Diagnóstico em Regressão L_1

Kévin Allan Sales Rodrigues

DISSERTAÇÃO APRESENTADA
AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA
OBTENÇÃO DO TÍTULO
DE
MESTRE EM CIÊNCIAS

Programa: Estatística

Orientadora: Prof^a. Dr^a. Silvia Nagib Elia

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro do CNPq

São Paulo, março de 2019

Diagnóstico em Regressão L_1

Esta versão da dissertação contém as correções e alterações sugeridas pela Comissão Julgadora durante a defesa da versão original do trabalho, realizada em 14/03/2019. Uma cópia da versão original está disponível no Instituto de Matemática e Estatística da Universidade de São Paulo.

Comissão Julgadora:

- Prof^a. Dr^a. Silvia Nagib Eliañ (orientadora) - IME-USP
- Prof. Dr. Gustavo Henrique de Araújo Pereira - UFSCAR
- Prof. Dr. Rinaldo Artes - INSPER

Agradecimentos

À minha mãe Verônica, pelo apoio dado durante o meu mestrado.

À USP pelo ambiente amigável e pela excelente estrutura oferecida. Ao CNPq e a USP pelas bolsas concedidas durante o mestrado. A todos os professores que participaram da minha formação acadêmica e que me receberam tão bem neste programa, especialmente aos professores Silvia Elian, Vanderlei Bueno e Márcia Branco. À Regiane pelo excelente serviço prestado na secretaria do programa e por estar sempre disposta a ajudar.

Aos professores do DEMA-UFC, em especial, José Aílton por ter orientado minha monografia durante a graduação; Rafael Farias, que me acompanhou durante minha trajetória na Estatística desde o primeiro ano do curso; João Maurício, não só por ser um excelente professor (o melhor na minha opinião) mas também por ser uma pessoa muito amigável, em especial, pelo apoio, carinho, respeito e amizade, pelos puxões de orelha e principalmente pela paciência que teve comigo durante o curso.

Um agradecimento especial a minha orientadora, Profa. Silvia Nagib Elian, pelo suporte prestado, pelas suas correções, incentivos e principalmente pela paciência durante a orientação deste trabalho.

À minha namorada Dafne, que jamais me negou apoio, carinho e incentivo. Obrigado, amor da minha vida, por ter bastante paciência comigo durante o período final do meu mestrado.

Aos meus amigos e colegas de pós-graduação, em especial, Alia Guirado, Rafael Oliveira, Alessandra, Fernando Henrique, Jônatas Alves, José Sartori, Lucas Prates, Simone Harnik, Daniel Oliveira e Renato Araújo, (por questão de limitação de espaço utilizei uma amostra de tamanho $n = 10$ para representar a população) que fizeram parte da minha formação e que continuarão presentes na minha vida quase certamente.

À Carmen Diva Saldiva de André pela sugestão do uso dos dados da poluição.

Imensamente a todos que contribuíram direta ou indiretamente para o meu desenvolvimento acadêmico e pessoal.

“Os que se encantam com a prática sem a ciência
são como os timoneiros que entram no navio sem
timão nem bússola, nunca tendo certeza do seu des-
tino.”

(Leonardo da Vinci)

Resumo

RODRIGUES, K. A. S. **Diagnóstico em Regressão L_1** . 2019. 100 f. Dissertação (Mestrado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2019.

Este texto apresenta um método alternativo de regressão que é denominado regressão L_1 . Este método é robusto com relação a *outliers* na variável Y enquanto o método tradicional, mínimos quadrados, não oferece robustez a este tipo de *outlier*. Neste trabalho reanalisaremos os dados sobre imóveis apresentados por Narula e Wellington (1977) à luz da regressão L_1 . Ilustraremos os principais resultados inferenciais como: interpretação do modelo, construção de intervalos de confiança e testes de hipóteses para os parâmetros, análise de medidas de qualidade do ajuste do modelo e também utilizaremos medidas de diagnóstico para destacar observações influentes. Dentre as medidas de influência utilizaremos a diferença de verossimilhanças e a diferença de verossimilhanças condicional.

Palavras-chave: Regressão L_1 , *Likelihood displacement*, Medidas de Influência, Métodos de Diagnóstico, Regressão Robusta.

Abstract

RODRIGUES, K. A. S. **Diagnostic in L_1 Regression**. 2019. 100 f. Dissertação (Mestrado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2019.

This text presents an alternative method of regression that is called L_1 regression. This method is robust to outliers in the Y variable while the traditional least squares method does not provide robustness to this type of outlier. In this work we will review the data about houses presented by Narula and Wellington (1977) in the light of the L_1 regression. We will illustrate the main inferential results such as: model interpretation, construction of confidence intervals and hypothesis tests for the parameters, analysis of quality measures of model fit and also use diagnostic measures to highlight influential observations. Among the measures of influence we will use the likelihood displacement and the conditional likelihood displacement.

Keywords: LAD Regression, *Likelihood displacement*, Influence Measures, Diagnostic methods, Robust Regression.

Sumário

Lista de Figuras	ix
Lista de Tabelas	xi
1 Introdução	1
1.1 Nomenclatura da regressão L_1	3
1.2 Objetivo	4
1.3 Organização do Trabalho	4
2 Regressão L_1 - Conceitos	5
2.1 Distribuição Laplace	7
2.2 Estimação dos parâmetros	8
2.3 Norma L_k	12
2.4 Inferência	12
2.4.1 Intervalos de confiança	13
2.4.2 Testes de hipóteses	14
2.5 Relações entre modelos lineares	15
3 Testes Para Verificar a Distribuição dos Erros	19
3.1 Testes baseados na função de distribuição empírica	19
3.2 Testes baseados no resíduo quantílico	23
4 Medidas de Influência e Detecção de Pontos Influentes	27
4.1 Ponto de ruptura e de ajuste perfeito	28
4.2 Algoritmo para detectar <i>outliers</i> em Y e em X	31
4.3 Medidas de influência baseadas na diferença de verossimilhanças	33
4.4 Outras medidas de influência	35
5 Aplicações	39
5.1 Dados sobre imóveis	39
5.2 Dados sobre poluição	50
5.2.1 Termos técnicos	50
5.2.2 Análise	51

6 Conclusão	63
A Algoritmos Para Ajuste E Inferência Na Regressão L_1	65
A.1 Algoritmo para ajuste da regressão L_1	65
A.2 Algoritmo para inferência na regressão L_1	67
A.3 Algoritmo para detectar <i>outliers</i>	71
A.4 Algoritmo para a diferença de verossimilhanças	75
A.5 Algoritmo para a diferença de verossimilhanças condicional	76
A.6 Algoritmo para a distância de Cook	77
A.7 Algoritmo para a medida FD_i	78
A.8 Algoritmo para a medida FD_{1i}	79
A.9 Algoritmo para calcular os resíduos quantílicos	80
Referências Bibliográficas	83

Lista de Figuras

2.1	Funções de distribuição da Laplace com parâmetros (0, 1), (1, 1) e (0, 2).	7
2.2	Gráfico com as distribuições Laplace(0, 1) e normal(0, 1).	8
5.1	Diagramas de dispersão dos dados sobre imóveis.	42
5.2	Gráfico QQ-Plot dos resíduos quantílicos do modelo ajustado para dados sobre imóveis.	44
5.3	Gráfico da diferença de verossimilhanças do modelo ajustado para dados sobre imóveis.	46
5.4	Gráfico da diferença condicional de verossimilhanças do modelo ajustado para dados sobre imóveis.	46
5.5	Gráfico da distância de Cook do modelo ajustado para dados sobre imóveis.	47
5.6	Gráfico da função FD_i do modelo ajustado para dados sobre imóveis.	47
5.7	Gráfico da função FD_{1i} do modelo ajustado para dados sobre imóveis.	48
5.8	Gráfico QQ-Plot dos resíduos quantílicos do modelo ajustado para dados sobre poluição.	56
5.9	Gráfico da diferença de verossimilhanças do modelo ajustado para variável resposta CVF.	59
5.10	Gráfico da diferença de verossimilhanças condicional do modelo ajustado para variável resposta CVF.	59
5.11	Gráfico da distância de Cook do modelo ajustado para variável resposta CVF.	60
5.12	Gráfico da função FD_i do modelo ajustado para variável resposta CVF.	60
5.13	Gráfico da função FD_{1i} do modelo ajustado para variável resposta CVF.	61

Lista de Tabelas

3.1	Pontos críticos da estatística U^2 (Watson).	21
3.2	Pontos críticos da estatística W^2 (Cramér-von Mises).	21
3.3	Pontos críticos da estatística A^2 (Anderson-Darling).	22
3.4	Pontos críticos da estatística $\sqrt{n}D$ (Kolmogorov-Smirnov).	22
3.5	Pontos críticos da estatística $\sqrt{n}V$ (Kuiper).	22
5.1	Dados sobre imóveis.	40
5.2	Estatísticas descritivas com dados sobre imóveis.	41
5.3	Matriz com as correlações entre as variáveis Y, X_1, X_2, X_3 e X_4 .	41
5.4	Estimativas, erros padrão das estimativas, valores-P e limites inferior e superior dos intervalos de confiança (ao nível de 95% de confiança) no modelo ajustado.	42
5.5	Dados sobre imóveis e valores preditos via regressão L_1 .	43
5.6	Análise confirmatória para modelo ajustado com dados sobre imóveis.	49
5.7	Estatísticas descritivas com dados sobre poluição.	52
5.8	Dados sobre poluição.	53
5.9	Matriz com as correlações entre as variáveis resposta CVF, VEF1, VEF1/CVF1, FEF25-75 e PFE e as variáveis explicativas X_1, X_2, X_3, X_4, X_5 e X_8 dos dados sobre poluição.	54
5.10	Tabela descritiva das variáveis X_6 e X_7 .	54
5.11	Estimativas, erros padrão das estimativas, valores-P e limites inferior e superior dos intervalos de confiança (ao nível de 95% de confiança) no modelo ajustado para variável resposta CVF.	55
5.12	Valores preditos via regressão L_1 para variável resposta CVF.	57
5.13	Observações influentes no modelo ajustado para a variável CVF.	58
5.14	Análise confirmatória para modelo ajustado com dados sobre poluição.	62

Capítulo 1

Introdução

Na ampla literatura estatística, podemos encontrar vários trabalhos sobre regressão linear usando o método dos mínimos quadrados para estimação dos parâmetros que foi proposto por Legendre (1805). Tal método apresentará resultado ótimo sob certas condições (homoscedasticidade, erros com caudas mesocúrticas, etc). Porém, em vários casos tais condições não são satisfeitas, como por exemplo: casos onde os erros não possuem distribuição normal, ou possuem caudas pesadas, ou temos a presença de valores aberrantes, nestas situações a regressão de mínimos quadrados não se mostra satisfatória. Neste último caso, mesmo em situações em que as condições sejam satisfeitas, necessitamos de algum procedimento robusto que acomode a presença desses valores aberrantes, ou de técnicas que nos permitam os identificá-los.

Ao investigarmos um pouco mais a literatura estatística que versa sobre o assunto, percebemos que o processo de estimação dos coeficientes de regressão via mínimos quadrados possui um antecedente histórico, que o precede em pelo menos meio século. Tal antecedente, conhecido como regressão da mínima soma dos erros absolutos ou simplesmente regressão L_1 , é considerado uma boa alternativa robusta (Rousseeuw e Leroy, 2003) inclusive para os casos onde as condições para a aplicação do método de mínimos quadrados são satisfeitas. Neste trabalho discorreremos sobre a regressão L_1 , inicialmente mostrando-a como um caso particular da regressão quantílica, posteriormente, apresentando a análise inferencial relativa a estimação intervalar e aos testes de hipóteses. Após apresentarmos os aspectos básicos da regressão L_1 finalmente focaremos no tema desta dissertação que é a análise do ajuste de modelos de regressão L_1 .

O modelo de regressão linear é dado por,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} + \epsilon_i, \quad i = 1, \dots, n \text{ e } p + 1 \leq n,$$

em que Y_i é a variável resposta associada à i -ésima observação, X_{ki} é a k -ésima variável explicativa da i -ésima observação, para $k = 1, 2, \dots, p$; β_0 é o intercepto do modelo, β_k é parâmetro do modelo associado a k -ésima variável explicativa e ϵ_i é o erro associado à

i -ésima observação.

A regressão L_1 é recomendada para casos em que a distribuição dos erros tem caudas pesadas (Bloomfield e Steiger, 1983). Diversos autores (Birkes e Dodge, 1993; Bloomfield e Steiger, 1983; Rousseeuw e Leroy, 2003) consideram-na uma boa alternativa robusta para a regressão de mínimos quadrados. Isto decorre do fato da regressão L_1 possuir menor sensibilidade às observações cujos resíduos absolutos sejam altos quando comparada à regressão de mínimos quadrados.

O método foi introduzido no século XVIII por Boscovich (1757), porém, desenvolveu-se de forma muito lenta quando comparada à regressão via mínimos quadrados, ficando a regressão L_1 por muitos anos em uma espécie de estado de dormência, por motivos que mais adiante serão explicados. Posteriormente, Boscovich (1760) também desenvolveu um algoritmo geométrico para resolver o problema (estimar os coeficientes da regressão L_1). Já a formulação algébrica do algoritmo foi elaborada por Laplace (1793).

Devido ao trabalho de Edgeworth (1887), o problema da regressão L_1 reapareceu e o problema proposto por Boscovich (1757) foi reformulado. Edgeworth (1888) discutiu sobre a não unicidade da solução da regressão L_1 , além de contribuir na demonstração de que os estimadores L_1 são equivalentes aos de máxima verossimilhança quando os erros têm distribuição Laplace, e na elaboração de um algoritmo para determinar a reta da regressão L_1 .

Além do empecilho computacional, a deficiência no ferramental estatístico para inferência se mostrou um problema para o avanço e aplicação da regressão L_1 , se comparado com a regressão de mínimos quadrados (pois os resultados inferenciais são quase sempre assintóticos, ao contrário dos resultados inferenciais exatos dos mínimos quadrados). Todavia, há mais de 30 anos, diversos estudos de Monte Carlo foram desenvolvidos sobre propriedades dos estimadores para pequenas amostras; resultados assintóticos sobre a distribuição dos estimadores e procedimentos para inferência também foram apresentados; algumas aplicações da regressão L_1 são: astronomia (Dodge, 1997); modelagem de diversos dados geofísicos (Afifah *et al.*, 2017); estimação de parâmetros farmacocinéticos (Bezerra, 1990).

Até aqui tecemos breves comentários sobre a história da regressão L_1 , mas para uma abordagem mais completa da história e desenvolvimento da regressão L_1 recomendamos a leitura de Dodge (2002). Uma revisão abrangente dos avanços mais recentes sobre o método de mínima soma dos erros absolutos é apresentada por Dielman (1984, 2005).

Os livros de Bloomfield e Steiger (1983) (*Least Absolute Deviations: Theory, Applications, and Algorithms*), Birkes e Dodge (1993) (*Alternative Methods of Regression*) e Rousseeuw e Leroy (2003) (*Robust Regression and Outlier Detection*) são ótimos para o estudo inicial da regressão L_1 pois apresentam as propriedades, inferência e algoritmos da regressão de mínima soma dos erros absolutos, além de fornecer o contexto histórico do desenvolvimento deste método de regressão. Em especial, a segunda referência,

Birkes e Dodge (1993), apresenta uma comparação bem abrangente entre os mais diversos métodos de regressão (mínimos quadrados, mínima soma de erros absolutos, regressão M, regressão não paramétrica, regressão *Ridge* e Regressão Bayesiana).

1.1 Nomenclatura da regressão L_1

Um detalhe importante da regressão L_1 é que enquanto no método de mínimos quadrados há um consenso a respeito da nomenclatura (a maioria dos livros e artigos modernos sobre mínimos quadrados se refere ao método como “mínimos quadrados”) no método de mínima soma dos erros absolutos há um dissenso (existem diversas nomenclaturas para o método). Felizmente Bloomfield e Steiger (1983) já fizeram uma lista com os termos mais recorrentes e nós transcrevemos e traduzimos tais termos a seguir.

Usamos o seguinte formato: nome em inglês (sigla em inglês) [nome em Português]

- *Absolute Deviation Curve-fitting* (AD) [Ajuste de curva via desvio absoluto];
- *Least Absolute Error* (LAE) [Mínimo erro absoluto];
- *Least Absolute Residual Regression* (LAR) [Regressão via mínimo resíduo absoluto];
- *Least Absolute Value Regression* (LAV) [Regressão via mínimo valor absoluto];
- *Least Deviations* (LD) [Desvios mínimos];
- *Least Total Deviation* (LTD) [Desvio total mínimo];
- *Minimum Absolute Deviation Regression* (MAD) [Regressão via desvio absoluto mínimo];
- *Minimum Deviation* (MD) [Desvio mínimo];
- *Minimum Sum of Absolute Errors* (MSAE) [Soma mínima de erros absolutos];
- *Sum of Absolute Deviations* (SAD) [Soma de desvios absolutos];
- *Sum of Absolute Errors* (SAE) [Soma de erros absolutos];
- *Sum of Absolute Value of Deviations* (SAV) [Soma de valores absolutos dos desvios].

É importante saber que existem várias denominações (e também quais são) para o método de mínima soma dos erros absolutos pois assim é possível fazer a pesquisa de artigos sobre regressão L_1 sem que nenhum artigo escape da busca. A tradução dos termos foi realizada pelo autor deste texto.

1.2 Objetivo

O objetivo deste trabalho é apresentar um método de regressão robusto a *outliers* (na variável Y) bastante intuitivo e com boas propriedades inferenciais como alternativa ao popular método de mínimos quadrados e principalmente incentivar o uso da regressão L_1 . Para atingir esse objetivo apresentamos a teoria inferencial (estimação pontual, intervalos de confiança, testes de hipóteses), testes para verificar se a suposição acerca da distribuição dos erros é adequada, métodos de diagnóstico (para pontos alavanca e observações influentes) e utilizamos as técnicas apresentadas em dois conjuntos de dados. Para tornar ainda mais fácil a aplicação dos métodos discutidos neste texto, criamos o pacote *LadR* (Rodrigues, 2019), que poupará o trabalho de implementação dos métodos apresentados.

1.3 Organização do Trabalho

O Capítulo 2 contém uma breve revisão do processo inferencial na regressão L_1 que inclui o método de estimação pontual dos parâmetros, intervalos de confiança e teste de hipóteses. No Capítulo 3 abordamos testes para verificar se a suposição dos erros seguirem distribuição Laplace foi atendida ou não. Os testes são baseados na função de distribuição empírica e no resíduo quantílico. Discutimos os conceitos de ponto de ruptura e ponto de ajuste perfeito, apresentamos medidas de influência baseadas na diferença de verossimilhanças, na distância de Cook, na exclusão de observações e o algoritmo para detectar observações aberrantes de Dodge (1997) no Capítulo 4. O Capítulo 5 reúne dois conjuntos de dados, um com dados sobre imóveis e outro com dados de poluição, e ilustra a aplicação dos métodos apresentados nos capítulos anteriores.

Os algoritmos dos métodos apresentados e posteriormente aplicados no Capítulo 5 encontram-se no Apêndice A e também no pacote *LadR* (Rodrigues, 2019), que foi desenvolvido simultaneamente com a elaboração desta dissertação. O pacote *LadR* está disponível no CRAN (*Comprehensive R Archive Network*) no endereço <https://cran.r-project.org/package=LadR>, do *software R* (R Core Team, 2018).

Outro pacote que usaremos é o *lawstat* (Gastwirth *et al.*, 2017) que contém uma função chamada *laplace.test* que calcula as cinco estatísticas presentes na Seção 3.1. Estas estatísticas serão úteis para alguns testes de hipóteses acerca da distribuição dos erros.

Capítulo 2

Regressão L_1 - Conceitos

Um dos tipos de modelos de regressão mais difundidos é o modelo linear, neste modelo assume-se que há uma relação linear entre a variável resposta e as variáveis explicativas. A forma do modelo linear é dada na Equação (2.1),

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} + \epsilon_i, \quad i = 1, \dots, n \text{ e } p + 1 \leq n \quad (2.1)$$

em que Y_i é a variável resposta relativa à i -ésima observação, X_{ki} é a k -ésima variável explicativa da i -ésima observação, β_0 é o intercepto do modelo, β_k é o parâmetro associado à k -ésima variável explicativa e ϵ_i é o erro relativo à i -ésima observação. Quando o modelo (2.1) só possui uma variável explicativa, isto é, $p = 1$, é denominado modelo de regressão linear simples.

Vale salientar a matriz $\mathbf{X}$ não precisa ter posto completo como no caso da estimação via mínimos quadrados. Podemos expressar este mesmo modelo na forma matricial, isto é,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

em que $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)^\top$ é o vetor de variáveis resposta com dimensão $n \times 1$, $\mathbf{X} = (\mathbf{X}_1^\top, \mathbf{X}_2^\top, \dots, \mathbf{X}_n^\top)^\top$, em que $\mathbf{X}_i = (1, X_{1i}, X_{2i}, \dots, X_{pi})$ é o vetor com as p variáveis explicativas associadas à i -ésima observação, é a matriz de especificação ou de planejamento do modelo, que contém as variáveis explicativas para as n observações, com dimensão $n \times (p+1)$ no modelo com intercepto (com dimensão $n \times p$ no modelo sem intercepto), $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^\top$ é o vetor com $p+1$ parâmetros de posição do modelo de regressão e $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \dots, \epsilon_n)^\top$ é o vetor $n \times 1$ dos erros aleatórios.

Geralmente também se supõe a homoscedasticidade dos erros (igualdade das variâncias dos erros). Além da homoscedasticidade também suporemos que os erros aleatórios são independentes e identicamente distribuídas com distribuição Laplace(0, λ) (note que λ é o parâmetro de escala do modelo de regressão, veja a Seção 2.5 para mais detalhes).

Existem várias formas de estimar os parâmetros de posição e escala de um modelo linear. Dentre os métodos de estimação se destacam o de mínimos quadrados, máxima verossimilhança e mínima soma dos erros absolutos. Em geral, estes estimadores são distintos e só coincidem em casos bem particulares. Por exemplo, os estimadores de mínimos quadrados coincidem com os estimadores de máxima verossimilhança quando os erros têm distribuição normal e os estimadores de mínima soma dos erros absolutos coincidem com os estimadores de máxima verossimilhança quando os erros têm distribuição exponencial dupla (Laplace).

Cada método de estimação tem suas peculiaridades e suposições próprias e consequentemente propriedades particulares. Por exemplo, o método de máxima verossimilhança fornece propriedades assintóticas úteis (sob as famosas condições de regularidade) mas requer a especificação da distribuição dos erros, ϵ_i , e além disso requer a maximização da função de verossimilhança. Já o método dos momentos não requer a maximização de nenhuma função e em geral é mais fácil de ser implementado, entretanto, de modo geral, não possui boas propriedades assintóticas.

O método de mínimos quadrados tem boas propriedades: estimadores não viesados, fácil resolução (com a suposição de posto completo e homoscedasticidade), não requer a definição da distribuição dos erros e foi amplamente estudado. Mas em geral, a solução de mínimos quadrados é sensível a observações aberrantes (observações aberrantes podem gerar grandes resíduos, em valor absoluto; ou gerar pequenos resíduos, em valor absoluto, porém deslocar a reta de regressão para próximo de si e assim influenciar as estimativas dos parâmetros do modelo e alterar completamente a inferência a partir do modelo ajustado). Já as estimativas via método de mínima soma dos erros absolutos não sofrem desse problema mas em contrapartida, em geral, requerem solução via algoritmos de programação linear.

