Universidade de Brasília

Instituto de Ciências Exatas

Departamento de Estatística

Estatística Robusta Aplicada ao Mercado de Ações e ao Índice Bovespa

por

Thiago de Lima Macêdo

Brasília

2014

THIAGO DE LIMA MACÊDO - RA:12/0140756

PROJETO SOBRE ESTATÍSTICA ROBUSTA APLICADA AO MERCADO DE AÇÕES E AO ÍNDICE BOVESPA

Relatório Final do Projeto Final de Estágio supervisionado obrigatório apresentado à Universidade de Brasília como requisito para a obtenção do título de bacharel em Estatística.

Orientador: Prof. Dr. Raul Yukihiro Matsushita (EST/UnB)

Brasília
2014

Dedicatória

À minha esposa, Kelly Cristina, por todo o período que esteve ao meu lado me apoiando nos estudos. Ao meu filho, Matheus, por toda energia e motivação que ele gera e me faz querer crescer.

Agradecimentos

Agradeço a Deus pela força que ele me proporcionou para terminar essa etapa da minha vida e por ter trilhado o caminho para que eu chegasse até aqui.

À minha família, em especial, minha esposa Kelly Cristina e meu filho Matheus Macêdo. Meus pais Francisco e Cleide, meus irmãos Carlos e Mariana, minha cunhada Karina Keli e aos meus sobrinhos, enfim não tenho palavras pra agradecer. Cada um de vocês de alguma forma colaborou para que eu chegasse hoje aqui. Sou muito grato a todos que de alguma forma me ajudaram.

Ao Professor Raul Yukihiro Matsushita, pela orientação, paciência e disposição em ajudar a todo momento. Um grande exemplo de profissional.

Ao departamento de estatística pela oportunidade concedida de estudar em alto nível.

Ao professor Alan Ricardo pelas sugestões feitas a este trabalho e sua participação na banca examinadora.
Agradeço em especial aos amigos que fiz no curso, Agda, Alex, Bianca, Felipe, Guilherme e Raíssa.

Aos funcionários do departamento, pelo a disposição em ajudar e pela preciosa atenção.

Aos amigos da estatística, pela amizade e solidariedade, todos vocês que um dia deram conselhos, dicas legais e sugestões de questões de provas.

Por fim, obrigado a todos que de alguma forma contribuiu para que eu concluísse esse curso.

Resumo

Neste trabalho abordamos estatística robusta aplicada ao mercado de ações, para entender o processo de estimação robusta num conjunto de dados que apresentam outliers. Consideramos os dados referentes as ações da Ambev e as do Bradesco relacionadas com o índice IBOVESPA e utilizamos os métodos de estimação robustos feitos pelos modelos m-estimador de Huber, Tukey bisquare e Hampel. Por fim, comparamos esses métodos de estimações robustos com o método de estimação por mínimos quadrados ordinários.

Palavras-chave: Estatística robusta, m-estimadores, ponto de ruptura, função de influencia.

Abstract

In this work, approach robust statistical applied to the stock market to understand the robust estimation process in a data set that have outliers. We consider data concerning the shares of Ambev and Bradesco related with the Ibovespa index and use the robust estimation methods made by the models m-estimator Huber, m-estimator Tukey bisquare and m-estimator Hampel. Finally, we compare these methods of robust estimation with the method of estimation by ordinary least squares.

Keywords: Robust statistical, m-estimator, breakdown point, influence function.

Índice

Introdução

Esta relatório final versa sobre o estudo de estatísticas robustas em modelos de regressão linear. Um modelo é considerado robusto quando ele não é sensível a pequenas alterações de suas hipóteses.

Os métodos robustos sugiram na década de 60, com o objetivo de suavizar a influência de valores extremos (outliers) no processo de estimação de parâmetros. Devido a grandes dificuldades nos cálculos de estimação desses parâmetros por meio dos métodos robustos, esses métodos foram deixados de lado. Porém, em meados dos anos 80, os métodos robustos voltam a ter ênfase nos estudos estatísticos, pois os trabalhosos cálculos matemáticos passam a ser resolvidos com a ajuda dos computadores.

Contudo, os métodos de estimação robusta ainda não são oferecidos em cursos de graduação em estatística, talvez isso ocorra por causa das dificuldades inerentes da teoria para esse nível de estudo.

No Capítulo 1, é apresentado o objetivo geral e também os objetivos específicos deste relatório final.

No Capítulo 2 é mostrado o desenvolvimento da literatura estatística para modelos de regressão linear, no qual apresentamos os passos que levaram o surgimento do con-

ceito de estimação robusta.

No Capítulo 3 apresentamos conceitos preliminares para estimação de parâmetros para o modelo de regressão linear simples. Apresentamos o conceito de regressão robusta e as definições de robustez qualitativa, função de influência e ponto de ruptura.

No Capítulo 4 definimos o que é um *M-Estimador*. Dentre um conjunto de *M-Estimadores*, nos restringimos as definições dos *M-Estimadores* de Huber, Tukey *bisquare* e Hampel, que são resistentes a pequenas alterações das hipóteses gaussianas em um modelo de regressão linear simples . Em cada um desses métodos, apresentamos suas funções ρ, φ e w.

A aplicação da metodologia apresentada nesse relatório é vista no Capítulo 5 no qual estudamos o modelo de regressão linear robusto em ações da AMBEV S/A ON e do BRADESCO PN N1.

No Apêndice, apresentamos um teorema e alguns resultados que foram utilizados no decorrer deste relatório final, bem como a programação utilizada no software estatístico R.

1 Objetivos

Neste capítulo apresentamos o objetivo geral e os objetivos específicos desta proposta de projeto. Os objetivos, aqui apresentados, servirão de norte para o desenvolvimento do projeto final.

1.1 Objetivo Geral

Apresentar motivações para o uso de métodos robustos de estimação para modelos de regressão e mostrar exemplos que possibilitem uma melhor compreensão desses estimadores robustos.

Também deseja-se apresentar algumas relações que existem entre métodos robustos de estimação com o método de estimação por mínimos quadrados.

1.2 Objetivo Específicos

Comparar os modelos de regressão robusto com os de regressão linear clássico através de exemplos aplicados em conjuntos de dados que possam representar de maneira convenientemente aos interesses desse estudo. Observar quais são as implicações que os *outliers* produzem nesses estimadores robustos e nos estimadores de mínimos quadrados.

Queremos observar as implicações dos *outliers* em conjunto de dados relacionados a algumas ações da bolsa de valores.

Além do mais, buscamos entender a importância dos conceitos relativos a curva de influencia e robustez qualitativa e quantitativa para os estimadores apresentados neste trabalho.

2

Revisão da Literatura

Neste capítulo, descrevemos resumo histórico do surgimento dos estimadores de mínimos quadrados, na literatura estatística, para estimação de parâmetros num modelo de regressão linear. Descrevemos também os passos que levaram ao surgimento da estimação robusta para esses modelos lineares.

2.1 Literatura

Neste trabalho estudaremos estimadores robustos, porém, recorremos a história para narrar os acontecimentos que sucederam até o surgimento dos conceitos inerentes a estimação robusta.

Em 1809 Gauss [5] assumiu para o modelo de regressão linear, $Y_i = \sum_{i=1}^{n} B_j X_{ij} + \epsilon_i$, que os erros possuem uma distribuição normal, com média zero e variância σ^2 comum. Em seus estudos no ano de 1821, Gauss [6] mostrou que entre todos os estimadores

não-viesados de B_j, os obtidos pelo método de mínimos quadrados são os que possuem a menor variância. Em 1912, Markov apresentou uma versão mais trivial desse teorema de Gauss e seu resultado é hoje conhecido como teorema de Gauss-Markov.

Com o teorema de Gauss-Markov e com as ideias do teorema do Limite Central (a soma de pequenos erros se aproximam assintoticamente de uma distribuição normal) fez com que demais autores usasse o método de mínimos quadrados para a estimação dos parâmetros B_j em modelos de regressão linear.

