ..., n,$$

para obter as estimativas de $\hat{\zeta}_0^{(s)}$, $\hat{\zeta}^{(s)}$, $\hat{\alpha}^{(s)}$, $\hat{\beta}^{(s)}$ e $\hat{\gamma}^{(s)}$.
7 Atualize as estimativas dos pontos de quebra por:

$$\hat{\psi}_k^{(s+1)} = \frac{\hat{\gamma}_k^{(s)}}{\hat{\beta}_k^{(s)}} + \hat{\psi}_k^{(s)}, \quad k = 1, ..., q.$$

8 Retorne ao passo 3.
9 **fim**
10 Até convergir

Para contornar a não diferenciabilidade da log-verossimilhança nos pontos de quebra, o que impede estimação da informação de Fisher, utilizou-se metodologia Bootstrap para aproximar os valores dos erros-padrões dos estimadores.

Especificamente, como usual na literatura, propomos utilizar os parâmetros estimados e resíduos do modelo ajustado para gerar reamostras do conjunto de dados em questão. As amostras bootstrap são então utilizadas para obter réplicas bootstrap do estimador. O processo é repetido e a distribuição bootstrap do estimador utilizada para calcular o erro padrão associado.

3 Simulação

Com o intuito de investigar o desempenho da metodologia proposta, foi realizado um extenso estudo de Monte Carlo sob diferentes cenários. Para efeito de comparação, consideou-se a metodologia clássica de Muggeo [2003]. O desempenho dos métodos de estimação foi avaliado via Viés e Raiz do Erro Quadrático Médio (REQM).

Os coeficientes dos modelos simulados são apresentados na Tabela 1. Com o intuito de simular cenários não contaminados e contaminados, o termo de erro foi gerado de uma N(0,1) e de uma $t_{(3)}$ padronizada, respectivamente.

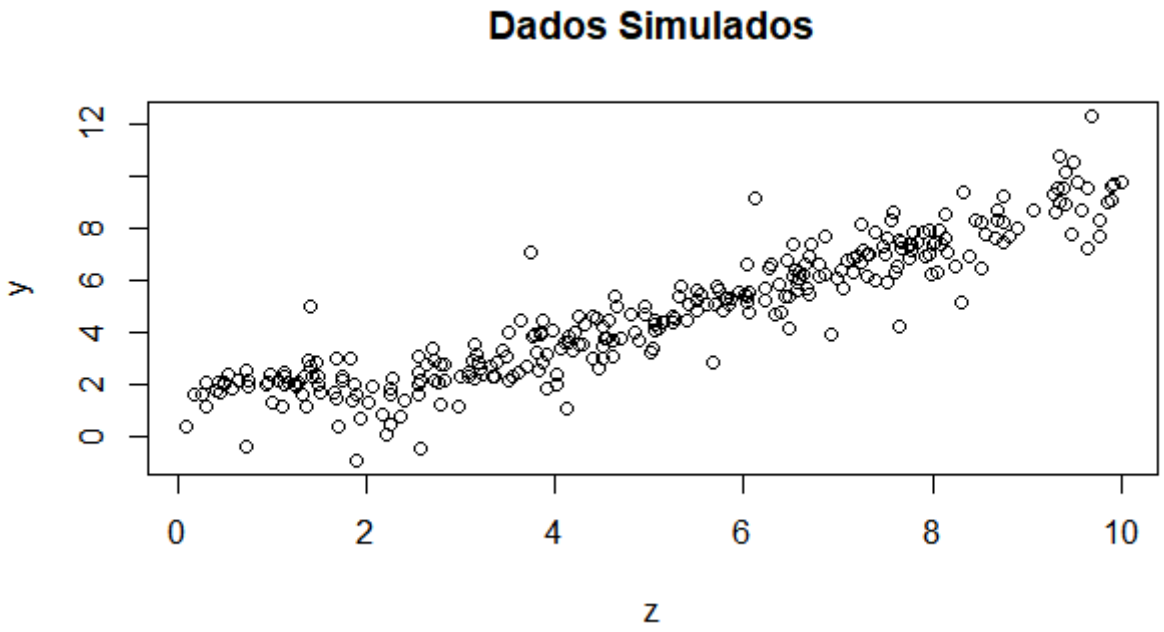

Figura 3: Uma das amostras utilizadas na simulação

Tabela 1: Modelos propostos

Modelo	ζ_0	β_1	ψ_1	α
1	2	1	2.5	0
2	2	1	5	0
3	2	1	7.5	0

As Tabelas 2, 3, 4 e 5 a seguir mostram o Viés e REQM dos estimadores clássico e o proposto neste estudo.