Além das diferenças nas suposições, no nível de dificuldade de cálculo e propriedades dos estimadores obtidos por estes métodos de estimação, existem as diferenças na interpretação das estimativas dos parâmetros do modelo. Por exemplo, é bem conhecida a interpretação dos $\hat{\beta}_i$'s no modelo ajustado via mínimos quadrados, estima-se que um incremento de uma unidade na variável explicativa X_i gera um incremento $\hat{\beta}_i$ (ou decremento se $\hat{\beta}_i$ for negativo) no valor esperado da variável resposta se as demais covariáveis forem mantidas fixadas. Isto é, $\hat{\beta}_i$ é a estimativa da taxa de variação na média da variável resposta por unidade da i -ésima variável explicativa. De modo análogo, utilizando o estimador de mínima soma de erros absolutos, $\hat{\beta}_i$ é a taxa de variação na mediana da variável resposta por unidade da i -ésima variável explicativa, já que, conforme veremos a seguir, o estimador de β na regressão L_1 coincide com o correspondente estimador no modelo de regressão quantílica para o quantil de ordem 0,5.

2.1 Distribuição Laplace

A função densidade de probabilidade (FDP) da distribuição Laplace é dada por

$$f(x) = \frac{1}{2\lambda} e^{-\frac{|x-\mu|}{\lambda}} \mathbb{I}_{\mathbb{R}}(x), \quad \mu \in \mathbb{R} \text{ e } \lambda > 0.$$

De acordo com [Narula e Wellington \(1977, 1985\)](#) o método de regressão L_1 oferece uma alternativa robusta ao procedimento de mínimos quadrados quando os erros têm distribuição Laplace.

A distribuição Laplace é também conhecida como exponencial dupla devido ao fato de que essa distribuição é ou pode ser obtida como a diferença entre duas variáveis aleatórias com distribuição exponencial com parâmetro $1/\lambda$. Além disso se W tem distribuição Laplace, $|W|$ tem distribuição exponencial. Dentre as principais características da Laplace destacamos: simetria em torno de μ , média, moda e mediana iguais a μ , variância dada por $2\lambda^2$ e curtose 6, ou seja, é leptocúrtica (possui caudas mais pesadas que a distribuição normal). Sua função geradora de momentos e função distribuição acumulada são dadas, respectivamente por

$$M_X(t) = \frac{e^{\mu t}}{1 - (\lambda t)^2}, \quad |t| < \frac{1}{\lambda}$$

e

$$F(x) = \begin{cases} \frac{1}{2} e^{\frac{x-\mu}{\lambda}}, & \text{se } x \leq \mu \\ 1 - \frac{1}{2} e^{-\frac{x-\mu}{\lambda}}, & \text{se } x > \mu \end{cases}.$$

A Figura 2.1 apresenta os gráficos das funções distribuição acumulada das distribuições Laplace(0,1), Laplace(1,1) e Laplace(0,2). O parâmetro de posição, μ , desloca o gráfico à direita (se $\mu > 0$) ou à esquerda (se $\mu < 0$).

Figura 2.1: Funções de distribuição da Laplace com parâmetros (0, 1), (1, 1) e (0, 2).

A Figura 2.2 apresenta as funções densidade de probabilidade das distribuições Laplace e normal nas suas respectivas formas padrão. A diferença entre estas distribuições é nítida, principalmente no ponto zero, que é a média de ambas, sendo que a densidade

da distribuição Laplace apresenta um pico e a densidade da normal uma curva suave. Também nota-se com clareza que as caudas da Laplace têm um decaimento mais lento que as caudas da normal, o que faz com que a distribuição Laplace tenha caudas mais pesadas que a distribuição normal.

Figura 2.2: Gráfico com as distribuições $Laplace(0, 1)$ e $normal(0, 1)$.

Do ponto de vista inferencial, um fato importante é que os estimadores de máxima verossimilhança dos parâmetros de posição, μ , e escala, λ , da Laplace são dados, respectivamente, por

$$\hat{\mu} = \begin{cases} X_{(\frac{n+1}{2})}, & \text{se } n \text{ é ímpar} \\ \frac{X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}}{2}, & \text{se } n \text{ é par} \end{cases}$$

e

$$\hat{\lambda} = \sum_{i=1}^n \frac{|X_i - \hat{\mu}|}{n},$$

em que $X_{(\frac{n+1}{2})}$, $X_{(\frac{n}{2})}$ e $X_{(\frac{n}{2}+1)}$ denotam as estatísticas de ordem $\frac{n+1}{2}$, $\frac{n}{2}$ e $\frac{n}{2} + 1$, respectivamente. Fica evidente que $\hat{\mu}$ é simplesmente a mediana amostral das observações da variável aleatória X .

2.2 Estimação dos parâmetros

Antes de efetivamente introduzirmos o problema de estimação dos parâmetros do modelo de regressão pelo procedimento de mínima soma dos erros absolutos, trataremos brevemente do modelo de regressão quantílico que foi proposto por [Koenker e Bassett \(1978\)](#). Conforme comentado anteriormente, o modelo de regressão L_1 pode ser visto como um caso particular do modelo quantílico pois é o caso quando o quantil de interesse é 0,5, isto é, a mediana.

Seja $Y_i, i = 1, \dots, n$ uma amostra aleatória da variável aleatória Y , com distribuição simétrica em torno de $\beta(\theta)$. Para $0 < \theta < 1$, o θ -ésimo quantil amostral pode ser definido como solução do problema de minimização dado por

$$\min_{\hat{\beta}(\theta) \in \mathbb{R}} \left[\sum_{i \in \{i: Y_i \geq \hat{\beta}(\theta)\}} \theta |Y_i - \hat{\beta}(\theta)| + \sum_{i \in \{i: Y_i < \hat{\beta}(\theta)\}} (1 - \theta) |Y_i - \hat{\beta}(\theta)| \right], \quad \theta \in (0, 1).$$

Seja $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$ o vetor de variáveis resposta do modelo linear, em que Y_i é dado pela Equação (2.1). O estimador de $\beta(\theta)$ no modelo de regressão quantílica de ordem θ é qualquer solução em $\hat{\beta}(\theta)$ do problema de minimização (2.2)

$$\min_{\hat{\beta}(\theta) \in \mathbb{R}^{p+1}} \left[\sum_{i \in \{i: Y_i \geq \mathbf{X}_i \hat{\beta}(\theta)\}} \theta |Y_i - \mathbf{X}_i \hat{\beta}(\theta)| + \sum_{i \in \{i: Y_i < \mathbf{X}_i \hat{\beta}(\theta)\}} (1 - \theta) |Y_i - \mathbf{X}_i \hat{\beta}(\theta)| \right], \quad \theta \in (0, 1), \quad (2.2)$$

em que $\mathbf{X}_i$ é a i -ésima linha da matriz $\mathbf{X}$.

O parâmetro θ pondera a regressão quantílica, ou seja, é tal que pelo menos de $100\theta\%$ dos valores dos Y_i encontram-se acima do hiperplano de regressão quantílica e pelo menos $100(1 - \theta)\%$ abaixo.

Quando $\theta = 1/2$, $\hat{\beta}(\theta) = \hat{\beta}(0, 5)$ é o estimador da regressão L_1 , que é um caso particular importante da regressão quantílica, tendo como medida de posição a mediana amostral. Doravante usaremos as notações β e $\hat{\beta}$ para denotar, respectivamente, $\beta(0, 5)$ e $\hat{\beta}(0, 5)$ devido a importância deste caso particular. Note que neste caso o problema de minimização dado em (2.2) se reduz a

$$\min_{\hat{\beta} \in \mathbb{R}^{p+1}} \left[\frac{1}{2} \sum_{i=1}^n |Y_i - \mathbf{X}_i \hat{\beta}| \right] = \min_{\hat{\beta} \in \mathbb{R}^{p+1}} \left[\sum_{i=1}^n |Y_i - \mathbf{X}_i \hat{\beta}| \right],$$

que é o problema de minimização da regressão L_1 . Eis, portanto, o objeto do nosso estudo, determinar o estimador $\hat{\beta}$ que minimiza a soma dos resíduos absolutos.

Motivados pelo avanço da computação e sua acessibilidade, [Charnes et al. \(1955\)](#) reformularam o método de estimação via mínima soma dos erros absolutos como um problema de programação linear e o solucionaram usando o método simplex. De fato, o método de estimação é equivalente ao seguinte problema de programação linear:

$$\min_{\hat{\beta}} [\mathbf{1}^\top \boldsymbol{\epsilon}^+ + \mathbf{1}^\top \boldsymbol{\epsilon}^-], \quad (2.3)$$

em que $\boldsymbol{\epsilon}^+ = (\epsilon_1^+, \epsilon_2^+, \dots, \epsilon_n^+)^\top$ e $\boldsymbol{\epsilon}^- = (\epsilon_1^-, \epsilon_2^-, \dots, \epsilon_n^-)^\top$, sendo

$$\epsilon_i^+ = \begin{cases} \epsilon_i, & \text{se } \epsilon_i > 0 \\ 0, & \text{se } \epsilon_i \leq 0 \end{cases}, \quad i = 1, 2, \dots, n$$

e

$$\epsilon_i^- = \begin{cases} |\epsilon_i|, & \text{se } \epsilon_i < 0 \\ 0, & \text{se } \epsilon_i \geq 0 \end{cases}, \quad i = 1, 2, \dots, n;$$

sujeito à condição

$$Y = X\hat{\beta} + \epsilon^+ - \epsilon^-,$$

com ϵ^+ e ϵ^- sendo vetores cujos elementos são todos não negativos.

A solução deste problema é o estimador L_1 de β e ϵ^+, ϵ^- são os erros positivos e negativos, respectivamente e $\mathbf{1}^\top = (1, \dots, 1)$ com dimensão $1 \times n$.

É importante ressaltar que este problema não possui solução analítica e portanto é necessário utilizar um método iterativo, diferentemente do método de mínimos quadrados que possui solução analítica. Alguns algoritmos foram propostos para proceder a esta minimização, dentre os algoritmos propostos se destacam os de [Barrodale e Roberts \(1973, 1974\)](#) que se baseiam no método simplex da programação linear. Para mais detalhes sobre programação linear veja [Goldbarg e Luna \(2005\)](#).

O estimador do parâmetro de escala, λ , que será utilizado é o recomendado por [Birkes e Dodge \(1993\)](#) e [Tavares \(1998\)](#), que é resistente a *outliers*, dado por

$$\hat{\lambda} = \sqrt{n^*}(\hat{\epsilon}_{(n^*-m+1)} - \hat{\epsilon}_{(m)})/4,$$

em que n^* é o número de resíduos não nulos (em um modelo com intercepto e p variáveis explicativas $n^* = n - p - 1$), $m = (n^* + 1)/2 - \sqrt{n^*}$ e $\hat{\epsilon}_{(1)}, \dots, \hat{\epsilon}_{(n^*)}$ são as estatísticas de ordem construídas a partir dos n^* resíduos não nulos. Este estimador é fracamente consistente para λ , ou seja, ele converge em probabilidade para λ .

Antes de concluirmos esta seção, mostraremos que de fato os estimadores de máxima verossimilhança de β quando os erros têm distribuição exponencial dupla coincidem com os estimadores de mínima soma dos erros absolutos.

A função densidade de probabilidade (FDP) da distribuição $Laplace(\mu, \lambda)$ é dada por

$$f(x) = \frac{1}{2\lambda} e^{-\frac{|x-\mu|}{\lambda}} \mathbf{I}_{\mathbb{R}}(x), \quad \mu \in \mathbb{R} \text{ e } \lambda > 0.$$

Então a função de verossimilhança é dada por

$$L(\beta, \lambda, \mathbf{Y}) = \prod_{i=1}^n \frac{1}{2\lambda} \exp\left(-\frac{|Y_i - \mathbf{X}_i\beta|}{\lambda}\right) = (2\lambda)^{-n} \exp\left(-\frac{1}{\lambda} \sum_{i=1}^n |Y_i - \mathbf{X}_i\beta|\right),$$

para obtermos os estimadores de máxima verossimilhança precisaremos maximizar esta função com relação aos parâmetros β e λ . Para facilitar este trabalho consideraremos o logaritmo natural da função de verossimilhança (perceba que como o logaritmo é uma

função monótona crescente, então o ponto de máximo da função de verossimilhança coincidirá com o ponto de máximo do logaritmo da função de verossimilhança). Nessas condições

$$l(\beta, \lambda, \mathbf{Y}) = \ln(L(\beta, \lambda, \mathbf{Y})) = -n \ln(2\lambda) - \frac{1}{\lambda} \sum_{i=1}^n |Y_i - \mathbf{X}_i \beta|.$$

Maximizando esta função com relação a β , vemos que o termo $-n \ln(2\lambda)$ é constante, portanto

$$\begin{aligned} \max_{\hat{\beta}} \left[-n \ln(2\lambda) - \frac{1}{\lambda} \sum_{i=1}^n |Y_i - \mathbf{X}_i \hat{\beta}| \right] \\ = \max_{\hat{\beta}} \left[-\frac{1}{\lambda} \sum_{i=1}^n |Y_i - \mathbf{X}_i \hat{\beta}| \right]. \end{aligned}$$

Como $\lambda > 0$, então devemos obter $\hat{\beta}$ de modo a maximizar

$$\left[-\sum_{i=1}^n |Y_i - \mathbf{X}_i \hat{\beta}| \right].$$

É importante perceber que queremos maximizar uma função estritamente negativa, consequentemente, podemos reescrever esse problema de maximização como um problema de minimização, isto é,

$$\min_{\hat{\beta}} \left[\sum_{i=1}^n |Y_i - \mathbf{X}_i \hat{\beta}| \right].$$

Logo vemos que o estimador de máxima verossimilhança de β supondo distribuição exponencial dupla para os erros é equivalente ao estimador de mínima soma de erros absolutos de β . Poderíamos seguir obtendo o estimador de máxima verossimilhança de λ , basta derivar a função $l(\beta, \lambda, \mathbf{Y})$ e igualá-la a zero e resolver a equação com respeito a λ . Posteriormente é necessário mostrar que a função $l(\beta, \lambda, \mathbf{Y})$ é côncava, isto é, mostrar que a segunda derivada do logaritmo da função de verossimilhança é uma função negativa. Assim o estimador de máxima verossimilhança de λ fica dado por

$$\hat{\lambda} = \sum_{i=1}^n \frac{|Y_i - \mathbf{X}_i \hat{\beta}|}{n}.$$

Um fato interessante acerca da estimação via mínima soma de erros absolutos é que, em geral, o número de resíduos nulos é $p + 1$ porque no ajuste via regressão L_1 , na maioria dos casos, exatamente $p + 1$ observações ficam sobre o hiperplano de regressão, existindo situações atípicas em que mais de $p + 1$ resíduos são nulos.

2.3 Norma L_k

Uma norma é uma função que associa cada vetor de um espaço vetorial a um número real não-negativo. O conceito de norma está relacionado à noção geométrica de comprimento. No contexto da regressão linear temos um vetor de variáveis respostas pertencente ao conjunto $\mathbb{R}^n$, em que n é o tamanho da amostra e estamos interessados em obter a reta que possua a menor “distância” possível do vetor de variáveis respostas. Para isso é necessário que esta “distância” a ser minimizada seja definida a princípio.

As normas L_k são usualmente utilizadas como medida de distância de um vetor com relação à origem e são definidas pela Equação (2.4),

$$\|\mathbf{Y}\|_k = \left(\sum_{i=1}^n |Y_i|^k \right)^{1/k}, \quad 1 \leq k, \quad (2.4)$$

em que Y_i é a i -ésima componente do vetor $\mathbf{Y}$. Também é possível definir uma métrica (medida de distância entre dois vetores de mesma dimensão) a partir de uma norma, sendo esta uma das mais importantes aplicações das normas.

Dois casos particulares importantes são as normas L_2 e L_1 que são dadas, respectivamente, por

$$\|\mathbf{Y}\|_2 = \left(\sum_{i=1}^n (Y_i)^2 \right)^{1/2} \quad \text{e}$$

$$\|\mathbf{Y}\|_1 = \sum_{i=1}^n |Y_i|.$$

A norma L_2 é conhecida como distância euclidiana. É importante ressaltar que a reta de mínimos quadrados minimiza a distância euclidiana pois

$$\operatorname{argmin}_{\hat{\mathbf{Y}}} \left(\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \right)^{1/2} = \operatorname{argmin}_{\hat{\mathbf{Y}}} \left(\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \right)$$

devido a monotonicidade da função potência. Enquanto a reta de mínima soma dos desvios absolutos minimiza a norma L_1 , daí é derivado o nome da regressão L_1 . Assim dependendo da norma escolhida como medida de distância, a reta ajustada e as propriedades inferenciais podem mudar. Mais detalhes sobre normas de vetores e matrizes podem ser vistos em [Harville \(1997\)](#), [Searle \(2006\)](#) ou [Seber \(2008\)](#).

2.4 Inferência

Nesta subseção, serão abordados os aspectos inferenciais da regressão L_1 tais como intervalos de confiança e testes de hipóteses para os parâmetros do modelo.

Vários autores analisaram, por meio de estudos de simulação, a eficiência relativa entre o estimador via mínima soma de erros absolutos e o estimador de mínimos quadrados, fazendo uso de diferentes distribuições. Dentre eles destacamos **Rosenberg e Carlson (1977)**. Verificou-se a superioridade (eficiência) do estimador L_1 em relação ao estimador de mínimos quadrados, considerando inclusive tamanhos amostrais reduzidos, quando os erros têm distribuição Cauchy e Laplace. Os autores supracitados obtiveram alguns resultados, via simulação, relativos a propriedades dos estimadores L_1 :

1. O estimador L_1 tem menor desvio padrão que os estimadores de mínimos quadrados para regressão quando os erros têm curtose elevada.
2. Os estimadores L_1 são assintoticamente normais para amostras grandes.
3. A diferença $(\hat{\beta} - \beta)$ em regressão L_1 é assintoticamente normal com média 0 e matriz de variâncias e covariâncias $\lambda^2(\mathbf{X}^\top \mathbf{X})^{-1}$, em que λ^2/n é a variância da mediana em uma amostra de tamanho n .

Tais resultados mostraram que para as distribuições dos ϵ_i para as quais a mediana é assintoticamente mais eficiente do que a média, o estimador L_1 fornece um elipsóide de confiança para β cujo volume é estritamente menor que o volume do elipsóide de confiança obtido pelo método de mínimos quadrados. Por meio da distribuição assintótica dos estimadores L_1 , podemos obter intervalos de confiança e testes de hipóteses no contexto de regressão L_1 .

2.4.1 Intervalos de confiança

Denotemos por $\hat{\beta}$ o estimador de regressão L_1 de β no problema da 0,5-ésima regressão quantílica definida em (2.2). Estamos interessados em um intervalo de confiança para o contraste $\mathbf{h}^\top \beta$, em que $\mathbf{h}$ é um vetor de constantes $(p+1)$ -dimensional e β é o vetor de parâmetros, então o estimador de $\mathbf{h}^\top \beta$ é $\mathbf{h}^\top \hat{\beta}$. Baseado nos resultados assintóticos, podemos construir um intervalo de $(1-\alpha)100\%$ de confiança para $\mathbf{h}^\top \beta$, que é dado pela expressão (2.5):

$$IC_{1-\alpha}(\mathbf{h}^\top \beta) = \left[\mathbf{h}^\top \hat{\beta} - z_{\alpha/2} \lambda \left[\mathbf{h}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{h} \right]^{\frac{1}{2}}; \mathbf{h}^\top \hat{\beta} + z_{\alpha/2} \lambda \left[\mathbf{h}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{h} \right]^{\frac{1}{2}} \right], \quad (2.5)$$

em que $z_{\alpha/2}$ é o quantil de ordem $1-\alpha/2$ da normal padrão e λ é o desvio padrão amostral da mediana dos resíduos.

Para um componente β_i de β , o intervalo de $(1-\alpha)100\%$ de confiança para β_i será dado por

$$IC_{1-\alpha}(\beta_i) = \left[\hat{\beta}_i - z_{\alpha/2} \lambda \left[(\mathbf{X}^\top \mathbf{X})_{ii}^{-1} \right]^{\frac{1}{2}}; \hat{\beta}_i + z_{\alpha/2} \lambda \left[(\mathbf{X}^\top \mathbf{X})_{ii}^{-1} \right]^{\frac{1}{2}} \right], \quad (2.6)$$

em que o $(\mathbf{X}^\top \mathbf{X})_{ii}^{-1}$ representa o i -ésimo elemento da diagonal principal de $(\mathbf{X}^\top \mathbf{X})^{-1}$.

Se λ for desconhecido, o estimaremos. Temos que $\lambda = \theta/2$, em que $\theta = [f(\eta)]^{-1}$ em que $f(\eta)$ é a densidade dos erros aleatórios aplicada na mediana (η é a mediana, neste contexto $\eta = 0$ porque a mediana dos erros é zero por suposição). Para estimarmos λ será necessário estimar $[f(\eta)]^{-1}$. Uma estimativa consistente de $[f(\eta)]^{-1}$, dada por [Cox e Hinkley \(1974\)](#), é:

$$\hat{\theta} = \frac{\hat{\epsilon}_{(i)} - \hat{\epsilon}_{(r)}}{(i - r)/n}, \quad i = 1, \dots, n \text{ e } r = 1, \dots, n,$$

em que denotamos os resíduos observados por $\hat{\epsilon}_i, i = 1, \dots, n$ sendo que $\hat{\epsilon}_i = Y_i - \mathbf{X}_i \hat{\beta}$ e $\hat{\epsilon}_{(i)}$ e $\hat{\epsilon}_{(r)}$ são os resíduos ordenados de ordens i e r , respectivamente.

Além disso, $i = \lceil n/2 \rceil + v$ e $r = \lceil n/2 \rceil - v$, em que $\lceil \cdot \rceil$ indica a função teto (menor inteiro maior ou igual a) e v é escolhido de tal forma que i e r fiquem simétricos em torno da mediana amostral dos resíduos. Recomenda-se que v seja pequeno e $\hat{\epsilon}_{(i)}, \hat{\epsilon}_{(r)} \neq 0$ ([Bezerra, 1990](#)). Verifica-se que $\hat{\theta}$ é um estimador consistente de 2λ , logo uma estimativa consistente para λ é dada por $\hat{\lambda} = \hat{\theta}/2$.

A partir daí, obtém-se o intervalo de $(1 - \alpha)100\%$ de confiança para $\mathbf{h}^\top \beta$ substituindo λ por seu estimador consistente, $\hat{\lambda}$, no intervalo de confiança (2.5).

Os intervalos de confiança apresentados são baseados na normalidade assintótica de $\hat{\beta}$, suposição esta que depende do tamanho amostral e da distribuição das fontes de variação. O método bootstrap tem sido sugerido na construção de intervalos de confiança e construção de testes de hipóteses para pequenas amostras ([Bezerra, 1990](#)). Detalhes acerca do procedimento bootstrap podem ser encontrados em [Efron \(1979\)](#).

2.4.2 Testes de hipóteses

Reproduziremos a seguir os testes de hipóteses assintóticos apresentados por [Bezerra \(1990\)](#).

