**Critérios robustos de seleção
de modelos de regressão e
identificação de pontos aberrantes**

Alia Garrudo Guirado

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

Programa: Estatística

Orientador: Profa. Dra. Silvia Nagib Elia

Durante o desenvolvimento deste trabalho a autora recebeu auxílio financeiro da CAPES

São Paulo, fevereiro de 2019

Critérios robustos de seleção de modelos de regressão e identificação de pontos aberrantes

Esta versão da dissertação contém as correções e alterações sugeridas pela Comissão Julgadora durante a defesa da versão original do trabalho, realizada em 8 de março de 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. João Ricardo Sato - UFABC
- Prof. Dr. Rinaldo Artes - INSPER

AGRADECIMENTOS

Agradeço à minha família que sempre esteve ao meu lado, apoiaram todas minhas decisões e ajudaram a me levantar sempre que cai.

A meu esposo, por ser meu companheiro nesta aventura fora de nosso país, por estar sempre ao meu lado, pelo amor, paciência, compreensão e apoio em todos os momentos.

À Professora Silvia Nagib Elian por ter-me orientado nesta etapa da minha vida, pelas conversas e ensinamentos.

Agradeço a todos os professores que ajudaram na minha formação como aluna do Programa e me ensinaram a amar cada dia mais a Estatística.

Agradeço à CAPES pelo suporte financeiro.

A todos meus amigos que me receberam de braços abertos quando cheguei no Brasil. Um agradecimento especial ao Raidiel e Darsys pelo apoio com os dados reais analisados na dissertação.

Aos que direta ou indiretamente fizeram possível a realização desta dissertação.

A todos muito obrigada!

RESUMO

GARRUDO GUIRADO, A. **Crterios robustos de seleo de modelos de regresso e identificao de pontos aberrantes**. 2019. Dissertao (Mestrado) - Instituto de Matemtica e Estatstica, Universidade de So Paulo, So Paulo, 2019.

A Regresso Robusta surge como uma alternativa ao ajuste por mnimos quadrados quando os erros so contaminados por pontos aberrantes ou existe alguma evidncia de violao das suposies do modelo. Na regresso clssica existem crterios de seleo de modelos e medidas de diagnstico que so muito conhecidos. O objetivo deste trabalho apresentar os principais crterios robustos de seleo de modelos e medidas de deteco de pontos aberrantes, assim como analisar e comparar o desempenho destes de acordo com diferentes cenrios para determinar quais deles se ajustam melhor a determinadas situaes.

Os crterios de validao cruzada usando simulaes de Monte Carlo e o Crterio de Informao Bayesiano so conhecidos por desenvolver-se de forma adequada na identificao de modelos. Na dissertao confirmou-se este fato e alm disso, suas alternativas robustas tm se destacado neste aspecto. A anlise de resduos constitui uma forte ferramenta da anlise diagnstica de um modelo, no trabalho detectou-se que a anlise clssica de resduos sobre o ajuste do modelo de regresso linear robusta, assim como a anlise das ponderaes das observaes, so medidas de deteco de pontos aberrantes eficientes.

Foram aplicados os crterios e medidas analisados ao conjunto de dados obtido da Estao Meteorolgica do Instituto de Astronomia, Geofsica e Cincias Atmosfricas da Universidade de So Paulo para detectar quais variveis meteorolgicas influem na temperatura mnima diria durante o ano completo, e ajustou-se um modelo que permite identificar os dias associados entrada de sistemas frontais.

Palavras-chave: regresso robusta, seleo de modelos, identificao de pontos aberrantes, AIC, BIC, C_p , validao cruzada, R^2 .

ABSTRACT

GARRUDO GUIRADO, A. **Robust model selection criteria in regression and outliers identification**. 2019. Dissertação (Mestrado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2019.

Robust Regression arises as an alternative to least squares method when errors are contaminated by outliers points or there are some evidence of violation of model assumptions. In classical regression there are several criteria for model selection and diagnostic measures that are well known. The objective of this work is to present the main robust criteria of model selection and outliers detection measures, as well as to analyze and compare their performance according to different stages to determine which of them fit better in certain situations.

The cross-validation criteria using Monte Carlo simulations and Bayesian Information Criterion are known to be adequately developed in model identification. This fact was confirmed, and in addition, its robust alternatives also stand out in this aspect. The residual analysis is a strong tool for model diagnostic analysis, in this work it was detected that the classic residual analysis on the robust linear model regression fit, as well as the analysis of the observations weights, are efficient measures of outliers detection points.

The analyzed criteria and measures were applied to the data set obtained from the Meteorological Station of the Astronomy, Geophysics and Atmospheric Sciences Institute of São Paulo University to detect which meteorological variables influence the daily minimum temperature during the whole year, and was fitted a model that allows identify the days associated with the entry of frontal systems.

Keywords: robust regression, model selection, outliers identification, AIC, BIC, C_p , cross-validation, R^2 .

SUMÁRIO

Agradecimentos

Resumo	i
Abstract	ii
1 Introdução	1
1.1 Estrutura da dissertação	4
2 Critérios clássicos de seleção de modelos na regressão	5
2.1 Critérios clássicos da Classe 1	5
2.1.1 Cp de Mallows	5
2.1.2 Coeficiente de Informação de Akaike, AIC	6
2.1.3 Coeficiente de Informação Bayesiano, BIC	7
2.1.4 Coeficiente de Determinação R^2	7
2.1.5 Validação cruzada	8
2.2 Critérios clássicos da Classe 2	9
2.2.1 Seleção do tipo Stepwise	10
2.2.2 Fator de Inflação da Variância, VIF	11
3 Critérios robustos de seleção de modelos	13
3.1 Critérios robustos da Classe 1	14
3.1.1 Cp Robusto (RCp)	14
3.1.2 AIC robusto (RAIC)	15
3.1.3 BIC robusto (RBIC)	15
3.1.4 R^2 robusto	16
3.1.5 Validação cruzada em regressão robusta	16
3.1.6 Outros Critérios da classe 1	17
3.2 Critérios Robustos da classe 2	17

3.2.1	Stepwise Robusto	17
3.2.2	VIF robusto	18
4	Técnicas de diagnóstico para Regressão Robusta	20
4.1	Técnicas de diagnóstico clássicas	21
4.1.1	Matriz Hat	21
4.1.2	Resíduos	22
4.1.3	Distância de Cook	23
4.1.4	Distância de Mahalanobis	24
4.2	Técnicas de diagnóstico robustas	25
4.2.1	Matriz Hat	25
4.2.2	Resíduos robustos	26
4.2.3	Distância de Cook robusta	26
4.2.4	Distância de Mahalanobis robusta	26
4.2.5	Classificação por peso	27
5	Simulação e Aplicações	28
5.1	Avaliação do desenvolvimento dos critérios de seleção de modelos	28
5.1.1	Desenho da simulação	30
5.1.2	Resultados da Simulação	31
5.1.2.1	Seleção de modelos	31
5.1.2.2	Diagnóstico do ajuste	39
5.2	Exemplos e aplicação em dados reais	42
5.2.1	Dados Stack loss	43
5.2.2	Dados Coleman	45
5.2.3	Dados reais	48
6	Conclusões	55
	Apêndices	57
A.1	Tabelas e Figuras complementares	57
A.2	Algumas funções e pacotes do R	65
	Referências Bibliográficas	68

1 INTRODUÇÃO

A análise dos critérios de seleção de modelos e das medidas de diagnóstico precisa da aplicação da regressão robusta. A seguir introduzimos a metodologia de estimação robusta para ajustar o modelos e logo compará-los segundo os critérios e identificar os pontos aberrantes pelas medidas que serão apresentadas.

As estatísticas obtidas pelo método de Mínimos Quadrados Ordinários (MQO) são altamente influenciáveis por observações com valores discrepantes, Figura (1.1).

Figura 1.1: Sensibilidade do método de MQO na presença de pontos discrepantes

Uma única observação pode modificar muito os estimadores dos parâmetros quando usado o método de mínimos quadrados em comparação com o ajuste original. Esta situação sugere o uso da Regressão Robusta como alternativa para solucionar o problema da alta sensibilidade da regressão por mínimos quadrados.

Seja considerado o seguinte modelo linear:

$$y_i = \sum_{j=1}^P x_{ij} \beta_j + \varepsilon_i \quad (1.1)$$

para $i = 1, 2, \dots, n$, em que n é o número de observações no modelo, $y_i \in \mathbb{R}$ é o valor da variável aleatória dependente y para a i -ésima observação, $x_{ij} \in \mathbb{R}$ é o valor correspondente a j -ésima variável explicativa para a i -ésima observação, considerando que $x_{i1} = 1$ corresponde ao termo constante. P é o número total de parâmetros sob o modelo completo; $\beta_j \in \mathbb{R}$ é o parâmetro desconhecido que desejamos estimar associado à variável explicativa x_j e $\varepsilon_i \in \mathbb{R}$ são os erros aleatórios associados a

i -ésima observação, que de forma geral são independentes e normalmente distribuídos com média zero e variância σ^2 .

Com o objetivo de obter estimativas que fossem pouco sensíveis aos desvios das suposições do modelo, [Huber et al. \(1964\)](#) propõe uma generalização da estimação por mínimos quadrados ordinários minimizando em β_j uma função de perda ρ conveniente que permita obter estimativas não viesadas e eficientes para os β_j , ou seja, minimizando

$$\sum_{i=1}^n \rho \left(y_i - \sum_{j=1}^P x_{ij} \beta_j \right).$$

Algumas propriedades razoáveis para esta função ρ seriam satisfazer à condição $\rho(0) = 0$, ser uma função não negativa, simétrica e monótona [Fox et al. \(2002\)](#). Se ρ for diferenciável com respeito aos parâmetros β_j , seja $\psi = \rho'$, os estimadores $\hat{\beta}_j$ para β_j podem ser obtidos facilmente encontrando a solução em β_j do sistema:

$$\sum_{i=1}^n \psi \left(y_i - \sum_{j=1}^P x_{ij} \beta_j \right) x_{ij} = 0, \quad \text{para } j = 1, \dots, P. \quad (1.2)$$

A estimativa robusta do parâmetro de escala σ , s , é introduzida no sistema para fazer com que o estimador $\hat{\beta}_j$ seja invariante em escala, se $r_i = y_i - \sum_{j=1}^P x_{ij} \beta_j$, o sistema (1.2) se transforma no sistema:

$$\sum_{i=1}^n \psi \left(\frac{r_i}{s} \right) x_{ij} = 0, \quad \text{para } j = 1, \dots, P. \quad (1.3)$$

Geralmente, a estimativa s da escala, quando os erros são normalmente distribuídos, é obtida pelo método do Desvio Absoluto da Mediana (MAD), [Gauss \(1816\)](#). Tal escolha é considerada boa, pois assim consegue-se garantir a robustez e consistência do estimador $\hat{\beta}_j$, [Montgomery et al. \(2012\)](#). [Sommer e Staudte \(1995\)](#) propõem trabalhar com o estimador s reajustado pelo desvio absoluto da mediana, por ser mais resistente a pontos aberrantes. O desvio absoluto da mediana pode ser calculado como:

$$MAD(\hat{\mathbf{y}}) = 1,483 \underset{i}{med} \{ |\hat{\mathbf{y}} - \underset{i}{med}(\hat{\mathbf{y}})|, i = 1, 2, \dots, n \}, \quad (1.4)$$

em que $\hat{\mathbf{y}}$ é o vetor das observações dependentes estimado e a constante $1,483 = 1/(\Phi^{-1}(3/4))$ é o fator de correção que faz com que MAD seja um estimador não viesado e consistente para σ , sendo Φ a função de distribuição acumulada da distribuição normal padrão, quando se assume distribuição normal para os erros.

Esta classe de estimadores é conhecida como estimadores M. A grande vantagem de considerar este estimador é sua flexibilidade e utilizando a estimativa do parâmetro de escala auxiliar, obtemos um estimador que pode ser generalizado para o caso multiparamétrico, [Huber et al. \(1964\)](#).

Mas, como escolher o modelo adequado dentre todas as variáveis explicativas disponíveis? Pela necessidade de encontrar o modelo que se ajuste melhor à maioria dos dados, foram desenvolvidos novos critérios que foram se aprimorando ao passar dos anos. Na regressão clássica já existiam critérios clássicos de seleção de modelos, tais como C_p de Mallows (1973), o Critério de Informação de Akaike (1973), AIC, o Critério de Informação Bayesiano Schwarz *et al.* (1978), BIC, o Coeficiente de Determinação R^2 , Rao *et al.* (1973); Wright (1921), Validação Cruzada CV, Larson (1931), entre outros. No entanto, estes critérios também podem ser altamente influenciados por valores discrepantes e assim, necessitamos procedimentos robustos para escolher o modelo que se ajusta melhor à maioria dos dados. Foi pensando nisso que inúmeros pesquisadores estudaram as bases das versões robustas destes critérios, que foram adaptadas e apresentadas para os estimadores robustos mais utilizados, incluindo o estimador M.

Ronchetti e Staudte (1994) apresentam a versão robusta do C_p baseada no estimador M clássico do tipo Huber generalizado e mostram que este método tem um desenvolvimento adequado na presença de observações discrepantes. Sommer e Staudte (1995) fornecem a aproximação do C_p quando usada a função de escore de Huber. A versão robusta do AIC foi introduzida por Ronchetti (1982), e uma versão similar do BIC robusto foi apresentada por Machado (1993). O coeficiente de explicação R^2 é um critério muito conhecido e também foram analisadas alternativas robustas para este critério, Croux e Dehon (2003); Maronna *et al.* (2006); Renaud e Victoria-Feser (2010). Outro dos métodos de seleção de variáveis mais usados é o procedimento de validação cruzada, mesmo tendo um alto custo computacional e de tempo quando as dimensões dos dados são elevadas, este procedimento é muito competente, por isso Ronchetti *et al.* (1997) introduz uma alternativa robusta para avaliar os modelos considerados neste procedimento.

Existem numerosos critérios de seleção de modelos de regressão. Eles podem ser divididos em duas classes, a primeira formada por aqueles critérios que consideram todas as possíveis combinações das variáveis explicativas. Estes avaliam os modelos e escolhem o modelo que melhor se adequa aos dados com base no critério. A segunda classe é formada por procedimentos de seleção sequencial em que uma variável explicativa é inserida (ou removida) no modelo, com base em medidas que podem mudar de um passo para outro e que são computadas para todas as possíveis variáveis até satisfazer uma condição específica. Os procedimentos da primeira classe podem não ser muito eficientes quando o número de parâmetros ou o tamanho de amostra forem muito grandes. Neste caso, é necessário construir $2^{(P-1)} - 1$ possíveis modelos e o custo computacional e de tempo pode ser muito maior que aqueles critérios que procedem de forma sequencial e determinam a influência das variáveis explicativas passo a passo. Alguns dos critérios da segunda classe *Stepwise*, *Forward* e *Backward selection*, VIF, Efroymson (1960); Lin *et al.* (2011) e outros foram estudados amplamente

nas ultimas décadas.

A maioria dos critérios da Classe 1 podem ser expressos como a soma de dois termos:

$$criterio_p = g(\mathbf{X}_p; \mathbf{y}) + C(n; p), \quad (1.5)$$

em que $\mathbf{X}_p$ é a matriz de planeamento com $(p - 1)$ variáveis explicativas. O primeiro termo $g(\cdot; \cdot)$ é uma função arbitrária que depende das variáveis explicativas consideradas no modelo e da variável dependente, e o segundo termo C é uma função constante que depende do tamanho de amostra n e do número de parâmetros p do modelo.

1.1 Estrutura da dissertação

Os critérios clássicos de seleção de modelos na regressão serão apresentados no Capítulo 2 e suas contrapartes robustas mais utilizadas no Capítulo 3, junto a outras alternativas. No Capítulo 4 são revisadas as principais técnicas de diagnóstico clássicas e robustas para a detecção de pontos discrepantes. Também apresentamos no Capítulo 5 um estudo de simulação para avaliar as vantagens e desvantagens dos critérios de seleção de modelos e as medidas de detecção de pontos aberrantes analisados nos capítulos anteriores. São avaliados e comparados os desempenhos dos critérios e das medidas em vários cenários criados computacionalmente e aplicados para alguns exemplos tradicionais e dados reais. Finalmente no Capítulo 6, são apresentadas as conclusões e recomendações e no Apêndice são apresentadas Tabelas e Figuras complementares, assim como os códigos no R utilizados durante o trabalho de dissertação apresentados de forma resumida.

2 CRITÉRIOS CLÁSSICOS DE SELEÇÃO DE MODELOS NA REGRESSÃO

A seleção de modelos é uma parte importante em toda análise estatística. Muitos autores têm se envolvido e estudado como selecionar o “melhor” modelo dentre todos os possíveis. Neste capítulo apresentaremos os principais critérios de seleção de modelos de regressão não robustos. De modo geral, o objetivo destes critérios é encontrar o “melhor modelo” minimizando a perda quadrática $\sum_{i=1}^n r_i^2$.

2.1 Critérios clássicos da Classe 1

Os critérios da Classe 1 são aqueles que constroem todos os $2^{P-1} - 1$ modelos possíveis para os dados, e podem ser expressos como a soma de dois termos (1.5). Estes critérios podem ter um alto custo computacional quando o número de parâmetros P é muito grande ($P > 20$). A seguir são estudados os principais critérios clássicos de seleção de modelos que pertencem a esta classe.

2.1.1 Cp de Mallows

Um dos métodos mais usados na seleção de modelos de regressão consiste em gerar todos os modelos possíveis e compará-los em função da estatística C_p , Mallows (1973). Para um modelo particular com p parâmetros a estatística C_p é definida como

$$C_p = \frac{SQR_p}{\hat{\sigma}^2} - n + 2p, \quad (2.1)$$

em que SQR_p é a soma de quadrados dos resíduos para o submodelo com p parâmetros, $\hat{\sigma}^2$ é o estimador da variância amostral dos erros σ^2 sob o modelo completo e n é o tamanho da amostra. Se o submodelo com p parâmetros é melhor, então o valor do critério C_p deveria ser menor ou igual ao número de parâmetros p do modelo e aquele com menor valor do critério será um modelo relativamente preciso.

O gráfico C_p contra p permite detectar facilmente o melhor modelo, desde que o número de variáveis explicativas não seja muito grande. Para o C_p de Mallows, o critério se divide em $g(\mathbf{X}_p; \mathbf{y}) = SQR_p/\hat{\sigma}^2$ e $C(n; p) = 2p - n$. Note-se que em geral para critérios da Classe 1, temos vários submodelos com um tamanho particular, temos $\binom{P-1}{p}$ submodelos com p parâmetros. Se temos vários submodelos de igual tamanho p que contém as variáveis explicativas importantes, é muito provável que aquele com menor valor do C_p satisfaça a condição $C_p < p$, [Gilmour \(1996\)](#), e como a estatística C_p é baseada na estimativa por mínimos quadrados, este procedimento não protege completamente contra o super-ajuste, pois quanto maior o número de variáveis no modelo, menor será o valor da SQR. Por isso, o modelo completo tem muita chance de estar entre os modelos candidatos a serem selecionados por este critério.

2.1.2 AIC

O Critério de Informação de Akaike é um dos critérios mais populares para seleção de modelos de regressão. É definido como:

$$AIC_p = -2\log(L_p) + 2p, \quad (2.2)$$

para L_p função de verossimilhança do modelo e p o número de parâmetros considerados. Este critério pode ser aplicado em vários contextos: regressão, análise fatorial, [Akaike \(1987\)](#), variáveis latentes, [Eid e Langeheine \(1999\)](#), entre outros. O objetivo da seleção do modelo pelo critério AIC é estimar a perda de informação, segundo este raciocínio, o “melhor” modelo será aquele com menor valor do AIC, pois vai maximizar o logaritmo da função de verossimilhança e minimizar a perda de informação, [Akaike \(1973\)](#).

Quando o tamanho da amostra é pequeno, é muito provável que o critério AIC selecione modelos com muitas variáveis explicativas, gerando super-ajuste, [Claeskens et al. \(2008\)](#). Para evitar o problema do tamanho da amostra, foi desenvolvida uma correção do AIC chamada AIC corrigido (AIC_c) [Burnham e Anderson \(2003\)](#); [Hurvich e Tsai \(1989\)](#),

$$AIC_c = AIC + \frac{2(p^2 + p)}{n - p - 1}. \quad (2.3)$$

[Bhansali e Downham \(1977\)](#) apresentam a generalização do critério considerando o efeito do tamanho de amostra como:

$$AIC_{\alpha,p} = -2\log(L_p) + \alpha p, \quad (2.4)$$

mantendo a seleção do modelo que minimiza o valor do $AIC_{\alpha,p}$ com respeito a p e sendo α o termo de correção, sem o qual estaríamos simplesmente maximizando a verossimilhança. [Stone \(1977\)](#) demonstrou que o critério de Akaike generalizado é assintoticamente equivalente à quantidade

$-2\log(L_p) + 2\text{tr}(L_2^{-1}L_1)$, em que $\text{tr}(\cdot)$ indica o traço da matriz, $L_{2(p \times p)}$ matriz das segundas derivadas do logaritmo da função de verossimilhança com respeito aos parâmetros β_j e $L_{1(p \times p)}$ é a matriz do produto das derivadas de primeira ordens; então $\alpha p \approx 2\text{tr}(L_2^{-1}L_1)$ é uma proposta adequada. De modo geral, esta penalidade reduz o efeito do super-ajuste. Como o critério $AIC_{\alpha,p}$ também pertence à Classe 1, podemos dividi-lo da seguinte forma: $g(\mathbf{X}_p; \mathbf{y}) = -2\log(L_p)$ e $C(n; p) = \alpha p$.