Alguns anos mais tarde, em 1960, Tukey [16] chama a atenção ao expor que a presença de *outliers* ocorrem com uma frequência considerável em distribuições de dados e faz a seguinte pergunta: "O que acontece se a distribuição verdadeira desvia levemente da distribuição normal assumida?"E essa resposta hoje é conhecida, pois sabe-se que a presença de um único outlier é capaz de influenciar uma estimativa feita pela média aritmética.

Baseado no trabalho de Tukey [16], Huber [9] apresenta, em 1964, uma forma de estimação dos parâmetros $B_j$nos quais os outliers passam a exercer pouca influencia no modelo estatístico assumido. Esse método de estimação é conhecida como estimação robusta.

Este trabalho de Huber [9] pode ser considerado como o início da busca por estimadores robustos sob o ponto de vista formal, no qual Huber [9] apresenta uma classe de estimadores denotada por M-estimadores.

Hampel [7], apresenta em 1968, classificações de estimadores robustos de acordo com os aspectos:

1. qualitativo: uma pequena pertubação deve causar pequenos efeitos;

2. ruptura: qual grande deve ser uma pertubação para que o modelo entre em colapso;

3. função de influência: sensibilidade do estimador a pertubações muito pequenas.

No decorrer dos anos, vários autores estudaram os M-estimadores de Huber e também apresentarem outras classes de estimadores robustos. E com o auxílio da

tecnologia computacional, os estimadores robustos passaram a ser utilizados em resoluções de problemas nas diversas áreas do conhecimento. Contudo, ainda existe um longo caminho para a implementação de vários métodos de estimação robustos nos softwares estatísticos como SAS e R.

3

Preliminares

Neste capítulo, apresentamos as definições de modelo de regressão linear e de regressão linear simples. Depois, apresentamos o método de estimação por mínimos quadrados para o modelo de regressão linear simples. E encerramos esse capítulo com as definições de robustez qualitativa, função de influência e ponto de ruptura, que fazem parte do conceito de regressão robusta.

3.1 Modelo de Regressão Linear

Em uma análise de regressão, busca-se encontrar uma relação entre duas ou mais variáveis, no qual uma variável pode ser explicada em função de outra(s).

Definição 3.1. *Seja $\boldsymbol{X}$ um vetor de variáveis explicativas e Y uma variável resposta em função de $\boldsymbol{X}$. Então a relação entre essas variáveis é dada por:*

$$Y = f(\boldsymbol{X}).$$

3.1.1 Modelo de Regressão Linear Simples

Um modelo de regressão linear, onde existe apenas uma variável explicativa X, é denominado simples se pode ser escrito da seguinte forma:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{3.1}$$

em que Y_i é o valor da variável resposta no i-ésimo termo, β_0 e β_1 são os parâmetros do modelo, X_i é a variável explicativa conhecida (constante) do modelo no i-ésimo termo, ϵ_i é o erro aleatório gerado pelo modelo, no qual $E(\epsilon_i) = 0$ e $\sigma^2(\epsilon_i) = \sigma^2$ e a covariância entre ϵ_i e ϵ_j é zero para todo $i \neq j$ e $i = 1, \ldots, n$.

Nesse modelo, assumiremos que os erros ϵ_i são independentes entre si e possuem uma variância constante σ^2.

Em diversas ocasiões os parâmetros populacionais β_0 e β_1 são desconhecidos e impossíveis serem encontrados com exatidão devido a gradeza de certas populações. Contudo, podemos estimar esses parâmetros através de respectivas amostras da população em estudo. Assim, apresentamos a seguir o método de estimação por mínimos quadrados.

3.2 Método de Mínimos Quadrados

Seja X a variável explicativa associada a variável resposta Y, denotamos por (X, Y) essas observações em que (X_1, Y_1) representa a primeira delas no nosso conjunto de dados, (X_2, Y_2) a segunda e (X_i, Y_i) a i-ésima observação, onde $i = 1, \ldots, n$.

Assim, o método de estimação por mínimos quadrados (MMQ), em um modelo de regressão linear simples, consiste em encontrar estimadores b_0 e b_1 para β_0 e β_1, respectivamente, em um conjunto de observações de tamanho n, que minimizam a

seguinte função Q:

$$Q = \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 X_i\right)^2. \tag{3.2}$$

Note que:

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 X_i\right)$$

e

$$\frac{\partial Q}{\partial \beta_1} = -2 \sum_{i=1}^{n} X_i \left(Y_i - \beta_0 - \beta_1 X_i\right).$$

Portanto, ao igualar essas derivadas parciais a zero, e usando b_0 e b_1 para representar, respectivamente, um valor particular de β_0 e β_1 que minimiza Q, obtemos o seguinte sistema de equações:

$$\begin{cases} -2 \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 X_i\right) = 0, \\ -2 \sum_{i=1}^{n} X_i \left(Y_i - \beta_0 - \beta_1 X_i\right) = 0. \end{cases}$$

Resolvendo esse sistema, temos que:

$$b_0 = \overline{Y} - b_1 \overline{X}$$

e

$$b_1 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}.$$

Observe que estamos apresentando o MMQ e a seguir enunciamos o Teorema de Gauss-Markov que explica o motivo desse método ser tão utilizado em processos de estimação.

Teorema 3.1. *Sob as hipóteses do modelo de regressão linear simples apresentado na equação (3.1) e sem especificar uma distribuição de probabilidade para os erros. Temos que os estimadores de mínimos quadrados, b_0 e b_1, dos parâmetros β_0 e β_1, respectivamente, são não-viesados e possuem a menor variância entre todos os estimadores lineares não-viesados.*

3.3 Regressão Robusta

Para os modelos de regressão linear, Huber [9] buscou estimadores que sofrem pouca influência por valores extremos (*outliers*), hoje conhecidos por estimadores robustos.

Hampel [7] classifica os estimadores robustos de acordo com os aspectos:

- qualitativo: uma pequena pertubação deve ter pequenos efeitos;
- romptura: quão grande deve ser uma pertubação antes que o modelo entre em colapso;
- infinitesimal (função de influência): efeito que cada observação causa individualmente no estimador.

3.3.1 Robustez Qualitativa

A primeira definição matemática de robustez foi formulada por Hampel [7]. Ele considerava que uma pequena mudança na distribuição por de trás dos dados deve causar somente uma pequena alteração no desempenho de um procedimento estatístico.

Seu conceito de robustez de uma função estatística é baseado na continuidade em uma vizinhança de uma distribuição de probabilidade considerada.

Definição 3.2. *Dizemos que uma sequência de estatísticas* $\{T_n\}$ *é robusta qualitativamente para uma distribuição de probabilidade F, se dado* $\epsilon > 0$*, existe um* $\delta > 0$ *e um* $n_0 \in Z_n^*$*, tal que* $\forall\ Q \in F$ *e* $n \geq n_0$*,*

$$d_p\left(F, Q\right) < \delta \Rightarrow d_p\left[L_F\left(T_n\right), L_Q\left(T_n\right)\right] < \epsilon,$$

em que $L_F\left(T_n\right)$ *e* $L_Q\left(T_n\right)$ *denotam a distribuição de probabilidade de* $\left(T_n\right)$ *sob* F *e* Q*, respectivamente.*

Contudo, essa definição de robustez é somente qualitativa, isto é, ela apenas relata se um estimador é ou não robusto, porém não consegue medir o nível de robustez. Como queremos verificar se determinado estimador é mais robusto do que outro, então é necessário usar alguma medida quantitativa de robustez. No nosso caso, usamos os conceitos de função de influência e ponto de ruptura para medir a robustez de determinados estimadores.

3.3.2 Função de Influência

A função de influência é uma medida de robustez, definida por Hampel [8], que mede o efeito de pertubações infinitesimais no estimador.