1. $H_0 : \mathbf{h}^\top \beta = c$ versus $H_1 : \mathbf{h}^\top \beta \neq c$ com nível de significância α . Rejeita-se H_0 se $|Z^*| > z_{1-\alpha/2}$ em que Z^* é dado por

$$Z^* = \frac{\mathbf{h}^\top \hat{\beta} - c}{\hat{\lambda} [\mathbf{h}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{h}]^{\frac{1}{2}}}.$$

2. $H_0 : \beta_i = 0$ versus $H_1 : \beta_i \neq 0, i = 0, \dots, p$ com nível de significância α . Rejeita-se H_0 se $|Z^*| > z_{1-\alpha/2}$ em que Z^* é dado por

$$Z^* = \frac{\hat{\beta}_i}{\hat{\lambda} [(\mathbf{X}^\top \mathbf{X})_{ii}^{-1}]^{\frac{1}{2}}}.$$

3. $H_0 : \beta^* = 0$ versus $H_1 : \exists i \in \{1, \dots, p\}$ tal que $\beta_i \neq 0$ com nível de significância α .
Rejeita-se H_0 se $W > \chi^2_{1-\alpha}(p)$ em que W é dado por

$$W = \frac{(\hat{\beta}^*)^\top \mathbf{X}^{*\top} \mathbf{X}^* (\hat{\beta}^*)}{\hat{\lambda}^2},$$

em que $\mathbf{X}^*$ significa que a primeira coluna de $\mathbf{X}$, que é composta de 1's, foi retirada da matriz e no caso do β^* significa que o primeiro elemento, β_0 , foi retirado do vetor. A estatística do teste, W , possui distribuição assintótica Qui-Quadrado com p graus de liberdade. Note que $\chi^2_{1-\alpha}(p)$ denota o quantil de ordem $(1 - \alpha)$ da distribuição χ^2_p .

Segundo [Bassett e Koenker \(1978\)](#) os três testes apresentados são equivalentes ao teste de Neyman-Pearson (detalhes podem ser vistos na fonte original [Neyman e Pearson \(1933\)](#)) e ao teste da razão de verossimilhanças.

2.5 Relações entre modelos lineares

Até agora, no contexto do modelo linear, comentamos que os estimadores de mínimos quadrados são idênticos aos estimadores de máxima verossimilhança quando se supõe distribuição normal para os erros. Também apresentamos o resultado de que os estimadores via mínima soma dos erros absolutos (L_1) são idênticos aos estimadores de máxima verossimilhança quando os erros têm distribuição Laplace. Ressaltamos ainda que a regressão L_1 é um caso particular da regressão quantílica, quando estamos interessados no quantil de ordem 0,5.

Agora apresentaremos mais uma relação entre modelos. Inicialmente, introduziremos a família de distribuições elípticas (no caso univariado também podem ser chamadas de distribuições simétricas) e então enquadraremos as distribuições normal e exponencial dupla como casos particulares desta família. Outra característica interessante das distribuições normal e exponencial dupla é que, enquanto a distribuição normal pertence à família exponencial, a distribuição exponencial dupla não pertence à família exponencial, veja [Shao \(2005\)](#). Posteriormente apresentaremos brevemente as ideias do modelo linear elíptico e mostraremos que tanto o modelo com erros normais quanto o modelo com erros Laplace pertencem a esta classe.

As densidades da distribuição $N(\mu, \sigma^2)$ e da $Laplace(\mu, \lambda)$ são dadas, respectivamente, por

$$f(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \text{ e}$$

$$f(x, \mu, \lambda) = \frac{1}{2\lambda} e^{-\frac{|x-\mu|}{\lambda}}.$$

Definição 1. (*Família de distribuições elípticas*) Diz-se que a variável aleatória Y pertence a família de distribuições simétricas, com parâmetro de posição μ , e parâmetro de escala θ , se sua densidade é da forma:

$$f(y; \mu, \theta, g(\cdot)) = \frac{1}{\sqrt{\theta}} g(u) \quad , \quad \theta > 0, \quad \mu \in \mathbb{R}, \quad \text{em que } u = \frac{(y - \mu)^2}{\theta}, \quad (2.7)$$

$g(\cdot)$ é a função geradora de densidades satisfazendo $\int_0^{+\infty} u^{-1/2} g(u) du = 1$.

Algumas distribuições elípticas possuem parâmetros adicionais que não são parâmetros de posição nem de escala, esses parâmetros são os parâmetros de forma (um exemplo de distribuição simétrica com parâmetro de forma é t-Student, que tem um parâmetro denominado graus de liberdade que afeta sua curtose). Quando a densidade da distribuição simétrica possuir parâmetro de forma v , o parâmetro afetará a função geradora de densidades, isto é, a função geradora de densidades dependerá do parâmetro de forma,

$$f(y; \mu, \theta, g(\cdot), v) = \frac{1}{\sqrt{\theta}} g(u, v) \quad , \quad \theta > 0, \quad \mu \in \mathbb{R} \text{ e } v > 0.$$

Note que as distribuições normal e Laplace pertencem à família de distribuições elípticas com funções geradoras dadas, respectivamente, por

$$g(u) = \frac{1}{\sqrt{2\pi}} e^{-u/2}, \quad \mu = \mu \text{ e } \theta = \sigma^2$$

e

$$g(u) = \frac{1}{2} e^{-\sqrt{u}}, \quad \mu = \mu \text{ e } \theta = \lambda^2.$$

Então as duas distribuições apresentadas pertencem à família de distribuições elípticas e conseqüentemente os dois modelos apresentados são casos particulares do modelo linear elíptico. Os parâmetros do modelo linear elíptico podem ser estimados via máxima verossimilhança, e a grande vantagem desta formulação mais geral de modelo linear é que não estamos mais restritos à distribuição normal, podemos por exemplo utilizar uma distribuição com caudas mais pesadas que a normal como a Cauchy para modelar os dados.

A abordagem de atribuir uma distribuição com caudas mais pesadas para os erros se torna particularmente útil quando ajustamos um modelo com suposição de normalidade mas os resíduos apresentam um alto grau de curtose. É sabidamente conhecido na literatura estatística que a distribuição normal é mesocúrtica e conseqüentemente a mesma não é adequada para modelar dados com caudas pesadas. Mais detalhes sobre o modelo linear elíptico podem ser vistos em [Cysneiros et al. \(2005\)](#) e para um aprofundamento nas técnicas de diagnóstico, veja [Villegas et al. \(2013\)](#).

A família de distribuições elípticas tem a importante propriedade de ser fechada por combinação linear, isto é, a combinação linear de uma distribuição elíptica também é elíptica. Ou seja, as distribuições elípticas possuem a seguinte propriedade:

se $Y \sim El(\mu, \theta, g(\cdot), \alpha)$, $b \in \mathbb{R}$ e $a \neq 0$ então

$$W = aY + b \sim El(a\mu + b, a^2\theta, g(\cdot), \alpha).$$

Portanto, no caso dos modelos lineares elípticos, a distribuição do erro e a distribuição da variável resposta condicionada às variáveis explicativas pertencem a mesma classe de distribuições. Note que no caso dos modelos lineares generalizados isto não ocorre, pois as distribuições do erro e da variável resposta condicionada às variáveis explicativas não pertencem necessariamente a mesma classe de distribuições. Por isso no contexto dos modelos lineares generalizados se atribui a distribuição diretamente na variável resposta condicionada às variáveis explicativas. Basicamente, dentro da classe de modelos lineares elípticos, a inferência é realizada de forma análoga ao caso do modelo linear normal, isto é, só troca-se a distribuição normal por outra distribuição elíptica qualquer.

Estabelecidas estas últimas relações entre os modelos, retornaremos ao foco deste texto, que é a regressão L_1 .

No próximo capítulo abordaremos métodos para analisar o ajuste de modelos de regressão L_1 . Mais especificamente serão abordados testes para verificar se as suposições acerca dos erros são adequadas, suposições como homoscedasticidade (variância constante dos erros para todas as observações), estrutura aditiva dos erros (isso é importante devido ao fato de supormos um modelo com estrutura aditiva durante todo este texto) e se os erros possuem distribuição Laplace. Esta última suposição pode ser relaxada se o tamanho da amostra for suficientemente grande devido a alguns resultados assintóticos acerca dos estimadores de mínima soma de erros absolutos.

Capítulo 3

Testes Para Verificar a Distribuição dos Erros

Uma das etapas mais importantes da modelagem estatística é verificar se as suposições para o uso do modelo são satisfeitas. No contexto da regressão L_1 , é importante verificar se os erros possuem distribuição Laplace, esta distribuição foi apresentada na Seção 2.1 do Capítulo 2. No capítulo anterior vimos que os estimadores L_1 são equivalentes aos de máxima verossimilhança quando os erros têm distribuição Laplace. Assim, se os erros podem ser modelados como provenientes de uma distribuição Laplace, o estimador de máxima verossimilhança será equivalente ao estimador L_1 , justificando o uso da regressão L_1 .

É importante destacar que a suposição de que os erros possuem distribuição Laplace está para a regressão L_1 da mesma forma que a suposição de normalidade para os erros está para a regressão via mínimos quadrados, em ambos os casos a suposição da distribuição dos erros não é necessária para o ajuste do modelo, entretanto, quando estas suposições são satisfeitas a inferência pode ser realizada com maior segurança (pois se os resíduos apontam que a suposição de distribuição Laplace é adequada para os erros então isto é um indício de que o método de estimação via mínima soma de erros absolutos é adequado). Por isso precisaremos de ferramentas que nos permitam identificar desvios da suposição de distribuição Laplace para os erros no contexto da regressão L_1 .

Nas próximas duas seções abordaremos, respectivamente, testes baseados na função de distribuição empírica (FDE) e no resíduo quantílico.

3.1 Testes baseados na função de distribuição empírica

Puig e Stephens (2000) apresentaram cinco testes para verificar se a distribuição dos dados é Laplace, todos baseados na FDE, que é definida por

$$F_n(k) = \frac{\#\{X_i \leq k\}}{n},$$

em que $\#\{X_i \leq k\}$ é o número de elementos amostrais menores ou iguais a k .

Os testes foram baseados nas estatísticas de qualidade do ajuste de: Cramér-von Mises, Watson, Anderson-Darling, Kolmogorov-Smirnov e Kuiper. Segundo os autores supracitados, o teste mais poderoso, nos restringindo às cinco estatísticas citadas, para testar se os dados têm distribuição Laplace contra a hipótese dos dados terem distribuição simétrica diversa é o baseado na estatística de Watson. Isto ocorre devido ao fato da estatística de Watson detectar melhor divergências com relação à variância da distribuição hipotética (é natural que isso ocorra, pois é esperado que o parâmetro de posição seja o mesmo para todas as distribuições simétricas consideradas, e então a comparação entre as distribuições recairá naturalmente sobre a variância).

O procedimento para realizar os cinco testes é apresentado em quatro passos a seguir:

- 1*) Calcular as estimativas pontuais de máxima verossimilhança dos parâmetros μ e λ da distribuição Laplace a partir da amostra $X_1, \dots, X_n$.
- 2*) Considerar as estatísticas de ordem $X_{(1)}, \dots, X_{(n)}$, em que $X_{(1)}$ é o mínimo e $X_{(n)}$ é o máximo e definir

$$z_{(i)} = F\left(x_{(i)}; \hat{\mu}, \hat{\lambda}\right),$$

em que a notação $F\left(x_{(i)}; \hat{\mu}, \hat{\lambda}\right)$ simboliza a função de distribuição de uma variável aleatória com distribuição Laplace $(\hat{\mu}, \hat{\lambda})$ calculada no ponto $x_{(i)}$.

- 3*) Escolher um teste (estatística do teste) dentre: Cramér-von Mises (W^2), Watson (U^2), Anderson-Darling (A^2), Kolmogorov-Smirnov (D) e Kuiper (V) e calcular a respectiva estatística do teste:

$$W^2 = \sum_{i=1}^n \left[\left[z_{(i)} - (2i-1)/(2n) \right]^2 \right] + \frac{1}{12n},$$

$$U^2 = W^2 - n \left(\bar{z} - \frac{1}{2} \right)^2,$$

em que $\bar{z}$ é a média dos $z_{(i)}$,

$$A^2 = -n - \frac{1}{n} \sum_{i=1}^n (2i-1) \left[\ln(z_{(i)}) + \ln(1 - z_{(n+1-i)}) \right],$$

$$D = \max(D^+, D^-) \text{ e}$$

$$V = D^+ + D^-,$$

em que

$$D^+ = \max_i (i/n - z_{(i)}) \text{ e } D^- = \max_i (z_{(i)} - (i-1)/n).$$

- 4*) Utilizando as Tabelas 3.1 a 3.5, fixando o nível de significância α , comparar a estatística calculada no passo anterior com o valor crítico fornecido pela respectiva tabela do teste escolhido. Caso o valor da estatística seja maior que o valor crítico, rejeita-se a hipótese nula, isto é, descarta-se a hipótese de que os dados são provenientes da distribuição Laplace.

Por exemplo, se escolhermos o teste baseado na estatística de Watson, U^2 , fixarmos o nível de significância em 5%, $\alpha = 0,05$, com o tamanho amostral igual a 20 e $U^2 = 0,085$, hipotéticos, obtemos $U^2 = 0,085 > 0,080 = U_*^2$ a partir da Tabela 3.1; então rejeitamos a hipótese nula.

Tamanho Amostral n	Nível de Significância					
	0,5	0,25	0,1	0,05	0,025	0,01
10	0,038	0,051	0,066	0,078	0,090	0,105
15	0,039	0,053	0,069	0,081	0,092	0,108
20	0,039	0,052	0,068	0,080	0,092	0,108
35	0,039	0,053	0,070	0,083	0,095	0,112
50	0,039	0,053	0,070	0,083	0,094	0,111
75	0,039	0,054	0,071	0,083	0,096	0,113
100	0,039	0,054	0,071	0,083	0,096	0,113
$n > 100$	0,039	0,054	0,072	0,084	0,097	0,114

Fonte: Puig e Stephens (2000).

Tabela 3.1: Pontos críticos da estatística U^2 (Watson).

Tamanho Amostral n	Nível de Significância					
	0,5	0,25	0,1	0,05	0,025	0,01
10	0,049	0,070	0,097	0,119	0,142	0,174
15	0,054	0,080	0,115	0,143	0,172	0,209
20	0,051	0,074	0,107	0,131	0,157	0,192
35	0,054	0,080	0,116	0,144	0,173	0,213
50	0,053	0,078	0,111	0,138	0,165	0,202
75	0,054	0,080	0,116	0,144	0,174	0,213
100	0,053	0,079	0,114	0,141	0,169	0,208
$n > 100$	0,054	0,080	0,116	0,144	0,173	0,214

Fonte: Puig e Stephens (2000).

Tabela 3.2: Pontos críticos da estatística W^2 (Cramér-von Mises).

Se o tamanho da amostra n não estivesse na tabela, bastaria interpolarmos para obter o ponto de corte U_*^2 . Recomenda-se que em caso de interpolação, sejam utilizados os

Tamanho Amostral n	Nível de Significância					
	0,5	0,25	0,1	0,05	0,025	0,01
10	0,366	0,513	0,714	0,869	1,023	1,234
15	0,397	0,570	0,807	0,991	1,160	1,415
20	0,381	0,541	0,760	0,930	1,103	1,336
35	0,396	0,568	0,797	0,987	1,179	1,438
50	0,390	0,554	0,783	0,961	1,137	1,373
75	0,395	0,567	0,797	0,984	1,178	1,442
100	0,394	0,561	0,792	0,972	1,156	1,408
$n > 100$	0,396	0,567	0,798	0,983	1,177	1,442

Fonte: Puig e Stephens (2000).

Tabela 3.3: Pontos críticos da estatística A^2 (Anderson-Darling).

Tamanho Amostral n	Nível de Significância					
	0,5	0,25	0,1	0,05	0,025	0,01
10	0,548	0,645	0,747	0,813	0,871	0,944
15	0,589	0,699	0,812	0,884	0,949	1,026
20	0,582	0,688	0,797	0,868	0,933	1,006
35	0,604	0,716	0,830	0,906	0,973	1,058
50	0,604	0,714	0,828	0,900	0,967	1,048
75	0,613	0,724	0,844	0,920	0,987	1,074
100	0,614	0,725	0,841	0,918	0,985	1,065

Fonte: Puig e Stephens (2000).

Tabela 3.4: Pontos críticos da estatística $\sqrt{n}D$ (Kolmogorov-Smirnov).

Tamanho Amostral n	Nível de Significância					
	0,5	0,25	0,1	0,05	0,025	0,01
10	0,897	1,016	1,133	1,206	1,282	1,373
15	0,921	1,045	1,170	1,251	1,322	1,411
20	0,929	1,054	1,179	1,261	1,333	1,423
35	0,952	1,080	1,210	1,295	1,374	1,467
50	0,959	1,089	1,222	1,305	1,383	1,473
75	0,970	1,101	1,235	1,319	1,395	1,489
100	0,974	1,106	1,240	1,324	1,404	1,501

Fonte: Puig e Stephens (2000).

Tabela 3.5: Pontos críticos da estatística $\sqrt{n}V$ (Kuiper).

tamanhos amostrais tabelados com a mesma paridade do seu n , isto é, se n é par, interpole usando os tamanhos amostrais pares tabelados, ignorando os elementos ímpares, caso n seja ímpar, utilize somente os elementos ímpares (Puig e Stephens, 2000). Segundo os autores, esse cuidado na interpolação decorre do fato das Tabelas de 3.1 a 3.5 terem sido obtidas via método de Monte Carlo (veja o artigo original, Metropolis e Ulam (1949), para mais detalhes) e a velocidade de convergência mudar dependendo da paridade do n , acarretando assim, diferenças para tamanhos amostrais pequenos ($n < 100$).

No Software livre *R* (R Core Team, 2018) existe um pacote voltado para Bioestatís-

tica, Políticas Públicas e Direito denominado *lawstat* (Gastwirth *et al.*, 2017) que contém uma função chamada *laplace.test* que calcula as cinco estatísticas, portanto basta definir o nível de significância desejado, usar esse comando e comparar os valores das estatísticas retornados pelo *R* com os respectivos valores críticos apresentados nas Tabelas 3.1 a 3.5 para obter o resultado do teste de hipóteses.

Como temos interesse em verificar se os erros do modelo de regressão possuem distribuição Laplace então podemos reescrever o procedimento para o nosso problema, os passos são apresentados a seguir.

- 1) Calcular os resíduos “studentizados” $\hat{\epsilon}_1^*, \dots, \hat{\epsilon}_n^*$, dados por

$$\hat{\epsilon}_i^* = \frac{\hat{\epsilon}_i}{\hat{\lambda}}, \quad i = 1, \dots, n,$$

em que $\hat{\epsilon}_1, \dots, \hat{\epsilon}_n$ são os resíduos da regressão L_1 e $\hat{\lambda} = \frac{1}{n} \sum_{i=1}^n |\hat{\epsilon}_i|$.

- 2) Considerar as estatísticas de ordem $\hat{\epsilon}_{(1)}^*, \dots, \hat{\epsilon}_{(n)}^*$, em que $\hat{\epsilon}_{(1)}^*$ é o mínimo e $\hat{\epsilon}_{(n)}^*$ é o máximo e definir

$$z_{(i)} = F(\hat{\epsilon}_{(i)}^*; 0, 1).$$

- 3) Executar o passo 3* como apresentado anteriormente.
- 4) Executar o passo 4* como apresentado anteriormente.

A obtenção da distribuição assintótica das estatísticas dos testes apresentados, sob a hipótese nula, é bastante complexa e pode ser encontrada na Seção 3 de Puig e Stephens (2000).

3.2 Testes baseados no resíduo quantílico

Nesta seção abordaremos o resíduo quantílico (Dunn e Smyth, 1996), que é um caso particular do *crude residual* (resíduo bruto) (Cox e Snell, 1968), e discutiremos suas aplicações na análise do ajuste de modelos de regressão L_1 . Podemos utilizar o resíduo quantílico para averiguar se as suposições a respeito dos erros são adequadas. Devemos checar se os resíduos possuem distribuição Laplace e se a suposição de homoscedasticidade dos erros é razoável.

Os resíduos quantílicos da regressão são calculados a partir dos resíduos “studentizados” apresentados na seção anterior. O cálculo dos resíduos quantílicos no contexto da regressão L_1 é fornecido por

$$r_i = \Phi^{-1}(F(\hat{\epsilon}_i^*; 0, 1)), \quad i = 1, \dots, n,$$

em que $F(\cdot, 0, 1)$ e $\Phi(\cdot)$ denotam, respectivamente, as funções de distribuição acumulada da Laplace(0, 1) e da Normal(0, 1). Antes de realizar esse processo, devem ser excluídos os resíduos nulos, que são $p + 1$ (como visto no capítulo anterior) resíduos num modelo com intercepto e p variáveis explicativas. Consequentemente, para utilizar os resíduos quantílicos, teremos efetivamente $n - p - 1$ resíduos e não n .

Se os resíduos “studentizados” de fato tiverem distribuição Laplace padrão então, assintoticamente, $F(\hat{\epsilon}_i^*; 0, 1) \sim U(0, 1)$ e consequentemente os resíduos quantílicos, $r_1, \dots, r_n$, terão distribuição normal assintótica padrão. Além disso, se os erros, $\epsilon_1, \dots, \epsilon_n$ forem independentes e identicamente distribuídos com distribuição Laplace(0, λ), então é razoável esperar que os resíduos quantílicos tenham, aproximadamente, distribuição normal padrão.

Portanto, a partir dos resíduos quantílicos, podemos utilizar gráficos para verificar se os erros realmente possuem distribuição Laplace a partir de um simples QQ-plot utilizando a distribuição normal padrão como referência. De modo semelhante, podemos verificar se o histograma dos resíduos quantílicos se aproxima da curva em forma de sino característica da distribuição normal. Outro gráfico útil é o dos r_i contra os valores preditos, $\hat{Y}$, que serve para verificar a suposição de homoscedasticidade.

Além das abordagens gráficas, é possível realizar testes de hipóteses para verificar se os resíduos quantílicos possuem distribuição normal padrão, fornecendo assim uma abordagem menos subjetiva do que o gráfico QQ-plot. Cabe frisar que qualquer teste de hipóteses para normalidade pode ser utilizado para testar a normalidade dos resíduos quantílicos. No capítulo de aplicações, utilizaremos o teste de Shapiro-Wilk (Shapiro e Wilk, 1965).

Nos casos em que a suposição de que os erros possuem distribuição Laplace não é satisfeita ainda há um modo de realizar a análise por meio da regressão L_1 . É possível utilizar uma transformação de Box-Cox (veja o artigo original, Box e Cox (1964), para mais detalhes) na variável resposta de modo que a variável resposta transformada tenha distribuição Laplace e o ajuste da regressão L_1 seguirá normalmente a partir desta variável resposta transformada (Parker, 1988). Além disso a transformação de Box-Cox pode ajudar a estabilizar as variâncias dos erros (homoscedasticidade) e também pode ajudar a conseguir a estrutura aditiva dos erros (isso é importante devido ao fato de supormos um modelo com estrutura aditiva), isto é, esta transformação permite resolver diversos problemas acerca das suposições dos erros, de modo análogo ao que ocorre no caso do modelo de regressão linear normal.

No Capítulo 4 serão introduzidos alguns conceitos importantes como: pontos alavanca, pontos influentes, ponto de ajuste perfeito e ponto de ruptura. Neste ponto do texto é importante ressaltar que embora o estimador de mínima soma de erros absolutos forneça robustez a *outliers* na variável resposta Y , ainda podem existir observações com potencial para afetar a inferência no contexto da regressão L_1 . Por isso são necessárias medidas que apontem se existem (e quais são) observações distorcendo os

resultados inferenciais. No próximo capítulo serão abordadas várias medidas de influência que serão úteis para a detecção de observações influentes, como afastamento de verossimilhança, distância de Cook e outras medidas.