2.1.3 BIC

A formulação do Critério de Informação Bayesiano (BIC) ou de Schwarz (SIC) [Schwarz et al. \(1978\)](#) é muito parecida com a do AIC, e pode ser considerada um caso particular da generalização (2.4),

$$BIC_p = -2\log(L_p) + p\log(n).$$

O BIC é outro dos critérios de seleção de modelos mais populares. Este critério também pode ser ineficiente quando aplicado em problemas de alta dimensionalidade no espaço paramétrico.

Sob a suposição de que os erros do modelo são independentes e identicamente distribuídos de acordo com uma distribuição normal, BIC pode ser expressado em função da SQR como:

$$BIC_p = n\log\left(\frac{SQR}{n}\right) + p\log(n). \quad (2.5)$$

O BIC é uma função crescente da soma de quadrados do resíduo e do número de parâmetros, portanto, um valor baixo de BIC implica em menos variáveis explicativas no modelo ou melhor ajuste ou ambos.

Este critério pode ser separado nos termos $g(\mathbf{X}_p; \mathbf{y}) = n\log\left(\frac{SQR}{n}\right)$ e $C(n; p) = p\log(n)$, note-se que BIC_p em (2.5) é similar ao AIC_p em (2.2), pois ambos tentam resolver o problema de super-ajuste agregando o termo constante, que para o BIC_p é maior que para o AIC_p . O modelo selecionado será aquele com menor valor do BIC_p .

2.1.4 Coeficiente de Determinação R^2

O Coeficiente de Determinação ou explicação R^2 é a proporção de variação da variável dependente $\mathbf{y}$ que pode ser explicada a partir das variáveis independentes [Rao et al. \(1973\)](#). Este critério apresenta valores entre 0 e 1 e pode ser considerado como um indicador da qualidade do ajuste para os modelos considerados. A formulação do R^2 para um modelo com p parâmetros é:

$$R_p^2 = 1 - \frac{SQR}{SQT},$$

para $SQR = \sum_{i=1}^n r_i^2$ soma de quadrados dos resíduos e $SQT = \sum_{i=1}^n (y_i - \bar{y})^2$ soma de quadrados totais, que é proporcional à variância dos dados. Quanto melhor for o ajuste de regressão linear dos dados em comparação com a média geral, mais próximo estará R^2 de 1. Então, $R^2 = 1$ indica que o modelo ajustado explica toda a variabilidade em $\mathbf{y}$, assim $R^2 = 0$ indica que não existe relação “linear” entre as variáveis explicativas $\mathbf{x}_j$ e a variável dependente $\mathbf{y}$.

O valor do coeficiente de determinação depende do número de observações n e das variáveis explicativas no modelo. Pode crescer quando n diminui e quando adicionamos uma nova variável explicativa, mas isso não significa que esse novo modelo seja melhor. Basta ver o caso em que $n = 2$, tem-se sempre $R^2 = 1$ pois a reta de regressão é a que passa pelos dois pontos e nesse caso o resíduo é zero para cada ponto. Como o valor de R^2 também aumenta com o número de variáveis explicativas incluídas no modelo, define-se alternativamente o coeficiente de explicação ajustado pelos graus de liberdade, R_a^2 , [Henri \(1961\)](#). Esta correção do R^2 considera uma penalidade de acordo com o número de variáveis explicativas incluídas no modelo,

$$R_a^2 = 1 - (1 - R^2) \frac{n-1}{n-p}, \quad (2.6)$$

com p o número de parâmetros do modelo analisado, incluindo o intercepto e n o tamanho de amostra. Algumas vantagens dessa nova medida são sua maior eficiência para pequenos tamanho de amostra e o fato de não aumentar com a inclusão de variáveis explicativas pouco importantes. Este critério também pode ser expressado como $g(\mathbf{X}_p; \mathbf{y}) + C(n; p)$ fazendo as transformações convenientes.

2.1.5 Validação cruzada

As primeiras ideias de validação cruzada foram usadas na área de Psicologia por [Larson \(1931\)](#) com o objetivo de obter uma estimativa mais confiável para o coeficiente de explicação R^2 . Este procedimento consiste em separar os dados em duas amostras, os dados de treinamento e os restantes dados de validação ou novos dados e avaliar o desempenho da estatística que precisamos analisar. A validação cruzada pode ser aplicada em quase qualquer área como na regressão, classificação, estimação da densidade [Bartlett et al. \(2002\)](#); [Geisser \(1975\)](#); [Stone et al. \(1984\)](#), entre outras.

Existem vários procedimentos de validação cruzada, um dos mais populares e conhecido é *Leave one out* (LOO) excluindo uma única observação de cada vez. Chamemos n_t o número de observações selecionadas para a amostra de treinamento e $n_v = n - n_t$ o número de observações restantes como número de observações para validação, no caso de LOO, $n_v = 1$. Logo para cada subconjunto de parâmetros, o processo de estimação e de validação será realizado n vezes. Outro procedimento muito comum é “Leave n_v out” (Ln_vO), onde n_v será uma quantidade maior que um, note se que

LOO é uma caso particular de Ln_vO quando $n_v = 1$. Neste caso realizaremos a análise $\binom{n}{n_v}$ vezes. Por último, mas não menos importante, temos o método k - *fold* que requer a partição da amostra em k amostras independentes de tamanhos aproximadamente iguais, e usar cada uma delas como amostras de validação de tamanho n_v . Note-se que se $k = n$, então o método k - *fold* se reduz ao LOO.

A taxa de erro da amostra de validação resultante em qualquer destes procedimentos de validação cruzada é geralmente avaliada usando o Erro Quadrático Médio $EQM = \frac{SQR}{n}$, que para a amostra de validação teria a seguinte expressão:

$$EQM = \frac{1}{n_v} \sum_{i=1}^{n_v} (y_i - \hat{y}_i)^2, \quad (2.7)$$

em que $\hat{y}_i$ é obtido a partir da amostra de treinamento, prevendo os valores da variável y na amostra de validação.

O critério de validação cruzada utilizado vai considerar uma mistura entre k - *fold* para $k = 2$ e o procedimento de Monte Carlo, [Rubinstein e Kroese \(2016\)](#). Levando em consideração que o objetivo é identificar o modelo mais adequado, [Arlot et al. \(2010\)](#) sugere utilizar um valor de n_v grande, e [Shao \(1993\)](#) sugere considerar $n_t \approx n^{3/4}$ para que o critério seja consistente no sentido de que a probabilidade de selecionar o modelo com melhor habilidade preditiva converge-se para um.

Este procedimento pode ser resumido nos seguintes passos:

1. Gerar m partições aleatórias dos dados da amostra para construção do modelo e treinamento de tamanho n_t e amostra para validação de tamanho $n_v = n - n_t$.
2. Para cada partição, construir os modelos com todas as combinações possíveis das $P - 1$ variáveis explicativas.
3. Para cada modelo calcular o EQM.
4. Repetir o terceiro passo para cada partição m ds dados.
5. Para cada modelo calcular a média dos valores do EQM obtidos em todas as partições e selecionar o modelo com menor valor da média do EQM.

Este procedimento de aplicar validação cruzada 2 - *fold* M vezes foi chamado Validação Cruzada Monte Carlo ($MCCV(n_v)$), [Shao \(1993\)](#).

2.2 Critérios clássicos da Classe 2

Por razões principalmente computacionais e por economia de tempo, os critérios de seleção do “melhor” modelo entre todos aqueles candidatos não devem ser aplicados quando a quantidade de parâmetros é muito grande. Este procedimento pode provocar um super - ajuste. Como alternativa

foram desenvolvidos outros métodos para selecionar o melhor subconjunto de variáveis explicativas de forma sequencial, adicionando ou removendo variáveis em cada passo.

Geralmente o critério para a adição ou remoção de variáveis é baseado numa estatística que geralmente é a estatística F ou t , comparando modelos com e sem as variáveis consideradas. O AIC, BIC, assim como outros critérios, também podem ser utilizados na decisão de adicionar ou remover variáveis.

Existem três abordagens principais da seleção *Stepwise*, a seleção *Forward* ou seleção antecipada, o método de eliminação *Backward*, e o procedimento bidirecional, que é uma mistura de ambos. Também existe um critério de seleção de variáveis que reduz o efeito da multicolinearidade baseado no fator de inflação da variância (*VIF*). Existem vários outros procedimentos de seleção de modelos de forma sequencial, nesta dissertação serão abordados os quatro procedimentos descritos acima por serem os mais populares.

2.2.1 Seleção do tipo Stepwise

Estes procedimentos partem de um modelo inicial que pode ser o modelo $\hat{y} = \bar{y}$, ou o modelo completo e em cada passo são adicionadas ou removidas variáveis explicativas em função da correlação existente e de um teste estatístico.

Seleção Forward

Este método de seleção inicia-se com o modelo $\hat{y} = \bar{y}$, ou seja, sem variáveis preditoras e vai testando a adição de cada variável usando um particular critério de ajuste de modelo. Se houver alguma variável que sua adição no modelo fornece uma melhoria estatisticamente significativa do ajuste, ela será incluída no modelo. Repete-se esse processo até que nenhuma outra variável explicativa produza uma melhora no ajuste segundo o critério escolhido.

Neste procedimento, a primeira variável incluída no modelo é aquela com maior correlação com a variável resposta. Uma vez adicionada a primeira variável (por exemplo $\mathbf{x}_1$), o próximo passo é determinar dentre as variáveis restantes ($\mathbf{x}_j$, $j = 2, \dots, P - 1$), qual é a mais correlacionada com a variável resposta $\mathbf{y}$ considerando a presença da primeira variável no modelo. Esta correlação é chamada correlação parcial ($r_{jy.1}$) da variável $\mathbf{x}_j$ e $\mathbf{y}$ eliminando o efeito de $\mathbf{x}_1$. É a correlação entre o resíduo do modelo atual ($\hat{\mathbf{y}} = \hat{\beta}_0 + \hat{\beta}_1 \mathbf{x}_1$) e o resíduo do modelo das novas variáveis explicativas em função da já incluída no modelo ($\hat{\mathbf{x}}_j = \hat{\alpha}_{0j} + \hat{\alpha}_{1j} \mathbf{x}_1$). Para determinar o ponto de parada, é usada a estatística F parcial. Esta estatística pode ser escrita em função do coeficiente de correlação parcial, Khan *et al.* (2007a), como :

$$F = \frac{(n - k - 1) \tilde{r}_{ky.1 \dots (k-1)}^2}{1 - r_{1y}^2 - \tilde{r}_{2y.1}^2 - \dots - \tilde{r}_{ky.1 \dots (k-1)}^2}$$

para $\tilde{r}_{ky,1\ldots(k-1)}^2$ coeficiente proporcional à correlação parcial da variável k - ésima e y em função das $k - 1$ variáveis já incluídas no modelo. Este processo é repetido adicionando a cada passo a variável com maior correlação parcial, até que a estatística F parcial não permita a inclusão de mais nenhuma variável explicativa no modelo.

A estatística F parcial associada à k - ésima variável explicativa testa a hipótese $H_0 : \beta_k = 0$ sendo que a rejeição ou não dessa hipótese implica na introdução ou manutenção da variável no modelo.

Seleção Backward

Enquanto o método Forward começa sem nenhuma variável explicativa no modelo e adiciona variáveis a cada passo, o método de Seleção Backward faz o caminho oposto; inicia-se com todas as variáveis candidatas no modelo e depois, a cada passo, testa a exclusão ou não de cada variável usando o critério de ajuste de modelo escolhido (usualmente o F parcial). Ele exclui a variável (se houver) que afeta de forma menos significativa o ajuste do modelo atual e repete esse processo até que outras variáveis possam ser excluídas sem uma perda de ajuste estatisticamente significativa.

Seleção bidirecional De modo geral, o procedimento *Stepwise* tem a opção de em cada passo selecionar uma nova variável a ser incorporada ao modelo, ou remover uma que já está presente neste, este é o caso do processo de seleção bidirecional. Este processo é uma combinação dos dois anteriores, testando em cada etapa a inclusão ou exclusão de variáveis no modelo atual.

A seleção bidirecional é uma modificação da seleção Forward em que a cada passo todas as variáveis do modelo são previamente avaliadas pelas suas estatísticas F parciais. Verifica-se, dado que as outras variáveis já estão no modelo, se essa nova variável melhora a explicação da variável y .

2.2.2 VIF

Na seleção de variáveis explicativas no modelo de regressão linear, [Lin et al. \(2011\)](#) apresentam o procedimento “VIF regression”, que se baseia em cálculos de estatísticas de teste que levam em conta a variância do estimador $\hat{\beta}_j$ associado à nova variável explicativa em função das variáveis presentes no modelo. Esta medida chamada “Fator de inflação da Variância” (VIF) avalia quanto a variância do estimador do coeficiente de regressão associado a cada variável explicativa é aumentada devido à existência de multicolinearidade.

Seja um particular modelo de regressão linear múltipla da forma

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \beta_j x_j + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I)$$

para β_j parâmetro e x_j é vetor contendo os valores da j - ésima variável explicativa para os n elemen-

tos da amostra, que desejamos incorporar no modelo, $j \in \{k+1; k+2; \dots; P\}$. Denote $\mathbf{X} = [\mathbf{1}_n \mathbf{x}_1 \mathbf{x}_2 \dots \mathbf{x}_k]$, $\tilde{\mathbf{X}} = [\mathbf{X} \mathbf{x}_j]$, $\boldsymbol{\beta} = (\beta_0; \beta_1; \dots; \beta_k)'$ e $\tilde{\boldsymbol{\beta}} = (\beta_0; \beta_1; \dots; \beta_k; \beta_j)'$.

Seja $\hat{\beta}_j$ o estimador por mínimos quadrados de β_j e $\mathbf{r} = (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y}$ o vetor de resíduos no modelo com as k primeiras variáveis explicativas. Nestas condições, temos o seguinte modelo hipotético, [Lin et al. \(2011\)](#):

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \beta_j \mathbf{x}_j + \boldsymbol{\varepsilon}, & \boldsymbol{\varepsilon} &\sim N(\mathbf{0}, \sigma^2 \mathbf{I}) \\ \mathbf{r} &= \gamma_j \mathbf{x}_j + \tilde{\boldsymbol{\varepsilon}}, & \tilde{\boldsymbol{\varepsilon}} &\sim N(\mathbf{0}, \tilde{\sigma}^2 \mathbf{I}). \end{aligned} \quad (2.8)$$

Os autores mostram que, quando β_j e γ_j são estimados por mínimos quadrados, vale a relação

$$\hat{\gamma}_j = \delta \hat{\beta}_j$$

para $\delta = \mathbf{x}_j'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{x}_j$. Logo, grandes valores de $\hat{\gamma}_j$ são indicativos de que a nova variável $\mathbf{x}_j$ é importante na explicação da variável $\mathbf{r}$, que pode ser encarada como a “parte da variável $\mathbf{y}$ ” que não é explicada pelas variáveis já incluídas no modelo.

A estatística

$$T_j = \frac{\delta^{1/2} \hat{\beta}_j}{\hat{\sigma}},$$

que depende de $\hat{\sigma}$, estimativa adequada de σ , [Lin et al. \(2011\)](#), é comparada com quantis da distribuição normal padrão para concluir sobre a inclusão de $\mathbf{x}_j$ no modelo. Grandes valores absolutos desta estatística sugerem a inclusão de $\mathbf{x}_j$ no modelo.

Este procedimento é denominado “VIF regression” devido ao fato que $\delta = 1 - R_{\mathbf{x}_j|\mathbf{X}_k}^2$, para $R_{\mathbf{x}_j|\mathbf{X}_k}^2$ coeficiente de explicação da regressão da variável $\mathbf{x}_j$, em função de todas as outras variáveis preditoras já incluídas no modelo. O valor do VIF associado à variável $\mathbf{x}_j$ é definido como:

$$VIF_j = \frac{1}{1 - R_{\mathbf{x}_j|\mathbf{X}_k}^2}. \quad (2.9)$$

Na existência de forte relação linear entre $\mathbf{x}_j$ e as variáveis $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k$, $R_{\mathbf{x}_j|\mathbf{X}_k}^2$ será próximo de 1, produzindo altos valores de VIF_j . Nestas condições, $\delta \approx 0$, implicando em pequenos valores de T_j . No caso ideal de inexistência de relação linear entre $\mathbf{x}_j$ e $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k$, $R_{\mathbf{x}_j|\mathbf{X}_k}^2 = 0$ e $VIF_j = 1$.

Este procedimento de avaliação é rápido e simples, pode ser adaptado utilizando algoritmos sequenciais como stepwise, veja Sessão (2.2.1) e evita o super-ajuste. De acordo com os resultados experimentais obtidos em [Lin et al. \(2011\)](#), o algoritmo VIF é substancialmente preciso, e é mais rápido que outros algoritmos para dados de alta dimensionalidade.

Neste capítulo apresentamos os principais critérios de seleção de modelos da regressão clássica, no capítulo próximo veremos os correspondentes destes critérios com um enfoque robusto.

3 CRITÉRIOS ROBUSTOS DE SELEÇÃO DE MODELOS

Huber *et al.* (1973) estende seus resultados de estimação robusta dos parâmetros de locação ao caso do modelo linear, o autor propõe analisar os estimadores de mínimos quadrados penalizados mediante determinados pesos (w_i) associados aos resíduos r_i . O procedimento de determinação do estimador M para β obtido pelo sistema (1.3) para ψ e a estimativa s para σ como em (1.4) também é conhecido como método de obtenção do estimador M de locação com estimativa preliminar da escala, Drumond (1989). Então para ψ função de escore $\psi = \rho'$, temos que $\hat{\beta}$ satisfaz o sistema de equações

$$\sum_{i=1}^n \psi(r_i^*) x_{ij} = 0, \quad \text{para } j = 1, \dots, P, \quad (3.1)$$

para $r_i^* = r_i/s$ resíduo padronizado pela estimativa robusta da escala obtida pelo MAD.

Uma maneira bastante comum de resolver o sistema (3.1) é usando mínimos quadrados reponderados iterativos. Desta forma obtemos a solução a partir do sistema de equações normais $\mathbf{X}'\mathbf{W}\mathbf{X}\beta = \mathbf{X}'\mathbf{W}\mathbf{y}$, para $\mathbf{X}_{(n \times p)}$ matriz de planejamento e $\mathbf{W}_{(n \times n)}$ matriz diagonal dos pesos cujo i -ésimo elemento da diagonal é dado por

$$w_i = \psi(r_i^*)/r_i^*, \quad i = 1, 2, \dots, n. \quad (3.2)$$

Usaremos neste trabalho a função escore ψ de Huber, pela sua vantagem de não eliminação total da influência de uma observação com resíduo muito grande, pois mantém todas as observações, limitando o efeito das mesmas, de acordo com seu resíduo,

$$\psi_c(r) = \begin{cases} r & |r| \leq c \\ c \sin(r) & |r| > c \end{cases}, \quad (3.3) \quad \text{em} \quad \rho_c(r) = \begin{cases} r^2/2 & |r| \leq c \\ c(|r| - c/2) & |r| > c \end{cases}. \quad (3.4)$$

que $\sin(r)$ toma valor 1 se $r \geq 0$, e -1 se $r < 0$. Os gráficos destas duas funções são apresentados na Figura (3.1) junto ao peso recebido para $c = 1,345$.

Para erros normalmente distribuídos, escolhemos o valor $c = 1,345$, este valor é determinado de tal forma que a eficiência dos testes de hipóteses $H_0 : \beta_j = 0, j = 2, 3, \dots, P$ do estimador M para β sejam de 95%, [Ronchetti \(1982\)](#).

Figura 3.1: Família de funções de Huber: função de perda ρ , função de escore $\psi = \rho'$ e peso $w = \psi(x)/x$ para $c = 1,345$

Nas seguintes seções são analisados os critérios robustos de seleção de modelos na regressão linear.

3.1 Critérios robustos da Classe 1

Nesta seção serão analisadas as versões robustas dos critérios clássicos vistos na Sessão 2.1, além de outras versões que forem desenvolvidas que também são consideradas critérios da Classe 1.

3.1.1 Cp Robusto (RCp)

O critério C_p (2.1) é baseado na estimação por mínimos quadrados através das estatísticas SQR e $\hat{\sigma}^2$; portanto é muito sensível à presença de pontos discrepantes e outras variações das suposições de normalidade. [Ronchetti e Staudte \(1994\)](#) apresentam uma versão robusta do critério C_p de Mallows (RC_p) que mostrou-se adequada quando se apresentam estas dificuldades, pois permite selecionar o modelo que se ajuste melhor à maioria dos dados. Esta quantidade é definida como

$$RC_p = \frac{W_p}{s^2} - (U_p - V_p), \quad (3.5)$$

para $W_p = \sum_i \hat{w}_i^2 r_i^2$ soma de quadrados dos resíduos ponderada calculada a partir do ajuste robusto do modelo com p parâmetros, $\hat{w}_i$ são os pesos (3.2) associados à i -ésima observação, $U_p = \sum_i \text{var}(\hat{w}_i r_i)$ e $V_p = \sum_i \text{var}(\hat{w}_i \delta_i)$ são constantes dependendo da função de peso e do número de parâmetros considerados no modelo e calculadas admitindo que $\sigma = 1$; além disso, $\delta_i = \hat{y}_i - E(y_i)$ e s^2 é o estimador

robusto para σ^2 no modelo completo obtido pelo MAD.