Definição 3.3. *Seja $\widehat{\beta}$ um estimador de β baseado nos dados completos e $\widehat{\beta}_0$ um estimador baseado nos dados após a retirada dos outliers. Então, $\widehat{\beta} - \widehat{\beta}_0$ é denominada curva de sensibilidade de $\widehat{\beta}$. A função de influência é uma versão assintótica da curva de sensibilidade. Assim, para uma pequena fração ϵ de contaminação de outliers idênticos, a função de influência é dado por*

$$IF_{\widehat{\beta}}(x_0, F) = \lim_{\epsilon \to 0^+} \frac{\widehat{\beta}_\infty \left[(1-\epsilon)\, F + \epsilon \delta_{x0}\right] - \widehat{\beta}_\infty (F)}{\epsilon},$$

em que x_0 é o outlier, δ_{x0} é um ponto de massa próximo de x_0, e $\widehat{\beta}_\infty (F)$ é o valor assintótico do estimador de F.

A função de influência mostra-nos o quanto um outlier influencia na estimativa feita. Para o caso de estimadores robustos, queremos garantir que a função de influencia não vá para o infinito quando x seja grande.

3.3.3 Ponto de Ruptura

Ponto de ruptura foi introduzido por Donoho e Huber [3] e atualmente é uma característica quantitativa de robustez bastante conhecida.

O ponto de ruptura mede, em proporção, a maior quantidade de contaminações que os dados podem conter antes de o estimador falhar.

Definição 3.4. *Seja uma amostra aleatória $Z_n = \{(x_i, y_i) : i = 1, 2, \ldots, n\}$ e o estimador $\widehat{\beta} = T(Z_n)$. Então o ponto de ruptura do estimador T para a amostra Z_n é definido por:*

$$\epsilon^* (T, Z_n) = \frac{m^* (Z_n)}{n},$$

em que $m^* (Z_n)$ é o menor inteiro m, para qual

$$\sup_{Z_m} \|T(Z_m) - T(Z_n)\| = \infty,$$

ou seja, a menor parte das observações que, substituídos por valores arbitrários, pode levar T para o infinito.

A fim de uma melhor compreensão didática, faremos uma exemplificação para a média aritmética.

Para média, temos que $T(Z_n) = \overline{Z}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i$. Note que $\epsilon^*\left(\overline{Z}_n, Z_n\right) = \frac{1}{n}$ e consequentemente $\lim_{n\to\infty} \epsilon^*\left(\overline{Z}_n, Z_n\right) = 0$, para qualquer amostra inicial de Z_n.

4

Metodologia

Neste capítulo, apresentamos a definição de um *M-Estimador*. E dentre um conjunto de *M-Estimadores*, nos restringimos aos *M-Estimadores* de Huber, Tukey *bisquare* e Hampel, que são resistentes a pequenas alterações das hipóteses gaussianas em um modelo de regressão linear simples . Em cada um desses métodos, apresentamos suas funções ρ, φ e w.

Aqui, também definimos o conceito de função *redescending* e explicamos qual sua importância no processo de estimação num modelo de regressão linear robusto com *outliers* nas caudas.

4.1 M-Estimadores

Definição 4.1. *Seja X1, ..., Xn amostra aleatória com distribuição comum F e seja o estimador* $T = Tn(X1, ..., Xn)$. *T é chamado de M-estimador quando:*

$$\sum_{i=1}^{n} \rho(x_i - T) := \min, \tag{4.1}$$

em que ρ *é uma função não constante.*

Diferenciando a expressão em (4.1) em relação a T, temos que

$$\sum_{i=1}^{n} \varphi\left(x_i - T\right) = 0. \tag{4.2}$$

Podemos escrever a equação (4.2) como

$$\sum_{i=1}^{n} w_i \left(x_i - T\right) = 0, \tag{4.3}$$

com

$$w_i = \frac{\varphi\left(x_i - T\right)}{x_i - T}, \tag{4.4}$$

note que temos uma representação formal de T como uma média ponderada, ou seja,

$$T = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}, \tag{4.5}$$

com o peso dependendo somente da amostra.

Pela definição (4.1), temos que:

- Se $\rho(t) = t^2$, então $T = \frac{\sum_{i=1}^{n} x_i}{n}$;
- Se $\rho(t) = |t|$, então $T = med(X1, ..., X2)$;
- Se $\rho(t) = -\log f(t)$, onde f é função densidade de F_T, então T é estimador de máxima verossimilhança.

Dentre os estimadores pertencentes a classe do M-estimadores, apresentamos a seguir M-estimadores que são resistentes a pequenas alterações (violações) das hipóteses gaussianas em um modelo de regressão linear simples. Estes estimadores, resistentes as essas violações, são chamados de robustos.

4.2 M-Estimador de Huber

Um escolha bastante popular para ρ é a função de Huber proposta por ele em 1994. Para esta função, temos

$$\rho\left(x_i - T\right) = \begin{cases} \frac{1}{2}\left(x_i - T\right)^2, & \text{para} \quad |x_i - T| \leq k, \\ k\left|x_i - T\right| - \frac{1}{2}k^2, & \text{para} \quad |x_i - T| > k. \end{cases} \tag{4.6}$$

E sua função φ é

$$\varphi\left(x_i - T\right) = \begin{cases} \left(x_i - T\right), & \text{se} \quad |x_i - T| \leq k, \\ k\,\text{sign}\left(x_i - T\right), & \text{se} \quad |x_i - T| > k, \end{cases} \tag{4.7}$$

em que $\text{sign}\left(x\right)$ retorna os valores -1, 0 e 1 para x negativo, zero e positivo, respectivamente.

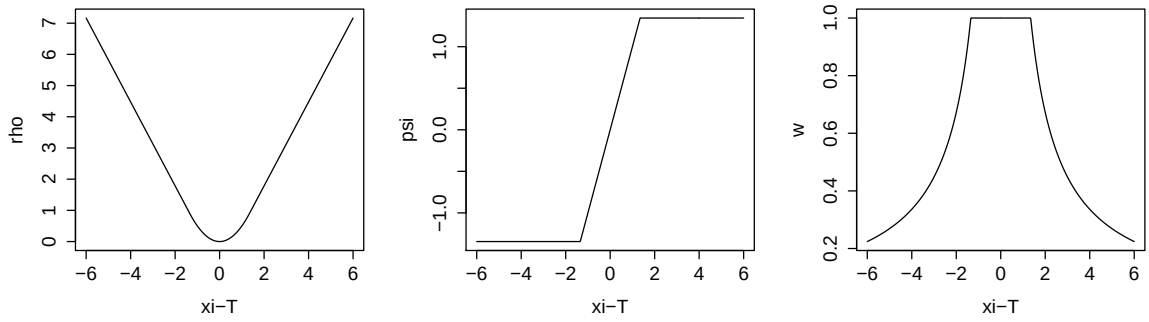
A sua função peso correspondente é

$$w_i\left(x_i - T\right) = \begin{cases} 1, & \text{para} \quad |x_i - T| \leq k, \\ \frac{k}{|x_i - T|}, & \text{para} \quad |x_i - T| > 0, \end{cases} \tag{4.8}$$

no qual k é chamado de constante de ajustamento. Pequenos valores de k produzem mais resistência a outliers, contudo trás consigo uma perda de eficiência sobre a distribuição normal. Temos para $k = 1,345$ uma eficiência de 95% no modelo Gaussiano e também proteção contra alguns *outliers* que possam estar presentes na amostra.

Observemos os gráficos das funções citadas.

As estimativas de Huber são robustas quando os valores atípicos têm baixa alavancagem, ou seja, não são discrepantes na direção do eixo x. Para obter estimativas que são robustos contra qualquer tipo de *outliers*, a função *bisquare* proposto por Tukey, ou a proposta por Hampel, pode ser preferível, elas são chamadas de funções *redescending*.



4.3 M-Estimador de Tukey bisquare

As funções *redescending*, pertencentes aos M-estimadores, são as φ que são não decrescentes próximas da origem, mas tendem 0 quando tende ao infinito. Isso implica que para x grande, a sua respectiva função ρ cresce mas lentamente que a ρ (4.6) de Huber.