Capítulo 4

Medidas de Influência e Detecção de Pontos Influentes

Neste capítulo abordaremos métodos de diagnóstico para detectar *outliers* em X , pontos alavanca e pontos influentes. Na Seção 4.1 abordaremos as definições de ponto de ruptura e de ponto de ajuste perfeito e também apresentaremos a medida para detectar *outliers* em X proposta por Ellis e Morgenthaler (1992). Na Seção 4.2 apresentaremos os algoritmos para detectar *outliers* em Y e em X propostos por Dodge (1997). Na Seção 4.3 apresentaremos as medidas de influência baseadas na diferença de verossimilhanças para a regressão L_1 propostas por Elia et al. (2000). Na última seção deste capítulo apresentaremos outras medidas de influência, que foram propostas por Sun e Wei (2004).

Um resultado importante demonstrado por Dodge (1997) é que o estimador L_1 permanece exatamente o mesmo ainda que se substitua o vetor de variáveis resposta, Y , pelo vetor Z desde que os sinais dos resíduos oriundos da regressão em Z (Z é qualquer vetor de dimensão n , que é o número de observações da amostra utilizada para ajustar o modelo) tenham os mesmos sinais (ou sejam resíduos nulos) que os correspondentes resíduos da regressão em Y , isto é, se $Y_i > \mathbf{X}_i \hat{\beta}$ então Z_i deve satisfazer $Z_i \geq \mathbf{Z}_i \hat{\beta}$ e se $Y_i < \mathbf{X}_i \hat{\beta}$ então Z_i deve satisfazer $Z_i \leq \mathbf{X}_i \hat{\beta}$. Este resultado é particularmente interessante pois nos dá uma noção do quanto o estimador L_1 é robusto em relação aos *outliers* em Y , isto é, podemos perturbar quase livremente (basta respeitar a condição acerca dos sinais dos resíduos) a variável resposta do modelo e ainda sim o estimador L_1 permanecerá inalterado. Perceba ainda que este resultado é válido para qualquer número de variáveis explicativas.

4.1 Ponto de ruptura e de ajuste perfeito

Os estimadores L_1 são robustos à *outliers* em Y ; *outlier* em Y é uma observação ou um conjunto de observações cujas variáveis respostas estão “afastadas” dos respectivos valores preditos a partir de um modelo ajustado sem sua respectiva observação. Entretanto os estimadores mencionados anteriormente não são robustos à *outliers* em X ; *outlier* em X é uma observação ou um conjunto de observações cujos vetores de variável explicativas estão “afastados” do subespaço gerado pela maioria dos vetores de variáveis explicativas.

Note que os *outliers* em X são conhecidos a princípio, pois a rigor as variáveis explicativas são fixas e não aleatórias. Na realidade os *outliers* em X podem ser até mesmo evitados por meio de um planejamento experimental adequado. Naturalmente, se o experimento é planejado previamente então é possível, e desejável, que não exista nenhuma observação cujo vetor de variáveis explicativas esteja “muito afastado” do subespaço gerado pelos vetores de variáveis explicativas associados às demais observações. Então é esperado que existam casos (planejamentos de experimentos) específicos que forneçam dados sem *outliers* em X e conseqüentemente permitam utilizar a regressão L_1 sem necessidade de preocupação de *outliers* em X ou Y .

Antes de prosseguirmos, apresentaremos duas definições importantes no contexto de regressão robusta, extraídas de [Rousseeuw e Leroy \(2003\)](#), ponto de ruptura e ponto de ajuste perfeito.

Definição 2 (Ponto de ruptura). *O ponto de ruptura é dado por*

$$\lim_{n \rightarrow \infty} \varepsilon_n^*(T, Z),$$

em que $\varepsilon_n^*(T, Z)$ é dado por

$$\varepsilon_n^*(T, Z) = \min \left\{ \frac{m}{n}; \text{ tal que } \text{viés}(m, T, Z) = \infty \right\},$$

com *viés* definido por

$$\text{viés}(m, T, Z) = \sup_{Z'} \|T(Z') - T(Z)\|,$$

em que $\|\cdot\|$ denota uma norma (esta norma pode ser escolhida arbitrariamente), Z é a amostra original com tamanho n , $\{(x_1, y_1), \dots, (x_n, y_n)\}$ e Z' é o conjunto de todas as amostras em que exatamente m observações da amostra Z são corrompidas, isto é, m das n observações de Z são trocadas por m observações arbitrárias gerando assim uma amostra corrompida. $T(\cdot)$ é a função que retorna os estimadores dos parâmetros de posição da regressão, ou seja,

$$T(Z) = \hat{\beta}.$$

□

Viés (m, T, Z) igual a infinito significa que m *outliers* podem ter um efeito arbitrariamente grande em T , provocando uma “quebra” no estimador.

Definição 3 (Ponto de ajuste perfeito). *O ponto de ajuste perfeito é dado por*

$$\lim_{n \rightarrow \infty} \delta_n^*(T, Z),$$

em que $\delta_n^*(T, Z)$ é dado por

$$\delta_n^*(T, Z) = \min \left\{ \frac{m}{n}; \exists Z' \text{ tal que } T(Z') \neq \beta \right\},$$

em que Z representa uma amostra de tamanho n , $\{(x_1, y_1), \dots, (x_n, y_n)\}$, tal que $y_i = x_i \beta$, $i = 1, \dots, n$ e Z' é o conjunto de todas as amostras em que exatamente m observações da amostra Z são corrompidas, isto é, m das n observações de Z são trocadas por m observações arbitrárias gerando assim uma amostra corrompida. $T(\cdot)$ é a função que retorna os estimadores dos parâmetros de posição da regressão, ou seja,

$$T(Z) = \hat{\beta}.$$

□

A interpretação do ponto de ruptura é a seguinte: se $100 \cdot \lim_{n \rightarrow \infty} \varepsilon_n^*(T, Z)$ % dos dados forem contaminados então a influência das observações oriundas da contaminação nos estimadores dos β 's é limitada. É bastante razoável esperar que um procedimento robusto tenha um ponto de ruptura alto. Perceba ainda que $\lim_{n \rightarrow \infty} \varepsilon_n^*(T, Z) \in [0, 0,5]$, isto é, o ponto de ruptura máximo é de 50%.

A interpretação do ponto de ajuste perfeito é clara: se $100 \cdot (1 - \lim_{n \rightarrow \infty} \delta_n^*(T, Z))$ % das observações seguir perfeitamente uma relação linear então uma técnica de regressão robusta deve “obter” a equação desta relação linear. É bem intuitivo e razoável esperar que procedimentos robustos sejam capazes de recuperar a equação da relação linear mesmo que a amostra esteja contaminada. Perceba ainda que $\lim_{n \rightarrow \infty} \delta_n^*(T, Z) \in [0, 0,5]$, ou seja, o ponto de ajuste perfeito máximo é de 50%.

Das definições apresentadas observamos que quanto maior o ponto de ruptura, mais robusto é o estimador dos parâmetros do modelo a *outliers* de modo geral. Analogamente, quanto maior o ponto de ajuste perfeito mais robusto é o estimador à presença de *outliers*. Um resultado importante, fornecido por [Rousseeuw e Leroy \(2003\)](#), é a relação

$$\lim_{n \rightarrow \infty} \delta_n^*(T, Z) \geq \lim_{n \rightarrow \infty} \varepsilon_n^*(T, Z),$$

ou seja, o ponto de ajuste perfeito nunca é menor que o ponto de ruptura. Sob condições bem gerais (invariância em regressão e em escala dos estimadores e a escolha de uma

métrica em $\mathbb{R}^n$ que preserve a continuidade uniforme do mapeamento de Y em $\hat{Y}$), Ellis e Morgenthaler (1992) demonstraram que vale a relação

$$\lim_{n \rightarrow \infty} \delta_n^*(T, Z) = \lim_{n \rightarrow \infty} \varepsilon_n^*(T, Z),$$

ou seja, o ponto de ruptura e o ponto de ajuste perfeito são iguais. Um resultado importante e que deve ser mencionado neste momento é que podemos induzir uma métrica por meio de uma norma qualquer, podemos por exemplo desenvolver uma métrica a partir da norma L_1 , isto é, partindo de uma norma qualquer podemos obter exatamente uma métrica. Vale ressaltar que uma métrica é simplesmente uma generalização da noção usual de distância para espaços mais gerais como o $\mathbb{R}^n$ por exemplo. Do resultado geral expresso na última equação apresentada, os autores supracitados verificaram que o estimador L_1 sai como caso particular, pois a métrica induzida pela norma L_1 preserva a continuidade uniforme do mapeamento de Y em $\hat{Y}$. Portanto no contexto da regressão linear com estimadores L_1 , o ponto de ruptura é exatamente igual ao ponto de ajuste perfeito.

Ellis e Morgenthaler (1992) propuseram uma medida capaz de indicar observações ou grupos de observações alavancas (*outliers* em X). Por meio desta medida de alavancagem os autores também foram capazes de ilustrar situações (planejamentos de experimentos) em que a regressão L_1 atinge o ponto de ruptura de 25%. A medida de alavancagem, h_{ii}^* , que os autores propuseram é dada pela Equação (4.1),

$$h(I) = \max_{f \in F} \frac{|f_{i_1}| + \cdots + |f_{i_m}|}{|f_1| + \cdots + |f_n|}, \quad (4.1)$$

em que $I = \{i_1, \dots, i_m\}$ é um conjunto com m números distintos pertencentes ao conjunto $\{1, \dots, n\}$ e $F = \langle X e_1, \dots, X e_p \rangle$ (onde e_i , $i = 1, \dots, p$, são os p vetores da base canônica do $\mathbb{R}^p$), ou seja, F é o subespaço gerado pelas colunas da matriz X ; f é um vetor real de dimensão n pertencente ao subespaço F , isto é, $f = \sum_{i=1}^p c_i X e_i$, com $(c_1, \dots, c_p) \neq 0$ e f_i é o i -ésimo elemento do vetor f . Caso seja necessário calcular a medida de alavancagem de uma única observação, a medida é simplificada para

$$h(i) = \max_{f \in F} \frac{|f_i|}{|f_1| + \cdots + |f_n|}.$$

Ellis e Morgenthaler (1992) destacaram que o estimador L_1 dos parâmetros de posição de um modelo de regressão linear múltiplo com intercepto, cujo planejamento de experimentos fornece p variáveis explicativas binárias assumindo valores em $\{-1, 1\}^p$ atinge ponto de ruptura de 25%. Isto significa dizer que é possível alterar arbitrariamente até um quarto das observações sem que as estimativas dos parâmetros do modelos se afastem das estimativas do modelo sem estas perturbações.

Além disso, os autores supracitados sugerem que estimadores L_1 ponderados sejam utilizados nos casos em que há um planejamento assimétrico (a ideia aqui é atribuir

pesos menores às observações cujas variáveis explicativas estejam mais afastadas do subespaço gerado pelas variáveis explicativas associadas à maioria das observações) de forma a aumentar o ponto de ruptura dos estimadores tornando-os mais robustos a *outliers* em X .

Ellis e Morgenthaler (1992) ainda salientam que podem existir planejamentos de experimentos mais favoráveis à regressão L_1 , fornecendo um ponto de ruptura ainda mais próximo de 50% (que é o ponto de ruptura máximo), portanto esta ainda é uma área aberta para futuras pesquisas.

4.2 Algoritmo para detectar *outliers* em Y e em X

Um método de detecção de *outliers* em Y e em X bem interessante foi proposto por Dodge (1997). Intuitivamente a ideia é trocar de lugar uma das variáveis explicativas com a variável resposta, ajustar a regressão como de costume e usar o fato de que a regressão L_1 é robusta aos *outliers* em Y e é capaz de detectá-los (por meio de altos valores absolutos atribuídos aos resíduos associados a estas observações), repete-se esse procedimento para cada variável explicativa. Então é possível detectar observações alavanca por meio deste procedimento. A seguir apresentaremos o algoritmo desse procedimento para o caso de uma regressão linear simples e depois o algoritmo mais geral que serve para um modelo com p variáveis explicativas.

Algoritmo 1 (modelo linear simples)

Passo 1 Ajuste a reta da regressão L_1 normalmente e estime o parâmetro de escala λ por meio de

$$\hat{\lambda} = 1,4826 \cdot MRADZ,$$

em que MRADZ é a mediana dos resíduos absolutos diferentes de zero. Este estimador é sugerido por Dodge (1997) pelo fato de que mesmo se os erros possuírem distribuição normal, $\hat{\lambda}$ ainda seja um estimador consistente de λ .

Passo 2 Para todas as n observações calcule o respectivo resíduo padronizado $\hat{\epsilon}_i/\hat{\lambda}$, em que $\hat{\epsilon}_i$ é o resíduo associado a i -ésima observação, e remova todas as observações cujo valor absoluto do respectivo resíduo padronizado seja maior ou igual a 2,5, isto é,

$$\left| \hat{\epsilon}_i/\hat{\lambda} \right| \geq 2,5.$$

Passo 3 Troque de lugar a variável explicativa e a variável resposta e ajuste a reta de regressão L_1 com as observações que não foram eliminadas no passo anterior e calcule novamente o estimador do parâmetro de escala.

Passo 4 Repita o procedimento do passo 2 eliminando da análise as observações cujo valor absoluto do respectivo resíduo padronizado seja maior ou igual a 2,5.

Passo 5 Com as observações que restaram após as sucessivas eliminações de *outliers* realizadas nos passos anteriores, ajuste a regressão L_1 usual utilizando a variável resposta e a variável explicativa nos seus respectivos lugares.

Uma observação importante é que as observações removidas no passo 2 são *outliers* em Y e as removidas no passo 4 são *outliers* em X , ou seja, além de discriminar os *outliers*, este método ainda permite a classificação dos mesmos. Outro detalhe fornecido pelo autor é que é possível adaptar o algoritmo de forma a torná-lo invariante com relação às observações excluídas do modelo final, ou seja, a ordem em que os *outliers* em Y e em X são excluídos não afetaria quantas e quais observações são eliminadas no modelo final. Para tanto é sugerido que os *outliers* em Y destacados no passo 2 sejam mantidos e ao final de passo 4 os mesmos sejam excluídos juntamente com os *outliers* em X destacados.

Algoritmo 2 (modelo linear múltiplo)

Passo 1 Ajuste a reta da regressão L_1 normalmente e estime o parâmetro de escala λ por meio de

$$\hat{\lambda} = 1,4826 \cdot MRADZ,$$

em que MRADZ é a mediana dos resíduos absolutos diferentes de zero.

Passo 2 Para todas as n observações calcule o respectivo resíduo padronizado $\hat{\epsilon}_i/\hat{\lambda}$, em que $\hat{\epsilon}_i$ é o resíduo associado a i -ésima observação.

Passo 3 Remova todas as observações cujo valor absoluto do respectivo resíduo padronizado seja maior ou igual a 2,5, isto é,

$$\left| \hat{\epsilon}_i/\hat{\lambda} \right| \geq 2,5.$$

Passo 4 Para $j = 1, \dots, p$ use a j -ésima variável explicativa do modelo original como a variável resposta e ajuste o hiperplano de regressão L_1 usando todas as demais variáveis (tanto a resposta quanto as demais variáveis explicativas do modelo original) com as observações que não foram eliminadas no passo 3 e calcule o estimador do parâmetro de escala $\hat{\lambda}_j$.

Passo 5 Para $j = 1, \dots, p$ calcule o i -ésimo resíduo da regressão em que a variável resposta é a j -ésima variável explicativa do modelo original, $\hat{\epsilon}_{ij}/\hat{\lambda}_j$. Defina D como

$$D = \left\{ i : \exists j \text{ tal que } \left| \hat{\epsilon}_{ij} / \hat{\lambda}_j \right| \geq 2,5 \right\},$$

ou seja, D é o conjunto de todas as observações cujo valor absoluto do seu respectivo resíduo padronizado foi maior ou igual a 2,5 em pelo menos uma das p regressões em que a variável resposta é a j -ésima variável explicativa do modelo original.

Passo 6 Remova todas as observações cujo índice esteja contido em D .

Passo 7 Com as observações que restaram após as sucessivas eliminações de *outliers* realizadas nos passos 3 e 6 ajuste a regressão L_1 usual utilizando a variável resposta e as variáveis explicativas nos seus respectivos lugares.

De forma análoga ao que ocorre no caso do algoritmo para o modelo linear simples, as observações removidas no passo 3 são *outliers* em Y e as removidas no passo 6 são *outliers* em X , ou seja, além de discriminar os *outliers*, este método ainda permite a classificação dos mesmos. Como no algoritmo para a regressão linear simples, é possível adaptar o algoritmo de forma a torná-lo invariante com relação às observações excluídas do modelo final, isto é, a ordem em que os *outliers* em Y e em X são excluídos não afetaria quantas e quais observações são eliminadas no modelo final. Para tanto é sugerido que os *outliers* em Y destacados no passo 3 sejam mantidos e ao final de passo 6 os mesmos sejam excluídos juntamente com os *outliers* em X destacados.

4.3 Medidas de influência baseadas na diferença de verossimilhanças

Nesta seção, *outlier* em X e ponto alavanca são sinônimos. Os pontos ainda podem ser classificados em duas categorias: bom ponto de alavanca e mau ponto de alavanca. Um bom ponto de alavanca é uma observação cujo vetor de variáveis explicativas está afastado do subespaço gerado pelas variáveis explicativas da maioria das observações e cuja variável resposta está próxima do seu respectivo valor predito pelo modelo.

Um mau ponto de alavanca é uma observação cujo vetor de variáveis explicativas está afastado do subespaço gerado pelas variáveis explicativas da maioria das observações e cuja variável resposta está longe do seu respectivo valor predito pelo modelo. Neste caso, um ponto alavanca pode chegar a ser tão ruim ao ponto em que a inferência a partir do modelo ajustado com todas as observações seja completamente diferente da inferência realizada com a exclusão da observação influente. Isto é, um ponto de alavanca ruim pode alterar fortemente intervalos de confiança para os parâmetros do modelo, alterar o resultado de testes de hipóteses, mudar drasticamente as estimativas

dos parâmetros, etc. Doravante denominaremos estes pontos nocivos que podem alterar a inferência (conclusões) a partir do modelo ajustado de pontos influentes.

Nesta seção abordaremos medidas de diagnóstico para destacar observações influentes. As medidas de influência apresentadas são baseadas na diferença de verossimilhanças e na diferença de verossimilhanças condicional que foram métodos de diagnóstico propostos por [Cook et al. \(1988\)](#).

Dentre as vantagens da diferença de verossimilhança como medida de influência pode-se destacar que esta medida não depende da estrutura linear do modelo, ou seja, pode ser utilizada no contexto de modelos não lineares; além disso é possível utilizar uma escala mais familiar como os percentis de uma distribuição Qui-quadrado com k graus de liberdade, em que k é o número de resíduos não nulos.

Uma medida de influência é uma função dos dados capaz de destacar observações influentes. [Elian et al. \(2000\)](#) propuseram a diferença de verossimilhanças ($LD_i(\theta)$) no contexto da regressão L_1 , que é uma medida de influência baseada na diferença entre o logaritmo da função de verossimilhança aplicada na estimativa de máxima verossimilhança (EMV) ($L(\hat{\theta}, \mathbf{Y})$) e o logaritmo da função de verossimilhança aplicada na EMV obtido sem a i -ésima observação ($L(\hat{\theta}_{(i)}, \mathbf{Y})$),

$$LD_i(\theta) = 2[L(\hat{\theta}, \mathbf{Y}) - L(\hat{\theta}_{(i)}, \mathbf{Y})]$$

que, no caso do modelo de regressão L_1 com variável resposta com distribuição Laplace se reduz a

$$2 \left(n \ln \left(\hat{\lambda}_{(i)} / \hat{\lambda} \right) + |Y_i - X_i \hat{\beta}_{(i)}| / \hat{\lambda}_{(i)} - 1 \right),$$

em que $\theta = (\beta^\top, \lambda)^\top$, λ é o parâmetro de escala da distribuição Laplace, $\hat{\beta}$ e $\hat{\lambda}$ são os estimadores de máxima verossimilhança de β e λ , além disso $\hat{\beta}_{(i)}$ e $\hat{\lambda}_{(i)}$ são os estimadores de máxima verossimilhança de β e λ , excluindo-se a i -ésima observação da amostra. Se a i -ésima observação não é influente então, a estimativa $\hat{\lambda}_{(i)}$, calculada sem a i -ésima observação, será próximo de $\hat{\lambda}$, de modo que $\hat{\lambda}_{(i)} / \hat{\lambda} \approx 1$. Além disso, $|Y_i - X_i \hat{\beta}_{(i)}|$ deve ser próximo da média dos erros absolutos, $\hat{\lambda}_{(i)}$, ou seja, $|Y_i - X_i \hat{\beta}_{(i)}| \approx \hat{\lambda}_{(i)}$ e como consequência, LD_i deve ser próxima de zero.

Além dessa medida, [Elian et al. \(2000\)](#) apresentaram o afastamento condicional das verossimilhanças ($LD_i(\beta|\lambda)$), que serve para avaliar a influência de uma observação na estimativa do vetor de parâmetros β ,

$$\begin{aligned} LD_i(\beta|\lambda) &= 2[L(\hat{\beta}, \hat{\lambda}, \mathbf{Y}) - \max_{\lambda} L(\hat{\beta}_{(i)}, \lambda, \mathbf{Y})] \\ &= 2n \ln \left(\frac{\sum_{j=1}^n |Y_j - X_j \hat{\beta}_{(i)}|}{\sum_{j=1}^n |Y_j - X_j \hat{\beta}|} \right). \end{aligned}$$

Como vimos nas seções anteriores, a regressão L_1 oferece robustez com relação aos

outliers na variável resposta, mas não fornece robustez aos pontos influentes. Por isso é importante utilizar métodos para detectar observações influentes como a diferença de verossimilhanças e a diferença de verossimilhanças condicional.

4.4 Outras medidas de influência

Nesta seção primeiramente apresentaremos as medidas de diagnóstico publicadas por Narula e Wellington (1985) e depois abordaremos algumas medidas de influência propostas por Sun e Wei (2004) no contexto da regressão L_1 como a distância de quase-verossimilhança e a distância de Cook. Cook (1977) apresenta esta medida para o modelo de regressão linear no contexto da estimativa de mínimos quadrados. Embora estas medidas de influência possam ser utilizadas tanto no caso de modelos lineares quanto não-lineares focaremos exclusivamente no contexto de modelos lineares devido ao escopo deste texto.

Narula e Wellington (1985) propuseram duas medidas para avaliar a influência da i -ésima observação no modelo. Uma que mensura a influência da i -ésima observação na estimativa do j -ésimo parâmetro de posição do modelo, $\hat{\beta}_j$, definida por

$$\hat{\beta}_j - \hat{\beta}_j^{(i)}, \quad j = 0, \dots, p, \quad i = 1, \dots, n$$

e outra que mede a influência da i -ésima observação no seu próprio ajuste, definida por

$$\hat{\epsilon}_i - \hat{\epsilon}_i^{(i)}, \quad i = 1, \dots, n,$$

em que $\hat{\epsilon}_i$ é o resíduo associado à i -ésima observação e $\hat{\epsilon}_i^{(i)}$ é o erro de predição de Y_i utilizando o modelo ajustado com todas as observações com exceção da i -ésima, isto é,

$$\hat{\epsilon}_i^{(i)} = Y_i - \mathbf{X}_i \hat{\beta}_{(i)},$$

sendo $\hat{\beta}_{(i)}$ o estimador de mínima soma dos erros absolutos de β excluindo a i -ésima observação da amostra.