Note-se que quando os pesos $\hat{w}_i$ são iguais a 1, RC_p se reduz ao C_p ; quando são diferentes de 1, o cálculo da quantidade $U_p - V_p$ pode se tornar complexo. Ronchetti e Staudte (1994) [Teorema 3 e 4] fornece uma aproximação assintótica para a quantidade $U_p - V_p$. No caso do estimador M clássico usando a função escore de Huber (3.3) com constante $c = 1,345$, obtem-se a aproximação da variância $V_p = 0,982p$ e $U_p - V_p = 0,710n - 1,294p$, Sommer e Staudte (1995).

3.1.2 AIC robusto (RAIC)

O critério AIC é baseado no método de máxima verossimilhança e pode ser também muito sensível à presença de pontos discrepantes nos dados. Por esse motivo, é preferível utilizar o estimador M e uma versão robusta do critério AIC.

Ronchetti (1982) define uma versão robusta do critério AIC como extensão do método de estimação por máxima verossimilhança do estimador M,

$$RAIC_{\alpha,p} = 2 \sum_i \rho(r_i^*) + \alpha p. \quad (3.6)$$

Existem vários métodos diferentes para calcular a quantidade α , mas a maioria dos autores propõem o uso de $\alpha p = 2 \text{tra}(L_2^{-1} L_1)$, sendo $L_2 = -E[\psi']$, e $L_1 = E[\psi \psi^t]$, Stone (1977), em que $\psi' = \partial \psi / \partial \beta$. No caso do estimador M do tipo Huber, α pode ser calculado como $\alpha = 2E[\psi^2(r^*)] / E[\psi'(r^*)]$ Hampel *et al.* (1986); Ronchetti (1982). Para ψ função de escore de Huber definida como em (3.3) temos a seguinte derivada para ψ :

$$\psi'(r) = \begin{cases} 1 & |r| \leq 1,345 \\ 0 & |r| > 1,345 \end{cases}.$$

3.1.3 BIC robusto (RBIC)

O BIC também depende da função de verossimilhança, e por isso é influenciável pelos pontos aberrantes; logo tem sentido considerar a versão robusta deste critério utilizando a função de perda, similar ao caso do RAIC pois estes dois critérios são equivalentes. A versão robusta do BIC, Machado (1993), é dada por

$$RBIC = 2 \sum_i \rho(r_i^*) + p \log(n). \quad (3.7)$$

3.1.4 R^2 robusto

Quando o interesse é determinar a qualidade do ajuste, o coeficiente de explicação R^2 é uma ferramenta simples e fácil de utilizar, mas como todas as ferramentas que se baseiam na estimação por mínimos quadrados, é sensível a pontos discrepantes e a abordagem robusta apresenta uma alternativa plausível.

A versão robusta do coeficiente de explicação R^2 considerado por [Croux e Dehon \(2003\)](#) é definida como

$$RR^2 = 1 - \frac{\sum_{i=1}^n \rho(r_i^*)}{\sum_{i=1}^n \rho\left(\frac{y_i - \bar{y}}{s}\right)}$$

Lembrando que $\frac{\sum_{i=1}^n (y_i - \hat{y})^2}{(n-p)}$ e $\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{(n-1)}$ são estimadores não viesados da variância dos erros para o modelo completo e o modelo contendo somente o intercepto, se ao invés de estimar β por mínimos quadrados utilizamos o estimador robusto M, [Maronna et al. \(2006\)](#) definem o coeficiente de explicação robusto ajustado para os graus de liberdade como

$$RR_a^2 = \frac{\sum_{i=1}^n \rho\left(\frac{y_i - \bar{y}_i}{s}\right)/(n-1) - \sum_{i=1}^n \rho(r_i^*)/(n-p)}{\sum_{i=1}^n \rho\left(\frac{y_i - \bar{y}_i}{s}\right)/(n-1)}. \quad (3.8)$$

Definições alternativas de RR_a^2 podem ser encontradas em [Renaud e Victoria-Feser \(2010\)](#) e [Croux e Dehon \(2003\)](#).

3.1.5 Validação cruzada em regressão robusta

Quando dispomos de uma amostra contaminada por pontos discrepantes, a aplicação do método de validação cruzada com uma função de perda não robusta pode ter resultados insatisfatórios, [Härdle \(1984\)](#). Por esse motivo, [Ronchetti et al. \(1997\)](#) decidem comparar o valor observado y_j com sua predição robusta $\hat{y}_j$ para $j \in \{1; 2; \dots; n_v\}$, substituindo o erro quadrado médio usual EQM (2.7) pela função de perda robusta

$$\sum_{j=1}^{n_v} \rho(r_j). \quad (3.9)$$

[Ronchetti et al. \(1997\)](#) resumem o procedimento de seleção de modelos usando o método de Monte Carlo da seguinte forma:

1. Gerar m partições aleatórias dos dados em amostra para construção do modelo de treinamento de tamanho n_t e amostra para validação de tamanho $n_v = n - n_t$.

Vários autores sugerem utilizar n_t aproximadamente $n^{3/4}$, [Ronchetti et al. \(1997\)](#); [Shao \(1993\)](#), pois escolher um valor $n_t \ll n$ permite que o processo seja consistente para a identificação do modelo correto, [Arlot et al. \(2010\)](#).

2. Para cada partição, construir os modelos com todas as combinações possíveis das $P - 1$ variáveis explicativas.
3. Para cada modelo calcular o critério como definido em (3.9).
4. Repetir o terceiro passo para cada partição m dos dados.
5. Para cada modelo calcular a média dos valores do critério de predição robusto (3.9) obtidos em todas as partições e selecionar o modelo com menor valor da média da função de perda.

Este procedimento pode ser afetado por uma quantidade grande de pontos aberrantes e se o número de parâmetros e o tamanho de amostra forem grandes, pode levar mais tempo que os outros critérios definidos na Sessão 3.1 em função do número M selecionado para o procedimento de Monte Carlo.

3.1.6 Outros Critérios da classe 1

Agostinelli (2002) desenvolveu extensões para os critérios C_p e AIC baseados na metodologia de máxima verossimilhança que foram implementados e incorporados no pacote estatístico “Weighted Likelihood Estimation” (WLE) do software R.

No entanto, as medidas são bastante complexas e no momento o pacote foi retirado do repositório de pacotes do R.

3.2 Critérios Robustos da classe 2

Os critérios robustos desenvolvidos até o momento envolvem a construção do ajuste de todos os submodelos possíveis com as $P - 1$ variáveis explicativas. Com o objetivo de evitar o ajuste de todos os submodelos, foram estudadas outras abordagens robustas. Como os procedimentos apresentados na Sessão 2.2 são computacionalmente adequados em situações de alta dimensionalidade, nesta sessão serão estudadas as propostas robustas destes procedimentos da classe 2.

3.2.1 Stepwise Robusto

Os procedimentos passo a passo descritos na Seção 2.2.1 dependem de quantidades que são altamente sensíveis a pontos aberrantes e contaminações dos erros, como o coeficiente de explicação. Por isso, uma forma simples de robustificar estes procedimentos é substituindo as medidas não robustas pelas suas correspondentes robustas, de modo que tenhamos procedimentos mais resistentes a perturbações no modelo, Khan *et al.* (2007a).

Uma solução é derivar as correlações entre pares de variáveis baseadas em um estimador de covariância bivariada que seja robusto. Para conjuntos de dados muito grandes e de alta dimensão, é necessário um estimador robusto que seja muito rápido, Pervin e Rahman (2018).

Huber (1981) sugere uma ideia de padronização dos dados denominada “Winsorização”, e o cálculo do coeficiente de correlação de Pearson usual para os dados winsorizados. Alqallaf *et al.* (2002) reanalisa esta abordagem para dados de alta dimensionalidade. Este processo consiste em utilizar uma matriz de correlação “inicial” dos dados padronizados pelo método de winsorização ajustado, Khan *et al.* (2007b) para torná-la mais robusta. Logo o método de Winsorização bivariada depende desta matriz de correlação inicial pela distância de Mahalanobis e a função de escore de Huber.

Finalmente para determinar o critério de parada, substituímos as correlações clássicas na estatística F parcial por suas alternativas robustas para obter uma estatística F parcial robusta.

Este critério teve um desempenho muito bom comparado com outras propostas e com suas versões clássicas não robustas, Pervin e Rahman (2018), além de apresentar eficiência na presença de pontos aberrantes.

3.2.2 VIF robusto

Conforme salientado por Dupuis *et al.* (2013), o procedimento VIF descrito na Sessão 2.2.2, que tem como objetivo evitar a presença de multicolinearidade no modelo selecionado, pode ser prejudicado com a existência de observações discrepantes. A consequência seria que, na tomada de decisão com base em testes de hipóteses, o nível de significância real pode ser muito diferente do nível de significância pré-fixado. Para superar esse problema, propõem o procedimento “VIF robusto”.

Seja $\hat{\beta} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}$ para $\mathbf{W} = \text{diag}\{w_i\}$ matriz diagonal dos pesos, que serão determinados de modo a reduzir ou até praticamente anular a influência de valores discrepantes no cálculo de $\hat{\beta}$. Sejam $\mathbf{X}^w = \text{diag}(\sqrt{w_i})\mathbf{X}$ e $\mathbf{y}^w = \text{diag}(\sqrt{w_i})\mathbf{y}$, o estimador $\hat{\beta}$ pode ser reescrito da seguinte forma: $\hat{\beta} = (\mathbf{X}^{w'}\mathbf{X}^w)^{-1}\mathbf{X}^w\mathbf{y}^w$. Os autores propõem utilizar os pesos w_i associados à função de escore de Huber (3.3) com constante 1,345,

$$w_i(r_i) = \begin{cases} 1 & |r_i| \leq 1,345 \\ \frac{1,345}{|r_i|} & |r_i| > 1,345 \end{cases}, \quad i = 1, 2, \dots, n. \quad (3.10)$$

Considere o modelo de regressão de $\mathbf{y}$ em função das k variáveis explicativas presentes na matriz de planejamento $\mathbf{X}$ e a variável $\mathbf{x}_j$, ou seja, no modelo com matriz de planejamento $\tilde{\mathbf{X}} = [\mathbf{X}^w \mathbf{x}_j^w]$. Após cálculos, verifica-se que $\hat{\beta}_j^w$, a estimativa do coeficiente de regressão associado a variável $\mathbf{x}_j$ é dada por:

$$\hat{\beta}_j^w = (\delta^w)^{-1} \hat{\gamma}_j^w$$

para $\hat{\gamma}_j^w = (\mathbf{x}_j^{w'} \mathbf{x}_j^w)^{-1} \mathbf{x}_j^{w'} \mathbf{r}^w$; $\mathbf{r}^w$ vetor de resíduos do modelo ponderado da variável $\mathbf{y}^w$ em função

das k primeiras variáveis e $\delta^w = 1 - \mathbf{x}_j^{w'} \mathbf{H}^w \mathbf{x}_j^w (\mathbf{x}_j^{w'} \mathbf{x}_j^w)^{-1}$ e $\mathbf{H}^w$ é a matriz *Hat* ou de projeção correspondente à matriz de planejamento $\mathbf{X}^w$, $\mathbf{H}^w = \mathbf{X}^w (\mathbf{X}^{w'} \mathbf{X}^w)^{-1} \mathbf{X}^{w'}$. Seja $R_j^{2w} = \mathbf{x}_j^{w'} \mathbf{H}^w \mathbf{x}_j^w (\mathbf{x}_j^{w'} \mathbf{x}_j^w)^{-1}$. Pode-se verificar que R_j^{2w} é o coeficiente de explicação da regressão de $\mathbf{x}_j^w$ em função de $\mathbf{X}^w$. Este valor é definido por [Dupuis et al. \(2013\)](#) como coeficiente de explicação robusto para o VIF. Como consequência, a medida $\frac{1}{\delta^w}$ é considerada um fator robusto de inflação da variância.

De maneira análoga a [Lin et al. \(2011\)](#), [Dupuis et al. \(2013\)](#) propõe um teste de hipóteses para concluir sobre a inclusão de $\mathbf{x}_j$ no modelo, com base no fator de inflação da variância robusto.

A estatística do teste é dada por

$$T_j^w = \frac{(\delta^w)^{-1/2} \hat{\gamma}_j^w}{\sqrt{\frac{s^2}{n} \left(\frac{1}{n} \sum_{i=1}^n x_{ij}^{w2} \right)^{-1} d}} \quad (3.11)$$

para d constante correspondente à eficiência do estimador do desvio, [Heritier et al. \(2009\)](#).

O procedimento de avaliação rápido e robusto proposto pelos autores é resumido pelos seguintes cinco passos:

1. Obter os resíduos $\mathbf{r}^w$.
2. Para $\mathbf{x}_j^w = \text{diag}(\sqrt{w_{ij}}) \mathbf{x}_j$, calcular $\hat{\gamma}_j^w$ e $s = \text{MAD}(\mathbf{r}^w - \mathbf{x}_j^w \hat{\gamma}_j^w)$.
3. Amostrar um pequeno subconjunto $I = \{i_1, \dots, i_l\} \subset \{1, \dots, n\}$ das observações e deixe $\mathbf{x}_I$ denotar a subamostra correspondente de $\mathbf{x}$.
4. Para $\mathbf{H}_I^w = \mathbf{X}_I^w (\mathbf{X}_I^{w'} \mathbf{X}_I^w)^{-1} \mathbf{X}_I^{w'}$, calcular $R_j^{2w} = \mathbf{x}_{jI}^{w'} \mathbf{H}_I^w \mathbf{x}_{jI}^w (\mathbf{x}_{jI}^{w'} \mathbf{x}_{jI}^w)^{-1}$.
5. Calcular a estatística T_j^w e compará-la com o quantil da distribuição normal padrão correspondente para decidir a inclusão ou não da variável $\mathbf{x}_j$ no modelo atual.

Nos Capítulos 2 e 3 forem definidos os principais critérios de seleção de modelos para regressão linear. Agora temos outro problema, depois de selecionado o modelo adequado, como identificar as observações discrepantes na amostra? Este tópico será abordado no capítulo seguinte.

4 TÉCNICAS DE DIAGNÓSTICO PARA REGRESSÃO ROBUSTA

A regressão robusta é uma ferramenta importante na análise de dados contaminados com pontos aberrantes. Além dos critérios de seleção de modelos, também estamos interessados em estudar o problema de detecção de pontos aberrantes.

Na Figura (4.1) destacam-se os pontos aberrantes mais comuns que podem ser encontrados numa regressão mediante um exemplo de regressão linear simples. Os pontos em branco correspondem a observações regulares nos dados, os pontos **A** e **C** são *outliers* em y , pois representam um deslocamento no valor de y . Os pontos **B** e **C** são pontos de alavanca, estes pontos, geralmente, têm um peso desproporcional no próprio valor ajustado. Note-se que o ponto **A** e **C** poderiam influenciar o modelo modificando de forma abrupta seu ajuste, estes pontos são chamados pontos influentes. O ponto **B**, mesmo deslocado dos demais, não parece diferir muito do padrão de ajuste seguido pelas observações regulares. Desta forma, pode-se definir uma classificação dos pontos de alavanca em bons pontos de alavanca (**B**) e maus pontos de alavanca (**C**).

Figura 4.1: Exemplo de regressão linear simples com observações regulares (pontos brancos), pontos aberrantes (**A**) e (**C**), bons pontos de alavanca (**B**) e maus pontos de alavanca (**C**)

Também percebe-se que um ponto influente pode provocar uma mudança substancial na análise de regressão veja-se Figura (1.1), nesse exemplo, são alterados os valores ajustados e/ou as estimativas dos coeficientes do modelo, o que pode levar a uma seleção errada das variáveis explicativas

importantes para um ajuste com várias variáveis explicativas.

Neste capítulo descrevemos vários procedimentos de detecção de pontos aberrantes. São estabelecidos pontos de cortes para estas medidas a fim de identificar da forma mais adequada nas simulações do Capítulo 5, na prática recomendamos identificar os pontos discrepantes de forma visual e pontual pelos gráficos correspondentes a cada medida e com o correspondente especialista dos dados em estudo.

4.1 Técnicas de diagnóstico clássicas

As técnicas de diagnóstico para modelos de regressão linear têm sido extensivamente estudadas, e tem-se desenvolvido várias medidas e métodos com o objetivo de determinar aquelas observações presentes na amostra que não apresentam um padrão comum às outras observações. Nesta seção apresentamos as principais medidas clássicas de identificação de pontos aberrantes.

4.1.1 Matriz $\mathbf{H}$

No modelo de regressão linear é muito útil determinar o quanto uma observação pode influenciar no seu valor predito. Esta informação está disponível na matriz de projeção conhecida como matriz *Hat*, [Hoaglin e Welsch \(1978\)](#), que permite expressar cada valor ajustado $\hat{y}_i$ como uma combinação linear que depende do valor observado y_i .

Lembrando que $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$, a matriz $\mathbf{H}$ é conhecida como matriz *Hat*, chapéu ou de projeção,

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

Esta matriz tem um papel fundamental para a análise diagnóstico do ajuste de um modelo, pois algumas quantidades importantes dependem desta matriz, tais como as covariâncias dos resíduos $\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$ para $\mathbf{I}_n$ matriz identidade de ordem n .

A matriz $\mathbf{H}$ satisfaz várias propriedades, [Draper e Smith \(2014\)](#), algumas delas são que é simétrica ($\mathbf{H}' = \mathbf{H}$) e idempotente ($\mathbf{H}^2 = \mathbf{H}$), e é a matriz de projeção ortogonal do vetor $\mathbf{y}$ no espaço linear gerado pelas colunas de $\mathbf{X}$.

Como $\mathbf{H}$ é idempotente e simétrica, temos que $\text{posto}(\mathbf{H}) = \text{tra}(\mathbf{H}) = \sum_{i=1}^n h_{ii} = p$ para h_{ii} i -ésimo elemento da diagonal de $\mathbf{H}$ e p o número de parâmetros do modelo. Este elemento h_{ii} desempenha um papel importante nas técnicas de diagnóstico [Cook e Weisberg \(1982\)](#) e satisfaz $\frac{1}{n} < h_{ii} < \frac{1}{b}$ para b constante que representa o número de vezes que é replicada a i -ésima linha de $\mathbf{X}$.

Devido às propriedades da matriz de projeção, pode-se verificar que:

$$\hat{y}_i = h_{ii}y_i + \sum_{j \neq i}^n h_{ji}y_j, \quad \text{e} \quad \sum_{j \neq i} h_{ij}^2 = h_{ii}(1 - h_{ii}).$$

Logo, h_{ii} é uma medida da influência da observação i - ésimas sobre seu próprio valor ajustador.

Sob a suposição de que todas as observações tem a mesma influência, podemos esperar que $h_{ii} \approx \frac{p}{n}$, [Hoaglin e Welsch \(1978\)](#) propõe analisar aquelas observações com valor $h_{ii} \geq \frac{2p}{n}$ como possíveis pontos de alavanca. Esta medida é fundamental para o cálculo de outras medidas. Existe uma ampla literatura sobre pontos de alavanca e sua influência no ajuste, [Hoaglin e Welsch \(1978\)](#); [Rousseeuw e Van Zomeren \(1990\)](#).

4.1.2 Resíduos

A análise de resíduos desempenha um papel importante na validação do modelo de regressão. Nesta seção veremos os principais resíduos que podem ser utilizados para detectar pontos discrepantes.

Resíduo padronizado r_i^*

O resíduo padronizado tradicional tem a seguinte expressão:

$$r_i^* = \frac{y_i - \hat{y}_i}{\hat{\sigma}} \quad (4.1)$$

para $\hat{\sigma}^2 = \sum_{i=1}^n \frac{r_i^2}{n-p}$. Quando os erros têm distribuição normal espera-se que o resíduo padronizado esteja na faixa $[-2;2]$, por isso, observações que cumpram a condição $|r_i^*| > 2$ serão consideradas pontos aberrantes.

Resíduo internamente studentizado t_i

O resíduo padronizado assume que suas variâncias são iguais, quando na verdade nem sempre isso acontece, $var(r_i) = \sigma^2(1 - h_{ii})$ e $cov(r_i, r_j) = -\sigma^2 h_{ij}$. Portanto, os resíduos r_i podem ter variâncias desiguais, daí a necessidade de padronizá-los dividindo-os pela estimativa do desvio padrão adequada,

$$t_i = \frac{r_i}{\hat{\sigma} \sqrt{1 - h_{ii}}}. \quad (4.2)$$

Este resíduo t_i recebe o nome de resíduo internamente studentizado [Montgomery et al. \(2012\)](#) e sob o modelo normal, estes devem permanecer na faixa $|t_i| \leq 2$, e qualquer observação com resíduo fora desse intervalo pode ser considerada como uma observação discrepante. O gráfico de t_i contra

o ordem das observações permite identificar visualmente aquelas observações candidatas a pontos aberrantes e possíveis perturbações das suposições do modelo normal linear.