Uma escolha popular para ρ e φ, que é uma função *redescending*, são as funções de Tukey *bisquare*. Os cálculos para essas funções são similares aos feitos anteriormente para o M-Estimador de Huber. Esse estimador é especialmente resistente a observações nos extremos das caudas, pois ρ é limitada e para esta função, temos que

$$\rho_B\left(x_i - T\right) = \begin{cases} \frac{k^2}{6}\left\{1 - \left[1 - \left(\frac{x_i - T}{k}\right)^2\right]^3\right\}, & \text{para} \quad |x_i - T| \leq k, \\ \frac{k^2}{6}, & \text{para} \quad |x_i - T| > k. \end{cases} \tag{4.9}$$

E sua função φ é

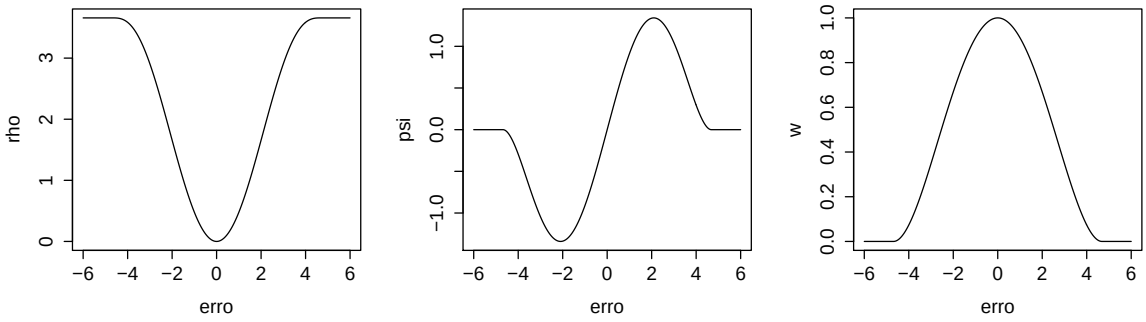
$$\varphi_B\left(x_i - T\right) = \begin{cases} \left(x_i - T\right)\left[1 - \left(\frac{x_i - T}{k}\right)^2\right]^2, & \text{se} \quad |x_i - T| \leq k, \\ 0, & \text{se} \quad |x_i - T| > k. \end{cases} \tag{4.10}$$

A sua função peso correspondente é

$$w_{Bi}\left(x_i - T\right) = \begin{cases} \left[1 - \left(\frac{x_i - T}{k}\right)^2\right]^2, & \text{para} \quad |x_i - T| \leq k, \\ 0, & \text{para} \quad |x_i - T| > 0, \end{cases} \tag{4.11}$$

e para $k = 4,685$, temos uma eficiência de 95% no modelo Gaussiano e ponto de ruptura de $0,5$.

Observemos os gráficos das funções citadas.



4.4 M-Estimador de Hampel

Hampel [8] definiu uma função *redescending* que protege o ajuste de maneira mais intensa contra observações que estão bem distantes do conjunto de dados. Para sua função, temos que

$$\rho_{Hp}(x_i - T) = \begin{cases} \frac{1}{2}(x_i - T)^2 & , \text{ para } |x_i - T| \leq a, \\ a|x_i - T| - \frac{1}{2}a^2 & , \text{ para } a < |x_i - T| \leq b, \\ \dfrac{a\left[c|x_i - T| - \frac{1}{2}(x_i - T)^2\right]}{c - b} - \dfrac{7}{6}a^2 & , \text{ para } b < |x_i - T| \leq c, \\ a(b + c - a) & , \text{ para } |x_i - T| > c. \end{cases} \tag{4.12}$$

E sua função φ é

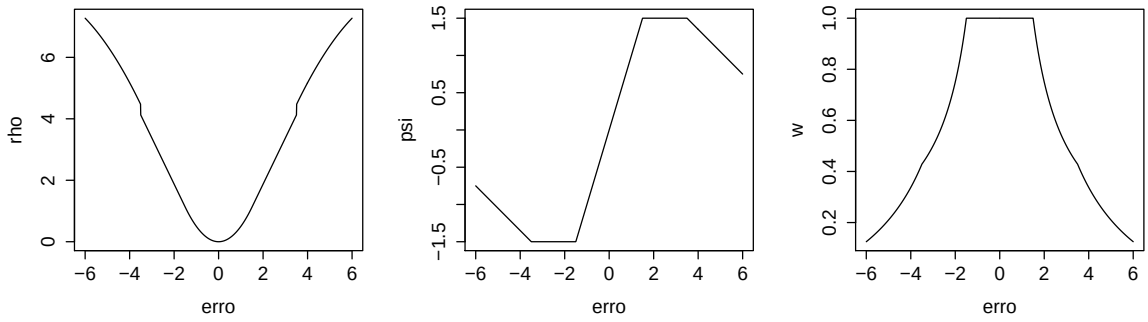
$$\varphi_{Hp}(x_i - T) = \begin{cases} (x_i - T) & , \text{ para } |x_i - T| \leq a, \\ a\,\text{sign}(x_i - T) & , \text{ para } a < |x_i - T| \leq b, \\ \dfrac{a\,\text{sign}(x_i - T)(c - |x_i - T|)}{c - b} & , \text{ para } b < |x_i - T| \leq c, \\ 0 & , \text{ para } |x_i - T| > c. \end{cases} \tag{4.13}$$

E finalmente, a sua função peso correspondente é

$$w_{Hp}\left(x_i - T\right) = \begin{cases} 1 & , \text{ para } |x_i - T| \leq a, \\ \dfrac{a}{|x_i - T|} & , \text{ para } a < |x_i - T| \leq b, \\ \dfrac{a\left(c - |x_i - T|\right)}{|x_i - T|\left(c - b\right)}, & \text{para } b < |x_i - T| \leq c, \\ 0 & , \text{ para } |x_i - T| > c, \end{cases} \tag{4.14}$$

e afim de obter uma melhor eficiência para esse estimador, os valores indicados para as constantes são a=1,5, b=3,5 e c=8,5. Assim, temos 95% de eficiência sobre o modelo Gaussiano e ponto de ruptura de $0,5$.

Observemos os gráficos das funções citadas.



5

Resultados

Neste Capítulo, apresentamos gráficos de ações da AMBEV S/A ON e do BRADESCO PN N1 que são negociadas na Bolsa de Valores Bovespa e relacionamos a variação dos preços dessas ações de um dia para o outro, referentes aos preços indicados no momento do fechamento da bolsa, com a variação do índice IBOVESPA, também de um dia para o outro e referente ao seu valor no horário de fechamento.

O índice IBOVESPA é um indicador do desempenho médio das cotações das ações de maior negociabilidade e representatividade na Bolsa de Valores de São Paulo e o seu resultado é referente a uma carteira teórica de ativos.

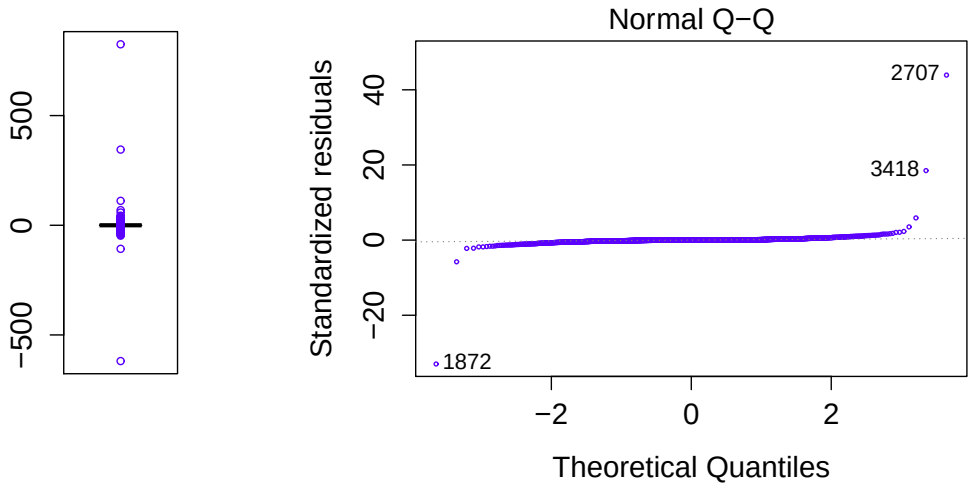
Os resultados obtidos nesse trabalho foram gerados com o auxílio do software estatístico R versão 3.1.3. Usamos os pacotes *MASS* e *robust*.