As medidas de influência propostas por Narula e Wellington (1985) são de simples compreensão, porém são pouco úteis nos casos em que existem muitas covariáveis ou muitas observações no modelo. Assim há a necessidade de uma única medida que leve em consideração a influência de uma observação em todo o ajuste do modelo, ou por exemplo, na estimativa do vetor β .

Abordaremos a seguir as medidas de influência propostas por Sun e Wei (2004), que permitem uma visão mais objetiva da influência de uma observação na estimativa dos parâmetros. Antes de apresentarmos as medidas propriamente ditas, destacaremos os dois requisitos necessários para que haja base teórica para as medidas de influência supracitadas, segundo Sun e Wei (2004):

1. $F(0) = \frac{1}{2}$, ou seja, mediana nula,
2. $F'(x) = f(x)$ é contínua na vizinhança de zero e $f(0) > 0$,

em que $F(\cdot)$ e $f(\cdot)$ são, respectivamente, a função de distribuição e a função densidade de probabilidade dos erros.

Naturalmente estas suposições são atendidas quando se atribui a distribuição Laplace para os erros pois esta distribuição é simétrica em torno do zero, consequentemente $F(0) = \frac{1}{2}$, e a densidade é contínua o que naturalmente implica na continuidade da função densidade de probabilidade na vizinhança de zero.

A distância de Cook para um modelo linear no contexto da regressão L_1 é dada por

$$D_i = \lambda^{-2} \left(\hat{\beta} - \hat{\beta}_{(i)} \right)^\top \mathbf{X}^\top \mathbf{X} \left(\hat{\beta} - \hat{\beta}_{(i)} \right),$$

em que $\lambda = (2f(0))^{-1}$, sendo $f(\cdot)$ a função densidade dos erros, e $\hat{\beta}_{(i)}$ é a estimativa de β sem a i -ésima observação na amostra.

Duas medidas de influência unidimensionais e intuitivas propostas pelos mesmos autores para a regressão L_1 são FD_i e FD_{1i} . Estas medidas representam a variação no vetor de valores ajustados pelo modelo com e sem a i -ésima observação. A diferença entre as duas medidas é que FD_i é construída com base na norma L_2 enquanto FD_{1i} se baseia na norma L_1 . A medida FD_i é definida por

$$FD_i = \lambda^{-2} \left\| \mathbf{X} \hat{\beta} - \mathbf{X} \hat{\beta}_{(i)} \right\|_2,$$

em que λ é o parâmetro de escala definido anteriormente.

Na análise de dados, [Sun e Wei \(2004\)](#) sugerem trocar λ por $\hat{\lambda} = (n - p)^{-1} \sum_{i=1}^n |\hat{\epsilon}_i|$ ou qualquer estimativa razoável de λ , pois o interesse é destacar as observações mais influentes da amostra e não o valor descritivo da medida de influência. Neste texto utilizaremos a estimativa robusta a *outliers* proposta por [Birkes e Dodge \(1993\)](#) dada por

$$\hat{\lambda} = \sqrt{n^*} (\hat{\epsilon}_{(n^*-m+1)} - \hat{\epsilon}_{(m)}) / 4,$$

em que n^* é o número de resíduos não nulos (em um modelo com intercepto e p variáveis explicativas $n^* = n - p - 1$), $m = (n^* + 1)/2 - \sqrt{n^*}$ e $\hat{\epsilon}_{(1)}, \dots, \hat{\epsilon}_{(n^*)}$ são as estatísticas de ordem construídas a partir dos n^* resíduos não nulos.

A medida FD_{1i} é dada por

$$FD_{1i} = \lambda^{-1} \left\| \mathbf{X} \hat{\beta} - \mathbf{X} \hat{\beta}_{(i)} \right\|_1.$$

De modo análogo à medida de influência FD_i é necessário utilizar uma estimativa para o parâmetro λ .

A medida FD_{1i} reflete a influência total da exclusão da i -ésima observação no ajuste do modelo. As medidas de influência FD_i e FD_{1i} são igualmente adequadas para problemas em que a predição é mais importante que a estimação e estas medidas são invariantes com relação a parametrização (Cook e Weisberg, 1982).

Sun e Wei (2004) também aplicaram o conceito de distância de quase-verossimilhança ao contexto da regressão L_1 . A medida de distância de quase-verossimilhança, DQV_i , é dada por

$$DQV_i = 4f(0) \left[\left\| \mathbf{Y} - \mathbf{X}\hat{\beta}_{(i)} \right\|_1 - \left\| \mathbf{Y} - \mathbf{X}\hat{\beta} \right\|_1 \right],$$

no caso da regressão L_1 a distância de verossimilhança é um importante caso particular da distância de quase-verossimilhança.

A ideia geral da distância de quase-verossimilhança é que a influência da i -ésima observação no estimador de $\hat{\beta}$ pode ser mensurada por $DQV_i = C \left[R(\hat{\beta}_{(i)}) - R(\hat{\beta}) \right]$, em que $R(\beta)$ denota uma função objetivo relacionada ao método de estimação baseada na amostra completa e C é uma constante positiva. Se a função objetivo $R(\beta)$ é o logaritmo da função de verossimilhança sob o método de estimação via máxima verossimilhança, então a distância de quase-verossimilhança se reduz à distância de verossimilhança.

Como observação final é importante salientar que devido ao interesse em detectar observações que se destacam nas medidas de diagnóstico, praticamente qualquer método de estimação de λ pode ser utilizado para calcular as medidas de diagnóstico abordadas em um conjunto de dados real.

No próximo capítulo apresentaremos a aplicação da regressão L_1 em dois conjuntos de dados. Essas aplicações ilustrarão as medidas de influência apresentadas neste capítulo além de ilustrar o que foi discutido nos capítulos anteriores.

Capítulo 5

Aplicações

Neste capítulo ilustraremos as técnicas que foram expostas nos capítulos anteriores, como: ajuste da regressão L_1 , intervalos de confiança e testes de hipóteses para os parâmetros do modelo, interpretação do modelo ajustado. Após a etapa inferencial será feita a análise do ajuste do modelo. Essa nova etapa é feita por meio das medidas de alavancagem e influência apresentadas anteriormente e também por meio de testes de hipóteses e gráficos para verificar se as suposições do modelo L_1 são satisfeitas. Essa análise será feita em dois conjuntos de dados que serão apresentados nas seções a seguir.

Como os estimadores de mínima soma dos erros absolutos não possuem solução analítica então são necessários métodos de otimização. Logo é inviável ajustar um modelo de regressão L_1 sem o auxílio de um computador. O *software* livre *R* (R Core Team, 2018) será utilizado para ajustar os modelos de regressão e realizar todos os procedimentos que foram apresentados nos capítulos anteriores. O *R* foi escolhido pois ele é gratuito e *open source* código aberto, isso significa que qualquer pessoa pode instalá-lo no computador sem pagar quaisquer tipos de taxa e ainda ter acesso à todos os algoritmos do programa e até mesmo construir seus próprios algoritmos e distribuí-los por meio de pacotes. Utilizaremos o pacote *LadR* (Rodrigues, 2019) para o ajuste e medidas de influência de modelos de regressão L_1 e o pacote *lawstat* (Gastwirth *et al.*, 2017) para os testes acerca da distribuição dos erros do modelo. Todos os algoritmos usados nesta dissertação estão disponíveis no Apêndice A e no pacote *LadR* (Rodrigues, 2019).

5.1 Dados sobre imóveis

Nesta seção reanalisaremos os dados sobre imóveis apresentados no trabalho de Narula e Wellington (1977) à luz da regressão L_1 . Os autores supracitados publicaram originalmente as 27 observações sobre imóveis contidas na Tabela 5.1. As variáveis explicativas X_1 , X_2 , X_3 e X_4 são, respectivamente, imposto (em centenas de dólares), área total do terreno (em milhares de pés quadrados), área construída (em milhares de pés

quadrados) e idade do imóvel (em anos). A variável resposta Y é o preço de venda do imóvel (em milhares de dólares). O objetivo do estudo é obter algum tipo de relação (caso a relação efetivamente exista) entre o preço de venda dos imóveis com suas respectivas características.

Antes de ajustarmos a regressão L_1 faremos uma breve análise descritiva dos dados. A Tabela 5.2 contém as estatísticas descritivas mínimo, primeiro quartil (Q_1), mediana, média, terceiro quartil (Q_3) e máximo da variável resposta Y (Preço) e das variáveis explicativas X_1 (Imposto), X_2 (Área), X_3 (Área Construída) e X_4 (Idade).

Observações	X_1	X_2	X_3	X_4	Y
1	4,9176	3,4720	0,9980	42	25,9
2	5,0208	3,5310	1,5000	62	29,5
3	4,5429	2,2750	1,1750	40	27,9
4	4,5573	4,0500	1,2320	54	25,9
5	5,0597	4,4550	1,1210	42	29,9
6	3,8910	4,4550	0,9880	56	29,9
7	5,8980	5,8500	1,2400	51	30,9
8	5,6039	9,5200	1,5010	32	28,9
9	15,4202	9,8000	3,4200	42	84,9
10	14,4598	12,8000	3,0000	14	82,9
11	5,8282	6,4350	1,2250	32	35,9
12	5,3003	4,9883	1,5520	30	31,5
13	6,2712	5,5200	0,9750	30	31
14	5,9592	6,6660	1,1210	32	30,9
15	5,0500	5,0000	1,0200	46	30
16	8,2464	5,1500	1,6640	50	36,9
17	6,6969	6,9020	1,4880	22	41,9
18	7,7841	7,1020	1,3760	17	40,5
19	9,0384	7,8000	1,5000	23	43,9
20	5,9894	5,5200	1,2560	40	37,5
21	7,5422	4,0000	1,6900	22	37,9
22	8,7951	9,8900	1,8200	50	44,5
23	6,0931	6,7265	1,6520	44	37,9
24	8,3607	9,1500	1,7770	48	38,9
25	8,1400	8,0000	1,5040	3	36,9
26	9,1416	7,3262	1,8310	31	45,8
27	12,0000	5,0000	1,2000	30	41

¹Fonte: Narula e Wellington (1977)

Tabela 5.1: Dados sobre imóveis.

Observamos inicialmente que o preço de venda dos imóveis em questão está aproximadamente entre 26 e 85 mil dólares, a mediana dos preços de venda está em torno de 37 mil dólares enquanto a média é 38500 dólares. O menor imposto que incide nos imóveis sob estudo é cerca de 390 dólares enquanto o maior é aproximadamente 1550 dólares, mas a média e a mediana são aproximadamente 725 e 610 dólares, respectiva-

Estatísticas	Mínimo	Q_1	Mediana	Média	Q_3	Máximo	Variância
Preço (Y)	25,90	29,95	36,90	38,50	40,75	84,90	204,70
Imposto (X_1)	3,891	5,180	6,093	7,245	8,304	15,42	8,31
Área (X_2)	2,275	4,722	5,850	6,348	7,563	12,80	5,78
Área C. (X_3)	0,975	1,188	1,488	1,512	1,658	3,42	0,31
Idade (X_4)	3	30	40	36,48	47	62	197,49

Tabela 5.2: Estatísticas descritivas com dados sobre imóveis.

mente.

A área total dos imóveis variou de 2275 a 12800 pés quadrados (lembre que um pé quadrado equivale a 0,09290304 metros quadrados, logo no sistema internacional de medidas temos, aproximadamente, uma variação de 211 a 1190 metros quadrados) enquanto a mediana da área total foi 5850 pés quadrados e a média foi cerca de 6350 pés quadrados. Já a área construída variou de 975 a 3420 pés quadrados com mediana e média dadas por 1488 e 1512 pés quadrados.

Com relação à idade dos imóveis, vemos que o mais recente foi construído há 3 anos enquanto o mais antigo foi construído há 62 anos. Sendo que em média as residências possuem 36 anos e meio e a mediana é 40 anos, ou seja, cerca de metade dos imóveis foi construída a menos de quatro décadas atrás enquanto a outra metade foi contruída há mais de 40 anos. As variâncias amostrais das variáveis são $\widehat{Var}(Y) = 204,70$, $S_1^2 = 8,31$, $S_2^2 = 5,78$, $S_3^2 = 0,31$, $S_4^2 = 197,49$.

Continuaremos a análise descritiva observando a matriz de correlação das variáveis Y, X_1, X_2, X_3 e X_4 expressa na Tabela 5.3. Observe que a variável idade está correlacionada negativamente com as demais variáveis e também note que as correlações entre as demais variáveis são sempre positivas. Em especial, vemos que as covariáveis, X_1, X_2, X_3 têm uma forte correlação positiva com a variável resposta enquanto X_4 tem uma fraca correlação com Y . A Figura 5.1 contém os gráficos de dispersão dos dados que reforça as observações destacadas a partir da Tabela 5.3.

	Preço (Y)	Imposto (X_1)	Área (X_2)	Área C. (X_3)	Idade (X_4)
Preço (Y)	1,0000	0,9153	0,7263	0,9288	-0,3103
Imposto (X_1)	0,9153	1,0000	0,6867	0,8324	-0,3711
Área (X_2)	0,7263	0,6867	1,0000	0,7033	-0,3803
Área C. (X_3)	0,9288	0,8324	0,7033	1,0000	-0,1781
Idade (X_4)	-0,3103	-0,3711	-0,3803	-0,1781	1,0000

Tabela 5.3: Matriz com as correlações entre as variáveis Y, X_1, X_2, X_3 e X_4 .

Agora iniciaremos a etapa de análise de regressão propriamente dita, utilizaremos um modelo linear com intercepto. A forma funcional do modelo que ajustaremos ao conjunto de dados sobre imóveis é dada pela Equação (5.1).

Figura 5.1: Diagramas de dispersão dos dados sobre imóveis.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \epsilon_i, \quad i = 1, \dots, 27. \quad (5.1)$$

As estimativas, os erros padrão das estimativas, os valores-P (níveis descritivos) e os limites inferior e superior dos intervalos de confiança (ao nível de 95% de confiança) do modelo ajustado através do método de regressão L_1 se encontram na Tabela 5.4. Somente a estimativa do parâmetro β_2 , que está ligado a covariável Área do imóvel, apresentou valor-P igual a 0,0501 que é um pouco maior que 0,05, devido a pouquíssima diferença continuaremos a análise com a covariável X_2 no modelo. A partir desses valores-P vemos que as quatro covariáveis são “significativas” para o modelo.

Parâmetros	Estimativa	Erro Padrão	valor-P	IC Li	IC Ls
β_0	9,6920	1,7909	< 0,0001	6,1819	13,2020
β_1	1,3834	0,2420	< 0,0001	0,9091	1,8576
β_2	0,4344	0,2217	0,0501	-0,0002	0,8689
β_3	14,1542	1,2720	< 0,0001	11,6612	16,6473
β_4	-0,1483	0,0290	< 0,0001	-0,2052	-0,0915
λ	1,784				

Tabela 5.4: Estimativas, erros padrão das estimativas, valores-P e limites inferior e superior dos intervalos de confiança (ao nível de 95% de confiança) no modelo ajustado.

A Tabela 5.5 apresenta os valores ajustados e resíduos via regressão L_1 ao lado das respectivas variáveis respostas. Observe que os resíduos das observações 1, 18, 22, 23 e 27 são todos nulos, isto é, estas são as denominadas observações definidoras e qualquer mudança nestas observações alterará a estimativa dos parâmetros do modelo. Note

ainda que temos cinco observações definidoras e cinco parâmetros β_i , $i = 0, 1, 2, 3, 4$, isto decorre do fato que o número de observações definidoras é sempre igual ao número de parâmetros de posição do modelo.

Observações	Y_i	$\hat{Y}_i$	$Y_i - \hat{Y}_i$
1	25,9	25,90000	0
2	29,5	30,20773	-0,7077344
3	27,9	27,66364	0,2363592
4	25,9	27,18507	-1,2850734
5	29,9	28,26451	1,6354851
6	29,9	22,68901	7,2109871
7	30,9	30,37972	0,5202758
8	28,9	38,07898	-9,1789827
9	84,9	77,45901	7,4409881
10	82,9	75,64122	7,2587757
11	35,9	33,14272	2,7572758
12	31,5	36,70911	-5,2091126
13	31,0	30,11617	0,8838310
14	30,9	31,95224	-1,0522400
15	30,0	26,46502	3,5349762
16	36,9	39,47406	-2,5740576
17	41,9	39,75289	2,1471052
18	40,5	40,50000	0
19	43,9	43,40363	0,4963726
20	37,5	32,22064	5,2793617
21	37,9	42,52091	-4,6209116
22	44,5	44,50000	0
23	37,9	37,90000	0
24	38,9	43,26562	-4,3656211
25	36,9	45,27038	-8,3703823
26	45,8	46,83921	-1,0392132
27	41,0	41,00000	0

Tabela 5.5: Dados sobre imóveis e valores preditos via regressão L_1 .

Agora faremos os testes para verificar se a suposição dos erros possuírem distribuição Laplace é razoável por meio dos cinco testes de hipóteses propostos na Seção 3.1 ao nível de 5% de significância. Os resultados dos testes são dados na lista a seguir, lembrando que o tamanho da amostra é 27 e que os valores críticos dos testes são dados nas Tabelas 3.1 a 3.5.

1. teste baseado na estatística de Watson: $U^2 = 0,0561 < 0,080 = U_*^2$

2. teste baseado na estatística de Cramér-von Mises: $W^2 = 0,0580 < 0,131 = W_*^2$
3. teste baseado na estatística de Anderson-Darling: $A^2 = 0,3845 < 0,930 = A_*^2$
4. teste baseado na estatística de Kolmogorov-Smirnov: $\sqrt{27}D = 0,6735 < 0,868 = \sqrt{27}D_*$
5. teste baseado na estatística de Kuiper: $\sqrt{27}V = 1,1104 < 1,261 = \sqrt{27}V_*$

Para os cinco testes a hipótese nula não foi rejeitada ao nível de 5% de significância, portanto não rejeitamos a hipótese de que os erros possuem distribuição Laplace.

Como discutido na Seção 3.2 do Capítulo 3 utilizaremos os resíduos quantílicos do modelo ajustado para verificar se a suposição dos erros possuírem distribuição Laplace é razoável. A Figura 5.2 ilustra o gráfico QQ-Plot dos resíduos quantílicos gerados a partir do modelo ajustado para os dados sobre imóveis. Esta figura corrobora a suposição de que os erros possuem distribuição Laplace.

Figura 5.2: Gráfico QQ-Plot dos resíduos quantílicos do modelo ajustado para dados sobre imóveis.

Adicionalmente, aplicamos o teste de Shapiro-Wilk (Shapiro e Wilk, 1965), considerando o nível de 5% de significância, nos resíduos quantílicos e obtemos o valor-P 0,438, ou seja, não rejeitamos a hipótese dos erros possuírem distribuição Laplace ao nível de 5% de significância.

Assim obtemos um consenso entre os cinco testes baseados na função distribuição empírica, no gráfico QQ-Plot dos resíduos quantílicos e no teste de Shapiro-Wilk executado nos resíduos quantílicos. Todos apontam que é razoável supor que os erros possuem distribuição Laplace.

Um coeficiente de determinação, R_2 , com propriedades análogas ao R^2 da regressão via mínimos quadrados foi proposto por André *et al.* (2000). Procederemos com outras medidas úteis à análise do modelo. O teste $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ indicou que o modelo é significativo. A Soma dos Erros Absolutos foi 77,8051 (quanto menor melhor); o Erro Absoluto Médio foi 2,8817 (quanto menor melhor); o Coeficiente de Determinação Múltipla - (R_2) foi 0,8833 (quanto maior melhor) e a Soma do Erros absolutos Preditivos foi 118,9192 (quanto menor melhor). O alto R_2 indica que o modelo está efetivamente explicando boa parte da variabilidade da variável resposta.

Agora analisaremos os gráficos da diferença de verossimilhanças e diferença de verossimilhanças condicional presentes, respectivamente, nas Figuras 5.3 e 5.4. Notamos que as observações 9 e 27 são influentes tanto na estimativa do parâmetro de escala quanto nas estimativas dos parâmetros de posição, pois estas observações se destacam na Figura 5.3. A observação 27 é destacada na Figura 5.4 enquanto a observação 9 não, isto é, a observação 9 não tem muita influência na estimativa do parâmetro β enquanto a observação 27 é influente nas estimativas de λ e β . As demais observações não parecem ser influentes nas estimativas dos parâmetros do modelo.

Vale ressaltar que os demais gráficos corroboram que a observação 27 é realmente influente no ajuste do modelo, como pode ser percebido no gráfico da distância de Cook na Figura 5.5, e nos gráficos das funções FD_i e FD_{1i} dados nas Figuras 5.6 e 5.7, respectivamente.

Note que o imóvel número 27 tem um imposto muito alto (X_1) embora tenha pouca área construída (X_3) e uma área total (X_2) de tamanho intermediário, ou seja, este imóvel claramente é atípico para essa amostra. Como a regressão L_1 não oferece robustez aos *outliers* em X esta observação acaba sendo influente no ajuste do modelo.

Após utilizarmos o algoritmo de Dodge (1997) para detectar *outliers* nos dados sobre imóveis detectamos os *outliers* em X : 3, 8, 21, 22, 24, 25, 27 e nenhum *outlier* em Y . Isso significa que as covariáveis associadas às observações com estes índices estão afastadas do subespaço gerado pelas colunas da matriz X . É interessante notar que as observações 22 e 27 são também observações definidoras.

Figura 5.3: Gráfico da diferença de verossimilhanças do modelo ajustado para dados sobre imóveis.

Figura 5.4: Gráfico da diferença condicional de verossimilhanças do modelo ajustado para dados sobre imóveis.

Figura 5.5: Gráfico da distância de Cook do modelo ajustado para dados sobre imóveis.

Figura 5.6: Gráfico da função FD_i do modelo ajustado para dados sobre imóveis.

Figura 5.7: Gráfico da função FD_{1i} do modelo ajustado para dados sobre imóveis.

Agora faremos uma breve análise confirmatória, isto é, examinaremos a influência da exclusão das observações 9 e 27 nas estimativas dos parâmetros do modelo. A Tabela 5.6 apresenta as estimativas dos parâmetros do modelo sem a observação 9, sem a observação 27, e sem as observações 9 e 27, entre parêntesis é apresentada a taxa de variação entre a estimativa do parâmetro do respectivo modelo e do modelo completo. É notável que a observação 27 é responsável por grandes mudanças nas estimativas dos parâmetros, há variação de 165% na estimativa de β_1 , a estimativa de β_2 chega a mudar de sinal e adicionalmente há um grande incremento na estimativa do parâmetro de escala do modelo da ordem de 90%, ou seja, a estimativa deste parâmetro quase dobra com a exclusão da observação 27, aumentando o comprimento dos intervalos de confiança e gerando uma inferência substancialmente distinta da inferência fornecida pelo modelo completo.

Com relação à observação 9 nota-se que a variação nas estimativas dos parâmetros decorrente de sua exclusão é de forma geral bem inferior à variação nas estimativas dos parâmetros decorrente da exclusão da observação 27, mas ainda assim a exclusão desta observação é responsável por inflar a estimativa de λ em 20%.