É notável que, num cenário sem contaminação, ambos os métodos têm desempenho quase similar e, como esperado, o método clássico fornece resultados pouco melhores. No cenário contaminado, é marcante a superioridade do método proposto, o que evidencia sua robustez.

Vale ressaltar que, de modo geral, para ambos estimadores, o intercepto (ζ_0) e o slope do primeiro segmento (α) são melhores estimados quando o ponto de quebra se aproxima do máximo da covariável Z_i, isto é, quando o primeiro segmento possui mais observações. Por outro lado, a diferença entre os slopes (β_1) e o ponto de quebra (ψ_1) são melhores estimados quando o esse último ocorre no meio do intervalo de valores de Z_i, de modo que ambos segmentos têm a mesma quantidade de observações.

Tabela 2: Viés dos modelos simulados - Sem contaminação

Método	Modelo	Parâmetro			
		ζ_0	α	β_1	ψ_1
Clássico	1	-0.0012	-0.0016	0.0007	0.0065
	2	-0.0002	0.0011	-0.0016	0.0027
	3	0.0063	-0.0015	0.0054	-0.011
Robusto	1	-0.0008	-0.0015	0.0009	0.0090
	2	-0.0007	0.0013	-0.0012	0.0055
	3	0.0068	-0.0016	0.0073	-0.009

Tabela 3: REQM dos modelos simulados - Sem contaminação

Método	Modelo	Parâmetro			
		ζ_0	α	β_1	ψ_1
Clássico	1	0.1659	0.1141	0.1161	0.1835
	2	0.1127	0.0391	0.0561	0.1604
	3	0.0943	0.0222	0.1079	0.1869
Robusto	1	0.1715	0.1166	0.1188	0.1854
	2	0.1150	0.0399	0.0576	0.1626
	3	0.0962	0.0225	0.1102	0.1886

Tabela 4: Viés dos modelos simulados - Com valores atípicos

Método	Modelo	Parâmetro			
		ζ_0	α	β_1	ψ_1
Clássico	1	0.0038	-0.0007	-0.0004	0.0137
	2	0.0037	-0.0011	0.0037	0.0058
	3	0.0063	-0.0014	-0.0059	-0.0272
Robusto	1	0.0070	-0.0027	0.0022	0.0063
	2	0.0014	-0.0004	0.0008	0.0017
	3	0.0022	-0.0008	-0.0020	-0.0144

Tabela 5: REQM dos modelos simulados - Com valores atípicos

Método	Modelo	Parâmetro			
		ζ_0	α	β_1	ψ_1
Clássico	1	0.1715	0.1187	0.1207	0.1936
	2	0.1024	0.0375	0.0567	0.1647
	3	0.0898	0.0205	0.1165	0.1876
Robusto	1	0.1236	0.0875	0.0889	0.1443
	2	0.0771	0.0278	0.0423	0.1166
	3	0.0683	0.0155	0.0893	0.1420

4 Aplicação

Como descrito em Abreu [2017], o Consumo de Oxigênio (VO_2) descreve o volume de oxigênio utilizado pelo corpo. Durante um teste de esforço máximo, em que o indivíduo é submetido à exercício físico, em geral corrida na esteira, com incrementos de carga com taxa constante em intervalos fixos de tempo, até a exaustão, é bem conhecido que a dinâmica do VO_2 pode apresentar até três regimes: o primeiro descreve a fase de adaptação do organismo; o segundo representa a progressão linear do VO_2 após adaptação; e o terceiro, que pode não ocorrer, é denominado platô e representa a fase em que o VO_2 se estabiliza mesmo sob incrementos de carga. Nesse cenário, a regressão segmentada se mostra uma opção atrativa para modelar a dinâmica dos dados.