5.1 Gráficos

Os gráficos a seguir, nos auxiliam a termos uma melhor compreensão sobre a importância dos *"outliers"*, num determinado conjunto de dados, durante o processo de estimação dos parâmetros para um modelo linear. Em cada gráfico, apresentamos as retas geradas pelos processos de estimação vistos nos capítulos 3 e 4, aplicados em valores de ações pertencentes a BOVESPA.

5.1.1 AMBEV S/A ON

A série histórica da ação da Ambev com o índice Ibovespa é referente ao período de 05 de Janeiro de 2000 até 27 de Março de 2015.



(a) Boxplot (b) Gráfico de Normalidade

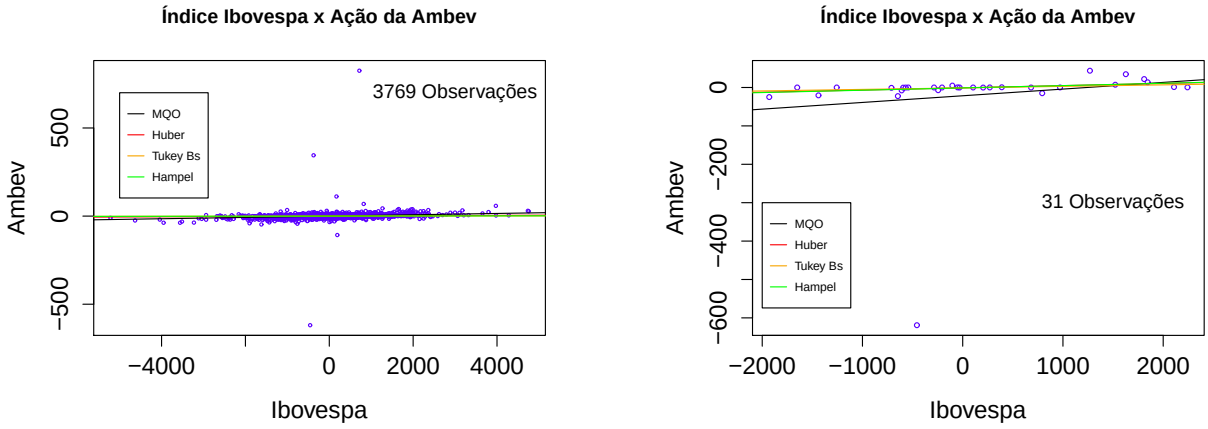
Figura 5.1: Gráficos de Distribuição dos Dados

Primeiramente, observa-se a existência de *outliers* no gráfico 5.1(a) e no gráfico 5.1(b) que a distribuição dos resíduos se afastam da normal. Portanto, devido a esses fatores, as hipóteses de normalidades não são todas satisfeitas e assim não podemos aplicar o teorema de Gaus-Markov para estimar os parâmetros do modelo através do processo MMQ. Faremos algumas comparações entre o método de estimação por MMQ

e alguns métodos robustos.

Assim, iniciamos as comparações a partir de dois gráficos, o primeiro 5.2(a) é referente a toda a série histórica contendo 3769 observações, o segundo 5.2(b) é refente aos valores em torno do *outlier* que ocorreu no dia 02-08-2007.

O motivo deste outlier foi devido ao fato que em 29 de junho de 2007, em Assembleia Geral Extraordinária, foi aprovado grupamento das ações em que se divide o capital social da Companhia, na proporção de 100 ações, então existentes, para 1 ação do capital após o grupamento, sem modificação do montante do capital social. A partir de 2 de agosto de 2007, as ações da Companhia, já grupadas, passaram a ser negociadas em cotação unitária e não mais em lote de 1.000 ações.



(a) Série Histórica até o Dia 27/03/2015

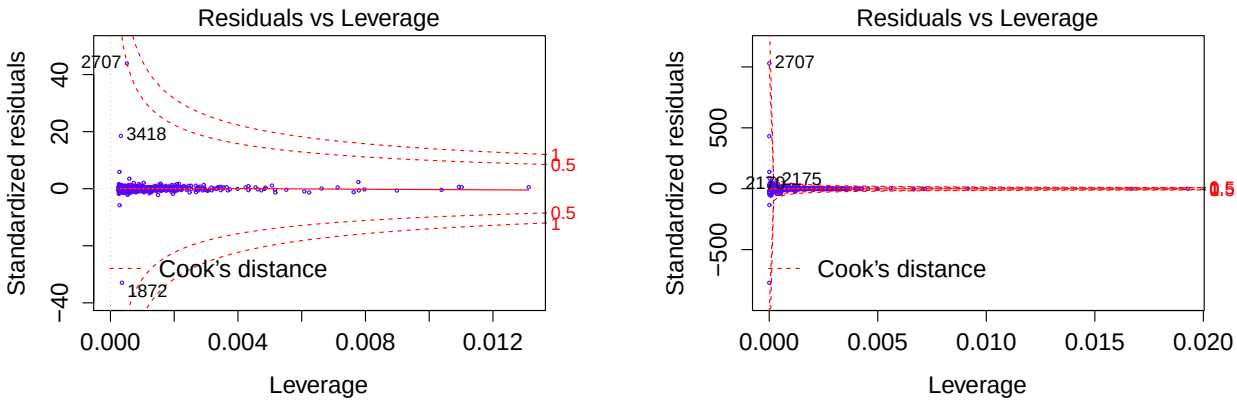
(b) Amostra em Torno do Outlier

Figura 5.2: Gráfico de Regressão

Note que este *outlier* é um caso isolado, mas possui um significado importante dentro do contexto. Assim, não podemos desconsiderá-lo durante o processo de estimação e portanto nosso processo deve ser robusto, pois o peso dado a ele deve ser pequeno.

Quando consideramos toda a série histórica, os métodos acima são bem semelhantes. Perceba que, mesmo o método de estimação por mínimos quadrados não sendo robusto, ele se aproximou bastante dos demais métodos robustos aqui apresentados. Isso ocorre porque o conjunto de dados possui 3769 observações e entre eles uma quantidade muito pequena de *outliers*, que não conseguem influenciar significativamente a estimação no modelo por MMQ e nem o de Huber, conforme pode ser observado nos gráficos 5.3(a)

e 5.3(b) da distância de Cook.