Note-se que t_i não é independente de $\hat{\sigma}^2$, logo t_i tem distribuição aproximada a t de student, mas não segue realmente esta distribuição, [Montgomery et al. \(2012\)](#). Com o objetivo de obter esta independência, podemos substituir $\hat{\sigma}^2$ por um estimador adequado que descrevemos na seguinte seção.

Resíduo externamente studentizado t_i^*

Vamos supor que podemos encontrar uma alternativa externa do estimador $\hat{\sigma}^2$ em (4.2) que não dependa da i - ésima observação. Se recalcularmos $\hat{\sigma}^2$ sem a i - ésima observação, chamemos de $\hat{\sigma}_{(i)}^2$ esta estimativa, obteremos um novo resíduo denominado resíduo externamente studentizado t_i^* , [Montgomery et al. \(2012\)](#). Este resíduo tem propriedades melhores que t_i , [Hoaglin e Welsch \(1978\)](#). Devido à independência entre numerador e denominador de t_i^* , prova-se que, válidas as suposições do modelo, t_i^* tem distribuição *t-student*, [teorema Fisher - Cochran, [Rao et al. \(1973\)](#)].

$$t_i^* = \frac{r_i}{\hat{\sigma}_{(i)} \sqrt{1 - h_{ii}}}, \quad (4.3)$$

para $\hat{\sigma}_{(i)}^2$ estimativa da variância σ^2 no modelo excluindo a i - ésima observação.

Note-se que $\hat{\sigma}_{(i)}^2$ pode ser calculada como $(n - p - 1)\hat{\sigma}_{(i)}^2 = \sum_{j=1}^n r_j^2 - \frac{r_i^2}{1 - h_{ii}}$, daí obtemos a seguinte expressão $\hat{\sigma}_{(i)}^2 = \hat{\sigma}^2 \left(\frac{n - p - t_i^2}{n - p - 1} \right)$ e o resíduo externamente studentizado (4.3) pode ser expressado em função do resíduo internamente studentizado como:

$$t_i^* = t_i \left(\frac{n - p - 1}{n - p - t_i^2} \right)^{1/2}. \quad (4.4)$$

Este resíduo segue uma distribuição *t* de *student* com $n - p - 1$ graus de liberdade. Como consequência, valores absolutos altos de t_i^* , $|t_i^*| > 2$, indicam que a i - ésima observação pode ser considerada ponto aberrante.

4.1.3 Distância de Cook

Para medir o efeito das observações no vetor de estimadores dos parâmetros β , [Cook \(1977\)](#) sugere uma função da distância entre $\hat{\beta}_{(i)} - \hat{\beta}$ em que $\hat{\beta}_{(i)}$ é a estimativa do vetor de parâmetros quando excluída a i - ésima observação. Esta nova medida é baseada em elipsoides de confiança. É desenvolvida para julgar a contribuição de cada observação na estimativa por mínimos quadrados do vetor de parâmetros em modelos de regressão linear quando a matriz de planejamento é de posto

completo, [Cook \(1977\)](#),

$$D_i = \frac{(\hat{\beta}_{(i)} - \beta)' X' X (\hat{\beta}_{(i)} - \beta)}{p \hat{\sigma}^2}. \quad (4.5)$$

O vetor $\hat{\beta}_{(i)}$ pode ser obtido como $\hat{\beta}_{(i)} = \hat{\beta} - \frac{r_i}{1 - h_{ii}} (X' X)^{-1} x_i$. Substituindo esta expressão em (4.5) obtemos a seguinte expressão simplificada para a distância de Cook, [Cook \(1977\)](#):

$$D_i = t_i^2 \frac{h_{ii}}{p(1 - h_{ii})}. \quad (4.6)$$

Esta medida pode ser usada para detectar pontos influentes nos dados, geralmente é usado como ponto de corte o valor 1, ou seja, analisar observações com valores grandes de D_i ($D_i > 1$ [Cook e Weisberg \(1982\)](#)). As observações detectadas por essas medidas podem ter altos valores do resíduo internamente studentizado t_i , de h_{ii} ou de ambos. Esta medida pode não ser adequada quando o resíduo for grande e h_{ii} pequeno, neste caso foi proposta uma medida que é supostamente mais apropriada, esta medida sera apresentada a seguir.

Distância de Cook modificada

A medida D_i pode ser aprimorada fazendo três transformações, [Atkinson \(1981\)](#), a primeira substituir $\hat{\sigma}$ por $\hat{\sigma}_{(i)}$ para estimar σ^2 que leva a usar t_i^* ao invés de t_i , a segunda, utilizar a raiz quadrada desta expressão, e a terceira é padronizar D_i tendo em conta o efeito do tamanho da amostra n

$$A_i = \left(\frac{n-p}{p} \frac{h_{ii}}{(1 - h_{ii})} \right)^{1/2} |t_i^*|. \quad (4.7)$$

Grandes valores de A_i significam que a observação é do fato um ponto influente para o modelo. Deixando de lado o termo que depende de n e p , A_i se reduz à medida que [Belsley et al. \(1980\)](#) definem como *DFFITs*. A medida A_i pode ser considerada uma versão modificada da Distância de Cook D_i e do *DFFITs* sem perder a essência de cada uma delas, para D_i é medir a influência das observações nos parâmetros de posição e para *DFFITs* é medir a influência nos parâmetros de posição e escala [Cook e Weisberg \(1982\)](#). Levando em consideração que o ponto de corte pra a Distância de Cook é $D_i = 1$, de acordo com as modificações feitas para definir A_i , usaremos como ponto de corte $A_i = \sqrt{n-p}$. Lembre-se que os pontos de corte estabelecidos são do interesse particular da avaliação das medias, na aplicação o usuário determinam o ponto de corte em função dos resultados obtidos.

4.1.4 Distância de Mahalanobis

Outra medida de identificação de pontos aberrantes é a Distância de [Mahalanobis \(1936\)](#). Esta medida permite quantificar quanto o ponto x_i se afasta do centro do conjunto de dados levando em

conta a estrutura de correlação entre as variáveis. A Distância de Mahalanobis é definida como:

$$MD_i = \sqrt{(\mathbf{x}_i - \mathbf{T})' \mathbf{V}^{-1} (\mathbf{x}_i - \mathbf{T})} \quad (4.8)$$

para $\mathbf{T}$ média aritmética dos vetores de dados e $\mathbf{V}$ matriz de variância e covariância amostral. Rousseeuw e Van Zomeren (1990) destacam o fato de que entre a Distância de Mahalanobis e os elementos da diagonal da matriz de Projeção $\mathbf{H}$ existe uma relação monótona dada por:

$$(MD_i)^2 = (n-1) \left(h_{ii} - \frac{1}{n} \right).$$

Sugere-se identificar como pontos aberrantes aquelas observações com distância de Mahalanobis superior a χ^2 , $\sqrt{\chi_{(p-1);0,975}^2}$, em que χ^2 , $\sqrt{\chi_{(p-1);0,975}^2}$ é o quantil de ordem 0,975 da distribuição qui-quadrado com $p-1$ graus de liberdade, sendo $p-1$ o número de variáveis explicativas no modelo.

Na seguinte seção serão apresentadas as alternativas robustas das medidas apresentadas nesta Seção e definimos outra medida que depende do peso das observações no modelo.

4.2 Técnicas de diagnóstico robustas

A identificação de pontos aberrantes pode ser difícil devido ao efeito de mascaramento. Detectar estes pontos não é trivial, especialmente quando existe mais de um. Os métodos de identificação clássicos discutidos na Seção 4.1 nem sempre encontram todos os pontos aberrantes, pois são baseados na média amostral e na matriz de covariância, os quais são afetados pelos mesmos pontos aberrantes. Com o objetivo de reduzir este efeito de mascaramento nas observações, nesta seção são apresentados algumas medidas robustas de identificação de pontos aberrantes.

4.2.1 Matriz Hat

A matriz *Hat* ou de projeção na regressão robusta é determinada da seguinte forma:

$$\mathbf{H}_r = \mathbf{X}(\mathbf{X}'\mathbf{V}_r^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}_r^{-1} \quad (4.9)$$

para $\mathbf{V}_r$ matriz diagonal cujos elementos da diagonal são os pesos associados às observações, Heiberger e Becker (1992); Sommer e Staudte (1995). Esta matriz diagonal também pode ser obtida utilizando estimativas robustas proporcionais às variâncias dos erros. Os elementos da diagonal de $\mathbf{H}_r$ serão representados como h_{ii}^r .

4.2.2 Resíduos robustos

As técnicas de diagnóstico robustas baseadas nos resíduos do modelo selecionado são as mesmas definidas na Seção (4.1.2), substituindo as correspondentes estimativas pelas suas alternativas robustas. Usaremos $\hat{y}_i$ obtido a partir da estimativa robusta $\mathbf{M}$ e s e $s_{(i)}$ serão as correspondentes estimativas robustas para os parâmetros de escala σ obtidos pelo MAD.

Para o resíduo padronizado tradicional usaremos

$$rr_i^* = \frac{y_i - \hat{y}_i}{s}, \quad (4.10)$$

e para as versões robustas do resíduo studentizado interno e externo t_i e t_i^* usaremos as seguintes definições

$$rt_i = \frac{y_i - \hat{y}_i}{s \sqrt{1 - h_{ii}^r}}, \quad (4.11)$$

$$rt_i^* = \frac{y_i - \hat{y}_i}{s_{(i)} \sqrt{1 - h_{ii}^r}}. \quad (4.12)$$

4.2.3 Distância de Cook robusta

As correspondentes robustas da distância de Cook e sua versão modificada proposta por Atkinson serão:

$$RD_i = (rt_i^2) \frac{h_{ii}^r}{p(1 - h_{ii}^r)}, \quad (4.13)$$

$$RA_i = \left(\frac{(n-p)}{p} \frac{h_{ii}^r}{(1 - h_{ii}^r)} \right)^{1/2} |rt_i^*|. \quad (4.14)$$

4.2.4 Distância de Mahalanobis robusta

Rousseeuw e Van Zomeren (1990) utilizam a Distância de Mahalanobis (1936) para definir uma distância robusta que consiga classificar de forma eficiente as observações de acordo com sua influência no ajuste.

$$RMD_i = \sqrt{(\mathbf{x}_i - \mathbf{T}_r)' \mathbf{V}_r^{-1} (\mathbf{x}_i - \mathbf{T}_r)}. \quad (4.15)$$

Eles sugerem substituir a média aritmética $\mathbf{T}$ e a matriz de covariância da amostra $\mathbf{V}$ na distância de Mahalanobis clássica (4.8), por medidas robustas, devido ao possível mascaramento que essas medidas clássicas podem sofrer.

O estimador que minimiza o determinante da matriz de covariância (MVD) é um dos mais usados para esta distância, em maior parte porque fornece um algoritmo rápido, Rousseeuw e Driessen (1999); Rousseeuw (1985). O objetivo deste método é encontrar o subconjunto de observações cuja matriz de covariância tenha o determinante mínimo. O estimador da escala é proporcional à matriz de covariância calculada para este conjunto de pontos e o estimador de localização é a média deles, Filzmoser *et al.* (2005). De acordo com Butler *et al.* (1993) o método MVD é assintoticamente normal,

o que faz que podamos comparar a distância de Mahalanobis robusta com o correspondente valor do quantil $\sqrt{\chi^2_{(p-1);0,975}}$ em que $(p-1)$ é a quantidade de variáveis explicativas no modelo.

4.2.5 Classificação por peso

Os dados contaminados são comumente encontrados em diversas situações, e os métodos robustos que identificam ou desvalorizam as observações discrepantes são ferramentas essenciais na estatística. Os pesos w_i associados às observações também podem ser utilizados como técnica de diagnóstico [Rousseeuw e Leroy \(2005\)](#). Neste caso, pode ser considerado o peso $w_i = \psi(r_i^*)/r_i^*$ associado ao modelo selecionado para diferenciar as observações discrepantes e pontos regulares usando o ponto de corte 0,25 como sugere [Filzmoser et al. \(2008\)](#). Logo, observações com $w_i < 0,25$ serão consideradas pontos aberrantes. Note-se que esta medida depende da função de escore ψ escolhida.

Para o caso particular da função de escore de Huber (3.3) com constante $c = 1,345$, a função do peso tem a seguinte expressão

$$w(r) = \begin{cases} 1 & |r| \leq 1,345 \\ \frac{1,345}{|r|} & |r| > 1,345 \end{cases}$$

que foi definida na equação (3.10).

Note-se que esta função de peso não elimina totalmente o efeito que pode ter uma observação no ajuste, pois quanto maior for o valor absoluto do resíduo menor será o peso, mas nunca vai assumir o valor zero, esta é a função de ponderação associada à família de funções de Huber ρ , ψ e w , veja Figura (3.1). Existem outras famílias de funções que geram uma ponderação que pode eliminar o efeito das observações, tais como a família de funções de [Hampel et al. \(1986\)](#), as funções de Tukey (*bisquare*), entre outras. Mais detalhes sobre estas famílias de funções podem ser encontrados em [Maronna et al. \(2006\)](#); [Rousseeuw e Leroy \(2005\)](#).

No capítulo seguinte será analisado o comportamento destas medidas de detecção de pontos aberrantes e os critérios de seleção de modelos da primeira classe vistos nos Capítulos 2 e 3.

5 SIMULAÇÃO E APLICAÇÕES

Neste capítulo faremos um estudo simulando diferentes cenários para avaliar e comparar os critérios de seleção de modelos da Classe 1 dos Capítulos 2 e 3 e as medidas de identificação de pontos aberrantes do Capítulo 4. Serão analisados os critérios e as medidas para os exemplos correspondentes aos dados de *Stack Loss* e *Coleman*, e ao banco de dados meteorológicos da temperatura mínima diária durante 2018.

5.1 Avaliação do desenvolvimento dos critérios de seleção de modelos

Para gerar as perturbações na distribuição dos erros, mantendo a normalidade e independência, serão modificadas algumas observações regulares. Usaremos uma combinação das propostas de perturbação propostas por [Dixon \(1950\)](#). As propostas consistem em misturar a distribuição dos erros $N(0, \sigma^2)$ com:

A - locação: $N(\mu_\epsilon, \sigma^2)$ e

B - inflação da variância: $N(0, K^2 \sigma^2)$, $K^2 > 1$.

Seja o modelo como definido em (1.1), a distribuição dos erros será perturbada da seguinte forma:

$$\epsilon_i \sim \begin{cases} N(0, \sigma^2) & \text{com probabilidade } (1 - \pi) \text{ observação regular} \\ N(\mu_\epsilon, K^2 \sigma^2) & \text{com probabilidade } \pi \text{ observação discrepante} \end{cases} \quad (5.1)$$

em que π , μ_ϵ e K são parâmetros fixos do modelo. π é a probabilidade de que uma observação seja ponto aberrante, μ_ϵ é o parâmetro de perturbação para a média zero e K é o parâmetro de inflação da variância para perturbar σ , [Hoeting et al. \(1996\)](#); [Taplin e Raftery \(1994\)](#).

De maneira similar ao estudo de simulação feito por [Ronchetti et al. \(1997\)](#), foram considerados seis fatores para avaliar o desempenho dos critérios de seleção de modelos:

- tamanho de amostra N ,
- tamanho dos vetores de parâmetros $\|\beta\|_1$,
- proporção de pontos aberrantes π ,

- perturbação dos erros (*pert*),
- classificação do modelo selecionado (correto, super-ajuste, sub-ajuste e outros) e
- o critério utilizado para selecionar o modelo

O algoritmo utilizado para o cálculo das estimativas $\hat{\beta}_j$ invariantes em relação a escala foi o método de mínimos quadrados ponderado iterativamente ou método dos pesos modificados. Este algoritmo converge independentemente do valor inicial desde que ψ seja monótona, [Huber \(1981\)](#), como acontece com a função de escore de Huber definida em (3.3). A estimativa robusta da escala é obtida pelo método MAD re-ajustada (1.4).

De acordo com [Beckman e Cook \(1983\)](#), é preciso estabelecer regras para julgar o desempenho das medidas de detecção de pontos aberrantes. Seja U o número de pontos aberrantes corretamente classificados como pontos aberrantes (verdadeiros positivos, VP), V o número de observações regulares erroneamente classificadas como pontos aberrantes (falsos positivos, FP) e n_π a quantidade de pontos aberrantes que depende do parâmetro fixado π . Serão utilizados quatro das seis regras propostas pelos autores, mais a precisão da classificação (*Accuracy*).

As regras utilizadas para julgar as medidas de detecção de pontos discrepantes são as estimativas das seguintes quantidades:

Acc(+) Precisão da classificação.

A.1(+) Probabilidade de não ter FP e todos os pontos aberrantes serem detectados,

$$Prob(U = n_\pi, V = 0).$$

A.2(+) Probabilidade de não ter FP e detectar pelo menos um ponto aberrante,

$$Prob(U > 0, V = 0)$$

A.4(−) A probabilidade de pelo menos um FP

$$Prob(V > 0).$$

A.6(−) A proporção de FP entre todas as observações que foram declaradas pontos aberrantes, esta quantidade é conhecida como “taxa de classificação falsa”,

$$E(V)/(E(U) + E(V)).$$

Os símbolos (+) e (−) significam que quanto maior o valor calculado na regra, a medida é melhor ou pior respectivamente.

De acordo com os valores na Matriz de Confusão da Tabela (5.1), as quantidades utilizadas nas regras para julgar as medidas de detecção de pontos discrepantes podem ser calculadas em função das seguintes quantidades.

- *VP*: Quantidade de observações classificadas corretamente como pontos aberrantes,
- *FP*: quantidade de observações classificadas erroneamente como pontos aberrantes,
- *VN*: quantidade de observações classificadas corretamente como pontos regulares,

- FN : quantidade de observações classificadas erroneamente como pontos regulares.

$$Acc = \frac{VP + VN}{(VP + FP + FN + VN)}, \quad (5.2)$$

$$A1 = \left(\frac{VP}{n_\pi} \right)^{n_\pi} \left(\frac{VN}{n - n_\pi} \right)^{(n - n_\pi)}, \quad (5.3)$$

$$A2 = \left[1 - \left(\frac{FN}{n_\pi} \right)^{n_\pi} \right] \left[\frac{VN}{n - n_\pi} \right]^{n - n_\pi}, \quad (5.4)$$

$$A4 = 1 - \left(\frac{VN}{n - n_\pi} \right)^{n - n_\pi}, \quad (5.5)$$

$$A6 = \frac{FP}{VP + FP}. \quad (5.6)$$

		Class. Predita		
		$p.aberr$	$p.reg$	Total
Class.	$p.aberr$	VP	FN	n_π
Actual	$p.reg$	FP	VN	$n - n_\pi$

Tabela 5.1: Matriz de confusão

Os algoritmos utilizados para as simulações foram baseados em funções e pacotes já existentes e funções criadas no programa R v3.5.2 e RStudio Version 1.1.463 em um computador Acer Aspire 5, Intel(R) CoreTM i5-7200U, CPU 2,5GHz ×4.

5.1.1 Desenho da simulação

Os tamanhos de amostra N foram escolhidos como pequenos, 25, 50 e 100, e grandes 500 e 1000 para observar o comportamento dos critérios para cada tamanho. Os critérios serão avaliados para modelos com dois e três variáveis explicativas importantes de um total de cinco que foram geradas mediante dois vetores de parâmetros, $\beta_p = (2; 0; 0; -1; 3; 0)$ e $\beta_g = (17; 10; 0; -15; 0; 13)$ sendo a primeira componente o intercepto. Foram considerados estos dois vetores para ter representatividade de parâmetros com valores absolutos grandes e pequenos. O objetivo é avaliar o desempenho dos critérios na identificação do modelo adequado quando são utilizados parâmetros pequenos $\|\beta_p\|_1 = 6$ e parâmetros grandes $\|\beta_g\|_1 = 55$, para $\|\cdot\|_1$ a norma 1 do vetor. Os parâmetros π , μ_ε e K foram fixados baseado nas sugestões e resultados obtidos por [Hoeting et al. \(1996\)](#); [Taplin e Raftery \(1994\)](#). Utilizaremos $\pi = (0,04; 0,08; 0,13; 0,2)$ para garantir pelo menos um ponto aberrante nas amostras pequenas e também, porque o estimador M é resistente na presença de no máximo 25% de pontos aberrantes, [Alma \(2011\)](#). O valor π pode ser associado ao nível de contaminação da amostra usando a proporção de pontos aberrantes correspondente em porcentagem. Os erros serão gerados de acordo com a distribuição Normal padrão, e serão fixados três cenários para os parâmetro de perturbação da locação e escala (pequena, moderada, alta) que dependerão dos tamanhos dos vetores de parâmetros

escolhidos. $pert_p = (\mu_\varepsilon; K^2)_p = [(1, 2^2); (2, 5^2); (3, 7^2)]$ para β_p , isto é, para o primeiro elemento, a distribuição dos erros $N(0, 1)$ sera perturbada com uma $N(1, 2^2)$, de igual forma para os outros dois elementos. $pert_g = (\mu_\varepsilon; K^2)_g = [(9, 3^2); (18, 7^2); (27, 10^2)]$ para β_g .