Como mostrado na Figura 1, além do comportamento segmentado dos dados, é notável a presença de alguns pontos discrepantes que podem influenciar o processo de inferência. Por esse motivo, a metodologia proposta neste artigo será aplicada aos dados em questão.

Conforme citado nas seções anteriores, o modelo de regressão segmentada necessita de um chute inicial para estimação dos parâmetros. A título de comparação e demonstração da importância da visualização dos dados antes da estimação foram utilizados três modelos para os dados com três chutes iniciais distintos para o ponto de quebra (ψ_1), todos eles mostraram tiveram bom desempenho mas optou-se utilizar $\psi_1 = 9$ pois este apresentou o melhor resultado em termos de erro padrão.

Os coeficientes estimados possuem pequenas diferenças, mas que serão validadas para o definição do modelo mais eficiente.

Tabela 6: Coeficientes estimados do modelo aplicado

Método	Parâmetro			
	ζ_0	α	β_1	ψ_1
Clássico	-18.336	5.262	-3.153	9.284
Robusto	-18.033	5.217	-3.197	9.427

A Tabela 7 define, então, o método mais eficiente, mostrando que o modelo Robusto (proposto) se saiu bem melhor em todos os cenários propostos.

Tabela 7: Erro padrão do modelo aplicado

Método	Parâmetro			
	ζ_0	α	β_1	ψ_1
Clássico	1.178	0.159	0.175	0.158
Robusto	0.909	0.123	0.140	0.127

A Figura 2 mostra os dois modelos estimados sobrepostos nos dados utilizados para $\psi_1 = 9$.

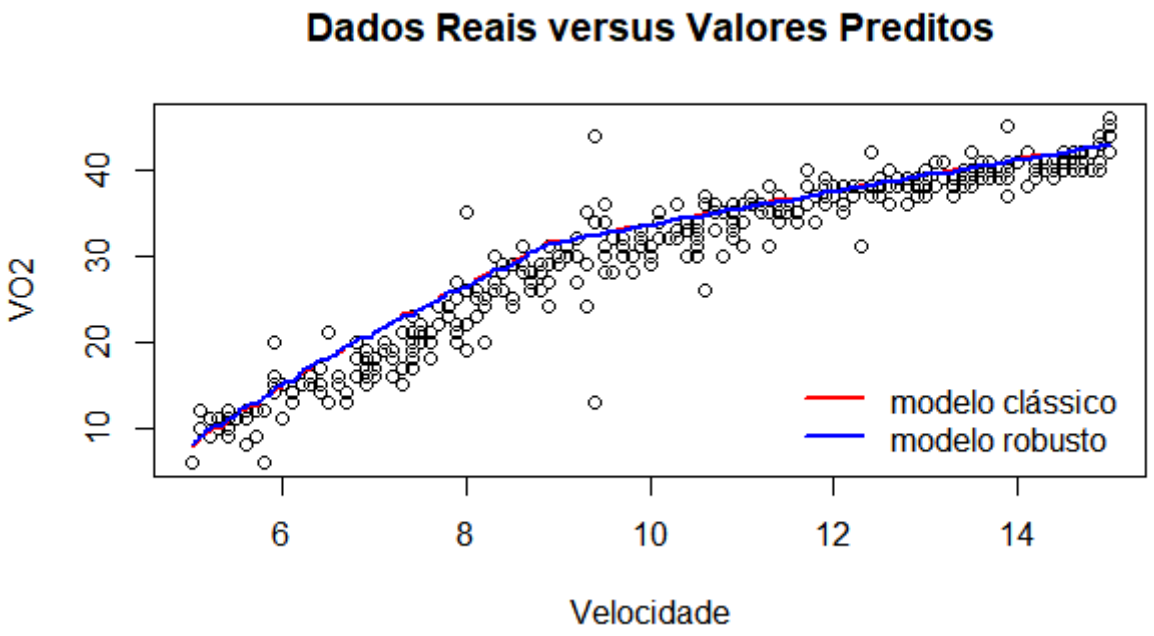

Figura 4: Comparação dos modelos

As Figuras 3 e 4 mostram a distribuição dos valores de ψ_1 estimados via *bootstrap* e suas médias dado chute inicial $\psi_1 = 9$.