Considerando a exclusão conjunta das observações 9 e 27 vemos que a influência conjunta da exclusão de ambas as observações não é tão grave quanto a exclusão da observação 27 isoladamente. Este último detalhe decorre do fato das medidas de influência aplicadas só levarem em conta a influência de uma observação por vez, então já era esperado que quando se excluísse as observações 9 e 27 não necessariamente teríamos uma influência superior do que a influência decorrente da exclusão de apenas uma das

observações. Como conclusão final, vemos que as medidas de influência foram capazes de destacar observações cuja exclusão realmente afeta as estimativas dos parâmetros do modelo.

Estimativas	modelo completo	sem obs 9	sem obs 27	sem obs 9 e 27
$\hat{\beta}_0$	9,69	11,42 (+18%)	1,55 (-84%)	11,39 (+18%)
$\hat{\beta}_1$	1,38	1,37 (-1%)	3,67 (+165%)	1,37 (-1%)
$\hat{\beta}_2$	0,43	0,72 (+67%)	-0,12 (-128%)	0,72 (+66%)
$\hat{\beta}_3$	14,15	11,34 (-20%)	8,75 (-38%)	11,32 (-20%)
$\hat{\beta}_4$	-0,15	-0,13 (-9%)	-0,05 (-68%)	-0,13 (-10%)
$\hat{\lambda}$	1,78	2,16 (+21%)	3,40 (+90%)	2,10 (+18%)

Tabela 5.6: *Análise confirmatória para modelo ajustado com dados sobre imóveis.*

5.2 Dados sobre poluição

Nesta aplicação reanalisaremos os dados sobre poluição apresentados por Santos *et al.* (2016) por meio da regressão L_1 . O objetivo desta análise de dados é mensurar os efeitos de diversas covariáveis em algumas medidas espirométricas, isto é, temos várias variáveis resposta e portanto faremos um modelo para uma dessas variáveis resposta para ilustrar os métodos propostos nos capítulos anteriores

Antes de prosseguirmos a análise dos dados esclareceremos os significados dos termos utilizados, uma vez que estes termos são todos técnicos.

5.2.1 Termos técnicos

Espirometria é uma área que estuda diversas medidas das características e performance do pulmão (Carey, 2010). Esse conjunto de medidas é realizado por médicos pneumologistas e fisioterapeutas. A espirometria permite o registro de vários volumes e dos fluxos de ar. A palavra espirometria vem do latim: *spirare* = respirar + *metrum* = medida. O termo foi criado no século XVIII quando cientistas investigavam uma forma de aferir o volume de oxigênio utilizado na respiração. Mais detalhes sobre espirometria podem ser vistos em Carey (2010).

De modo geral, a espirometria mede a velocidade e a quantidade de ar que um indivíduo é capaz de colocar para dentro e para fora dos pulmões. Trata-se de um exame não invasivo e indolor e é realizado respirando-se pela boca através de um tubo conectado a um aparelho chamado espirômetro, que é capaz de registrar o volume e a velocidade do ar respirado. A interpretação deste exame exige conhecimento de fisiologia e da mecânica respiratória humana e de doenças relacionadas ao pulmão.

As variáveis mais comuns na espirometria são a capacidade vital forçada (CVF), volume expiratório forçado (VEF) em intervalos cronometrados de 0,5, 1 (VEF1), 2 e 3 segundos, razão VEF1/CVF1, fluxo expiratório forçado 25-75% (FEF25-75) e Pico de fluxo expiratório (PFE).

1. Capacidade vital forçada (CVF): é o volume máximo de ar exalado com esforço máximo a partir do ponto de máxima inspiração.
2. Volume expiratório forçado em 1 segundo (VEF1): é o volume de ar exalado no primeiro segundo a partir do ponto de máxima inspiração. Os valores médios dependem do sexo e da idade da pessoa. Valores entre 80 e 120% da média são considerados normais.
3. VEF1/CVF1: razão entre volume expiratório forçado, cronometrado, e capacidade vital forçada no intervalo de um segundo. Em adultos saudáveis, o valor deve ser de aproximadamente 70 a 85%, diminuindo com a idade. Em doenças obstrutivas,

como a asma, bronquite crônica e enfisema, o valor de VEF1 diminui devido ao aumento da resistência ao fluxo expiratório.

4. Fluxo expiratório forçado 25-75% (FEF25-75): é o fluxo ou velocidade do ar que é exalado dos pulmões durante o ponto intermediário de uma expiração forçada.
5. Pico de fluxo expiratório (PFE): (ou débito expiratório máximo instantâneo) é o débito máximo de ar conseguido durante uma expiração forçada máxima.

Outro termo técnico é o PM, que vem do inglês e significa *particulate matter*, que são partículas muito finas de sólidos ou líquidos suspensos no ar. Em Português significa material particulado. Há fontes naturais e antropogênicas de materiais particulados. Entre as fontes naturais estão os vulcões, poeira e incêndios em florestas, por exemplo. Uma das maiores fontes antropogênicas de particulados são a queima de combustíveis fósseis em motores de combustão interna de veículos. É bem conhecido que esses materiais particulados finos constituem um elemento de poluição atmosférica, podendo causar doenças no sistema respiratório de seres humanos e animais em geral.

Se define uma gradação para esse material particulado como $PM_{2,5}$ e PM_{10} . Esse número da gradação é simplesmente uma medida em micrômetros (μm), por exemplo, PM_{10} são as partículas cujo diâmetro é menor que $10 \mu m$ e de forma semelhante $PM_{2,5}$ são as partículas cujo diâmetro é menor que $2,5 \mu m$.

O objetivo principal da análise do conjunto de dados de Santos *et al.* (2016) é verificar o efeito da exposição a $PM_{2,5}$ nas medidas espirométricas apresentadas anteriormente. Com os termos técnicos já definidos podemos seguir para a análise dos dados.

5.2.2 Análise

Analisaremos o conjunto de dados sobre poluição, que foi apresentado por Santos *et al.* (2016), ajustando um modelo de regressão L_1 para a variável resposta Capacidade vital forçada (CVF), que é uma das medidas espirométricas discutidas na subseção anterior.

O conjunto de dados possui 99 observações sobre trabalhadores não fumantes de três ocupações (motoristas de táxi, controladores de tráfego e guardas florestais) da cidade de São Paulo. Após uma avaliação clínica, os participantes fizeram quatro visitas semanais consecutivas, nas quais receberam um amostrador pessoal de $PM_{2,5}$ de 24 horas e tiveram os testes de função pulmonar medidos no dia seguinte. O objetivo é estudar a associação entre a variável CVF e os níveis médios de $PM_{2,5}$, idade, circunferência da cintura, tempo de trabalho, jornada de trabalho diária, presença de diabetes ou hipertensão arterial e hábitos tabágicos anteriores.

A Tabela 5.8 contém as variáveis resposta CVF, VEF1, VEF1/CVF1, FEF25-75 e PFE, além das covariáveis X_1 ($PM_{2,5}$ medida em μm), X_2 (Idade em anos), X_3 (Circunferência

da cintura medido em centímetros), X_4 (Tempo no mesmo trabalho em anos), X_5 (Número de horas trabalhadas por dia), X_6 (Exposição à fumaça de tabaco em casa), X_7 (Ter hipertensão ou diabetes) e X_8 (Monóxido de carbono exalado medido em ppm).

A Tabela 5.7 contém as estatísticas descritivas mínimo, primeiro quartil (Q_1), mediana, média, terceiro quartil (Q_3) e máximo das variáveis resposta (CVF, VEF1, VEF1/CVF1, FEF25-75 e PFE) e das variáveis explicativas $X_1, X_2, X_3, X_4, X_5, X_8$. A variável com a maior variância é a X_1 ($PM_{2,5}$) e a covariável com a menor é a X_5 (Número de horas trabalhadas por dia). Quase todas as variáveis apresentam média e mediana muito próximas, com exceção da variável X_8 (Monóxido de carbono exalado medido em ppm).

Estatísticas	Mínimo	Q_1	Mediana	Média	Q_3	Máximo	Variância
CVF	2,340	3,970	4,410	4,379	4,775	7,440	0,5574
VEF1	1,170	3,165	3,460	3,439	3,740	5,630	0,3959
VEF1/CVF1	50,00	75,35	79,80	78,54	82,70	98,50	43,711
FEF25-75	58,00	74,30	81,90	85,16	95,35	144,90	221,44
PFE	3,200	7,300	8,400	8,442	9,450	16,90	3,23
X_1	8,50	24,96	34,93	36,96	43,77	89,70	253,9
X_2	28	42	48,10	47,94	54	66	82,35
X_3	73,30	92,55	100	101,35	108,05	135	149,7
X_4	0	5	9	11,14	15,50	40	73,85
X_5	6	7,5	12	11,09	14	18	11,83
X_8	0	0	1	2,525	4	9	8,8233

Tabela 5.7: Estatísticas descritivas com dados sobre poluição.

Observações	CVF	VEF1	VEF1/CVF1	FEF25-75	PFE	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	X ₈
1	2,34	1,17	50	144,90	3,20	44,45	51,60	126,50	3	13	0	1	0
2	4,82	3,45	71,50	68,70	8,50	48,05	53,50	98,20	3	14	0	0	1
3	4,37	3,74	85,60	72,80	9,20	49,15	56,70	91	15	12	0	0	0
4	2,68	2,29	85,30	108,20	6,60	39,20	64,90	95,10	21	12	0	1	0
5	3,07	2,53	82,20	93,60	7,30	70,80	65,70	100,90	26	16	1	0	1
6	2,96	2,21	74,60	104,70	5,50	53,63	58,80	114,10	35	15	0	1	2
7	3,86	2,91	75,30	77,30	6,40	47,45	62,20	126	13	12	0	0	0
8	3,09	2,27	73,50	93,80	6,70	48,23	65,20	108,10	5	15	0	0	1
9	3,56	2,88	80,80	88,40	5,30	56,85	57,60	82,90	2	14	0	0	0
10	3,26	2,01	61,50	96,20	5,70	60,53	58,30	107,10	4	13	1	1	1
11	3,61	2,96	81,80	86,30	6,70	43,80	58,30	81,10	3	15	0	0	2
12	3,41	2,76	81	89,60	6,40	55,03	60,30	115,50	15	12	0	1	1
13	4,51	3,89	86,30	71,30	9,50	36,23	55,80	98,10	1	14	0	0	1
14	4,29	3,35	78,10	79,30	7,50	51,25	51,30	96,50	10	14	0	0	1
15	3,90	3,30	84,60	121	5,90	62,35	31,90	100,60	5	14	0	0	0
16	4,84	3,76	77,80	66,20	7,30	37,75	56,10	111,20	14	11	0	1	0
17	4,55	3,77	82,80	71,50	8,30	54,58	54,90	107	11	15,50	0	1	0
18	4,57	3,52	77	89	8,80	38,53	39,60	127	15	16	0	1	0
19	4,17	3,26	78,30	79,60	7,30	43,23	53,30	90,60	21	11	0	0	2
20	4,68	3,74	80	71,40	7,30	34,30	52,90	96,50	8	14	0	0	1
21	3,75	3,12	83,20	95,60	9,50	35,35	47,60	90,40	12	13	0	0	2
22	4,52	3,49	77,10	83	7,10	35,95	44,60	87,40	11	15	0	0	1
23	4,84	4,04	83,40	78,40	11,90	45,20	43,90	98,20	10	12	0	0	1
24	4,46	3,47	77,70	83,20	8	41,20	45,40	101,10	5	12	0	0	1
25	4,45	3,71	83,20	100,80	9,30	37,55	34,30	102,50	12	14	0	0	2
26	4,33	3,04	70,30	82,10	8,10	30,73	48,10	82,80	11	10	0	0	1
27	4,30	3,43	79,90	113	8,90	48,90	30,60	95	9	12	0	0	1
28	4,41	3,63	82,30	70,60	9,60	22,85	58,50	114,90	20	11	0	1	1
29	4,53	3,74	82,70	95,20	8,80	16,70	36,40	109,90	15	16	0	0	1
30	4,34	3,24	74,70	76,30	8,80	32,08	63,30	114,70	27	14	1	1	4
31	4,16	3,42	82,40	82	8,30	34,93	36	88,10	5	15	0	0	9
32	5,02	3,59	71,80	78,70	7,70	34,47	55,90	98,50	2	12	0	1	4
33	4,42	3,56	80,60	85,20	7,70	36,30	44,40	99,20	8	17	0	0	4
34	4,19	3,46	82,60	103,20	7,10	42,53	46,20	106,50	14	13	0	0	9
35	3,79	2,65	70	103,50	6,60	64,38	42	111,90	11	16	0	0	9
36	4,27	3,39	79,30	70,70	9,50	34,50	61,10	119,60	18	16	0	1	3
37	5,12	4,14	80,80	93,90	10,10	43,75	31,20	91,40	4	14	0	0	2
38	4,67	3,87	82,80	82,40	7,90	38,30	42,90	92,50	10	16	0	0	9
39	4,30	3,48	81	78,80	9,20	32,13	51,80	93,40	12	14	0	0	1
40	7,44	5,63	75,70	58	10,50	81,70	36,30	86,60	10	13	0	0	1
41	3,65	2,92	79,90	95,50	7	36,50	50	104,80	20	14	0	1	0
42	4,92	3,63	73,90	63,70	9,40	43,13	58,10	95	40	15	0	0	0
43	5,14	4,12	80,10	70,90	9,80	28,78	46,40	112,60	4	13	0	0	0
44	4,61	3,81	82,60	74,40	8,50	32,35	51	85,50	26	15	0	0	0
45	4,97	3,89	78,30	71,80	9,80	27,08	47,90	114,40	15	13	0	0	0
46	3,59	2,95	82,20	89,30	9,80	72,30	56,20	99,20	0	10	0	0	0
47	4,34	3,21	74	107	7,60	89,70	32,60	111,10	7	16	0	0	0
48	4,70	3,92	83,50	67,90	9	36,90	56,30	110,90	4	12	0	1	9
49	4,69	3,59	76,60	78,40	9	36,95	45,90	105,70	4	14	0	0	5
50	4,60	3,67	79,80	99	8,90	21,48	33,70	101,10	3	18	0	0	0
51	3,71	3,06	82,60	98,10	8,30	20,05	46,50	103	15	18	1	1	0
52	3,53	2,94	83,60	96,90	7,70	22,40	51,20	100,30	6	14	0	1	1
53	4,51	3,16	70	74,90	6,80	26,13	51,80	105,10	18	8	0	0	1
54	3,85	3,02	78,50	98	8	33,98	44,30	131,90	8	12	0	1	0
55	4,95	3,65	73,80	76,90	11,50	25,13	43,60	106,50	5	15	0	1	1
56	5,42	4,20	77,50	68,70	10,60	22,80	45	92,60	2	6	0	0	0
57	4,49	3,48	77,50	74,20	8,70	14,67	53	85,70	16	9	0	0	2
58	3,65	3,34	91,50	96,40	9,40	34,83	49	89,60	17	9	0	0	1
59	4,25	3,25	76,30	81,60	11,60	18,20	50	93,20	15	9	1	0	1
60	4,48	2,92	65	77,30	6,80	18,33	50	105	29	9	0	1	1
61	4,51	3,57	79,10	72,90	9,10	23,03	54	112,60	30	9	0	1	0
62	5,59	4,70	84	76,20	10,20	25,13	37	100	6	9	0	0	0
63	4,18	3,60	86,10	83,10	9,40	40,40	50	105,60	6	9	1	1	1
64	4,89	3,71	75,90	67,20	8,40	19,70	54	82,70	18	9	0	0	1
65	5,19	4,29	82,70	71,90	10,40	21,80	45	97	19	9	0	0	2
66	4,09	3,38	82,50	88,40	8,80	29,13	47	116	6	9	0	1	0
67	2,87	2,11	73,50	100	6,70	29,48	66	129	35	9	0	1	1
68	4,28	3,23	75,40	68,50	7,40	29,43	64	135	17	9	0	1	0
69	4,13	3,48	84,30	125,10	7,80	24,68	28	73,30	1	6	0	0	0
70	5,11	3,84	75,10	79,10	10,60	19,30	40	90	15	11	0	0	1
71	4,35	3,38	77,50	81	16,90	19,30	49	96,10	23	9	0	1	1
72	3,04	2,59	85,40	115,80	9,50	19,30	49	104	26	9	0	1	0
73	3,89	3,25	83,70	79,60	8,20	19,27	59	106,30	17	9	0	1	1
74	5,37	3,91	72,80	72,90	10,50	19,43	42	88,70	17	8	0	0	0
75	4,11	3,28	79,90	80	9	30,90	54	116,20	6	9	0	0	0
76	4,55	3,98	87,50	97,30	7,50	18,10	35	90	8	12	0	0	1
77	4,57	3,47	75,80	81,50	8,20	43,58	45	99,30	19	6	0	0	6
78	5,73	4,51	78,70	66,70	8,90	52,57	43,50	94,30	2	6	0	0	9
79	5,44	4,31	79,30	73,90	10,50	43,10	40,30	98,10	4	6	0	0	2
80	4,35	3,42	78,60	67,70	6,60	69,10	48,10	78,80	1	6	0	0	9
81	3,58	2,90	81	89,20	9,60	63,35	56,30	108,60	18	7	0	1	4
82	4	3,18	79,40	95	8,30	82,55	43,70	107,40	0	7	1	0	6
83	5,38	4,34	80,40	85,30	9,60	36,20	33,50	98,60	6	7	1	0	5
84	4,54	3,36	74,10	86,20	6,70	29,17	41,60	108	11	7	0	0	5
85	4,34	3,24	74,70	76,30	8,80	42,38	53,70	125,50	6	6	0	1	5
86	5,58	4,63	83	81,90	9,70	31,67	50,90	116,40	8	7	0	0	7
87	4,08	3,50	85,80	94	8,40	36,13	43	94,50	2	7	0	0	9
88	5,67	3,47	61,30	64,70	7	47,85	40,50	107	1	7	0	0	5
89	3,94	3,39	86	102,70	7,90	25,23	40,10	101,80	5	7	0	0	3
90	5,02	3,17	63,10	64,40	6,90	31,83	45,60	103,90	1	7	0	0	5
91	3,52	2,92	82,90	108	8,50	30,53	43,70	92,50	9	7	0	0	5
92	4,73	3,32	70,10	81,40	8,10	22,60	39,50	99,60	5	7	0	0	9
93	4,22	3,29	78,10	95,60	9,60	38,25	41,30	94,50	8	7	0	0	9
94	4,41	3,58	81,30	80,70	7,10	21,08	52,60	93,50	1	7	0	1	3
95	5,14	4,18	81,30	90,90	8,20	27,40	33,40	90	2	7	0	0	6
96	4,40	3,67	83,40	75,80	9,40	17,30	53,40	88	15	7	0	0	4
97	5,16	3,47	67,20	75,60	4,40	24,80	42	101	5	7	0	1	9
98	5,24	5,16	98,50	87,30	12,90	8,50	33,40	92	9	7	0	0	8
99	4,43	3,09	69,80	97,60	7,90	19,40	36,20	86	5	7	0	0	6

²Fonte: Santos *et al.* (2016)

Tabela 5.8: Dados sobre poluição.

	CVF	VEF1	VEF1/CVF1	FEF25-75	PFE	X_1	X_2	X_3	X_4	X_5	X_8
CVF	1,0000	0,8967	-0,0075	-0,6538	0,4394	-0,1373	-0,4594	-0,2352	-0,2256	-0,2153	0,1915
VEF1	0,8967	1,0000	0,4215	-0,5435	0,5762	-0,1839	-0,4714	-0,3084	-0,2101	-0,1727	0,1655
VEF1/CVF1	-0,0075	0,4215	1,0000	0,0479	0,4045	-0,1265	-0,1072	-0,2586	0,0136	0,0339	-0,0173
FEF25-75	-0,6538	-0,5435	0,0479	1,0000	-0,3167	0,1111	-0,2600	0,0680	-0,0579	0,1385	-0,0667
PFE	0,4394	0,5762	0,4045	-0,3167	1,0000	-0,2806	-0,2209	-0,1715	0,0793	-0,1535	-0,0621
X_1	-0,1373	-0,1839	-0,1265	0,1111	-0,2806	1,0000	0,0962	0,0829	-0,1668	0,2351	0,0486
X_2	-0,4594	-0,4714	-0,1072	-0,2600	-0,2209	0,0962	1,0000	0,3242	0,4220	0,1311	-0,2487
X_3	-0,2352	-0,3084	-0,2586	0,0680	-0,1715	0,0829	0,3242	1,0000	0,1811	0,1250	-0,1288
X_4	-0,2256	-0,2101	0,0136	-0,0579	0,0793	-0,1668	0,4220	0,1811	1,0000	0,1477	-0,2867
X_5	-0,2153	-0,1727	0,0339	0,1385	-0,1535	0,2351	0,1311	0,1250	0,1477	1,0000	-0,3320
X_8	0,1915	0,1655	-0,0173	-0,0667	-0,0621	0,0486	-0,2487	-0,1288	-0,2867	-0,3320	1,0000

Tabela 5.9: Matriz com as correlações entre as variáveis resposta CVF, VEF1, VEF1/CVF1, FEF25-75 e PFE e as variáveis explicativas X_1 , X_2 , X_3 , X_4 , X_5 e X_8 dos dados sobre poluição.

Continuaremos a análise descritiva observando a matriz de correlação das variáveis CVF, VEF1, VEF1/CVF1, FEF25-75, PFE, X_1 , X_2 , X_3 , X_4 , X_5 e X_8 que é expressa na Tabela 5.9. Observe que a variável X_2 (Idade) está correlacionada negativamente com as variáveis CVF e VEF1, além disso esta é uma das correlações mais fortes da matriz de correlação. Só a variável X_8 (Monóxido de carbono exalado medido em ppm) apresenta correlação positiva com a variável CVF. Também existe uma correlação moderada entre as variáveis X_2 e X_4 .

Com relação às variáveis X_6 (Exposição à fumaça de tabaco em casa), X_7 (Ter hipertensão ou diabetes), temos a Tabela 5.10. É notável que a maioria das pessoas que comporam a amostra não possui diabetes nem hipertensão e não estão expostos à fumaça do tabaco em casa (64 observações), além disso um número considerável de pessoas possuem diabetes ou hipertensão (31 observações).

X_6/X_7	0	1	Total
0	64	27	91
1	4	4	8
Total	68	31	99

Tabela 5.10: Tabela descritiva das variáveis X_6 e X_7 .

Agora iniciaremos a etapa de análise de regressão propriamente dita, utilizaremos um modelo linear com intercepto para estudar a relação entre a variável CVF e as covariáveis X_1 (PM_{2,5} medida em μm), X_2 (Idade em anos), X_3 (Circunferência da cintura medido em centímetros), X_4 (Tempo no mesmo trabalho em anos), X_5 (Número de horas trabalhadas por dia), X_6 (Exposição à fumaça de tabaco em casa), X_7 (Ter hipertensão ou diabetes) e X_8 (Monóxido de carbono exalado medido em ppm). A forma funcional do modelo que ajustaremos ao conjunto de dados é dada pela Equação (5.2).

$$CVF_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_8 X_{8i} + \epsilon_i, \quad i = 1, \dots, 99. \quad (5.2)$$

As estimativas, os erros padrão das estimativas, os valores-P e os limites inferior e superior dos intervalos de confiança (ao nível de 95% de confiança) do modelo ajustado se encontram na Tabela 5.11. Somente as estimativas dos parâmetros β_0 , β_1 e β_2 , apresentaram valor-P menores que 0,05, portanto estes parâmetros são estatisticamente

distintos de zero. A partir desses valores- P vemos que as covariáveis X_1 ($PM_{2,5}$ medida em μm) e X_2 (Idade em anos) são “significativas” para o modelo.