Para cada tamanho de amostra foi fixada a matriz de planejamento $\mathbf{X}$ com cinco variáveis independentes e identicamente distribuídas $U[-1; 1]$. Os erros foram gerados e perturbados como definido em (5.1), para cada tamanho de amostra (N), vetor de parâmetros (β), proporção de pontos aberrantes (π), e perturbação fixada ($N(\mu_\pi; K^2)$). Depois foi gerada a variável dependente para cada simulação. Em cada cenário com perturbação dos erros ($\frac{5}{N} \times \frac{2}{\beta} \times \frac{4}{\pi} \times \frac{3}{pert} = 120$ cenários) e sem perturbação ($\pi = 0$; temos $\frac{5}{N} \times \frac{2}{\beta} = 10$ cenários) foram realizadas 100 réplicas para avaliar a eficiência dos critérios de seleção de modelos e das medidas de diagnóstico.

5.1.2 Resultados da Simulação

Nesta seção vamos analisar o desempenho dos critérios de seleção de modelos e das medidas de detecção de pontos discrepantes. A seleção de modelos será avaliada de acordo com a porcentagem de classificação nas seguintes quatro categorias, de maneira similar ao estudo realizado em Ronchetti *et al.* (1997). A identificação do modelo correto (só com as variáveis explicativas importantes) será classificada como **Correto** e será a quantidade principal de nosso estudo. **Super-ajuste** significa que foi selecionado o modelo correto mais apenas uma variável, **Sub-ajuste** que foi selecionado um modelo que tem uma variável menos das variáveis do modelo correto e **Outro** para qualquer outra seleção de modelos. A detecção de pontos discrepantes será avaliada em função das regras Acc, A.1, A.2, A.4 e A.6 definidas na Seção 5.1.

As tabelas do presente trabalho foram geradas usando o pacote do R “stargazer”, Hlavac, Marek (2018).

5.1.2.1 Seleção de modelos

Segundo o desenho da simulação apresentado na sessão anterior, foram geradas cinco matrizes de planejamento fixas para cada tamanho de amostra, veja Figura A.2 do Apêndice. Para cada um dos cenários foram realizados os ajustes correspondentes a 100 simulações e selecionado o modelo adequado de acordo com o critério utilizado. Baseada nessa seleção, foram geradas tabelas com as médias das porcentagens de classificação para cada uma das quatro possíveis categorias, correto, super-ajuste, sub-ajuste e outros. Os valores correspondentes aos cenários sem e com perturbação são apresentados brevemente nas Tabelas A.1 A.2 do Apêndice. Devido à quantidade de casos são exibidas somente as primeiras e últimas 6 linhas das tabelas.

As Figuras 5.1 e 5.2 apresentam as médias da porcentagem de classificação de acordo com os

critérios clássicos e robustos quando os erros não são perturbados. Note-se que ambos os grupos de critérios têm comportamentos similares. O RC_p foi mais adequado do que o C_p para a classificação Super-ajuste, o que resultou em uma particular redução na categoria outras, que corresponde às classificações incorretas. Os critérios BIC e $MCCV$ e suas versões robustas tiveram uma média da porcentagens de classificação correta bem maior que os outros critérios, o que está de acordo com as ideias discutidas no trabalho de [Arlot et al. \(2010\)](#), os autores destacam estes dois critérios como bons critérios de identificação do modelo correto.

As medidas resumo dos dados correspondentes às Figuras 5.1 e 5.2 são apresentados na Tabela A.3 do Apêndice.

Figura 5.1: Média das Porcentagens de classificação dos critérios clássicos para os cenários sem perturbação dos erros.

Figura 5.2: Média das Porcentagens de classificação dos critérios robustos para os cenários sem perturbação dos erros.

Figura 5.3: Média das Porcentagens de classificação dos critérios clássicos para os cenários com perturbação na distribuição dos erros.

Figura 5.4: Média das Porcentagens de classificação dos critérios robustos para os cenários com perturbação na distribuição dos erros.

Para conferir se de forma geral houve alguma mudança substancial entre os critérios clássicos e robustos, apresentamos também os gráficos das médias de classificação dos critérios quando são perturbados os erros.

A Figura 5.3 mostra como os critérios clássicos apresentam menor porcentagem de classificação correta na presença de pontos aberrantes, sendo o AIC o melhor deles com uma média de porcenta-

gem de classificação correta de 54%. Também podemos perceber que aparentemente a presença de observações discrepantes faz com que os critérios clássicos tendam a selecionar modelos com uma variável explicativa importante a menos, ou seja, a porcentagem de classificação na categoria Subajuste aumenta.

De modo geral, os critérios robustos na presença de pontos aberrantes têm um desempenho similar ao caso em que não há perturbação alguma, Figura 5.4 contra Figura 5.2. Apenas o critério RR_a^2 teve um aumento da média de classificação correta em quase 40%. Comparando com respeito à média de classificação correta, os critérios robustos têm melhor desenvolvimento que os critérios clássicos na presença de pontos aberrantes. Nesta dissertação, exceto menção em caso contrário, nos referiremos à média da porcentagem de classificação correta simplesmente como classificação correta.

Os dados correspondentes às Figuras 5.3 e 5.4 são apresentados na Tabela A.4 do Apêndice.

As medidas resumo correspondentes à classificação correta e o tempo de cálculo de cada critério são apresentadas nas Tabelas 5.2 e 5.3 para os cenários sem e com perturbação dos erros respectivamente. Serão sinalados os critérios robustos com o símbolo (*) correspondente, quando houver diferenças estatisticamente significantes entre a média destes critérios e suas versões clássicas. Foi utilizado o teste *t – student* sob normalidade e na ausência desta o teste da soma dos postos de Wilcoxon.

Tabela 5.2: Medidas resumo da porcentagem de classificação correta e o tempo de cálculo para cada critério no cenário sem perturbação dos erros.

Critério	Média	Classificação correta (%)					Tempo (s)					
		DP	Mín.	Q1	Q3	Máx.	Média	DP	Mín.	Q1.	Q3.1	Máx.
BIC	87,70	12,93	57	82,50	96	99	0,082	0,01	0,07	0,08	0,08	0,10
MCCV	88,80	12,36	57	85,25	96	99	4,202	0,24	3,90	4,06	4,31	4,64
R_a^2	32,20	10,88	18	23,25	41,75	47	0,082	0,01	0,07	0,08	0,08	0,10
AICc	69,40	8,68	56	65,25	74	82	0,084	0,01	0,07	0,08	0,09	0,10
Cp	29,80	5,33	21	26	33	38	0,081	0,01	0,07	0,08	0,08	0,10
RBIC	89,30	15,22	51	88,75	99	100	0,080	0,01	0,07	0,08	0,08	0,10
RMCCV	88	13,09	53	86,50	95,50	98	4,227	0,23	3,94	4,07	4,32	4,67
RR_a^2	35,40	9,43	17	29,50	41,75	49	0,082	0,01	0,07	0,08	0,08	0,10
RAIC	56,30*	12,94	30	50,25	64,50	73	0,083	0,01	0,07	0,08	0,09	0,10
RCp	27,50	9,41	12	20,50	32,75	44	0,080	0,01	0,07	0,07	0,08	0,09

Nota:

* $p < 0,1$; ** $p < 0,05$; *** $p < 0,01$

Note-se que o desvio padrão das porcentagens de classificação dos critérios clássicos aumenta bastante na presença de pontos aberrantes, enquanto que os critérios robustos apresentam pequenas mudanças nos desvios padrões, exceto o RR_a^2 e $RMCCV$ que tiveram uma dispersão maior quando os erros foram perturbados, Tabela 5.3.

A classificação correta para os critérios clássicos assim como os robustos quando são aplicados

Tabela 5.3: Medidas resumo da porcentagem de classificação correta e o tempo de cálculo para cada critério no cenário com perturbação dos erros.

Critério	Média	Classificação correta (%)					Média	Tempo (s)				
		DP	Mín.	Q1	Q3	Máx.		DP	Mín.	Q1.	Q3.1	Máx.
BIC	50,29	41,66	0	4,75	98	100	0,0715	0,01	0,06	0,07	0,07	0,09
MCCV	48,52	25,00	1	29	72	98	4,4447	0,34	4,00	4,21	4,67	5,63
R_a^2	46,85	30,95	0	20,75	72,25	100	0,0714	0,01	0,06	0,07	0,07	0,09
AICc	54,04	36,25	0	15,50	90,25	100	0,0840	0,004	0,08	0,08	0,09	0,09
Cp	27,43	38,90	0	0	61,75	100	0,0847	0,01	0,07	0,08	0,09	0,11
RBIC	85,27***	16,56	35	80	97	100	0,0707	0,01	0,06	0,06	0,07	0,09
RMCCV	85,22***	21,61	2	86,75	97	100	4,5394***	0,32	4,05	4,32	4,72	5,64
RR_a^2	74,03***	25,69	6	52,75	98,25	100	0,0718	0,01	0,06	0,07	0,07	0,09
RAIC	57,91	12,02	29	53	66	82	0,0713***	0,01	0,06	0,07	0,07	0,09
RCp	29,88***	10,88	11	20,75	38	55	0,0697***	0,01	0,06	0,06	0,07	0,09

Nota: * p<0,1; ** p<0,05; *** p<0,01

na presença de pontos aberrantes varia muito, variando entre 0 e 100% para os critérios clássicos e entre 2 e 100% para os robustos. Mesmo a classificação correta máxima atingindo 100%, alguns cenários apresentam porcentagem de classificação correta baixos.

O tempo de cálculo (em segundos) dos critérios depende da implementação realizada e da capacidade de processamento do computador onde foram analisados, não entanto, um detalhe importante é perceber que o tempo de cálculo para os critérios baseados em Monte Carlo, *MCCV*, e *RMCCV* são muito maiores que os tempos de cálculo para os outros critérios. Mesmo tendo um tempo de cálculo maior, 4 segundos é um tempo aceitável para a seleção de modelos, pois nesse tempo são realizadas as $M = 50$ partições da amostra e ajustados os $2^{P-1} - 1$ submodelos possíveis para determinar o mais adequado. Levando em consideração este procedimento, e que os critérios baseados em Monte Carlo têm porcentagens de classificação correta elevada para amostras perturbadas, pode-se sugerir analisar mais a fundo este critério. O *RBIC* também é um candidato a análise para dados com perturbações na distribuição dos erros por ter a mais alta porcentagem de classificação correta.

A seguir serão analisados os critérios de seleção de modelos quando os erros são perturbados para os diferentes cenários, ($5 \times 2 \times 4 \times 3 = 120$ cenários).

A Figura 5.5 mostra os valores de classificação correta para os critérios clássicos e robustos baseado no tamanho de amostra e no vetor de parâmetros. Sugere-se que os critérios de seleção de modelos tendem a melhorar a identificação do modelo correto na medida que aumenta o tamanho de amostra, veja também a Tabela 5.4. No Apêndice são mostradas as medidas resumo da classificação correta segundo o tamanho de amostra e o tipo de critério, Tabela A.5.

É de esperar que a seleção de modelos seja melhor quando os parâmetros têm valores absolutos altos. Na Figura 5.5 e a Tabela 5.5 podemos comprovar que os critérios de seleção de modelos são realmente mais eficientes quando os verdadeiros parâmetros são grandes. Levando em consideração

o tipo de critério e os vetores de parâmetros, podemos dizer que os critérios clássicos têm desempenho ruim quando os parâmetros são pequenos, pois quase todos apresentam valores de classificação correta abaixo do 50%. A média de classificação correta para este cenário (critérios clássicos em modelos com parâmetros pequenos com perturbação dos erros) é 31,92% e o quantil do 75% é de 52,75%, estes valores são inviáveis na seleção de modelos, Tabela 5.5.

Figura 5.5: Médias das porcentagens de classificação correta dos critérios considerando a combinação do tamanho de amostra N e do tamanho do vetor de parâmetros $\|\beta\|$, com perturbação dos erros.

Tabela 5.4: Medidas resumo da média da porcentagem de classificação correta pelo tamanho de amostra quando a distribuição dos erros é perturbada.

Amostra	Média	DP	Mín.	Q1	Q2	Q3	Máx.
25	40,15	29,59	0	16,75	36,50	60,50	100
50	52,88	32,71	0	22	55	81,25	100
100	56,85	33,41	0	27	63,50	88,25	100
500	63,93	33,85	0	32	71	96	100
1000	65,92	33,47	0	38	73	98	100

Tabela 5.5: Medidas resumo da média da porcentagem de classificação correta pelo tipo de vetor de parâmetros e o tipo de critério clássico e robusto, com perturbação dos erros.

Tipo crit.	betas	Média	DP	Mín.	Q1	Q2	Q3	Máx.
Clássicos	17,10,0,-15,0,13	58,93	36,08	0	28,75	65	94	100
	2,0,0,-1,3,0	31,92	30,85	0	3	23	52,75	100
Robustos	17,10,0,-15,0,13	74,80	23,39	25	55	80,50	97	100
	2,0,0,-1,3,0	58,12	29,12	2	30,50	59	88	100

Com o objetivo de avaliar o desempenho dos critérios de forma mais profunda, foram gerados os

correspondentes gráficos de médias para cada cenário segundo o tamanho de amostra. Os resultados obtidos estão apresentados para dois valores de N , tamanhos de amostra pequenos $N = 25$, Figura 5.6 e os tamanhos de amostra grandes, $N = 500$, Figura 5.7. Os três casos restantes, correspondentes aos tamanhos de amostra $N = 50, 100$ e 1000 encontram-se no Apêndice, Figuras A.3, A.4 e A.5. As figuras mostram o Nível de Contaminação em porcentagem no eixo horizontal inferior, a porcentagem de classificação correta no eixo vertical à esquerda, a perturbação da distribuição dos erros com o correspondente vetor de parâmetros é estabelecido pelo eixo vertical à direita. O termo “beta_letra $N(\#1, \#2)$ ” se refere ao vetor de parâmetros de tamanho “letra” (**p**equeno ou **g**rande) perturbando os erros com uma Normal com média $\#1$, e variância $\#2$. Pelo eixo horizontal superior são separados os critérios clássicos e os robustos para facilitar a comparação.

De acordo com as Figuras 5.6, A.3, A.4, 5.7 e A.5, podemos dizer que os critérios são um pouco instáveis para amostras e parâmetros pequenos. Confirma-se de modo geral que os critérios robustos são melhores que os critérios clássicos, são mais eficientes para os modelos gerados a partir de parâmetros grandes, e na medida que se incrementa o tamanho de amostra, alguns critérios robustos mesmo para pequenos valores dos parâmetros também se tornam eficientes. Pode-se perceber em todos os gráficos um comportamento atípico para alguns cenários. Quando os modelos são gerados a partir do vetor de parâmetros pequenos com perturbação alta ($betas_p N(3;49)$) e a partir dos parâmetros grandes com perturbação pequena ($betas_g N(9;9)$) obtemos resultados discordantes. A classificação correta do critério RR_a^2 piora muito nestes cenários, e o C_p melhora de forma absurda, chegando a competir com os melhores critérios robustos.

Os critérios clássicos tendem a diminuir a classificação correta quando aumenta a proporção de pontos aberrantes em amostras pequenas, e para amostras grandes, tendem a estabilizar seu comportamento, exceto para o grupo de cenários discordantes.

Para o tamanho de amostra $N = 25$, RR_a^2 e $RBIC$ são os melhores critérios para os parâmetros pequenos e grandes respectivamente, exceto para os cenários discordantes, neste caso o melhor critério é o $RMCCV$. Já, a partir do tamanho de amostra $N = 50$, à medida que aumenta o tamanho de amostra, a classificação destes três critérios melhora consideravelmente, primeiro para os parâmetros pequenos e depois para qualquer vetor de parâmetros. Note-se que para os tamanhos de amostra pequenos e parâmetros grandes, o critério $RMCCV$ diminui a classificação correta para os níveis de contaminação maiores.

Resumidamente podemos dizer que os critérios $RBIC$, $RMCCV$ e RR_a^2 foram os critérios com melhor desempenho. Nos cenários discordantes RR_a^2 não apresenta a mesma trajetória, pois foi um dos piores critério em classificar corretamente. Nestes cenários podemos encontrar alguns critérios clássicos, como o BIC e outros que são tão eficientes quanto os critérios robustos, e ainda melhores

Figura 5.6: Gráfico das médias da porcentagem de classificação correta em cada cenário, tamanho de amostra $N = 25$

Figura 5.7: Gráfico das médias da porcentagem de classificação correta em cada cenário, tamanho de amostra $N = 500$

que alguns deles. O critério clássico C_p e sua alternativa robusta RC_p tiveram o pior desempenho no estudo de simulação, só no caso dos cenários discordantes, C_p foi um dos critérios mais eficientes.

5.1.2.2 Diagnóstico do ajuste

No presente trabalho também é de interesse avaliar as técnicas de diagnóstico que foram analisadas no Capítulo 4. É importante levar em conta que a detecção de pontos aberrantes foi realizada considerando pontos de corte fixos. Em dados reais é recomendável fazer a detecção de acordo com os correspondentes valores das medidas e de maneira visual em seus gráficos. Outro detalhe importante é que segundo o desenho das simulações, os pontos aberrantes foram gerados segundo uma distribuição normal padrão perturbada com outra $N(\mu; K^2\sigma^2)$, para $\mu \neq 0$ e $K > 1$, logo, serão fixados os índices das observações correspondentes a esta distribuição, mas pode acontecer que seus valores estejam dentro da amplitude considerada como ponto regular. Esta situação faz com que os critérios tenham uma taxa de detecção mais baixa da verdadeira.

A Tabela 5.6 apresenta a média e o desvio padrão das cinco regras consideradas para avaliar as medidas de detecção de pontos aberrantes. Lembrando que a precisão do ajuste é uma regra positiva, $Acc.(+)$, podemos verificar na tabela que as medidas robustas tiveram melhor precisão que as clássicas. Dentre as medidas clássicas, o resíduo padronizado e os resíduos interno e externamente studentizado forem os de melhor precisão. No caso das medidas robustas, a distância de Cook e de Atkinson tiveram precisões elevadas, ($Acc = 0,956$), assim como o peso w das observações, ($Acc = 0,969$) com baixo desvio padrão. Esta última foi a melhor de todas as medidas segundo esta regra. Da mesma maneira podem ser analisadas as demais regras.

Tabela 5.6: Média e desvio padrão das regras de avaliação das medidas de detecção de pontos aberrantes.

Medidas	Acc	A1	A2	A4	A6	Acc_sd	A1_sd	A2_sd	A4_sd	A6_sd
h_{ii}	0,888	0,000	0,014	0,061	0,092	0,060	0,000	0,081	0,143	0,222
$r_i/\hat{\sigma}$	0,941	0,150	0,912	0,050	0,023	0,043	0,255	0,173	0,162	0,071
t_i	0,942	0,155	0,913	0,054	0,025	0,042	0,257	0,175	0,166	0,076
t_i^*	0,943	0,160	0,913	0,058	0,028	0,042	0,260	0,178	0,171	0,082
D_i	0,890	0,019	0,039	0,0001	0,0001	0,060	0,097	0,107	0,001	0,001
A_i	0,895	0,056	0,177	0,0004	0,001	0,063	0,181	0,251	0,003	0,004
MD_i	0,889	0,000	0,000	0,000	0,000	0,060	0,000	0,000	0,000	0,000
h_{ii}^r	0,947	0,190	0,855	0,080	0,030	0,045	0,288	0,229	0,176	0,056
r_i/s	0,954	0,154	0,206	0,786	0,279	0,020	0,183	0,213	0,220	0,149
rt_i	0,952	0,141	0,187	0,807	0,292	0,022	0,170	0,196	0,202	0,156
rt_i^*	0,948	0,120	0,159	0,837	0,310	0,023	0,144	0,168	0,172	0,161
RD_i	0,956	0,283	0,882	0,012	0,009	0,045	0,341	0,231	0,032	0,028
RA_i	0,956	0,258	0,885	0,024	0,017	0,043	0,321	0,212	0,058	0,046
RMD_i	0,873	0,0002	0,026	0,212	0,252	0,062	0,001	0,102	0,308	0,368
w	0,969	0,324	0,925	0,003	0,002	0,040	0,342	0,176	0,009	0,008

Como são cinco regras para avaliar as medidas de diagnóstico, poderíamos identificar facilmente

a medida mais adequado segundo cada uma destas, mas o objetivo é encontrar a medida que seja globalmente mais adequada. Para isto será considerada a área do pentágono correspondente a cada medida construído a partir do valor das 5 regras, e assim, a medida com maior área será considerada a mais adequada. Para calcular as áreas dos pentágonos foram invertidas as direções dos eixos correspondentes às regras negativas, em que o extremo inferior de cada eixo se corresponde com o mínimo global dos 75 valores correspondentes as 5 regras em cada uma das 15 medidas de diagnóstico analisadas. O extremo superior de cada eixo se corresponde com o máximo global de todos os valores.

Os eixos dos pentágono para as medidas serão:

Acc Proporção de pontos corretamente classificados.

A1 Probabilidade de classificar corretamente todos os pontos.

A2 Probabilidade de classificar corretamente todos os pontos regulares e identificar pelo menos um ponto aberrante.

1-A4 Probabilidade de classificar corretamente todos os pontos regulares.

1-A6 Efeito dos pontos aberrantes corretamente classificados sobre o total de pontos classificados como aberrantes.

Figura 5.8: Gráfico de radar ou de aranha para os pentágonos das medidas de identificação de pontos aberrantes.