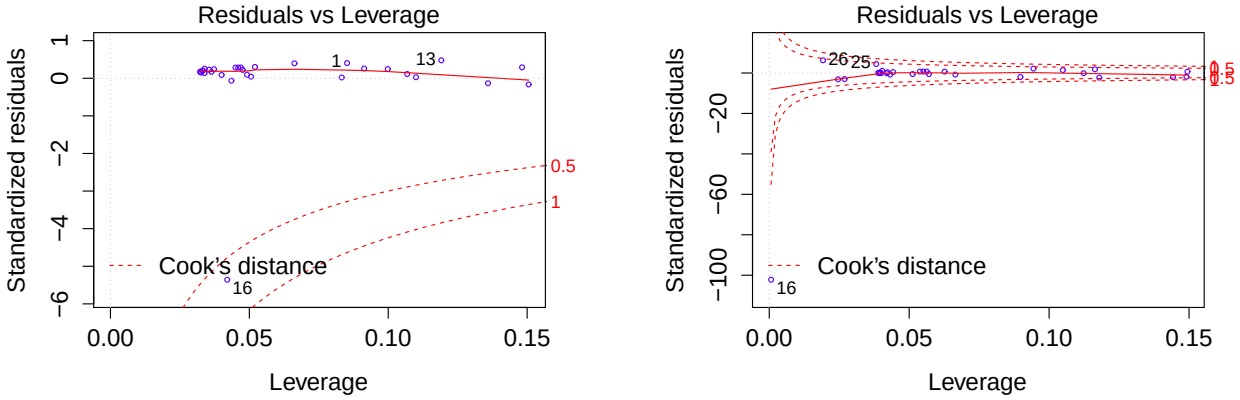
(a) Modelo de Mínimos Quadrados Ordinários

(b) Modelo Robusto de Huber

Figura 5.3: Gráfico da Distância de Cook para as 3769 Observações

Porém, quando selecionamos uma amostra contendo 31 observações, no qual o *outlier* de 02-08-2007 também faz parte dessas observações e as demais estão ao seu redor, percebemos que ele sozinho foi capaz de interferir no processo de estimação por MMQ, conforme mostra o gráfico 5.4(a) da distância de Cook, mas os demais métodos robustos suavizam a interferência desse *outlier* no processo de estimação dos parâmetros e para o caso do modelo de Huber é possível visualizar essa baixa interferência no gráfico 5.4(b). Também podemos notar no gráfico 5.2(b) a diferença entre o método não robusto por MMQ e os métodos robustos de Huber, Tukey bisquare e Hampel.

A diferença entre as estimativas para os dados da Ambev é apresentado na tabela (5.1.1). Note que para a situação no qual temos 31 observações e existe 1 (um) *outlier* entre elas, a estimativa b_0 feita por MMQ é aproximadamente 21 vezes maior do que as demais estimativas b_0's feitas pelos demais métodos e que o valor de b_1 produzido por MMQ é aproximadamente 3 vezes maior do que as demais estimativas b_1's produzidas pelos demais métodos.



(a) Modelo de Mínimos Quadrados Ordinários (b) Modelo Robusto de Huber

Figura 5.4: Gráfico da Distância de Cook para as 31 Observações

Tabela 5.1: Estimativas Para a Reta de Regressão Linear

Estimativas dos Modelos de Regressão Linear dos Dados da Ambev								
	3769 Observações				31 Observações			
	MQO	Huber	TukeyBs	Hampel	MQO	Huber	TukeyBs	Hampel
b_0	0.0266	-0.0399	-0.0133	-0.0191	-21.5443	-0.9154	-1.0397	-1.0886
b_1	0.0037	0.0010	0.0002	0.0003	0.0174	0.0057	0.0040	0.0060

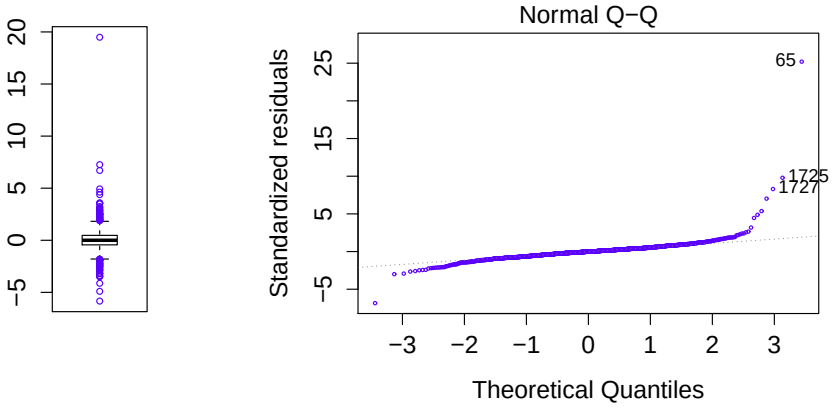
5.1.2 BRADESCO PN N1

A série histórica das ações do Bradesco relacionadas com o índice Ibovespa é referente ao período de 02 de Janeiro de 2008 até 27 de Março de 2015.

Primeiramente, observa-se a existência de *outliers* no gráfico 5.5(a) e no gráfico 5.5(b) e que a distribuição dos resíduos se afasta da normal. Assim, fazemos uma comparação entre o método de estimação por MMQ e métodos robustos com o objetivo de averiguar a existência de diferenças nos resultados.

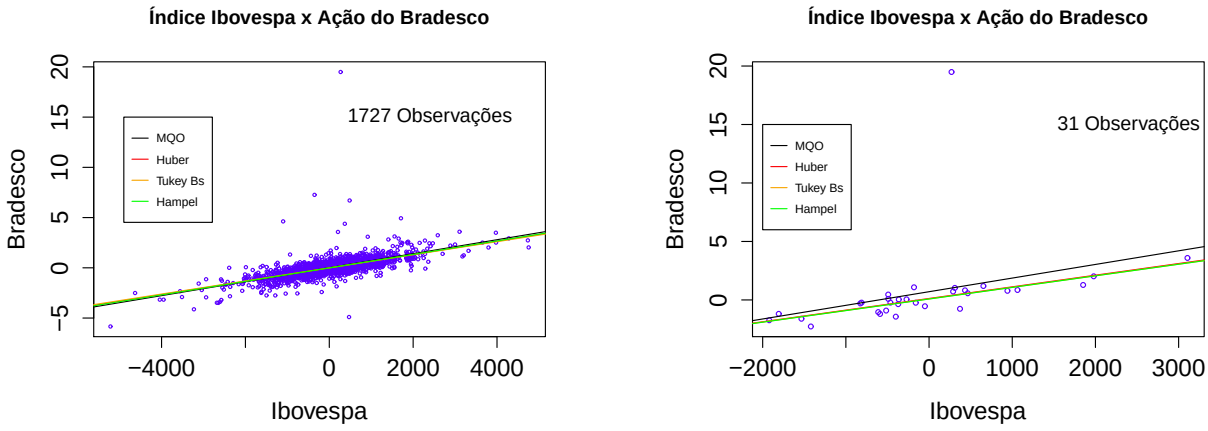
Iniciamos as comparações com dois gráficos, o primeiro 5.6(a) é referente a toda a série histórica contendo 1727 observações, o segundo 5.6(b) é refente aos valores em torno do *outlier* que ocorreu no dia 07-04-2008. Esse *outlier* ocorreu pelo fato de que em 07-04-2008, para celebrar o centenário da imigração japonesa no Brasil, o Bradesco

preparou uma série de ações de comunicação que ressaltam os vínculos estabelecidos entre os japoneses, o país e a histórica parceria entre o banco e a comunidade nipo-brasileira, solidificada no decorrer dos anos.



(a) Boxplot (b) Gráfico de Normalidade

Figura 5.5: Gráficos de Distribuição dos Dados

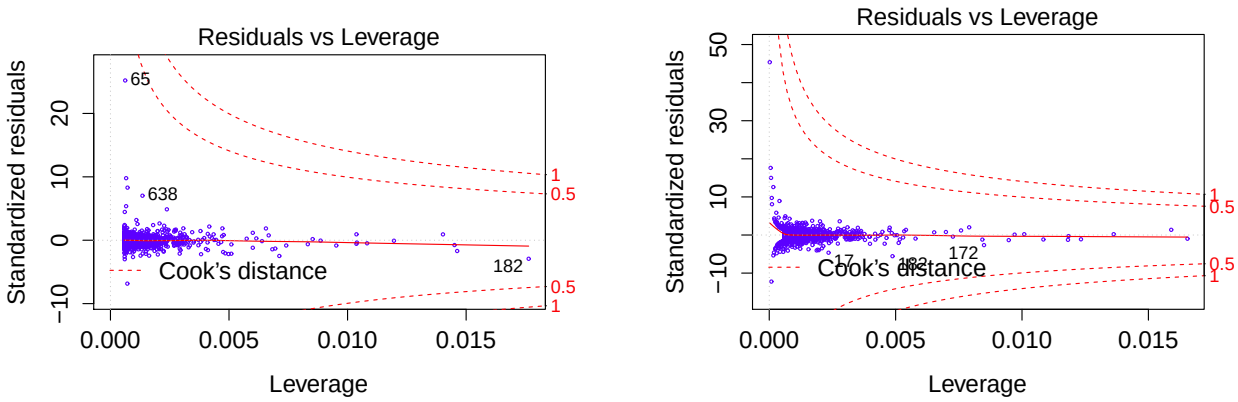


(a) Série Histórica até o Dia 07-04-2008 (b) Amostra em Torno do Outlier

Figura 5.6: Gráfico de Regressão

Sabemos que esse *outlier* possui um significado importante dentro do contexto e também não podemos desconsiderá-lo durante o processo de estimação e portanto nosso processo deve ser robusto, pois não podemos dar tanto peso a ele, principalmente por se tratar de um caso isolado.