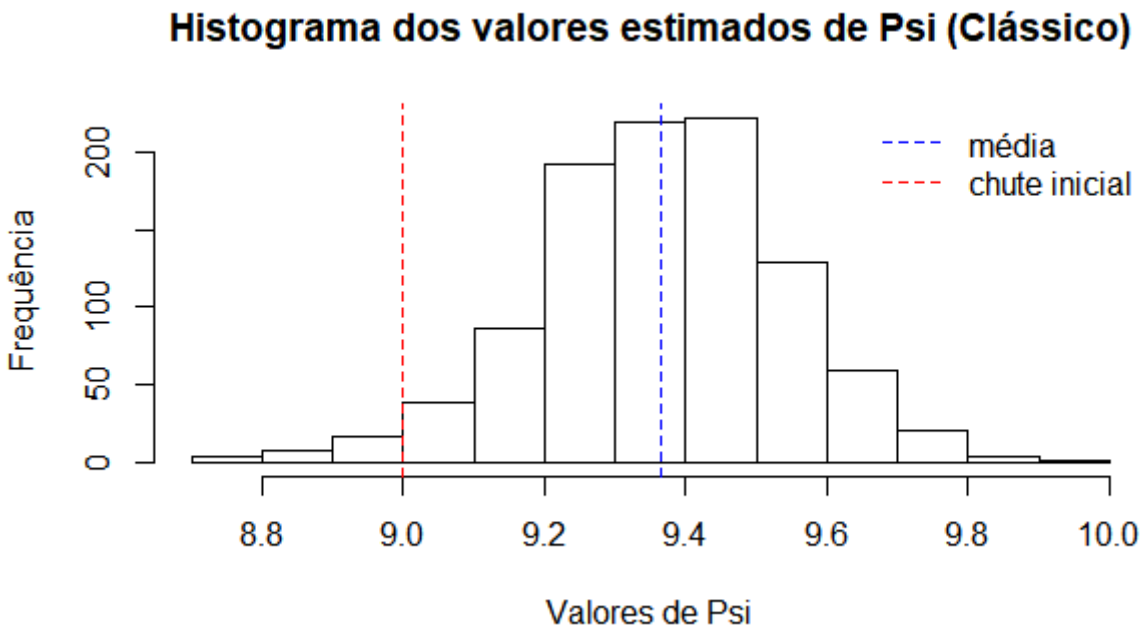

Figura 5: Histograma dos valores estimados de ψ_1 para o método Clássico

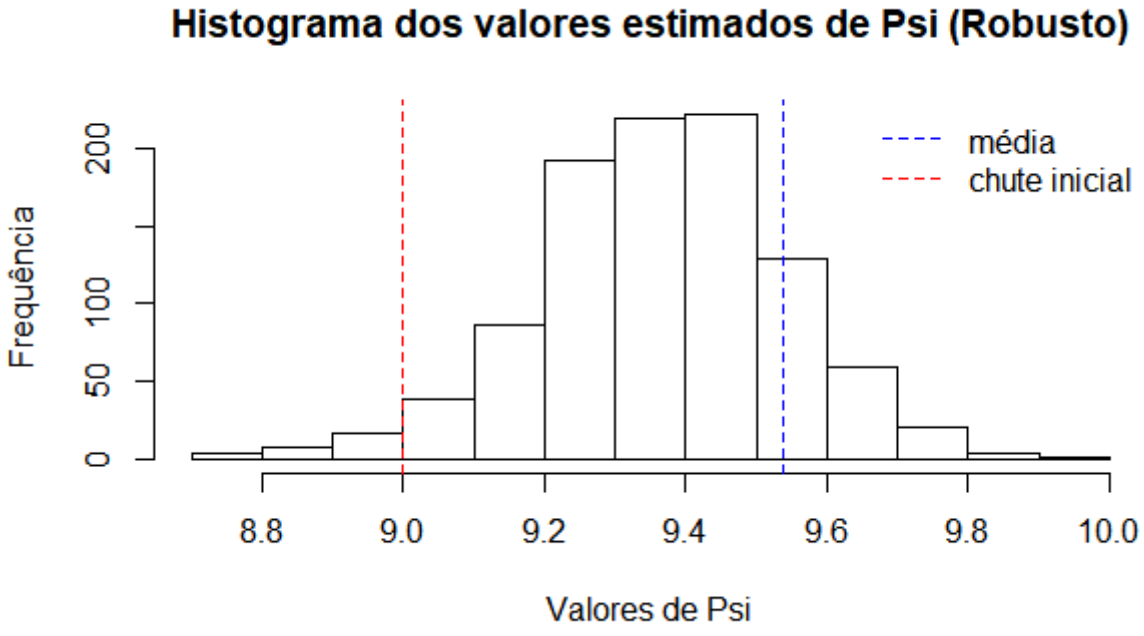

Figura 6: Histograma dos valores estimados de ψ_1 para o método Robusto

A Figura 5 mostra que os resíduos parecem não atender ao pressuposto de normalidade do modelo de regressão linear [Poole and O'Farrell, 1971].

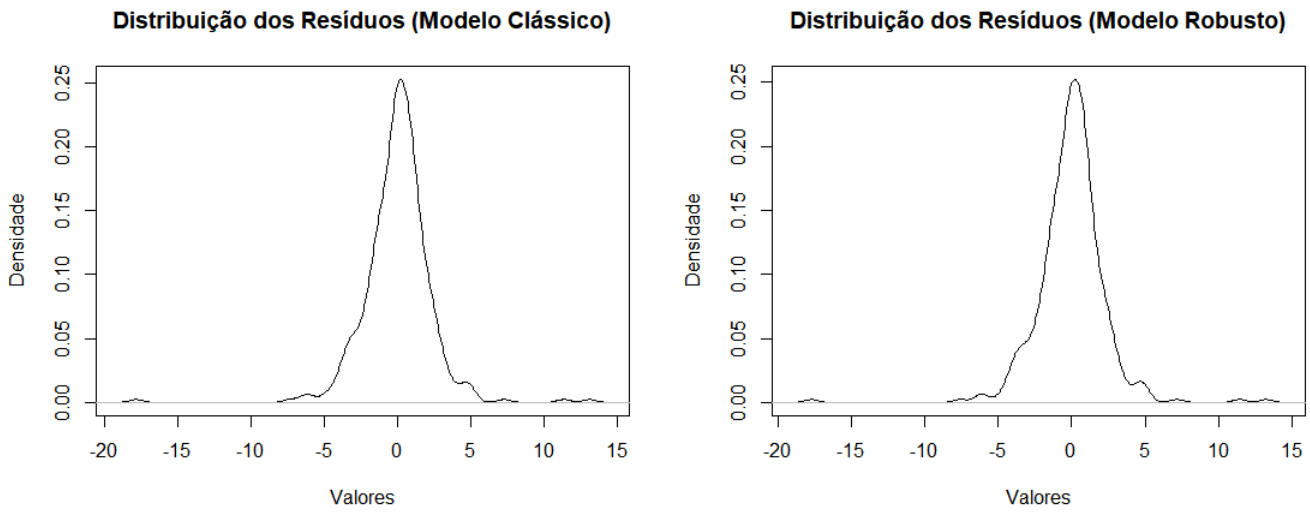

Figura 7: Distribuição dos Resíduos

Foram realizados os testes de Shapiro-Wilk e Jarque-Bera para confirmação do diagnóstico de não normalidade dos resíduos e a hipótese de normalidade foi rejeitada em ambos os modelos (clássico e proposto).

5 Conclusão

Esse artigo apresentou uma abordagem de estimação do modelo de regressão segmentada afim de fornecer estimativas robustas em amostras com outliers. O algoritmo apresentado estende o método apresentado por Muggeo [2003], utilizando Estimadores-M ao invés dos convencionais na estimativa dos parâmetros da regressão. Foram simulados dados com observações discrepantes e realizados diversos experimentos em condições diferentes, comparando os dois métodos confrontados (o clássico e o proposto neste artigo). Logo após, foram realizadas aplicações práticas utilizando dados da fisiologia do exercício inicialmente analisados por Abreu [2017]. Os dados analisados aparentam possuir *outliers* e conseguimos demonstrar empiricamente que o modelo sugerido apresenta melhores resultados quando comparados com o método clássico proposto por Muggeo [2003].