Segundo a Figura 5.8 e sua Tabela correspondente, Tabela 5.7, a medida do peso w , que identifica observações discrepantes quando seu peso é inferior a 0,25, foi a mais adequada para detectar

pontos aberrantes com uma área de 1,663. Sendo que a área máxima de uma medida é 2,38. Pode-se conferir também que as medidas Distancia de Cook robusta, RD_i com uma área de 1,578 e sua modificação de Atkinson RA_i com área 1,539, assim como os três resíduos clássicos, resíduo padronizado $r_i/\hat{\sigma}$, resíduo interna e externamente studentizado t_i e externo t_i^* com áreas 1,423, 1,423 e 1,422 respectivamente tiveram um bom desempenho. Todas as medidas clássicas têm quase a mesma probabilidade de classificar corretamente os pontos regulares e quase a mesma proporção de pontos aberrantes corretamente classificados com respeito às observações classificadas como pontos aberrantes (valores de 1-A4 e 1-A6 similares). Este comportamento não está presente de forma uniforme para todas as medidas robustas. As medidas dos resíduos interna e externamente studentizado robustas, e o resíduo padronizado robusto são as medidas com o pior desempenho entre as medidas clássicas e robustas, sendo que suas versões clássicas foram as melhores entre estas.

Tabela 5.7: Média das regras e área das medidas de detecção de pontos aberrantes clássicas e robustas

Medidas	Acc	A1	A2	1-A4	1-A6	Área
h_{ii}	0,888	0,000	0,014	0,939	0,908	0,795
$r_i/\hat{\sigma}$	0,941	0,150	0,912	0,950	0,977	1,423
t_i	0,942	0,155	0,913	0,946	0,975	1,423
t_i^*	0,943	0,160	0,913	0,942	0,972	1,422
D_i	0,890	0,019	0,039	1,000	1,000	0,926
A_i	0,895	0,056	0,177	1,000	0,999	1,013
MD_i	0,889	0,000	0,000	1,000	1,000	0,898
h_{ii}^r	0,947	0,190	0,855	0,920	0,970	1,399
r_i/s	0,954	0,154	0,206	0,214	0,721	0,506
rt_i	0,952	0,141	0,187	0,193	0,708	0,479
rt_i^*	0,948	0,120	0,159	0,163	0,690	0,440
RD_i	0,956	0,283	0,882	0,988	0,991	1,578
RA_i	0,956	0,258	0,885	0,976	0,983	1,539
RMD_i	0,873	0,0002	0,026	0,788	0,748	0,601
w	0,969	0,324	0,925	0,997	0,998	1,663

Para verificar se o performance de estas medidas depende do tamanho de amostra ou da proporção de pontos aberrantes, foram comparadas as áreas das medidas correspondentes a cada cenário, Figura 5.9. Esta análise revelou uma pequena diminuição uniforme das áreas quando o tamanho de amostra aumenta, e as medidas se estabilizam para proporções grandes de erros perturbados. Mediante esta análise confirmou-se que o critério w do peso é o mais adequado.

A fim de detectar alguma relação entre a eficiência das medidas de diagnóstico e o critério de seleção de modelos utilizados foram obtidas as áreas correspondentes às medidas segundo os modelos selecionados por cada critério, Tabela 5.8. Note-se que a detecção de pontos aberrantes para os dados simulados não depende do critério de seleção escolhido, já que as medidas tiveram o mesmo

Tabela 5.8: Média das áreas das medidas correspondentes aos modelos selecionados por cada critério de seleção de modelos

Medidas	RAIC	RBIC	RC_p	RMCCV	RR_a^2
h_{ii}	0,789	0,805	0,778	0,808	0,797
$r_i/\hat{\sigma}$	1,422	1,424	1,418	1,425	1,425
t_i	1,422	1,425	1,418	1,425	1,424
t_i^*	1,421	1,424	1,417	1,425	1,423
D_i	0,931	0,925	0,929	0,919	0,923
A_i	1,020	1,012	1,021	1,005	1,009
MD_i	0,898	0,898	0,898	0,898	0,898
h_{ii}^r	1,417	1,397	1,403	1,383	1,394
r_i/s	0,518	0,521	0,450	0,512	0,533
rt_i	0,489	0,496	0,419	0,488	0,506
rt_i^*	0,452	0,455	0,386	0,446	0,463
RD_i	1,583	1,581	1,575	1,571	1,579
RA_i	1,544	1,547	1,516	1,540	1,549
RMD_i	0,576	0,618	0,552	0,638	0,622
w	1,664	1,662	1,672	1,658	1,660

desempenho independentemente do critério de seleção de modelos utilizado.

A identificação de pontos aberrantes também resultou ser independente do tipo de perturbação na distribuição dos erros e do tamanho dos parâmetros.

Na seguinte seção aplicaremos os critérios de seleção de modelos e as medidas de detecção de pontos aberrantes a dois exemplos de dados conhecidos, os dados de *Stack Loss* e os de *Coleman*, e também para uma parte do banco de dados obtido no Instituto de Astronomia, Geofísica e Ciências Atmosféricas da Universidade de São Paulo (IAG-USP) para identificar quais variáveis meteorológicas influem na temperatura mínima diária.

5.2 Exemplos e aplicação em dados reais

Nesta seção serão aplicados os critérios de seleção de modelos e as medidas de identificação de pontos aberrantes estudadas nas Sessões 2.1, 3.1 e no Capítulo 4. Vamos considerar dois exemplos baseados nos dados de *Stack Loss* e *Coleman*, e uma aplicação em dados reais. Os dados reais correspondem a uma parte dos dados coletados pela Estação Meteorológica do Instituto de Astronomia, Geofísica e Ciências Atmosféricas da Universidade de São Paulo (IAG-USP) referentes à temperatura mínima e variáveis meteorológicas para cada hora, de cada dia, durante os anos de 2000 a 2018.

Figura 5.9: Área das medidas clássicas e robustas de acordo com os cenários determinados pelo tamanho de amostra e o nível de contaminação

5.2.1 Dados Stack loss

Os dados *Stack loss* foram apresentados por [Brownlee \(1965\)](#) e foram utilizados como referência por muitos autores na regressão robusta e detecção de pontos aberrantes. Estes dados são constituídos por três variáveis explicativas e 21 observações, obtidas a partir de medições de uma planta que oxida amônia em ácido nítrico em 21 dias consecutivos. As variáveis explicativas são *Air Flow*, que representa o fluxo de resfriamento do ar, *Water Temp* é a temperatura da água de resfriamento circulada e *Acid Conc* é a concentração do ácido. A variável dependente é *stack.loss*, uma medida (inversa) da eficiência global da planta.

O consenso geral sobre o ajuste do modelo linear para os dados *Stack Loss* é que a variável *Acid*

Conc deve ser descartada do modelo, ou seja, o modelo adequado seria obtido a partir das variáveis explicativas *Air Flow* e *Water Temp*, além de estabelecer as observações 1, 3, 4 e 21 como pontos aberrantes potenciais, [Atkinson e Riani \(2012\)](#); [Hoeting et al. \(1996\)](#). Os dados são apresentados na Tabela 5.9

Tabela 5.9: *Dados Stack Loss*

	Air.Flow	Water.Temp	Acid.Conc.	stack.loss
1	80	27	89	42
2	80	27	88	37
3	75	25	90	37
4	62	24	87	28
5	62	22	87	18
6	62	23	87	18
7	62	24	93	19
8	62	24	93	20
9	58	23	87	15
10	58	18	80	14
11	58	18	89	14
12	58	17	88	13
13	58	18	82	11
14	58	19	93	12
15	50	18	89	8
16	50	18	86	7
17	50	19	72	8
18	50	19	79	8
19	50	20	80	9
20	56	20	82	15
21	70	20	91	15

De acordo com os critérios de seleção de modelos, Tabela 5.10, cinco dos 10 critérios analisados selecionam o modelo mais adequado (com *Air Flow* e *Water Temp*), note-se que o critério *RBIC* foi um deles. O modelo completo foi selecionado por três critérios, dois deles robustos incluindo o *RMCCV* e o modelo com a única variável explicativa *Air Flow* foi selecionado duas vezes pelos critérios C_p e RC_p .

A fim de identificar os pontos aberrantes nos dados *Stack Loss*, foram analisadas as medidas estudadas no Capítulo 4 ao correspondente modelo selecionado com as duas primeiras variáveis explicativas. Na Tabela 5.11 são apresentados os índices das observações que foram identificadas como pontos aberrantes pelo menos por uma medida, na terceira e quarta coluna são apresentadas as frequências com que foram detectadas cada observação pelas medidas clássicas (terceira coluna) e robustas (quarta coluna).

As medidas robustas identificaram as quatro observações consideradas pontos aberrante. A Dis-

Tabela 5.10: Frequência de seleção das variáveis explicativas dos dados *Stack Loss* pelos critérios de seleção de modelos

Critério	constante	Air.Flow	Water.Temp	Acid.Conc.
Cp	1	1	0	0
AICc	1	1	1	0
R_a^2	1	1	1	0
BIC	1	1	1	0
MCCV	1	1	1	1
RCp	1	1	0	0
RAIC	1	1	1	1
RR_a^2	1	1	1	0
RBIC	1	1	1	0
RMCCV	1	1	1	1
Total	10	10	8	3

Tabela 5.11: Frequência de pontos aberrantes detectados segundo as medidas clássicas e robustas

Índice das obs. detectadas	Contagem	Medidas clássicas	Medidas robustas
1	4	0	4
2	1	0	1
3	5	0	5
4	8	2	6
15	1	0	1
16	1	0	1
17	1	0	1
18	1	0	1
19	1	0	1
21	14	6	8

tância de Mahalanobis robusta identificou além dos pontos 1, 3 e 21 as observações 2, e da 15 até 19 como observações discrepantes, sendo que sua versão clássica não identificou nenhum ponto aberrante. Os três resíduos robustos r_i/s , rt_i e rt_i^* foram as únicas medidas que identificaram conjuntamente os 4 pontos aberrantes 1, 3, 4 e 21 e a medida do peso das observações só identificou as observações 3, 4 e 21. Já as medidas clássicas indentificaram exclusivamente as observações 4 e 21.

5.2.2 Dados Coleman

O exemplo a seguir vem das ciências sociais. O conjunto de dados é baseado no estudo de Coleman *et al.* (1966) para 20 escolas, e têm-se cinco variáveis explicativas: média dos salários dos professores por aluno, (*salaryP*) *staff salaries per pupil*, a porcentagem média dos pais de colarinho branco, (*fatherWc*), *percent of white-collar fathers*, média do status socioeconômico, (*sstatus*), média da pontuação do teste verbal dos professores, (*teacherSc*) e média do nível educacional das mães, (*motherLev*).

A variável dependente de interesse é a média da pontuação do teste verbal de cada escola, Tabela 5.12. A proposta defendida por Rousseeuw e Leroy (2005) sobre estes dados se refere a assumir que as observações 3, 17 e 18 são pontos aberrantes. Os autores utilizam o modelo completo mesmo na existência de correlação elevada entre as variáveis explicativas, Figura 5.10.

Tabela 5.12: *Dados Coleman*

	salaryP	fatherWc	sstatus	teacherSc	motherLev	Y
1	3,83	28,87	7,20	26,60	6,19	37,01
2	2,89	20,10	-11,71	24,40	5,17	26,51
3	2,86	69,05	12,32	25,70	7,04	36,51
4	2,92	65,40	14,28	25,70	7,10	40,70
5	3,06	29,59	6,31	25,40	6,15	37,10
6	2,07	44,82	6,16	21,60	6,41	33,90
7	2,52	77,37	12,70	24,90	6,86	41,80
8	2,45	24,67	-0,17	25,01	5,78	33,40
9	3,13	65,01	9,85	26,60	6,51	41,01
10	2,44	9,99	-0,05	28,01	5,57	37,20
11	2,09	12,20	-12,86	23,51	5,62	23,30
12	2,52	22,55	0,92	23,60	5,34	35,20
13	2,22	14,30	4,77	24,51	5,80	34,90
14	2,67	31,79	-0,96	25,80	6,19	33,10
15	2,71	11,60	-16,04	25,20	5,62	22,70
16	3,14	68,47	10,62	25,01	6,94	39,70
17	3,54	42,64	2,66	25,01	6,33	31,80
18	2,52	16,70	-10,99	24,80	6,01	31,70
19	2,68	86,27	15,03	25,51	7,51	43,10
20	2,37	76,73	12,77	24,51	6,96	41,01

Figura 5.10: *Matriz de correlação correspondente ao conjunto de dados Coleman*

Estes dados foram analisados para determinar quais variáveis explicativas são importantes. A Tabela 5.13 apresenta os resultados correspondentes a cada critério de seleção de modelos. Pode-se verificar que todos os critérios concordam com a presença da variável *sstatus* no modelo. Depois desta

variável, as mais frequentes são *teacherSc* (frequência sete) e *salaryP* (frequência quatro). Note-se que não há um consenso de acordo com a frequência sobre os modelos selecionados para critérios clássicos ou robustos, os clássicos selecionam o modelo com as variáveis *sstatus* e *teacherSc*, ou só a variável *sstatus*, este último modelo também é selecionado pelo critério robusto *RMCCV*. Em particular o *RBIC* e *RAIC* selecionaram o modelo completo, de acordo com a proposta de Rousseeuw e Leroy (2005).

Tabela 5.13: Frequência de seleção das variáveis explicativas dos dados de Coleman pelos critérios de seleção de modelos.

Critério	constante	salaryP	fatherWc	sstatus	teacherSc	motherLev
Cp	1	0	0	1	0	0
AICc	1	0	0	1	1	0
R_a^2	1	1	0	1	1	0
BIC	1	0	0	1	1	0
MCCV	1	0	0	1	0	0
RCp	1	0	1	1	1	1
RAIC	1	1	1	1	1	1
RR_a^2	1	1	0	1	1	0
RBIC	1	1	1	1	1	1
RMCCV	1	0	0	1	0	0
Total	10	4	3	10	7	3

De acordo com a seleção de variáveis, acreditamos que há dois possíveis modelos a serem considerados, o modelo que contém como única variável explicativa *sstatus*, (x_3) e o modelo que contém *sstatus* e *teacherSc*, (x_3, x_4), Tabela 5.14.

Tabela 5.14: Ajuste robusto correspondente aos dois modelos selecionados para os dados de Coleman

	Regressão Robusta	
	(x_3)	(x_3, x_4)
Constante	33,130*** (0,505)	13,516* (8,190)
sstatus	0,600*** (0,051)	0,575*** (0,045)
teacherSc		0,788** (0,327)
Erro padrão do resíduo	1,518 (gl = 18)	1,593 (gl = 17)
Nota:	* p<0,1; ** p<0,05; *** p<0,01 níveis descritivos obtidos pelo R	

A identificação de pontos aberrantes foi realizada para os dois modelos propostos, sendo que para ambos modelos foram selecionadas as observações 3 e 18 como observações discrepantes e as medidas robustas identificam para ambos modelos estas mesmas observações. A identificação destas

duas observações pode ser explicada pelos alto e baixo valores do status socioeconômico destas duas escolas que também têm uma média de pontuação do teste verbal baixo e alto respectivamente.

No modelo que contém só a variável *sstatus* as medidas clássicas identificam unicamente a observação 18, as robustas identificam além das observações 3 e 18, as observações 6, 10 e 17 com uma frequência de cinco medidas robustas. A medida do peso das observações identificou estas cinco observações, considerando além das três observações aberrantes propostas por [Rousseeuw e Leroy \(2005\)](#) as observações 6 e 10.

No modelo com as variáveis *sstatus* e *teacherSc*, as medidas robustas identificaram além das observações 3 e 18, a observação 17 com frequência de 4 medidas. Neste modelo a identificação realizada pela medida do peso w coincide com a proposta dos autores.

5.2.3 Dados reais

Esta seção visa testar os algoritmos anteriormente discutidos na determinação das variáveis meteorológicas com maior influência no comportamento da temperatura mínima diária durante um ano, assim como na identificação de valores anômalos que estejam provavelmente associados a erros de medição ou a ocorrência de fenômenos meteorológicos significativos como passagem de frentes frias, ondas de calor ou outros.

A proposta da análise destes dados surgiu do problema apresentado por pesquisadores da área da meteorologia para testar os critérios e medidas analisados nos capítulos anteriores.

Descrição dos dados

Será analisada a influência das variáveis meteorológicas na temperatura mínima do dia. A seleção das variáveis explicativas utilizadas para estimar a temperatura mínima esteve baseada no trabalho de [Klein e Glahn \(1974\)](#). Os dados foram coletados na estação meteorológica do IAG-USP no ano de 2018. Na Tabela 5.15 são descritas as sete variáveis explicativas e a variável dependente $t_{\min}$ considerada. Os valores de temperaturas máximas e mínimas no dia foram obtidas a partir de medições horárias feitas com termômetros. A temperatura do ponto de orvalho foi calculada a partir de registros dos termômetros de temperatura seca e temperatura do bulbo úmido (psicrômetro), esta variável mede a temperatura mínima a partir da qual inicia a queda de rocío, que geralmente acontece no horário das 6 UTC, (este horário representa de acordo com a estação, as três ou quatro horas da manhã aproximadamente). A pressão atmosférica e as componentes do vento derivaram-se de medições com barômetro e anemógrafo, respectivamente.

A Tabela 5.16 apresenta as variáveis descritas mostrando os primeiros e últimos seis dias do ano de 2018. Este banco de dados não contém observações incompletas. Os dados não serão disponibilizados pois eles só podem ser solicitados ao IAG-USP. Para mais informação solicitar dados

Tabela 5.15: Descrição das variáveis do banco de dados do IAG-USP

Variável	Descrição	Unidade
t_min	Temperatura mínima do ar no dia D	°C
t_min_ant	Temperatura mínima do ar no dia D-1	°C
t_max_ant	Temperatura máxima do ar no dia D-1	°C
t_orvalho	Temperatura do ponto de orvalho às 06 UTC no dia D. Esta variável é uma medida de umidade.	°C
dif_pres	Diferença entre as pressões atmosféricas dos dias D e D-1 às 06 UTC.	hPa
ff	Velocidade média do vento entre as 05 UTC e as 06 UTC no dia D.	km/h
u	Componentes da projeção de ff nos paralelos.	km/h
v	Componentes da projeção de ff nos meridianos.	km/h

Nota:

D = 1,2,...,365 dias

http://www.estacao.iag.usp.br/sol_dados.php

Tabela 5.16: Dados coletados da Estação Meteorológica do IAG - USP dos 365 dias do ano de 2018

	t_min	t_min_ant	t_max_ant	t_orvalho	dif_pres	ff	u	v
1	19,80	19,60	26,30	18,90	-1,67	6	-5,50	-2,30
2	21,20	19,80	30,10	20,23	3,05	10	-9,20	-3,80
3	18,20	21,20	28,20	18,97	1,84	11	-10,20	-4,20
4	16,20	18,20	26,00	16,59	-7,24	4	-3,70	-1,50
5	15,30	16,20	28,20	15,60	0,14	6	-5,50	-2,30
6	18,40	15,30	29,00	17,85	0,78	1	-0,90	-0,40
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
360	15,80	18,50	23,80	15,99	-0,38	4	-3,70	-1,50
361	17,70	15,80	26,50	17,74	2,03	6	-5,50	-2,30
362	17,00	17,70	27,60	18,28	-0,05	2	-1,80	-0,80
363	16,50	17,00	29,20	16,85	0,59	0	0,00	0,00
364	16,50	16,50	30,10	16,42	0,42	5	-4,60	-1,90
365	16,60	16,50	31,40	17,24	-0,69	2	-1,80	-0,80

Análise Descritiva

As medidas resumo correspondentes aos dados do IAG-USP são apresentados na Tabela 5.17. A temperatura mínima durante o ano 2018 esteve entre 5,9 e 21,6 °C, para uma média de 15,54 °C e desvio padrão 3,124 °C. Note-se também que a temperatura máxima tem uma variabilidade maior comparada com as outras variáveis, no entanto, sua média também é maior, de modo que os coeficientes de variação são próximos.

Seleção de modelos

Muitas das variáveis meteorológicas possuem um comportamento sazonal, fato relacionado com a variação anual da quantidade de radiação solar recebida na superfície da terra. Por exemplo, a temperatura é particularmente susceptível a estas mudanças, fazendo com que seus valores médios

Tabela 5.17: Medidas resumo do banco de dados da Estação Meteorológica do IAG - USP

Statistic	N	Média	DP	Mínimo	Q1	Q3	Máximo
t_min	365	15,540	3,124	5,900	13,300	17,900	21,600
t_min_ant	365	15,548	3,131	5,900	13,300	18,000	21,600
t_max_ant	365	26,134	4,452	13,800	23,200	29,100	35,000
t_orvalho	365	15,503	3,126	5,267	13,049	17,981	21,111
dif_pres	365	-0,004	2,048	-7,239	-1,140	1,300	6,454
ff	365	4,419	2,901	0,000	2,000	6,000	14,000
u	365	-4,073	2,684	-12,900	-5,500	-1,800	0,000
v	365	-1,687	1,106	-5,400	-2,300	-0,800	0,000

anuais apresentam uma forma aproximadamente sinusoidal com máximos e mínimos nos meses de verão (i.e. dezembro, janeiro e fevereiro) e inverno (i.e. junho, julho e agosto), respectivamente. Levando em conta que a base de dados apresenta medições diárias durante um ano, é necessário diminuir o efeito da sazonalidade nas variáveis de temperatura. Para transformar esses dados foi ajustada uma função sinusoidal e os valores ajustados foram subtraídos dos valores originais.