Quando consideramos toda a série histórica, os métodos são bem semelhantes. Nota-se que, mesmo o método de estimação por mínimos quadrados não sendo robusto, suas estimativas são próximas das outras geradas pelos métodos robustos aqui apresentados. Podemos observar no gráfico 5.7 da distância de Cook, que os *outliers* possuem pouco influência na estimação por MQO e também pouca influência no processo de estimação robusto de Huber.



(a) Modelo de Mínimos Quadrados Ordinários (b) Modelo Robusto de Huber

Figura 5.7: Gráfico da Distância de Cook para as 1727 Observações

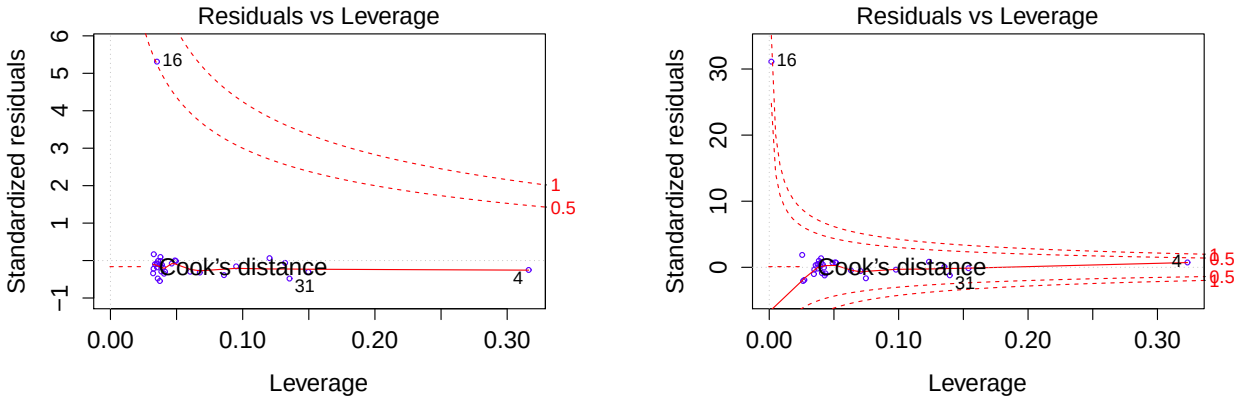
Essa justificativa se é similar à explicada anteriormente na seção 5.1.1.

Quando selecionamos uma amostra contento 31 observações, no qual o *outlier* de 07-04-2008 também faz parte dessas observações e as demais estão ao seu redor, percebemos que apenas este *outlier* foi capaz de interferir no processo de estimação por mínimos quadrados, como pode-se observar no gráfico 5.8, mas os demais métodos robustos suavizam a sua interferência no processo de estimação dos parâmetros. Também pode-se notar no gráfico 5.6(b) a diferença entre o método não robusto de estimação MMQ e os métodos robustos de Huber, Tukey bisquare e Hampel.

As diferenças entre as estimativas para o dados do Bradesco é apresentado na tabela 5.1.2.

Podemos observar na tabela 5.1.2, no caso para 31 observações, que o valor da estimativa b_0 estimada por MQO é aproximadamente 7 vezes maior do que os demais b_0, estimados por Huber, Tukey bisquare e Hampel, e que a estimativa de b_1, no caso

MMQ, é 0.0002 maior que as outras estimativas b_1, que possuem o mesmo valor.



(a) Modelo de Mínimos Quadrados Ordinários (b) Modelo Robusto de Huber

Figura 5.8: Gráfico da Distância de Cook para as 31 Observações

Tabela 5.2: Estimativas Para a Reta de Regressão Linear

Estimativas dos Modelos de Regressão Linear dos Dados do Bradesco								
	3769 Observações				31 Observações			
	MQO	Huber	TukeyBs	Hampel	MQO	Huber	TukeyBs	Hampel
b_0	0.0129	-0.0069	-0.0066	-0.0103	0.7036	0.1254	0.0899	0.0773
b_1	0.0007	0.0007	0.0007	0.0007	0.0012	0.0010	0.0010	0.0010

5.1.3 Considerações Finais

Observa-se tanto nos dados referentes as ações AMBEV S/A ON, quanto nos referentes as ações BRADESCO PN N1, que as análises de regressão feitas num conjunto restrito de dados, vizinhos a um *outlier*, produz resultados diferentes para as estimativas (b_0 e b_1) feitas pela regressão feita por MMQ e as outras feitas pelos métodos robustos propostos por Huber, Tukey e Hampel.

Assim, percebemos a importância de considerar os métodos de regressão linear robustos em conjuntos de dados que possuem *outliers*. Como os métodos robustos não são tão influenciados por uma determinada quantidade de valores atípicos, isso faz com

que os investimentos baseados nas retas de regressão linear robusta sejam considerados conservadores, pois os investidores não são influenciados a realizar compra ou venda de ações quando ocorrem valores atípicos (extremos) no seu preço.

Contudo, por que simplesmente não excluímos os *outliers*? Porque neste caso de ações do IBOVESPA, temos um enorme banco de dados que é atualizado diariamente e precisamos de um método de estimação automatizado e que saiba lidar de maneira eficiente com a presença desses valores atípicos, pois fica inviável analisar, caso a caso, cada *outliers* que surge no banco de dados.

A realização desse trabalho completou uma lacuna existente sobre os conhecimentos referentes à estimação robusta, pois esse conteúdo não foi ofertado durante o curso de graduação por ser considerado complexo. Geralmente a parte de estimação robusta é estudada em cursos de pós-graduação, mas acredito ser possível ofertar uma disciplina introdutória desse conteúdo a nível de graduação.

Durante o estudo, nota-se que diversos métodos de análise robusta ainda não foram implantados no software R e portanto, para próximos trabalhos, pode-se tentar implementar nesse software métodos robustos que ainda não fazem parte de seus pacotes.