Parâmetros	Estimativa	Erro Padrão	valor-P	IC Li	IC Ls
β_0	5,9660	0,7131	< 0,0001	4,5685	7,3636
β_1	-0,0102	0,0045	0,0235	-0,0190	-0,0014
β_2	-0,0273	0,0087	0,0017	-0,0444	-0,0103
β_3	0,0055	0,0064	0,3926	-0,0071	0,0181
β_4	-0,0060	0,0087	0,4920	-0,0231	0,0111
β_5	-0,0275	0,0207	0,1851	-0,0681	0,0132
β_6	-0,3402	0,2404	0,1571	-0,8115	0,1310
β_7	-0,2243	0,1811	0,2156	-0,5792	0,1307
β_8	0,0019	0,0241	0,9381	-0,0454	0,0492
λ	0,6369				

Tabela 5.11: Estimativas, erros padrão das estimativas, valores- P e limites inferior e superior dos intervalos de confiança (ao nível de 95% de confiança) no modelo ajustado para variável resposta CVF.

A Tabela 5.12 apresenta os valores ajustados e resíduos via regressão L_1 ao lado das respectivas variáveis respostas. Observe que os resíduos das observações 5, 33, 59, 71, 77, 82, 85, 94 e 98 são todos nulos, ou seja, estas são as denominadas observações definidoras e qualquer mudança nestas observações alterará a estimativa dos parâmetros do modelo. Note ainda que temos nove observações definidoras e nove parâmetros β_i , $i = 0, 1, \dots, 8$, isto decorre do fato que o número de observações definidoras é sempre igual ao número de parâmetros de posição do modelo.

Agora faremos os testes para verificar se a suposição dos erros possuírem distribuição Laplace é razoável por meio dos cinco testes de hipóteses propostos na Seção 3.1 ao nível de 5% de significância. Os resultados dos testes são dados na lista a seguir, lembrando que o tamanho da amostra é 99 e que os valores críticos dos testes são dados nas Tabelas 3.1 a 3.5.

1. teste baseado na estatística de Watson: $U^2 = 0,0811 < 0,083 = U_*^2$
2. teste baseado na estatística de Cramér-von Mises: $W^2 = 0,0811 < 0,141 = W_*^2$
3. teste baseado na estatística de Anderson-Darling: $A^2 = 0,5068 < 0,972 = A_*^2$
4. teste baseado na estatística de Kolmogorov-Smirnov: $\sqrt{99}D = 0,7033 < 0,918 = \sqrt{99}D_*$
5. teste baseado na estatística de Kuiper: $\sqrt{99}V = 1,2632 < 1,324 = \sqrt{99}V_*$

Para os cinco testes a hipótese nula não foi rejeitada ao nível de 5% de significância, portanto não rejeitamos a hipótese de que os erros possuem distribuição Laplace.

Como discutido no Capítulo 3, Seção 3.2, usaremos os resíduos quantílicos do modelo ajustado para verificar se a suposição dos erros possuírem distribuição Laplace é razoável. A Figura 5.8 ilustra o gráfico QQ-Plot dos resíduos quantílicos gerados a partir do modelo ajustado para os dados sobre imóveis. Esta figura favorece fracamente a suposição de que os erros possuem distribuição Laplace, pois as caudas nas estão bem ajustadas à reta de referência.

Figura 5.8: Gráfico QQ-Plot dos resíduos quantílicos do modelo ajustado para dados sobre poluição.

Adicionalmente, aplicamos o teste de Shapiro-Wilk (Shapiro e Wilk, 1965), considerando o nível de 5% de significância, nos resíduos quantílicos e obtemos o valor-P 0,0829, isto é, não rejeitamos a hipótese dos erros possuírem distribuição Laplace ao nível de 5% de significância. Este conjunto de dados ilustrou que às vezes o QQ-Plot não fornece subsídios para uma conclusão clara e assim é interessante ter um método mais objetivo, como um valor-P por exemplo, para tomar a decisão.

Assim obtemos um consenso entre os cinco testes baseados na função distribuição empírica, no gráfico QQ-Plot dos resíduos quantílicos e no teste de de Shapiro-Wilk executado nos resíduos quantílicos. Todos apontam que é razoável supor que os erros possuem distribuição Laplace.

Procederemos com algumas medidas úteis à análise do modelo. O teste $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ indicou que o modelo é significativo. A Soma dos Erros Absolutos foi 77,8051 (quanto menor melhor); o Erro Absoluto Médio foi 2,8817 (quanto menor melhor); o Coeficiente de Determinação Múltipla - (R_2) foi 0,8833 (quanto maior melhor) e a Soma do Erros absolutos Preditivos foi 118,9192 (quanto menor melhor). O alto R_2 indica que o modelo está efetivamente explicando boa parte da variabilidade da variável resposta.

Observações	Y_i	$\hat{Y}_i$	$Y_i - \hat{Y}_i$
1	2,34	4,20	-1,86
2	4,82	4,15	0,67
3	4,37	4,00	0,37
4	2,68	3,64	-0,96
5	3,07	3,07	0
6	2,96	3,60	-0,64
7	3,86	4,07	-0,21
8	3,09	3,85	-0,76
9	3,56	3,87	-0,31
10	3,26	3,40	-0,14
11	3,61	3,95	-0,34
12	3,41	3,75	-0,34
13	4,51	4,22	0,29
14	4,29	4,13	0,16
15	3,90	4,60	-0,70
16	4,84	4,05	0,79
17	4,55	3,78	0,77
18	4,57	4,43	0,14
19	4,17	4,14	0,03
20	4,68	4,27	0,41
21	3,75	4,38	-0,63
22	4,52	4,38	0,14
23	4,84	4,46	0,38
24	4,46	4,50	-0,04
25	4,45	4,76	-0,31
26	4,33	4,45	-0,12
27	4,30	4,77	-0,47
28	4,41	4,12	0,29
29	4,53	4,88	-0,35
30	4,34	3,44	0,90
31	4,16	4,68	-0,52
32	5,02	4,07	0,95
33	4,42	4,42	0
34	4,19	4,43	-0,24
35	3,79	4,29	-0,50
36	4,27	3,84	0,43
37	5,12	4,76	0,36
38	4,67	4,43	0,24
39	4,30	4,28	0,02
40	7,44	4,20	3,24
41	3,65	4,07	-0,42
42	4,92	3,81	1,11
43	5,14	4,64	0,50
44	4,61	4,14	0,47
45	4,97	4,56	0,41
46	3,59	3,96	-0,37
47	4,34	4,29	0,05
48	4,70	4,10	0,60
49	4,69	4,52	0,17
50	4,60	4,87	-0,27
51	3,71	3,91	-0,20
52	3,53	4,25	-0,72
53	4,51	4,54	-0,03
54	3,85	4,53	-0,68
55	4,95	4,44	0,51
56	5,42	4,84	0,58
57	4,49	4,50	-0,01
58	3,65	4,42	-0,77
59	4,25	4,25	0
60	4,48	4,35	0,13
61	4,51	4,22	0,29
62	5,59	4,96	0,63
63	4,18	3,92	0,26
64	4,89	4,39	0,50
65	5,19	4,69	0,50
66	4,09	4,51	-0,42
67	2,87	3,89	-1,02
68	4,28	4,08	0,20
69	4,13	5,18	-1,05
70	5,11	4,78	0,33
71	4,35	4,35	0
72	3,04	4,37	-1,33
73	3,89	4,17	-0,28
74	5,37	4,79	0,58
75	4,11	4,53	-0,42
76	4,55	4,94	-0,39
77	4,57	4,57	0
78	5,73	4,60	1,13
79	5,44	4,78	0,66
80	4,35	4,23	0,12
81	3,58	3,86	-0,28
82	4,00	4,00	0
83	5,38	4,66	0,72
84	4,54	4,88	-0,34
85	4,34	4,34	0
86	5,58	4,66	0,92
87	4,08	4,75	-0,67
88	5,67	4,77	0,90
89	3,94	4,96	-1,02
90	5,02	4,78	0,24
91	3,52	4,73	-1,21
92	4,73	5,00	-0,27
93	4,22	4,74	-0,52
94	4,41	4,41	0
95	5,14	5,08	0,06
96	4,40	4,54	-0,14
97	5,16	4,69	0,47
98	5,24	5,24	0
99	4,43	5,04	-0,61

Tabela 5.12: Valores preditos via regressão L_1 para variável resposta CVF.

O teste $H_0 : \beta_1 = \dots = \beta_7 = \beta_8 = 0$ indicou que o modelo é significativo. A Soma dos Erros Absolutos foi 46,4791 (quanto menor melhor); o Erro Absoluto Médio foi 0,4695 (quanto menor melhor); o Coeficiente de Determinação Múltipla - (R_2) foi $R_2 = 0,1907$ (quanto maior melhor) e a Soma dos Erros absolutos Preditivos foi 51,1719 (quanto menor melhor). O baixo R_2 indica que o modelo está explicando apenas cerca de um quinto da variabilidade da variável resposta.

Agora analisaremos os gráficos da diferença de verossimilhanças e diferença de verossimilhanças condicional presentes, respectivamente, nas Figuras 5.9 e 5.10. Notamos que as observações 1 e 40 são influentes tanto na estimativa do parâmetro de escala quanto nas estimativas dos parâmetros de posição, pois estas observações se destacam na Figura 5.9. Já na Figura 5.10, as observações 18 e 51 se sobressaem, sendo assim mais influentes que as demais nas estimativas dos parâmetros de posição do modelo. Vale ressaltar que os demais gráficos corroboram que essas observações são realmente influentes no ajuste do modelo, como pode ser percebido no gráfico da distância de Cook na Figura 5.11, e nos gráficos das funções FD_i e FD_{1i} dados nas Figuras 5.12 e 5.13, respectivamente. Entretanto a escala dos gráficos é diferente, percebe-se que embora as observações 18 e 51 se destaquem estas observações possuem baixo valor nas medidas de influência apresentadas. Por exemplo, na Figura 5.10 o valor do afastamento condicional de verossimilhanças associado a observação 51 é aproximadamente 0,6 enquanto na Figura 5.9 o valor do afastamento de verossimilhanças associado a observação 40 é aproximadamente 14.

A Tabela 5.13 apresenta os dados das observações influentes destacadas pela Figura 5.9. Note que o indivíduo número 1 tem medida $PM_{2,5}$ acima do terceiro quartil (inala uma grande quantidade de partículas finas), tem aproximadamente 52 anos (está próximo do terceiro quartil neste quesito), tem 126 centímetros de cintura (a segunda maior medida de cintura da amostra), além de trabalhar 13 horas por dia, ser diabético ou hipertenso, não é exposto à fumaça do tabaco em casa. Ou seja, esse indivíduo claramente é atípico para essa amostra. A observação 40 indica uma altíssima exposição a partículas finas porém uma baixa quantidade de monóxido de carbono exalada e além disso ele possui apenas 86 centímetros de cintura (essa medida está abaixo do primeiro quartil), essa observação apresenta outra peculiaridade que é ter FEF25-75 58 (o valor mínimo dessa variável na amostra foi 58), portanto essa observação também é atípica. Como a regressão L_1 não oferece robustez aos *outliers* em X esta observação acaba sendo influente no ajuste do modelo.

Observações	CVF	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
1	2,34	44,45	51,60	126,50	3	13	0	1	0
40	7,44	81,70	36,30	86,60	10	13	0	0	1

Tabela 5.13: Observações influentes no modelo ajustado para a variável CVF.

Figura 5.9: Gráfico da diferença de verossimilhanças do modelo ajustado para variável resposta CVF.

Figura 5.10: Gráfico da diferença de verossimilhanças condicional do modelo ajustado para variável resposta CVF.

Vale ressaltar que os métodos empregados no ajuste e diagnóstico do modelo para a variável resposta CVF pode ser utilizado na análise e diagnóstico das demais variáveis resposta: Volume expiratório forçado em 1 segundo (VEF1), VEF1/CVF1, Fluxo expiratório forçado (FEF) e Pico de fluxo expiratório (PFE). Não ilustramos a análise dos modelos para as outras quatro variáveis resposta devido ao fato de ser necessário realizar exatamente os mesmos procedimentos ilustrados nesta seção.

Figura 5.11: Gráfico da distância de Cook do modelo ajustado para variável resposta CVF.

Figura 5.12: Gráfico da função FD_i do modelo ajustado para variável resposta CVF.

Figura 5.13: Gráfico da função FD_{1i} do modelo ajustado para variável resposta CVF.

Agora faremos uma breve análise confirmatória, ou seja, examinaremos a influência da exclusão das observações 1 e 40 nas estimativas dos parâmetros do modelo. A Tabela 5.14 apresenta as estimativas dos parâmetros do modelo sem a observação 1, sem a observação 40, e sem as observações 1 e 40, entre parêntesis é apresentada a taxa de variação entre a estimativa do parâmetro do respectivo modelo e do modelo completo.

A exclusão da observação 1 é responsável por mudanças significativas nas estimativas dos parâmetros, em especial, há variação de -174% na estimativa de β_8 . Com relação à observação 40 nota-se que a variação na estimativa do parâmetro β_8 chega a ser superior a 500% , ou seja, a estimativa sextuplica. Considerando a exclusão conjunta das observações 1 e 40, mais uma vez nota-se uma grande variação na estimativa de β_8 .

Nos três casos estudados percebe-se que a exclusão destas variáveis é responsável por diminuir a estimativa de λ , isto é, o comprimento dos intervalos de confiança diminui quando estas observações são excluídas. Como conclusão final, vemos que as medidas de influência foram capazes de destacar observações cuja exclusão realmente afeta as estimativas dos parâmetros do modelo.

Estimativas	modelo completo	sem obs 1	sem obs 40	sem obs 1 e 40
$\hat{\beta}_0$	5,97	5,99 (+0%)	5,86 (-2%)	5,83 (-2%)
$\hat{\beta}_1$	-0,01	-0,01 (-1%)	-0,01 (-0%)	-0,01 (+2%)
$\hat{\beta}_2$	-0,03	-0,03 (+3%)	-0,03 (-8%)	-0,03 (-6%)
$\hat{\beta}_3$	0,01	0,01 (+4%)	0,00 (-15%)	0,01 (-2%)
$\hat{\beta}_4$	-0,01	-0,01 (+5%)	-0,00 (-18%)	-0,01 (-13%)
$\hat{\beta}_5$	-0,03	-0,03 (-4%)	-0,03 (-8%)	-0,03 (-9%)
$\hat{\beta}_6$	-0,34	-0,35 (+4%)	-0,32 (-6%)	-0,32 (-7%)
$\hat{\beta}_7$	-0,22	-0,23 (+3%)	-0,21 (-7%)	-0,20 (-9%)
$\hat{\beta}_8$	0,00	-0,00 (-174%)	0,01 (+511%)	0,01 (+458%)
$\hat{\lambda}$	0,64	0,54 (-15%)	0,54 (-16%)	0,60 (-5%)

Tabela 5.14: *Análise confirmatória para modelo ajustado com dados sobre poluição.*

Capítulo 6

Conclusão

Os modelos ajustados aos conjuntos de dados sobre imóveis e sobre poluição ficaram bem ajustados. Em ambos os casos a hipótese dos erros terem distribuição Laplace pareceu razoável, o que justificou o uso da regressão L_1 pois o estimador de mínima soma dos erros absolutos é equivalente ao de máxima verossimilhança quando os erros têm distribuição Laplace. O detalhe mais importante dessas análises de dados é que com exceção da medida baseada na diferença de verossimilhanças, todas as demais medidas concordam entre si.

Algumas sugestões para trabalhos futuros são: avaliar a eficiência das medidas de influência discutidas em casos onde a suposição de erros com distribuição Laplace não é satisfeita, especialmente com distribuições leptocúrticas (caudas pesadas); analisar novamente o conjunto de dados sobre poluição, mas desta vez fazendo um modelo (usando o mesmo conjunto de variáveis explicativas) para cada uma das cinco variáveis resposta e verificar se as observações influentes nos cinco modelos são as mesmas, pois como a regressão L_1 não oferece robustez aos *outliers* em X então é razoável esperar que os mesmos pontos sejam influentes em todos os modelos.

Apêndice A

Algoritmos Para Ajuste E Inferência Na Regressão L_1

A.1 Algoritmo para ajuste da regressão L_1

Algoritmo para ajuste de modelos de regressão L_1 relacionado aos Capítulos 1 e 2.

```
#### algoritmo para ajuste da regressão  $L_1$ 
```

```
#' Fitting LAD Models
```

```
ladfit = function (x, y, intercept = TRUE) {  
  tolerance = 1e-07  
  print.it = TRUE  
  xn = if (is.matrix(x))  
    dimnames(x)[[2]]  
  else "X"  
  x = as.matrix(x)  
  dx = dim(x)  
  vars = dx[2]  
  if (length(xn) == 0)  
    xn = paste("X", 1:vars, sep = "")  
  if (intercept) {  
    vars = vars + 1  
    xn = c("Intercept", xn)  
    x = cbind(1, x)  
  }  
  obs = dx[1]  
  if (obs != length(y))
```

```

    stop("Y and X have different number of observations!")
  if (obs <= vars)
    stop("Less observations than variables")
  storage.mode(y) <- "double"
  z <- matrix(0, nrow = obs + 2, ncol = vars + 2)
  z[1:obs, 1:vars] <- x
  storage.mode(z) <- "double"
  fit <- .Fortran("l1", n = as.integer(obs), p = as.integer(vars),
    n2 = as.integer(obs + 2), p2 = as.integer(vars + 2),
    z = z, y = y, tol = as.double(tolerance), cf = double(vars),
    resid = double(obs), work = integer(obs))
  minimum <- fit$z[obs + 1, vars + 1]
  rank = fit$z[obs + 1, vars + 2]
  info = fit$z[obs + 2, vars + 1]
  numIter = fit$z[obs + 2, vars + 2]
  fitted = drop(x %*% fit$cf)
  msg = character(0)
  if (info == 0) {
    msg = "Non-unique solution possible!"
    if (print.it)
      warning(msg)
  }
  else if (info != 1)
    stop("Premature termination!!")
  if (rank != vars) {
    msg = c(msg, paste("Matrix not of full rank, apparent rank",
      rank))
    if (print.it)
      warning(msg[length(msg)])
  }
  out = list(coefficients = fit$cf, minimum = minimum, fitted.values = fitted,
    residuals = fit$resid, rank = rank, numIter = numIter,
    info = info)
  names(out$coefficients) <- xn
  if (length(msg))
    out$message = list(msg)
  out
}

```


A.2 Algoritmo para inferência na regressão L_1

Algoritmo para ajuste de modelos de regressão L_1 relacionado ao Capítulos 2. Baseado no algoritmo da dissertação de [Tavares \(1998\)](#).

```
#### algorithm to make inference in LAD models

#' Fitting LAD Models

#Variáveis a serem definidas pelo usuário:

#intercepto: recebendo o valor 0, ajusta o modelo sem intercepto e recebendo o
  valor 1, ajusta o modelo com intercepto.

#imprime: recebendo o valor 1, imprime os valores observados, ajustados e os
  resíduos do modelo e recebendo 0, não imprime tais valores.

#Verifica se o usuario pediu modelo com ou sem intercepto

##colocando valores "default"
ladreg = function(y,X, intercepto =1, alfa=0.05, imprime=1){

  if(intercepto == 0){
    fit = ladfit(X, y, intercept = F)
    Xmat =X
    SEAm = sum(abs(y))
    fator=0
  }

  if(intercepto == 1){
    fit = ladfit(X, y, intercept = T)
    Xmat =cbind(1,X)
    SEAm = sum(abs(y-median(y)))
    fator=1
  }

  #Calculo da Estimativa de tao
```

```

r= cbind(residuals(fit))
coef1 = cbind(coef(fit))
aju = y-r
n = nrow(X) - ncol(X) -fator
m = (n+1)/2 -sqrt(n)
res = cbind(sort(r))
res1 = matrix(0, nrow(res) -nrow(coef1),1)
def = matrix(0, nrow(coef1),1)
j=1
for(i in 1:nrow(r)) if(r[i] == 0) ((def[j] = i) & (j = j+1))
j=1
for(i in 1:nrow(res)) if(res[i] !=0) ((res1[j] = res[i]) & (j = j+1))
pos1 = n-m +1
pos2 = m
e1 = round(pos1)
e2 = round(pos2)
if (e1 > n) (e1 = n)
if(e2 == 0) (e2 = 1)
tao = sqrt(n)*(res[e1] -res1[e2])/4

#Calculos dos Intervalos de Confianca, Estatisticas de Testes e Niveis
descritivos

SEA = sum(abs(res))
XX = solve(t(Xmat)%*%Xmat)
DX = diag(XX)
IC = matrix(0, nrow(coef1),2)
TH = matrix(0, nrow(coef1),2)
#DP = matrix(0, nrow(coef1),1)
aux = cbind(sqrt(DX))
DP = tao*aux
IC[,1] = coef1 -qnorm(0.975)*tao*aux
IC[,2] = coef1 +qnorm(0.975)*tao*aux
E = matrix(0, nrow(coef1), 1)
for(i in 1:ncol(Xmat)) E[i] = coef1[i]/(tao*aux[i])
E = abs(E)
TH[,1] = E
TH[,2] = 2*(1-pnorm(E,0,1))
EReg = matrix(0,1,2)
reg = (SEAm -SEA)/(tao/2)

```

```

EReg[1,1] = reg
EReg[1,2] = 1-stats::pchisq(reg, ncol(X))

#Calculos das Estatisticas para Selecao de Variaveis

EAM = SEA/length(y)
RSEA = SEAm -SEA
R2 = RSEA/(RSEA +(nrow(X) -ncol(X) -fator)*tao/2)
R3 = RSEA/(RSEA +(nrow(X) -ncol(X) -fator)*EAM/2)
Raju = 1 -(1-R3)*(nrow(y)/(nrow(X) -ncol(X) -fator))
SEAP = 0
for(i in 1:nrow(X)) {
  X1 = matrix(0, nrow(X) -1, ncol(X))
  y1 = matrix(0, nrow(X) -1, 1)
  k=1
  for(j in 1:nrow(X)) if (j != i) (X1[k,] = X[j,]) & (y1[k] = y[j]) & (k=k+1)
  if(intercepto == 0) (fit1 =ladfit(X1,y1, intercept = F))
  if(intercepto == 1) (fit1 =ladfit(X1,y1, intercept = T))
  coefaux = cbind(coef(fit1))
  yaju = Xmat[i,]%*%coefaux
  parc = abs(y[i] -yaju)
  SEAP = SEAP +parc
}

#Observacoes Definidoras
print(def)

#Verifica se o usuario pediu a impressao dos valores observados, ajustados e
residuos
if(imprime == 1) {
  OAR = cbind(y, aju, r)
  #Valores Observados, Ajustados e Residuos
  print(OAR)
}

#Estimativa de tao
tao = round(tao, digits=4)

#Estimativas dos Parametros do Modelo
coef1 = round(coef1, digits=4)