Sabe-se que neste caso existe autocorrelação entre os erros, devido ao fato de que as variáveis são tomadas sequencialmente no tempo. No entanto, o procedimento de seleção é descritivo, não utilizando testes de significância, e portanto, este fato não é um grave problema. A situação em que existe autocorrelação dos erros no procedimento robusto é um tópico a ser estudado.

Considerou-se inicialmente o modelo completo, com todas as variáveis explicativas e a partir deste modelo foi aplicado o método de mínimos quadrados ordinários. Percebeu-se a presença de observações discrepantes, Figura 5.11, neste caso, é adequado utilizar o método da regressão robusta como alternativa.

Foram aplicados os critérios de seleção de modelos clássicos e robustos ao ajuste linear robusto. Os resultados da seleção de modelos são apresentados na Tabela 5.18. Nesta tabela podemos conferir que as variáveis selecionadas com mais frequência foram a temperatura do ponto de orvalho (frequência 10), a diferença de pressão (frequência 10), a temperatura mínima do dia anterior (frequência 9) e a projeção u do vento (frequência 8). Logo, o modelo adequado para estes dados será aquele que contém estas variáveis. Todos os critérios coincidem no sentido que as variáveis explicativas temperatura de orvalho e diferença de pressão devem ser consideradas. O modelo mais adequado foi o mesmo selecionado pelos “melhores” critérios robustos RBIC e RMCCV e suas versões clássicas, o critério robusto RAIC também selecionou este modelo.

De acordo com a seleção realizada, o modelo final tem a seguinte equação:

$$\widehat{t_min}_i = -0,642 + 0,215 \cdot t_min_ant_i + 0,693 \cdot t_orvalho_i + 0,084 \cdot dif_pres_i - 0,079 \cdot u_i, \quad (5.7)$$

Figura 5.11: Análise de diagnóstico do R para o ajuste pelo método de MQO do modelo completo dos dados do IAG-USP

Tabela 5.18: Frequência de seleção das variáveis explicativas dos dados do IAG-USP pelos critérios de seleção de modelos.

Critério	constante	t_min_ant	t_max_ant	t_orvalho	dif_pres	ff	u	v
Cp	1	1	1	1	1	1	1	1
AICc	1	1	0	1	1	1	1	0
R_a^2	1	1	0	1	1	0	0	0
BIC	1	1	0	1	1	0	1	0
MCCV	1	1	0	1	1	0	1	0
RCp	1	0	0	1	1	0	0	1
RAIC	1	1	0	1	1	0	1	0
RR_a^2	1	1	0	1	1	1	1	0
RBIC	1	1	0	1	1	0	1	0
RMCCV	1	1	0	1	1	0	1	0
Total	10	9	1	10	10	3	8	2

A Tabela 5.19 mostra o resumo do ajuste do modelo selecionado pelo método de Mínimos Quadrados Ordinários e Regressão Robusta, com as correspondentes estimativas clássicas e robustas dos coeficientes e os desvios padrões entre parêntesis (). Note-se que as estimativas dos parâmetros correspondentes ao método de Regressão Robusta têm desvios padrões menores que as estimativas obtidas pelo método Mínimos Quadrados Ordinários.

As variáveis meteorológicas mais significativas coincidiram com algumas das variáveis que melhor explicam a física do processo de diminuição da temperatura do ar. Especificamente, a temperatura mínima do dia anterior está fortemente relacionada com a variável dependente devido à frequente persistência de condições atmosféricas similares em dias consecutivos. A temperatura do

Tabela 5.19: *Análise de regressão do ajuste do modelo da variável temperatura mínima diária em função das variáveis temperatura mínima do dia anterior, temperatura de orvalho às 06 UTC, diferença de pressão com o dia anterior às 06 UTC e a componente de projeção u da velocidade do vento pelo método de Mínimos Quadrados Ordinários e Regressão Robusta.*

	Variável dependente:	
	$y = t_min$	
	<i>MQO</i>	<i>Reg. Rob.</i>
Constante	-0,604*** (0,097)	-0,642*** (0,093)
t_min_ant	0,230*** (0,035)	0,215*** (0,033)
t_orvalho	0,657*** (0,034)	0,693*** (0,032)
dif_pres	0,079*** (0,026)	0,084*** (0,025)
u	-0,059*** (0,020)	-0,079*** (0,019)
Erro padrão do resíduo (gl = 360)	0,999	0,819
Nota: * $p < 0,1$; ** $p < 0,05$; *** $p < 0,01$ níveis descritivos obtidos pelo R		

ponto de orvalho limita a diminuição da temperatura e a relevância da variável diferença de pressão nas últimas 24 horas está relacionada com o fornecimento de informação sobre a entrada de novos sistemas meteorológicos que fazem com que aconteça uma mudança no valor mínimo da temperatura. Embora a magnitude da componente u do vento esteja menos relacionada com a temperatura mínima, pode-se apontar que o seu valor está associado à temperatura, por tanto influência no comportamento da temperatura mínima.

Identificação de pontos aberrantes

Os resultados da análise de detecção de pontos aberrantes são apresentados na Tabela 5.20.

Foram detectadas ao todo 50 observações discrepantes, algumas com maior e outras com menor frequência. Analisando as medidas de cada tipo, vamos considerar aquelas observações que foram selecionadas mais de quatro vezes pelas medidas clássicas e mais de cinco pelas medidas robustas, sendo que as medidas clássicas são sete e as robustas oito. Segundo este raciocínio, as observações detectadas com maior frequência pelas medidas clássicas seriam 13, 15, 19, 29, 191, 194, 237, 257, 277, 292, 300, 327. Pelas medidas robustas foram detectados os mesmos dias que as clássicas, adicionando os dias 3, 323 e 356.

Levando em consideração a medida que obteve os melhores resultados na simulação, w , os seguintes dias foram detectados como observações discrepantes usando o ponto de corte 0,5: 13, 19, 29,

Tabela 5.20: *Frequência de todos os pontos aberrantes detectados segundo as medidas clássicas e robustas para os dados do IAG-USP.*

Índice das observações detectadas	Freq. total	Freq. clássica	Freq. robusta	Índice das Observações detectadas	Freq. total	Freq. clássica	Freq. robusta
3	8	2	6	238	4	2	2
4	4	2	2	246	7	3	4
13	12	5	7	247	4	2	2
15	11	5	6	248	4	2	2
19	12	5	7	249	1	0	1
29	11	5	6	250	4	2	2
41	6	3	3	257	11	5	6
75	7	3	4	263	1	1	0
140	6	4	2	266	8	3	5
141	4	2	2	269	1	0	1
142	4	2	2	277	12	5	7
143	3	1	2	292	11	5	6
154	8	4	4	293	1	0	1
165	8	4	4	300	15	7	8
191	13	5	8	301	4	2	2
194	13	6	7	316	6	3	3
201	6	3	3	322	2	1	1
210	3	0	3	323	9	3	6
212	3	0	3	327	12	5	7
222	4	2	2	340	5	2	3
223	2	0	2	342	1	0	1
224	1	0	1	343	3	1	2
226	4	2	2	344	4	2	2
228	2	1	1	346	7	3	4
237	11	5	6	356	8	2	6

Nota 1: primeiras 25 detectadas

Nota 2: últimas 25 detectadas

75, 191, 237, 246, 257, 277, 292, 300, 323, 327, 346, Figura 5.12

Note-se que esta medida selecionou apenas quatro dias diferentes dos selecionados com maior frequência pelas medidas robustas, identificando dois dias a mais, e dois dias a menos.

Os dias detectados como observações discrepantes pelo peso das observações w foram associados com fenômenos meteorológicos de interesse. Observou-se que nos meses de inverno e primavera, os dias detectados são consequência da chegada de um sistema frontal sobre o ponto de medição, isto significa que os dias identificados usando o peso inferior a 0,5 estiveram associados com a chegada de sistemas frontais.

Desta forma, observa-se que as medidas de diagnóstico robustas foram eficientes no sentido de detectar observações realmente atípicas segundo a natureza do problema.

Finalizando, no capítulo seguinte discutiremos as conclusões mais relevantes da dissertação.

Figura 5.12: Diagnóstico para os dados do IAG-USP da medida w com ponto de corte 0,5.

6 CONCLUSÕES

Nesta dissertação foram apresentados vários critérios de seleção de modelos clássicos e robustos na análise de regressão. Foi visto que este tipo de critérios são úteis para decidir qual modelo se adequa melhor à maioria dos dados, sendo os critérios robustos mais adequados na presença de pontos aberrantes. Também foram apresentadas as técnicas de diagnóstico clássicas mais utilizadas e suas correspondentes alternativas robustas, além de introduzir uma nova medida de detecção de pontos aberrantes.

Aplicaram-se os critérios e medidas analisadas a um banco de dados reais e realizou-se a comparação entre as metodologias clássicas e robustas.

Analisando os gráficos e tabelas geradas conjuntamente, pode-se concluir que os critérios robustos de seleção de modelos identificam com maior frequência o modelo correto, em particular os critérios RBIC e RMCCV tiveram as melhores porcentagens de acerto na presença de pontos aberrantes. Para dados sem evidência da existência de pontos aberrantes, as versões clássicas destes critérios também mostram-se competentes. Para as medidas de identificação de pontos aberrantes, a medida correspondente ao peso w associado a cada observação mostrou-se a mais adequada, juntamente com a distância de Cook robusta e sua correspondente modificação de Atkinson robusta, assim como os resíduos clássicos. Note-se que estas últimas medidas referem-se à abordagem clássica aplicada num ajuste de regressão robusta.

As figuras e tabelas apresentadas nesta dissertação foram geradas no *software* R, algumas rotinas podem ser encontradas no Apêndice [A.2](#).

Para continuação deste trabalho sugere-se a implementação dos critérios da Classe 2 e a possibilidade de considerar combinações entre os critérios da Classes 1 e 2. Também sugere-se analisar a natureza e eficiência dos critérios de seleção de modelos em dados com estruturas de correlação pré-estabelecidas.

A seguir é apresentado o esquema sugerido para o análise de dados em situações similares às analisadas durante as simulações, tamanho de amostra até 1000, quantidade de variáveis não muito maior do que cinco e que sejam uniformemente distribuídas.

Figura 6.1: Esquema sugerido para a análise de dados

APÊNDICES

A.1 Tabelas e Figuras complementares

Tabela A.1: Porcentagens de classificação e tempo de cálculo das 100 simulações para cada cenário sem perturbação dos erros, $N \times \underset{5}{betas} \times \underset{2}{classe} \times \underset{4}{critérios} = 400$

	Porc. (%)	Tempo (s)	N	$(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6)$	Classe	Critério
1	58	0,081	25	2,0,0,-1,3,0	Correto	AICc
2	20	0,081	25	2,0,0,-1,3,0	Super-ajuste	AICc
3	12	0,081	25	2,0,0,-1,3,0	Sub-ajuste	AICc
4	10	0,081	25	2,0,0,-1,3,0	Outros	AICc
5	81	0,086	25	17,10,0,-15,0,13	Correto	AICc
6	19	0,086	25	17,10,0,-15,0,13	Super-ajuste	AICc
⋮	⋮	⋮	⋮	⋮	⋮	⋮
395	0	0,098	1000	2,0,0,-1,3,0	Sub-ajuste	RR_a^2
396	38	0,098	1000	2,0,0,-1,3,0	Outros	RR_a^2
397	42	0,095	1000	17,10,0,-15,0,13	Correto	RR_a^2
398	36	0,095	1000	17,10,0,-15,0,13	Super-ajuste	RR_a^2
399	0	0,095	1000	17,10,0,-15,0,13	Sub-ajuste	RR_a^2
400	22	0,095	1000	17,10,0,-15,0,13	Outros	RR_a^2

Tabela A.2: Porcentagens de classificação e tempo de cálculo das 100 simulações para cada cenário quando perturbada a distribuição dos erros, $N \times \underset{5}{betas} \times \underset{2}{\pi} \times \underset{4}{pert} \times \underset{3}{classe} \times \underset{4}{critérios} = 4800$

	Porc. (%)	Tempo (s)	N	$(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6)$	π (%)	Pert. (μ_K^2)	Classe	Critério
1	49	0,082	25	2,0,0,-1,3,0	4	1_4	Correto	AICc
2	14	0,082	25	2,0,0,-1,3,0	4	1_4	Super-ajuste	AICc
3	23	0,082	25	2,0,0,-1,3,0	4	1_4	Sub-ajuste	AICc
4	14	0,082	25	2,0,0,-1,3,0	4	1_4	Outros	AICc
5	16	0,084	25	2,0,0,-1,3,0	4	2_25	Correto	AICc
6	6	0,084	25	2,0,0,-1,3,0	4	2_25	Super-ajuste	AICc
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
4795	0	0,094	1000	17,10,0,-15,0,13	20	18_49	Sub-ajuste	RR_a^2
4796	0	0,094	1000	17,10,0,-15,0,13	20	18_49	Outros	RR_a^2
4797	100	0,092	1000	17,10,0,-15,0,13	20	27_100	Correto	RR_a^2
4798	0	0,092	1000	17,10,0,-15,0,13	20	27_100	Super-ajuste	RR_a^2
4799	0	0,092	1000	17,10,0,-15,0,13	20	27_100	Sub-ajuste	RR_a^2
4800	0	0,092	1000	17,10,0,-15,0,13	20	27_100	Outros	RR_a^2

(a) $N = 25$

(b) $N = 50$

(c) $N = 100$

(d) $N = 500$

(e) $N = 1000$

Figura A.2: Matriz de correlação entre as variáveis explicativas.

Tabela A.3: Medidas resumo das porcentagens de classificação de cada critério em cada classe. Tabela para dados sem perturbação, correspondente às Figuras (5.1) e (5.2).

Critério	Clase	Média	DP	Mín.	Q1	Q2	Q3	Máx.
BIC	Correto	87,70	12,93	57	82,50	92	96	99
	Super-ajuste	9,60	7,55	1	4	7,50	16	21
	Sub-ajuste	1,40	4,43	0	0	0	0	14
	Outros	1,30	3,13	0	0	0	0,75	10
MCCV	Correto	88,80	12,36	57	85,25	93	96	99
	Super-ajuste	7,80	5,12	1	4	7	11,50	16
	Sub-ajuste	1,90	6,01	0	0	0	0	19
	Outros	1,50	3,75	0	0	0	0,75	12
R_a^2	Correto	32,20	10,88	18	23,25	32	41,75	47
	Super-ajuste	42,40	5,82	30	41	43	47	48
	Sub-ajuste	1,70	3,47	0	0	0	0,75	10
	Outros	23,70	10,93	11	13,50	24,50	28	41
AICc	Correto	69,40	8,68	56	65,25	69	74	82
	Super-ajuste	25,30	6,13	16	21	24,50	28,75	36
	Sub-ajuste	1,20	3,79	0	0	0	0	12
	Outros	4,10	3,38	0	2	2,50	6,75	10
Cp	Correto	29,80	5,33	21	26	30	33	38
	Super-ajuste	8,20	5,18	3	4,25	5,50	12,75	16
	Sub-ajuste	0,60	1,90	0	0	0	0	6
	Outros	61,40	8,69	50	52	64	69,75	70
RBIC	Correto	89,30	15,22	51	88,75	94,50	99	100
	Super-ajuste	7,60	8,42	0	1	5	10,25	23
	Sub-ajuste	1,80	5,69	0	0	0	0	18
	Outros	1,30	3,13	0	0	0	0,75	10
RMCCV	Correto	88	13,09	53	86,50	92	95,50	98
	Super-ajuste	8,10	4,25	2	4,50	8	11	15
	Sub-ajuste	2,20	6,96	0	0	0	0	22
	Outros	1,70	4,37	0	0	0	0,75	14
RR_a^2	Correto	35,40	9,43	17	29,50	37,50	41,75	49
	Super-ajuste	42,60	3,53	36	40,50	43	44,75	48
	Sub-ajuste	0,30	0,95	0	0	0	0	3
	Outros	21,70	9,07	11	13,50	21	28,50	38
RAIC	Correto	56,30	12,94	30	50,25	57,50	64,50	73
	Super-ajuste	33	5,25	24	31	34,50	36,75	39
	Sub-ajuste	0,40	1,26	0	0	0	0	4
	Outros	10,30	9,27	2	4	8,50	13	33
RCp	Correto	27,50	9,41	12	20,50	28	32,75	44
	Super-ajuste	44	8,12	26	42,25	45	48,25	55
	Sub-ajuste	0,90	1,91	0	0	0	0	5
	Outros	27,60	12,80	14	19	27	31	58

Tabela A.4: Medidas resumo das porcentagens de classificação de cada critério em cada classe. Tabela para dados com perturbação, correspondente às Figuras (5.3) e (5.4).

Critério	Classe	Média	DP	Mín.	Q1	Q2	Q3	Máx.
BIC	Correto	50,29	41,66	0	4,75	48,50	98	100
	Super-ajuste	1,30	3,03	0	0	0	1	17
	Sub-ajuste	32,61	33,35	0	0	23,50	57	99
	Outros	15,80	24,24	0	0	2,50	20,25	98
MCCV	Correto	48,52	25,00	1	29	47	72	98
	Super-ajuste	21,56	13,48	0	11	19,50	31	52
	Sub-ajuste	9,97	12,25	0	0	6	16	50
	Outros	19,94	19,79	0	3	11,50	36	76
R_a^2	Correto	46,85	30,95	0	20,75	45,50	72,25	100
	Super-ajuste	17,17	15,86	0	1	15	32,25	50
	Sub-ajuste	16,33	22,15	0	0	3	31	74
	Outros	19,65	19,70	0	3	12	34	92
AICc	Correto	54,04	36,25	0	15,50	61	90,25	100
	Super-ajuste	4,74	8,42	0	0	0	5,25	31
	Sub-ajuste	26,17	27,76	0	0	17,50	47	82
	Outros	15,05	22,07	0	0	5	20	98
Cp	Correto	27,43	38,90	0	0	1	61,75	100
	Super-ajuste	1,40	3,45	0	0	0	0	19
	Sub-ajuste	30,98	38,53	0	0	4	66,50	100
	Outros	40,19	41,45	0	1	22	96	100
RBIC	Correto	85,27	16,56	35	80	91	97	100
	Super-ajuste	8,71	6,97	0	3	8	11	34
	Sub-ajuste	3,83	8,58	0	0	0	0,25	32
	Outros	2,20	4,86	0	0	0	2	23
RMCCV	Correto	85,22	21,61	2	86,75	94	97	100
	Super-ajuste	5,27	4,70	0	2	4	8	26
	Sub-ajuste	6,93	17,17	0	0	0	1	81
	Outros	2,58	8,45	0	0	0	1	56
RR_a^2	Correto	74,03	25,69	6	52,75	84	98,25	100
	Super-ajuste	14,33	15,49	0	0	9	27	53
	Sub-ajuste	7,06	18,54	0	0	0	0	82
	Outros	4,58	7,12	0	0	1	6,25	26
RAIC	Correto	57,91	12,02	29	53	61	66	82
	Super-ajuste	31,92	5,71	17	28	31	35	48
	Sub-ajuste	1,07	2,95	0	0	0	0	12
	Outros	9,10	8,47	0	3	7	12,25	33
RCp	Correto	29,88	10,88	11	20,75	28	38	55
	Super-ajuste	42,08	7,78	22	38	43	47	61
	Sub-ajuste	1,34	2,69	0	0	0	1	13
	Outros	26,70	12,82	6	15	25,50	37,25	56

Tabela A.5: Medidas resumo das médias das porcentagens de classificação correta pelo tamanho de amostra e para o tipo de critério clássico e robusto quando perturbados os erros.