ANEXO A – Códigos do R

```
    Ind<-function(cond=TRUE){
  return(as.numeric(cond))
}

M_segmented<-function(Y, X=NULL, Z, psi0, method='classic'){
  require(lmtest)
  if(method!='classic') require(MASS)
  if(is.null(X)){
    g<-.01
    q<-length(psi0)
    n<-length(Y)
    psi<- psi0
    s<-0
    while((g<.1)&(s<50)){
      s<-s+1
      U<-V<-matrix(nrow = n,ncol = q)
      for(i in 1:n){
        for(k in 1:q){
          U[i,k]<- (Z[i] - psi[k])*Ind(Z[i] > psi[k])
          V[i,k]<- -Ind(Z[i] > psi[k])
        }
      }

      if(method=='classic') fit<-lm(Y ~ Z + U + V) else
      fit<-rlm(Y ~ Z + U + V, psi = psi.huber)
      coeficientes<-fit$coefficients
      beta<-coeficientes[3:(q+2)]
      gama<-coeficientes[(q+3):(2*q+2)]
      zeta0<- coeficientes[1]
      alfa<-coeficientes[2]
      psinovo<- (gama/beta) + psi
      g<-min(coeftest(fit)[(q+3):(2*q+2),4])
      psi <- psinovo
    }
    w<-0
    if(s==50) {warning('Did not converge!'); w<-1}
    result<-c(zeta0,alfa,beta,psi)
    names(result)<-c("zeta_0","alfa",
    kronecker('beta_',1:q,paste,sep=''), kronecker('psi_',1:q,paste,sep=''))
    return(list(result=result, w=w))
  }
  else{
    stop('Not implemented yet!')
  }
}
```

Referências

V. Abreu. Identificação do platô do consumo máximo de oxigênio: revisão dos métodos apresentados na literatura e proposta de uma nova metodologia. Master's thesis, Universidade Federal do Espírito Santo, 2017.

S. Chatterjee and A. S. Hadi. Influential observations, high leverage points, and outliers in linear regression. *Statistical Science*, 1(3):379–393, 1986.

P. I. Feder et al. On asymptotic distribution theory in segmented regression problems–identified case. *The Annals of Statistics*, 3(1):49–83, 1975.

R. Hoffmann and S. Vieira. Análise de regressão. *Uma introdução à econometria*, 2, 1977.

P. J. Huber. *Robust statistics*. Springer, 2011.

V. E. McZgee and W. T. Carleton. Piecewise regression. *Journal of the American Statistical Association*, 65(331):1109–1124, 1970.

M. Muggeo. Segmented mixed models with random changepoints in r. Technical report, Working Paper. 2016. https://www. researchgate. net/publication/292629179 . . . , 2016.

V. M. Muggeo. Estimating regression models with unknown break-points. *Statistics in Medicine*, 22(19):3055–3071, 2003.

V. M. Muggeo. Segmented: an R package to fit regression models with broken-line relationships. *R news*, 8(1):20–25, 2008.

S. C. Patrício and A. J. Q. Sarnaglia. Estimação do modelo de regressão segmentada em dados autocorrelacionados. Available upon request to alessandro.sarnaglia@ufes.br, 2019.

M. A. Poole and P. N. O'Farrell. The assumptions of the linear regression model. *Transactions of the Institute of British Geographers*, pages 145–158, 1971.

P. J. Rousseeuw, S. Van Aelst, K. Van Driessen, and J. A. Gulló. Robust multivariate regression. *Technometrics*, 46(3):293–305, 2004.

S. Serneels, C. Croux, P. Filzmoser, and P. J. Van Espen. Partial robust m-regression. *Chemometrics and Intelligent Laboratory Systems*, 79(1-2):55–64, 2005.

C. Stuart. Robust regression. *Department of Mathematical Sciences, Durham University*, 169, 2011.

H. L. Taylor, E. Buskirk, and A. Henschel. Maximal oxygen intake as an objective measure of cardio-respiratory performance. *Journal of applied physiology*, 8(1):73–80, 1955.