```

```

print(coef1)

#Estimativas dos Erros Padrao dos Estimadores do Modelo
DP = round(DP, digits=4)
print(DP)

#Intervalos de Confianca (95%)
IC = round(IC, digits =4)
print(IC)

#Estatistica do Teste Individual dos Parametros e Nivel Descritivo
TH = round(TH, digits =4)
print(TH)

#Estatistica do Teste do Efeito de Regressao e Nivel Descritivo
EReg = round(EReg, digits =4)
print(EReg)

###Estatistica para Selecao de Variaveis

#Soma dos Erros Absolutos (SEA)
SEA = round(SEA, digits=4)

#Erro Absoluto Medio (EAM)
EAM = round(EAM, digits =4)

#Coeficiente de Determinacao Multipla - R2
R2 = round(R2, digits=4)

#Soma do Erros absolutos Preditivos (SEAP)
SEAP =round(SEAP, digits=4)

FIT = list("OAR" = OAR,  "tao" = tao, "coef" = coef1,
          "DP" = DP, "IC" = IC,  "TH" = TH, "EReg" = EReg,
          "SEA" = SEA, "EAM" = EAM, "R2" = R2, "SEAP" = SEAP)

return(FIT)

}

```

A.3 Algoritmo para detectar *outliers*

Algoritmo da Seção 4.2.

```
#' Algorithm To Detect Outliers

##colocando valores "default"
detectOutliers= function(y,X, intercepto =1){

  residuos =  ldfit(X,y)$residuals
  #print(ldfit(X,y)$coef)

  ##número de observações
  n = length(y)
  ##numero de parametros
  p = ncol(as.matrix(X)) + intercepto

  ###ordenando resíduos absolutos em ordem decrescente
  orde =sort(abs(residuos),decreasing = TRUE)

  ###excluindo resíduos nulos
  result =NULL
  for(i in 1:(n-p)){
    result[i]=orde[i]
  }

  ##MRADZ é a mediana dos resíduos absolutos diferentes de zero
  ##estimativa do lambda
  lambda = 1.4826*stats::median(result)

  #print(lambda)

  ##calculando resíduos "padronizados"
  residuos2 = residuos/lambda

  ##checando se os resíduos "padronizados" > 2,5
  residuos2 = residuos/lambda
  #print(residuos2)
```

```

##criando conjunto com as observações excluídas da análise
excluir =c()

for(i in 1:length(residuos2)){
  if(abs(residuos2[i]) >= 2.5){
    excluir = c(excluir,i)
  }
}
###etapa 1 encerrada (excluindo outliers em Y)

###etapa 2 (excluindo outliers em X)
if(!is.null(excluir)){
  n = length(y[-excluir])
}else{
  n = length(y)
}

##criando conjunto com as observações excluídas da análise
excluir2 =c()

for(i in 1:ncol(as.matrix(X))){

  if(!is.null(excluir)){
    residuos = ladfit(cbind(y[-excluir], as.matrix(X)[-excluir,-i]),
      as.matrix(X)[-excluir,i])$residuals
  }else{
    residuos = ladfit(cbind(y, as.matrix(X)[,-i]),
      as.matrix(X)[,i])$residuals
  }

#  if(!is.null(excluir)){
#    print(ladfit(cbind(y[-excluir], as.matrix(X)[-excluir,-i])
#      ,as.matrix(X)[-excluir,i])$coef)
#  }else{
#    print(ladfit(cbind(y, as.matrix(X)[,-i]),as.matrix(X)[,i])$coef)
#  }

###ordenando resíduos absolutos em ordem decrescente
orde =sort(abs(residuos),decreasing = TRUE)

```

```

###excluindo resíduos nulos
result =NULL
for(j in 1:(n-p)){
  result[j]=orde[j]
}

##MRADZ é a mediana dos resíduos absolutos diferentes de zero
##estimativa do lambda
lambda = 1.4826*stats::median(result)

#print(lambda)

##calculando resíduos "padronizados"
residuos2 = residuos/lambda

##checando se os resíduos "padronizados" > 2,5
residuos2 = residuos/lambda
#print(residuos2)

#####criando conjunto com as observações excluídas da análise (outliers em X)

for(k in 1:length(residuos2)){
  if(abs(residuos2[k]) >= 2.5){
    excluir2 = c(excluir2,k)
  }
}
#print(excluir2)

} ### fim da etapa 2 (excluindo outliers em X)

##checando quais observações restaram no modelo
indices = 1:length(y)

##testando se o conjunto não é nulo
if(!is.null(excluir)){
  indices =indices[-excluir]
}

```

```
##testando se o conjunto não é nulo
if(!is.null(excluir2)){
  excluir2 =unique(excluir2)
  indices =indices[-unique(excluir2)]
}

###determinando outilers
outliers=setdiff(1:length(y),indices)

outliers = list("all" = outliers, "y" = excluir, "x" = setdiff(outliers,excluir))

###esse vetor contém as observações que restaram
return(outliers)
}
```


A.4 Algoritmo para a diferença de verossimilhanças

Algoritmo da Seção 4.3.

```
#### medida de influencia

#' Calculate Likelihood Displacement

likelihoodD = function(y,X){

  n = length(y)

  ###Calculando taoi que e o tao sem a i-esima observacao
  ###e calculando os betai que sao betai[,i]

  betai = coef(ladfit(as.matrix(X)[-1,], y[-1]))
  taoi = sum(abs(y-cbind(1,X)%*%as.matrix(betai)[,1]))

  for (i in 2:n){
    betai = cbind(betai, coef(ladfit(as.matrix(X)[-i,], y[-i])))
    taoi = c(taoi, sum(abs(y-cbind(1,X)%*%as.matrix(betai)[,i])))
  }
  taoi = taoi/(n-1)

  ###Calculando LD(betai, taoi)

  tao = sum(abs(y-cbind(1,X)%*%as.matrix(coef(ladfit(X, y)))[,1]))/n

  LD = 2*(n*log(taoi[1]/tao) + abs(y[1]
-t(as.matrix(cbind(1,X)[1,]))%*%as.matrix(betai)[,1])/taoi[1] -1)

  for (i in 2:n){
    LD = c(LD, 2*(n*log(taoi[i]/tao) + abs(y[i]
-t(as.matrix(cbind(1,X)[i,]))%*%as.matrix(betai)[,i])/taoi[i] -1))
  }

  plot(LD, main="Likelihood Displacement", xlab="Indices",
ylab="Likelihood Displacement")
  return(LD)
}
```

A.5 Algoritmo para a diferença de verossimilhanças condicional

Algoritmo da Seção 4.3.

```
#### medida de influencia condicional

#’ Calculate Conditional Likelihood Displacement

likelihoodDC = function(y,X){

  n = length(y)
  betai = coef(ladfit(as.matrix(X)[-1,], y[-1]))
  taoi = sum(abs(y-cbind(1,X)%*%as.matrix(betai)[,1]))

  for (i in 2:n){
    betai = cbind(betai, coef(ladfit(as.matrix(X)[-i,], y[-i])))
    taoi = c(taoi, sum(abs(y-cbind(1,X)%*%as.matrix(betai)[,i])))
  }
  taoi = taoi/(n-1)

  ###Calculando LD(betai| taoi)
  beta = coef(ladfit(X, y))

  tao = sum(abs(y-cbind(1,X)%*%as.matrix(coef(ladfit(X, y)))[,1]))/n

  LD_condicional = 2*n*log(sum(abs(y
-as.matrix(cbind(1,X))%*%as.matrix(betai)[,1]))/sum(abs(y
-as.matrix(cbind(1,X))%*%as.matrix(beta))))

  for (i in 2:n){
    LD_condicional = c(LD_condicional, 2*n*log(sum(abs(y
-as.matrix(cbind(1,X))%*%as.matrix(betai)[,i]))/sum(abs(y -
as.matrix(cbind(1,X))%*%as.matrix(beta))))))
  }

  plot(LD_condicional, main="Likelihood Displacement Condicional",
    xlab="Indices", ylab="Likelihood Displacement")
  return(LD_condicional)
}
```

A.6 Algoritmo para a distância de Cook

Algoritmo da Seção 4.4.

```
#' Calculate Cook Distance
```

```
CookDistance = function(y,X){

  n = length(y)

  ###Calculando taoi que e o tao sem a i-esima observacao
  ###e calculando os betai que sao betai[,i]

  beta = coef(ladfit(as.matrix(X), y))

  betai = coef(ladfit(as.matrix(X)[-1,], y[-1]))
  taoi = sum(abs(y-cbind(1,X)%*%as.matrix(betai)[,1]))

  for (i in 2:n){
    betai = cbind(betai, coef(ladfit(as.matrix(X)[-i,], y[-i])))
    taoi = c(taoi, sum(abs(y-cbind(1,X)%*%as.matrix(betai)[,i])))
  }
  taoi = taoi/(n-1)

  ###Calculando LD(betai, taoi)

  tao = sum(abs(y-cbind(1,X)%*%as.matrix(coef(ladfit(X, y)))[,1]))/n

  CD = tao^(-2)* t(beta-betai[,1])%*%(t(cbind(1,X)) %*%cbind(1,X))%*% (beta-betai[,1])

  for (i in 2:n){
    CD = c(CD, tao^(-2)* t(beta-betai[,i])%*%(t(cbind(1,X))
      %*%cbind(1,X))%*% (beta-betai[,i]))
  }

  plot(CD, main="Cook Distance", xlab="Indices",ylab="Cook Distance")
  return(CD)

}
```

A.7 Algoritmo para a medida FD_i

Algoritmo da Seção 4.4.

```
#main author: Kévin Allan Sales Rodrigues
```

```
#### uma medida de influencia
```

```
#' F Distance
```

```
FD = function(y,X){
```

```
  n = length(y)
```

```
  ###Calculando taoi que e o tao sem a i-esima observacao
```

```
  ###e calculando os betai que sao betai[,i]
```

```
  beta = coef(ladfit(as.matrix(X), y))
```

```
  betai = coef(ladfit(as.matrix(X)[-1,], y[-1]))
```

```
  taoi = sum(abs(y-cbind(1,X)%*%as.matrix(betai)[,1]))
```

```
  for (i in 2:n){
```

```
    betai = cbind(betai, coef(ladfit(as.matrix(X)[-i,], y[-i])))
```

```
    taoi = c(taoi, sum(abs(y-cbind(1,X)%*%as.matrix(betai)[,i])))
```

```
  }
```

```
  taoi = taoi/(n-1)
```

```
  ###Calculando LD(betai, taoi)
```

```
  tao = sum(abs(y-cbind(1,X)%*%as.matrix(coef(ladfit(X, y)))[,1]))/n
```

```
  FF = tao^(-2)*sqrt(sum((cbind(1,X)%*%beta -cbind(1,X)%*%betai[,1])^(2)))
```

```
  for (i in 2:n){
```

```
    FF = c(FF, tao^(-2)*sqrt(sum((cbind(1,X)%*%beta -cbind(1,X)%*%betai[,i])^(2))))
```

```
  }
```

```
  plot(FF, main="F Distance", xlab="Indices",ylab="F Distance")
```

```
  return(FF)
```

```
}
```

A.8 Algoritmo para a medida FD_{1i}

Algoritmo da Seção 4.4.

#main author: Kévin Allan Sales Rodrigues

uma medida de influencia

#' F1 Distance

FD1 = function(y,X){

 n = length(y)

 ###Calculando taoi que e o tao sem a i-esima observacao

 ###e calculando os betai que sao betai[,i]

 beta = coef(ladfit(as.matrix(X), y))

 betai = coef(ladfit(as.matrix(X)[-1,], y[-1]))

 taoi = sum(abs(y-cbind(1,X)%*%as.matrix(betai)[,1]))

 for (i in 2:n){

 betai = cbind(betai, coef(ladfit(as.matrix(X)[-i,], y[-i])))

 taoi = c(taoi, sum(abs(y-cbind(1,X)%*%as.matrix(betai)[,i])))

 }

 taoi = taoi/(n-1)

 ###Calculando LD(betai, taoi)

 tao = sum(abs(y-cbind(1,X)%*%as.matrix(coef(ladfit(X, y)))[,1]))/n

 FF = tao^(-1)*sum(abs(cbind(1,X)%*%beta -cbind(1,X)%*%betai[,1]))

 for (i in 2:n){

 FF = c(FF, tao^(-1)*sum(abs(cbind(1,X)%*%beta -cbind(1,X)%*%betai[,i])))

 }

 plot(FF, main="F1 Distance", xlab="Indices",ylab="F1 Distance")

 return(FF)

}

A.9 Algoritmo para calcular os resíduos quantílicos

Algoritmo voltado para a Seção 3.2.

```
#main author: Kévin Allan Sales Rodrigues
```

```
#### a test to error distribution
```

```
quantileResiduals = function(model){
```

```
#### resíduos do modelo
```

```
#model$OAR[,3]
```

```
residuos =model$OAR[,3]
```

```
#### lambda do modelo
```

```
#model$tao
```

```
tao= model$tao
```

```
#### resíduos padronizados
```

```
#model$OAR[,3]/model$tao
```

```
###Indicator Function
```

```
# ifelse((x>=0),1,0)
```

```
##número de observações
```

```
n = length(residuos)
```

```
##numero de parametros
```

```
p = dim(model$coef)[1]
```

```
###excluindo resíduos nulos
```

```
result =NULL
```

```
for(i in 1:n){
```

```
  if(residuos[i]!=0){
```

```
    result=c(result,residuos[i])}
```

```
}
```

```
###ordenando resíduos em ordem decrescente
```

```

result =sort(result,decreasing = TRUE)

### Laplace Cumulative Distribution Function
FDA = function(x){ifelse((x>=0),(1-(1/2)*exp(-x)),(1/2)*exp(x))}

###Quantile Residual
quantileResidual =qnorm(FDA(result/tao))

#stats::qqnorm(quantileResidual, main="Quantile Residuals",
ylab="Quantile Residuals", ylim=c(-4,4))
#stats::qqline(quantileResidual)

return(quantileResidual)
}

```


Referências Bibliográficas

- Afifah, R., Andriyana, Y. e Jaya, I. G. N. M. (2017). Robust geographically weighted regression with least absolute deviation method in case of poverty in java island. *AIP Conference Proceedings*, **1827**(1), 020023. doi: 10.1063/1.4979439. URL <https://aip.scitation.org/doi/abs/10.1063/1.4979439>. Citado na pág. 2
- André, C. D. S., Elian, S. N., Narula, S. C. e Tavares, R. A. (2000). Coefficients of determinations for variable selection in the msae regression. *Communications in Statistics - Theory and Methods*, **29**, 623–642. Citado na pág. 45
- Barrodale, I. e Roberts, F. D. K. (1973). An improved algorithm for discrete L_1 linear approximation. *SIAM Journal on Numerical Analysis*, **10**, 839–848. Citado na pág. 10
- Barrodale, I. e Roberts, F. D. K. (1974). Solution of an overdetermined system of equations in the L_1 norm. *Communications of the ACM*, **17**, 319–320. Citado na pág. 10
- Bassett, G. e Koenker, R. (1978). Asymptotic theory of least absolute error regression. *Journal of the American Statistical Association*, **73**, 618–622. Citado na pág. 15
- Bezerra, M. I. S. (1990). Métodos para análise de regressão L_1 . Dissertação de Mestrado, IMECC/UNICAMP, Brasil. Citado na pág. 2, 14
- Birkes, D. e Dodge, Y. (1993). *Alternative Methods of Regression*. New York, John Wiley & Sons. Citado na pág. 2, 3, 10, 36
- Bloomfield, P. e Steiger, W. L. (1983). *Least Absolute Deviations: Theory, Applications, and Algorithms*. Birkhauser, Boston. Citado na pág. 2, 3
- Boscovich, R. J. (1757). De litteraria expeditione per pontificiam ditionem, et synopsis amplioris operis, ac habentur plura ejus ex exemplaria etiam sensorum impressa. *Bononiensi Scientiarum et Artum Instituto Atque Academia Commentarii*, **4**, 353–396. Citado na pág. 2
- Boscovich, R. J. (1760). De recentissimis graduum dimensionibus, et figura, ac magnitudine terrae inde derivanda. *Philosophiae Recentioris, a Benedicto Stay in Romano Archigynasis Publico Eloquentare Professore, vesibus traditae, Libri X, cum adnotianibus et Supplementis P. Rogerii Joseph Boscovich, S. J.*, **2**, 406–426. Citado na pág. 2
- Box, G. E. P. e Cox, D. R. (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society, Series B*, **26**, 211–252. Citado na pág. 24
- Carey, W. D. (2010). *Current Clinical Medicine, 2nd Edition*. Saunders. Citado na pág. 50
- Charnes, A., Cooper, W. W. e Ferguson, R. O. (1955). Optimal estimation of executive compensation by linear programming. *Management Science*, **1**, 138–151. Citado na pág. 9

- Cook, R. D. e Weisberg, S. (1982). *Residuals and Influence in Regression*. New York, Chapman & Hall. Citado na pág. 37
- Cook, R. D., Peña, D. e Weisberg, S. (1988). The likelihood displacement: a unifying principle for influence measures. *Communications in Statistics - Theory and Methods*, **17**, 623–640. Citado na pág. 34
- Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, **19**(1), 15–18. Citado na pág. 35
- Cox, D. R. e Snell, E. J. (1968). A general definition of residuals (with discussion). *Journal of the Royal Statistical Society, Series B*, **30**, 248–275. Citado na pág. 23
- Cox, R. D. e Hinkley, D. V. (1974). *Theoretical statistics*. London, Chapman and hall Ltd. Citado na pág. 14
- Cysneiros, F. J. A., Paula, G. A. e Galea, M. (2005). *Modelos Simétricos Aplicados*. São Paulo: ABE. 9a Escola de Modelos de Regressão. Citado na pág. 16
- Dielman, T. E. (1984). Least absolute value estimation in regression models: An annotated bibliography. *Communications in Statistics- Theory and Methods*, **4**, 513–541. Citado na pág. 2
- Dielman, T. E. (2005). Least absolute value regression: recent contributions. *Journal of Statistical Computation and Simulation*, **75**(4), 263–286. Citado na pág. 2
- Dodge, Y. (1997). Lad regression for detecting outliers in response and explanatory variables. *Journal of Multivariate Analysis*, **61**, 144–158. Citado na pág. 2, 4, 27, 31, 45
- Dodge, Y. (2002). *Statistical Data Analysis Based on the L_1 -norm and Related Methods*. Birkhäuser. Citado na pág. 2
- Dunn, P. K. e Smyth, G. K. (1996). Randomised quantile residuals. *Randomised quantile residuals. Journal of Computational and Graphical Statistics*, **5**, 236–244. Citado na pág. 23
- Edgeworth, F. Y. (1887). A new method of reducing observations relating to several quantities. *Philosophical Magazine (5th Series)*, **24**, 222–223. Citado na pág. 2
- Edgeworth, F. Y. (1888). On a new method of reducing observations relating to several quantities. *Philosophical Magazine (5th Series)*, **25**, 184–191. Citado na pág. 2
- Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, **7**, 1–26. Citado na pág. 14
- Elian, S. N., André, C. D. S. e Narula, S. C. (2000). Influence measure for the L_1 regression. *Communications in Statistics - Theory and Methods*, **29**, 837–849. Citado na pág. 27, 34
- Ellis, S. P. e Morgenthaler, S. (1992). Leverage and breakdown in L_1 regression. *Journal of the American Statistical Association*, **87**, 143–148. Citado na pág. 27, 30, 31
- Gastwirth, J. L., Gel, Y. R., Hui, W. L. W., Lyubchich, V., Miao, W. e Noguchi, K. (2017). *lawstat: Tools for Biostatistics, Public Policy, and Law*. URL <https://CRAN.R-project.org/package=lawstat>. R package version 3.2. Citado na pág. 4, 23, 39

- Goldbarg, M. C. e Luna, H. P. L. (2005). *Otimização combinatória e programação linear: modelos e algoritmos*. 2. ed. rev. e atual. Rio de Janeiro: Elsevier. Citado na pág. 10
- Harville, D. A. (1997). *Matrix algebra from a statistician's perspective*. New York: Springer. Citado na pág. 12
- Koenker, R. e Bassett, G. (1978). Regression quantiles. *Econometrica*, **46**, 33–50. Citado na pág. 8
- Laplace, P. S. (1793). Sur quelques points du système du monde. mémoires de l'académie royale des sciences de paris. páginas 1–87. Reprinted in his Oeuvres, Vol. 11, Gauthier et Fils, Paris, 1895, pp. 477-558. Citado na pág. 2
- Legendre, A. M. (1805). *Nouvelles Méthodes pour la Détermination des Orbites des Comètes*. Firmin Didot. Citado na pág. 1
- Metropolis, N. e Ulam, S. (1949). The monte carlo method. *Journal of the American Statistical Association*, **44**, 335–341. Citado na pág. 22
- Narula, S. C. e Wellington, J. F. (1977). Prediction, linear regression and minimum sum of absolute errors. *Technometrics*, **19**, 185–190. Citado na pág. 7, 39, 40
- Narula, S. C. e Wellington, J. F. (1985). Interior analysis for the minimum sum of absolute errors regression. *Technometrics*, **27**, 181–188. Citado na pág. 7, 35
- Neyman, J. e Pearson, E. S. (1933). On the problems of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London*, **231A**, 289–338. Citado na pág. 15
- Parker, I. (1988). Transformations and influential observations in minimum sum of absolute errors regression. *Technometrics*, **30**, 215–220. Citado na pág. 24
- Puig, P. e Stephens, M. A. (2000). Tests of fit for the laplace distribution, with applications. *Technometrics*, **42**, 417–424. Citado na pág. 19, 22, 23
- R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Citado na pág. 4, 22, 39
- Rodrigues, K. A. S. (2019). *LadR: Routines for Fit, Inference and Diagnostics in LAD Models*. version 1.0.7. CRAN, <https://CRAN.R-project.org/package=LadR>. Citado na pág. 4, 39
- Rosenberg, B. e Carlson, D. (1977). A simplex approximation of the sampling distribution of least absolute residual regression estimates. *Communications in Statistics - Simulation and Computation*, **6**, 421–437. Citado na pág. 13
- Rousseeuw, P. J. e Leroy, A. M. (2003). *Robust Regression and Outlier Detection*. Wiley. Citado na pág. 1, 2, 28, 29
- Santos, U. P., Garcia, M. L. S. B., Braga, A. L. F., Pereira, L. A. A., Lin, C. A., de André, P. A., de André, C. D. S., da Motta Singer, J. e Saldiva, P. H. N. (2016). Association between traffic air pollution and reduced forced vital capacity: A study using personal monitors for outdoor workers. *PLoS ONE*, **11**(10), e0163225. doi: 10.1371/journal.pone.0163225. Citado na pág. 50, 51, 53

- Searle, S. R. (2006). *Matrix Algebra Useful for Statistics*. New York, John Wiley & Sons. Citado na pág. 12
- Seber, G. A. F. (2008). *A matrix handbook for statisticians*. Wiley. Citado na pág. 12
- Shao, J. (2005). *Mathematical Statistics: Exercises and Solutions*. Springer. ISBN 9780387249704. URL <https://books.google.com.br/books?id=kEqNJXOxaCoC>. Citado na pág. 15
- Shapiro, S. S. e Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, **52**, 591–611. Citado na pág. 24, 44, 56
- Sun, R.-B. e Wei, B.-C. (2004). On influence assessment for lad regression. *Statistics & Probability Letters*, **67**, 97–110. Citado na pág. 27, 35, 36, 37
- Tavares, R. A. (1998). Seleção de variáveis em regressão L_1 . Dissertação de Mestrado, IME/USP, Brasil. Citado na pág. 10, 67
- Villegas, C., Paula, G. A., Cysneiros, F. J. A. e Galea, M. (2013). Influence diagnostics in generalized symmetric linear models. *Computational Statistics & Data Analysis (Print)*, **59**, 161–170. Citado na pág. 16