Tipo de critério	Amostra	Média	DP	Mín.	Q1	Q2	Q3	Máx.
Clássicos	25	30,57	30,33	0	4	21,50	48,25	98
	50	39,86	33,87	0	5,50	36	66,75	99
	100	43,67	35,87	0	6,75	41,50	78,25	100
	500	54,62	37,16	0	20	56,50	94,25	100
	1000	58,42	36,54	0	26,50	63	97	100
Robustos	25	49,73	25,57	2	33	45	70	100
	50	65,90	25,69	11	46	68,50	89	100
	100	70,02	24,61	16	53	85	90	100
	500	73,25	27,30	18	58,75	87	97	100
	1000	73,42	28,29	14	56,75	86,50	98	100

Tamanho de amostra N = 50

Figura A.3: Gráfico das médias da porcentagem de classificação correta em cada cenário do tamanho de amostra N = 50

Figura A.4: Gráfico das médias da porcentagem de classificação correta em cada cenário do tamanho de amostra $N = 100$

Tamanho de amostra $N = 1000$

Figura A.5: Gráfico das médias da porcentagem de classificação correta em cada cenário do tamanho de amostra $N = 1000$

A.2 Algumas funções e pacotes do R

```

1 #####
2 #          LIBRARIAS ----
3 #####
4 library(corrplot)
5 library(GGally)
6 library(caret)
7 library(dplyr)
8 library(ggplot2)
9 library(gtools)
10 library(MASS)
11 library(rel)
12 library(robustbase)
13 library(rlist)
14 library(stringr)
15 library(stargazer)
16 library(tcltk)
17 library(R.utils)
18 library(plyr)
19 library(ggplot2)
20 library(FSA)
21 require(RColorBrewer)
22 library(chemometrics)
23 #####
24 #          CORES E SIMBOLOS ----
25 #####
26 # cores dos critério de selecao de modelos
27 pal5 = brewer.pal(10, "Paired")[c(6,2,4,8,10)]
28 # forma dos simbolos dos criterios de selecao de modelos
29 form = c(17,15,16,18,8)
30 # cores das medidas de identificacao de potnos aberrantes
31 pal = c("magenta", "#228B22", "#CC0000", "#00008B", "#606060", "#8A2BE2", "#FF6F01", "black")
32 #####
33 #          CONSTRUCCAO DE TABELAS ----
34 #####
35 # media e desvio padrao da variavel X1 para cada grupo das variavels X2 e X3 da tabela tab
36 data.frame(tab %>%
37   group_by(X2, X3) %>%
38   summarize_at(vars(X1),funs(Media = mean, DP = sd)))
39 #####
40 #          SAIDA TABELAS EM STARGAZER ----
41 #####
42 # gerar tabela dos dados
43 stargazer(tab,summary = FALSE,type = "latex",decimal.mark = ",")
44 # gerar medidas resumo das tabelas de dados
45 stargazer(tab,summary = TRUE,type = "latex",decimal.mark = ",")
46 # gerar tabela resumo de ajustes de modelos

```



```

47 stargazer(fit1,fit2,...,fitk,type = "latex",decimal.mark = ",")
48 #####
49 #          GRAFICO DA PORCENTAGEM DE CLASSIFICACAO DOS CRITERIOS CLASSICOS E ROBUSTOS PARA
          CADA TAMANHO DE AMOSTRA, E TIPO DE PERTURBACAO DOS ERROS ----
50 #####
51 #definir endereco para salavra ao grafico em formato pdf
52 pdf("endereco_artigo.pdf", width = 7.08, height = 8.67)
53 #criar grafico
54 fig = ggplot(data, aes(x=N.cont, y=Porcentagem, group = Critérios))
55 fig + facet_grid(data$beta_pert ~ data$Tipo_crit) +
56   geom_line(aes(linetype = Critérios,color = Critérios), size = 1) +
57   geom_point(aes(shape = Critérios, color = Critérios), size = 4) +
58   scale_shape_manual(values = rep(form,2)) +
59   scale_color_manual(values = rep(pal5,2))+
60   scale_linetype_manual(values = rep(line,2)) +
61   theme(panel.grid = element_line(color = "grey"),
62         legend.position = "right",
63         legend.direction = "vertical",
64         strip.text = element_text(face = "bold"),
65         panel.background = element_rect(fill = "white"),
66         legend.text = element_text(size = 10),
67         legend.title = element_text(size = 12),
68         axis.text = element_text(color="black"),
69         axis.text.y = element_text(size = 12),
70         axis.text.x = element_text(size = 12),
71         axis.title.y = element_blank(),
72         axis.title.x = element_text(face = "bold"),
73         text = element_text(color="black",size = 13),
74         panel.border = element_rect(linetype = 1,fill = NA)) +
75   labs(title = "Tamanho de amostra N = n", x = "Níveis de Contaminação") +
76   scale_y_continuous(labels = function(x){paste0(x,"%")}) +
77   scale_x_discrete(labels = function(x){paste0(x,"%")})
78 #fechar imagem
79 dev.off()
80 #####
81 #          AREA DOS PENTAGONOS ASSOCIADOS A CADA MEDIDA DE DETECAO DE PONTOS ABERRANTES ----
82 #####
83 #calcula da area de uma medida para as 5 regras de avalcao
84 area = sin(2*pi/5)*0.5*(Acc*A1 + A1*A2 + A2*(1 - A4) + (1 - A4)*(1 - A6) + (1 - A6)*Acc
85 #####
86 #          GRAFICO DAS AREAS DAS MEDIDAS CLASSICAS E ROBUSTAS DE ACORDO COM O TAMANHO DE
          AMOSTRA E A PROPORCAO DE PONTOS ABERRANTES ----
87 #####
88 ggplot(DADOS, aes(x = Amostra, y = area, group = Medidas, color = Medidas, shape = Medidas)
      ) +
89   facet_grid(N.cont ~ Tipo_medida) +
90   geom_point(size=2,stroke = 1)+
91   scale_shape_manual(values = rep(c(1:7,1:8),2)) +
92   scale_color_manual(values = rep(c(pal[1:7],pal),2)) +

```

```

93 scale_y_continuous(breaks = c(0,0.5,1,1.5,2,2.38),
94                     limits = c(0,2.38)) +
95 theme(panel.grid = element_line(color = "grey"),
96        legend.position = "right",
97        legend.direction = "vertical",
98        strip.text = element_text(face = "bold"),
99        panel.background = element_rect(fill = "white"),
100        legend.text = element_text(size = 10),
101        legend.title = element_text(size = 12),
102        axis.text = element_text(color="black"),
103        axis.text.y = element_text(size = 12),
104        axis.text.x = element_text(size = 12),
105        axis.title.y = element_blank(),
106        axis.title.x = element_text(face = "bold"),
107        text = element_text(color="black",size = 13),
108        panel.border = element_rect(linetype = 1,fill = NA))+
109 labs(x = "Tamanho de Amostra", y = "Area do pentágono")
110 #####
111 # GRAFICO DAS MEDIDAS RESIDUO PADRONIZADO E W ----
112 #####
113 # residuo padronizado
114 rpad_cond = abs(rpad) > 2
115 ggplot(rpad,aes(x=Indice,y=rpad)) +
116   geom_point(colour = ifelse(rpad_cond,'red',"#00000080"),shape = ifelse(rpad_cond,15,16))
117   +
118   geom_text(aes(label = ifelse(rpad_cond,Indice,"")), vjust = -0.5, size = ifelse(rpad_cond
119     ,3,1)) +
120   geom_hline(yintercept = 2, lty = 2) +
121   geom_hline(yintercept = -2, lty = 2) +
122   scale_y_continuous(limits = c(-4,3)) +
123   labs(y = "Resíduo padronizado",title = "Pontos aberrantes")
124 # w peso das observacoes
125 w_cond = w < w_p.corte
126 ggplot(dw,aes(x=Indice,y=w)) +
127   geom_point(colour = ifelse(w_cond,'red',"#00000080"),shape = ifelse(w_cond,15,16)) +
128   geom_text(aes(label = ifelse(w_cond,Indice,"")),vjust = -0.5, size = ifelse(w_cond,3,1))
129   +
130   geom_hline(yintercept = w_p.corte, lty = 2)+
131   labs(subtitle = paste("Ponto de corte: peso inferior a ",round(w_p.corte,digits=5),sep =
132     " "),
133        x = "Indice",
134        y = "Peso das observações")

```

apendice/codigo.txt

REFERÊNCIAS BIBLIOGRÁFICAS

- Agostinelli(2002)** Claudio Agostinelli. Robust model selection in regression via weighted likelihood methodology. *Statistics & probability letters*, 56(3):289–300. Citado na pág. [17](#)
- Akaike(1973)** Hirotugu Akaike. Information theory and an extension of the maximum likelihood principle. *Proc. 2nd International Symposium on Information Theory (B. N. Petrov and F. Csaki eds.) Akademiai Kiado, Budapest*, páginas 267–281. Citado na pág. [3](#), [6](#)
- Akaike(1987)** Hirotugu Akaike. Factor analysis and aic. Em *Selected Papers of Hirotugu Akaike*, páginas 371–386. Springer. Citado na pág. [6](#)
- Alma(2011)** Özlem Gürnlü Alma. Comparison of robust regression methods in linear regression. *Int. J. Contemp. Math. Sciences*, 6(9):409–421. Citado na pág. [30](#)
- Alqallaf et al.(2002)** Fatemah A Alqallaf, Kjell P Konis, R Douglas Martin e Ruben H Zamar. Scalable robust covariance and correlation estimates for data mining. Em *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, páginas 14–23. ACM. Citado na pág. [18](#)
- Arlot et al.(2010)** Sylvain Arlot, Alain Celisse et al. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79. Citado na pág. [9](#), [16](#), [32](#)
- Atkinson e Riani(2012)** Anthony Atkinson e Marco Riani. *Robust diagnostic regression analysis*. Springer Science & Business Media. Citado na pág. [44](#)
- Atkinson(1981)** Anthony C Atkinson. Two graphical displays for outlying and influential observations in regression. *Biometrika*, 68(1):13–20. Citado na pág. [24](#)
- Bartlett et al.(2002)** Peter L Bartlett, Stéphane Boucheron e Gábor Lugosi. Model selection and error estimation. *Machine Learning*, 48(1-3):85–113. Citado na pág. [8](#)
- Beckman e Cook(1983)** Richard J Beckman e R Dennis Cook. Outlier..... s. *Technometrics*, 25(2):119–149. Citado na pág. [29](#)
- Belsley et al.(1980)** David A Belsley, Edwin Kuh e Roy E Welsch. Regression diagnostics. j, 1980. Citado na pág. [24](#)
- Bhansali e Downham(1977)** Rajendra J Bhansali e David Y Downham. Some properties of the order of an autoregressive model selected by a generalization of akaike's epf criterion. *Biometrika*, 64(3):547–551. Citado na pág. [6](#)
- Brownlee(1965)** Kenneth Alexander Brownlee. *Statistical theory and methodology in science and engineering*, volume 150. Wiley New York. Citado na pág. [43](#)
- Burnham e Anderson(2003)** Kenneth P Burnham e David R Anderson. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media. Citado na pág. [6](#)
- Butler et al.(1993)** RW Butler, PL Davies e M Jhun. Asymptotics for the minimum covariance determinant estimator. *The Annals of Statistics*, páginas 1385–1400. Citado na pág. [26](#)
- Claeskens et al.(2008)** Gerda Claeskens, Nils Lid Hjort et al. Model selection and model averaging. *Cambridge Books*. Citado na pág. [6](#)
- Coleman et al.(1966)** James S Coleman et al. Equality of educational opportunity. Citado na pág. [45](#)
- Cook(1977)** R Dennis Cook. Detection of influential observation in linear regression. *Technometrics*, 19(1):15–18. Citado na pág. [23](#), [24](#)
- Cook e Weisberg(1982)** R Dennis Cook e Sanford Weisberg. *Residuals and influence in regression*. New York: Chapman and Hall. Citado na pág. [21](#), [24](#)

-
- Croux e Dehon(2003)** Cristophe Croux e Catherine Dehon. Estimators of the multiple correlation coefficient: Local robustness and confidence intervals. *Statistical Papers*, 44(3):315–334. Citado na pág. 3, 16
- Dixon(1950)** Wilfred J Dixon. Analysis of extreme values. *The Annals of Mathematical Statistics*, 21(4):488–506. Citado na pág. 28
- Draper e Smith(2014)** Norman R Draper e Harry Smith. *Applied regression analysis*, volume 326. John Wiley & Sons. Citado na pág. 21
- Drumond(1989)** M.F.B. Drumond. Métodos robustos em análise de regressão. Dissertação de Mestrado, USP. Citado na pág. 13
- Dupuis et al.(2013)** Debbie J Dupuis, Maria-Pia Victoria-Feser et al. Robust vif regression with application to variable selection in large data sets. *The Annals of Applied Statistics*, 7(1):319–341. Citado na pág. 18, 19
- Efroymson(1960)** MA Efroymson. Multiple regression analysis. *Mathematical methods for digital computers*, páginas 191–203. Citado na pág. 3
- Eid e Langeheine(1999)** Michael Eid e Rolf Langeheine. The measurement of consistency and occasion specificity with latent class models: A new model and its application to the measurement of affect. *Psychological Methods*, 4(1):100. Citado na pág. 6
- Filzmoser et al.(2005)** Peter Filzmoser, Robert G Garrett e Clemens Reimann. Multivariate outlier detection in exploration geochemistry. *Computers & geosciences*, 31(5):579–587. Citado na pág. 26
- Filzmoser et al.(2008)** Peter Filzmoser, Ricardo Maronna e Mark Werner. Outlier identification in high dimensions. *Computational Statistics & Data Analysis*, 52(3):1694–1711. Citado na pág. 27
- Fox et al.(2002)** John Fox et al. Robust regression. *An R and S-Plus companion to applied regression*, página 91. Citado na pág. 2
- Gauss(1816)** Carl Friedrich Gauss. Bestimmung der genauigkeit der beobachtungen. *Abhandlungen zur Methode der kleinsten Quadrate*, 1887. Citado na pág. 2
- Geisser(1975)** Seymour Geisser. The predictive sample reuse method with applications. *Journal of the American statistical Association*, 70(350):320–328. Citado na pág. 8
- Gilmour(1996)** Steven G Gilmour. The interpretation of mallows’s cp-statistic. *The Statistician*, páginas 49–56. Citado na pág. 6
- Hampel et al.(1986)** Frank R Hampel, Elvezio M Ronchetti, Peter Rousseeuw e Werner A Stahel. *Robust Statistics: the Approach based on Influence Functions*. Wiley-Interscience. Citado na pág. 15, 27
- Härdle(1984)** Wolfgang Härdle. How to determine the bandwidth of some nonlinear smoothers in practice. Em *Robust and Nonlinear Time Series Analysis*, páginas 163–184. Springer. Citado na pág. 16
- Heiberger e Becker(1992)** Richard M Heiberger e Richard A Becker. Design of an s function for robust regression using iteratively reweighted least squares. *Journal of Computational and Graphical Statistics*, 1(3):181–196. Citado na pág. 25
- Henri(1961)** Theil Henri. Economic forecasts and policy. *Holland, Amsterdam: North*. Citado na pág. 8
- Heritier et al.(2009)** Stephane Heritier, Eva Cantoni, Samuel Copt e Maria-Pia Victoria-Feser. *Robust methods in Biostatistics*, volume 825. John Wiley & Sons. Citado na pág. 19
- Hlavac, Marek(2018)** Hlavac, Marek. stargazer: Well-formatted regression and summary statistics tables. r package version 5.2.2., 2018. URL <https://CRAN.R-project.org/package=stargazer>. Citado na pág. 31
- Hoaglin e Welsch(1978)** David C Hoaglin e Roy E Welsch. The hat matrix in regression and anova. *The American Statistician*, 32(1):17–22. Citado na pág. 21, 22, 23
- Hoeting et al.(1996)** Jennifer Hoeting, Adrian E Raftery, David Madigan et al. A method for simultaneous variable selection and outlier identification in linear regression. *Computational Statistics and Data Analysis*, 22(3):251–270. Citado na pág. 28, 30, 44
- Huber(1981)** P. J. Huber. *Robust Statistics*. John Wiley & Sons. Citado na pág. 17, 29

-
- Huber et al.(1964)** Peter J Huber et al. Robust estimation of a location parameter. *The annals of mathematical statistics*, 35(1):73–101. Citado na pág. [2](#)
- Huber et al.(1973)** Peter J Huber et al. Robust regression: asymptotics, conjectures and monte carlo. *The Annals of Statistics*, 1(5):799–821. Citado na pág. [13](#)
- Hurvich e Tsai(1989)** Clifford M Hurvich e Chih-Ling Tsai. Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307. Citado na pág. [6](#)
- Khan et al.(2007a)** Jafar A Khan, Stefan Van Aelst e Ruben H Zamar. Building a robust linear model with forward selection and stepwise procedures. *Computational Statistics & Data Analysis*, 52(1):239–248. Citado na pág. [10](#), [17](#)
- Khan et al.(2007b)** Jafar A Khan, Stefan Van Aelst e Ruben H Zamar. Robust linear model selection based on least angle regression. *Journal of the American Statistical Association*, 102(480):1289–1299. Citado na pág. [18](#)
- Klein e Glahn(1974)** William H Klein e Harry R Glahn. Forecasting local weather by means of model output statistics. *Bulletin of the American Meteorological Society*, 55(10):1217–1227. Citado na pág. [48](#)
- Larson(1931)** Selmer C Larson. The shrinkage of the coefficient of multiple correlation. *Journal of Educational Psychology*, 22(1):45. Citado na pág. [3](#), [8](#)
- Lin et al.(2011)** Dongyu Lin, Dean P Foster e Lyle H Ungar. Vif regression: a fast regression algorithm for large data. *Journal of the American Statistical Association*, 106(493):232–247. Citado na pág. [3](#), [11](#), [12](#), [19](#)
- Machado(1993)** Jose AF Machado. Robust model selection and m-estimation. *Econometric Theory*, 9(3):478–493. Citado na pág. [3](#), [15](#)
- Mahalanobis(1936)** Prasanta Chandra Mahalanobis. On the generalized distance in statistics. Em *Proceedings of the National Institute of Science of India* 12, páginas 49–55. Citado na pág. [24](#), [26](#)
- Mallows(1973)** Colin L Mallows. Some comments on c p. *Technometrics*, 15(4):661–675. Citado na pág. [3](#), [5](#)
- Maronna et al.(2006)** Ricardo A Maronna, Douglas R Martin e Victor J Yohai. *Robust Statistics: Theory and Methods*. John Wiley and Sons. Citado na pág. [3](#), [16](#), [27](#)
- Montgomery et al.(2012)** Douglas C Montgomery, Elizabeth A Peck e G Geoffrey Vining. *Introduction to linear regression analysis*, volume 821. John Wiley & Sons. Citado na pág. [2](#), [22](#), [23](#)
- Pervin e Rahman(2018)** Mst Pervin e Md Rahman. Classical and robust forward selection: A simulation study and real data application. *Journal of Biometrics & Biostatistics*, 09. doi: 10.4172/2155-6180.1000404. Citado na pág. [17](#), [18](#)
- Rao et al.(1973)** Calyampudi Radhakrishna Rao, Calyampudi Radhakrishna Rao, Mathematischer Statistiker, Calyampudi Radhakrishna Rao e Calyampudi Radhakrishna Rao. *Linear statistical inference and its applications*, volume 2. Wiley New York. Citado na pág. [3](#), [7](#), [23](#)
- Renaud e Victoria-Feser(2010)** Olivier Renaud e Maria-Pia Victoria-Feser. A robust coefficient of determination for regression. *Journal of Statistical Planning and Inference*, 140(7):1852–1862. Citado na pág. [3](#), [16](#)
- Ronchetti(1982)** Elvezio Ronchetti. *Robust testing in linear models: the infinitesimal approach*. Tese de Doutorado, ETH Zurich. Citado na pág. [3](#), [14](#), [15](#)
- Ronchetti e Staudte(1994)** Elvezio Ronchetti e Robert G Staudte. A robust version of mallows's cp. *Journal of the American Statistical Association*, 89(426):550–559. Citado na pág. [3](#), [14](#), [15](#)
- Ronchetti et al.(1997)** Elvezio Ronchetti, Christopher Field e Wade Blanchard. Robust linear model selection by cross-validation. *Journal of the American Statistical Association*, 92(439):1017–1023. Citado na pág. [3](#), [16](#), [28](#), [31](#)
- Rousseeuw e Driessen(1999)** Peter J Rousseeuw e Katrien Van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223. Citado na pág. [26](#)
- Rousseeuw e Leroy(2005)** Peter J Rousseeuw e Annick M Leroy. *Robust regression and outlier detection*, volume 589. John wiley & sons. Citado na pág. [27](#), [46](#), [47](#), [48](#)
- Rousseeuw e Van Zomeren(1990)** Peter J Rousseeuw e Bert C Van Zomeren. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical association*, 85(411):633–639. Citado na pág. [22](#), [25](#), [26](#)
- Rousseeuw(1985)** PJ Rousseeuw. A regression diagnostic for multiple outliers and leverage points. *Abstracts in IMS Bull*, 14:399. Citado na pág. [26](#)

-
- Rubinstein e Kroese(2016)** Reuven Y Rubinstein e Dirk P Kroese. *Simulation and the Monte Carlo method*, volume 10. John Wiley & Sons. Citado na pág. [9](#)
- Schwarz et al.(1978)** Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464. Citado na pág. [3](#), [7](#)
- Shao(1993)** Jun Shao. Linear model selection by cross-validation. *Journal of the American statistical Association*, 88(422):486–494. Citado na pág. [9](#), [16](#)
- Sommer e Staudte(1995)** Suzanne Sommer e Robert G Staudte. Robust variable selection in regression in the presence of outliers and leverage points. *Australian Journal of Statistics*, 37(3):323–336. Citado na pág. [2](#), [3](#), [15](#), [25](#)
- Stone et al.(1984)** Charles J Stone et al. An asymptotically optimal window selection rule for kernel density estimates. *The Annals of Statistics*, 12(4):1285–1297. Citado na pág. [8](#)
- Stone(1977)** Mervyn Stone. An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *Journal of the Royal Statistical Society. Series B (Methodological)*, páginas 44–47. Citado na pág. [6](#), [15](#)
- Taplin e Raftery(1994)** Ross H Taplin e Adrian E Raftery. Analysis of agricultural field trials in the presence of outliers and fertility jumps. *Biometrics*, páginas 764–781. Citado na pág. [28](#), [30](#)
- Wright(1921)** Sewall Wright. Correlation and causation. *Journal of agricultural research*, 20(7):557–585. Citado na pág. [3](#)