A

Programação no Software R

```
################################################
Dados do Yahoo Finanças - Banco de Dados Ambev
################################################

ambv <- read.csv("c:/Pasta/ABEV3SA.csv",sep=",",dec=".",header=TRUE)
ibo <- read.csv("c:/Pasta/ibovespa.csv",sep=",",dec=".",header=TRUE)
ibo <- ibo[1:3903,]
################################################
Colocar a Variável Data no Formato de Data
################################################
Sys.setlocale('LC_TIME','English')
ambv$Date=as.Date(ambv$Date,'%Y-%m-%d')
```

```
ibo$Date=as.Date(ibo$Date,'%Y-%m-%d')

################################################
Diferenca dos Preços na Variável Close
################################################
#Ambev

d1 <- ambv$Close
d2 <- d1[2:3903]
dif <- d2-d1[1:3902]
difClose <- c(dif,NA)
difClose
ambvA <- cbind(ambv,difClose)

#Ibovespa

c1 <- ibo$Close
c2 <- c(c1[2:3903],0)
dif <- c2[1:3902]-c1[1:3902]
difClose <- c(dif,NA)
difClose
iboA <- cbind(ibo,difClose)


################################################
Grafico de Série Temporal Para os Dados
Considerar a Variável "close"
################################################

plot(ambvA$difClose~ambvA$Date, cex=0.01, col="blue", type="l")
```

```
plot(iboA$difClose~iboA$Date, cex=0.01, col="blue", type="l")

################################################
Juntar Banco de Dados
################################################

Jamb<-merge(iboA, ambvA, by="Date")

################################################
Gráfico de Ibo x Ambv e Boxplot
################################################

plot(Jamb$difClose.y~Jamb$difClose.x, cex=0.4, col="blue",
     main=list("Índice Ibovespa x Ação da Ambev", cex=0.8),
     xlab="Ibovespa", ylab="Ambev")
text(3000,700,"3769 Observações",cex=0.8)
legend(-5000,700,c("MQO        ","Huber        ","Tukey Bs        ",
"Hampel        "), col=c(1,"red","orange","green"),
lty=c(1,1,1,1),cex=0.5)

boxplot(Jamb$difClose.y,cex=0.7,outcol="blue")

################################################
Modelos
################################################

require(MASS)
require(robust)

model <-lm(Jamb$difClose.y~Jamb$difClose.x)
```

```
model2<-rlm(Jamb$difClose.y~Jamb$difClose.x) ##Huber
model3<-rlm(Jamb$difClose.y~Jamb$difClose.x,psi=psi.bisquare) ##Tukey
model4<-rlm(Jamb$difClose.y~Jamb$difClose.x,psi=psi.hampel)

abline(model)
abline(model2, col="red")
abline(model3, col="orange")
abline(model4, col="green")
coef(model)
coef(model2)
coef(model3)
coef(model4)

################################################
Modelos Com Dados Próximos ao Outlier
################################################

MambvA <- ambvA[1962:1992,]
plot(MambvA$difClose~MambvA$Date, cex=0.01, col="blue")

Mamb <-merge(iboA, MambvA, by="Date")
plot(Mamb$difClose.y~Mamb$difClose.x, cex=0.6, col="blue",
     main=list("Índice Ibovespa x Ação da Ambev", cex=0.8),
     xlab="Ibovespa", ylab="Ambev")
text(1500,-300,"31 Observações",cex=0.8)
legend(-2000,-300,c("MQO        ","Huber        ","Tukey Bs        ",
"Hampel        "),col=c(1,"red","orange","green"),
lty=c(1,1,1,1),cex=0.5)

modl <-lm(Mamb$difClose.y~Mamb$difClose.x) #Mínimos Qudrados#
```

```
modl2<-rlm(Mamb$difClose.y~Mamb$difClose.x) #Huber#
modl3<-rlm(Mamb$difClose.y~Mamb$difClose.x,psi=psi.bisquare) #Tukey#
modl4<-rlm(Mamb$difClose.y~Mamb$difClose.x,psi=psi.hampel)

abline(modl)
abline(modl2, col="red")
abline(modl3, col="orange")
abline(modl4, col="green")
coef(modl)
coef(modl2)
coef(modl3)
coef(modl6)

MQO=c(coef(model))
Huber=c(coef(model2))
TukeyBs=c(coef(model3))
Hampel=c(coef(model4))
E=data.frame(MQO,Huber,TukeyBs,Hampel)
xtable(E, digits=4)

MQO1=c(coef(modl))
Huber2=c(coef(modl2))
TukeyBs3=c(coef(modl3))
Hampel4=c(coef(modl4))
F=data.frame(MQO1,Huber2,TukeyBs3,Hampel4)
xtable(F, digits=4)
G=data.frame(MQO,Huber,TukeyBs,Hampel,MQO1,Huber2,TukeyBs3,Hampel4)
xtable(G, digits=4)

plot(model,sub="",col="blue",cex=0.4)
```

```
plot(model2,sub="",col="blue",cex=0.4)

plot(modl,sub="",col="blue",cex=0.6)
plot(modl2,sub="",col="blue",cex=0.6)

##############################################
Funçoes $\rho$, $\varphi$ e w
Huber, Tukey bisquare e Hampel
##############################################
par(mfrow=c(1,3))
erro=seq(-6,6,0.01)

#####Huber#####################################
k=1.345
rho=(abs(erro)<=k)*erro^2/2+(abs(erro)>k)*(k*abs(erro)-k^2/2)
plot(erro,rho,type="l",xlab = "xi-T")
psi=(abs(erro)<=k)*erro+(abs(erro)>k)*sign(erro)*k
plot(erro,psi,type="l",xlab = "xi-T")
w=(abs(erro)<=k)+(abs(erro)>k)*k/abs(erro)
plot(erro,w,type="l",xlab = "xi-T")

#####Tukey#####################################
kt=4.685
a=1-(erro/kt)^2
b=1-a^3
c=kt^2/6
rho=(abs(erro)<=kt)*b*c+(abs(erro)>kt)*kt^2/6
plot(erro,rho,type="l")
psi=(abs(erro)<=kt)*erro*a^2+(abs(erro)>kt)*0
w=(abs(erro)<=kt)*a^2+(abs(erro)>kt)*0
```

```
plot(erro,psi,type="l")
plot(erro,w,type="l")

#####Hampel#####################################
a=1.5
b=3.5
c=8.5
d=a*(c*abs(erro)-(erro^2)/2)/(c-b)
e=(7*a^2)/6
rho=(abs(erro)<=a)*(erro^2)/2+(abs(erro)>a)*(abs(erro)<=b)*(a*abs(erro)-
(a^2)/2)+(abs(erro)>b)*(abs(erro)<=c)*(d-e)+(abs(erro)>c)*a*(b+c-a)
plot(erro,rho,type="l")
psi=(abs(erro)<=a)*erro+(abs(erro)>a)*(abs(erro)<=b)*a*sign(erro)+
(abs(erro)>b)*(abs(erro)<=c)*a*sign(erro)*
(c-abs(erro))/(c-b)+(abs(erro)>c)*0
w=(abs(erro)<=a)*1+(abs(erro)>a)*(abs(erro)<=b)*a/abs(erro)+(abs(erro)>b)*
(abs(erro)<=c)*a*(c-abs(erro))/(abs(erro)*(c-b))+(abs(erro)>c)*0
plot(erro,psi,type="l")
plot(erro,w,type="l")

##############################################
Observação
##############################################
Para os dados do Bradesco, a programação segue de maneira análoga.
```

Referências Bibliográficas

[1] Bustos, O., *Algumas Ideias de Robustez Aplicada à Estimação Paramétrica em Séries Temporais*, vol.1, IMPA, São Paulo, 1986.

[2] Casella, G. e Berger, R. L., *Statistical Inference*, 2nd, Duxbury Press, Califórnia, 2002.

[3] Donoho, D. L. and Huber, P. J. , *The Notation of Break-down Point*, in A Festschrift for E. L. Lehmann, Wadsworth, 1983.

[4] Michael H. Kutner,M.H., Nachtsheim, C.J., Neter, J., Li, W., *Applied Linear Statistical Models*, 5nd, McGraw-Hili Irwin, New York, 2005.

[5] Gauss, C.F.(1890) *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*. Perthes et Besser, Hamburg. Werke, **7**, 1-128. Translated by C. H. Davis as *Theory of The Motion of the Heavenly Bodies Moving about The Sun in Conic Sections*. Little, Brown, Boston, 1857, Reprinted by Dover, New York, 1963.

[6] Gauss, C. F. (1821, 1823, 1826) *Theoria combinationis observationum erroribus minimis obnoxiae*. Werke, vol. 4, pp. 1-94, translated by Stewart, G. W., as *Theory of the Combination of Observations Least Subject to Errors:* Part One, Part Two, Supplement Pt. 1 & Pt. 2, New York: SIAM.

[7] Hampel, F. *Contributions to the Theory of Robust Estimation*, PhD thesis, University of California, Berkeley, 1968.

[8] Hampel, F. R. *The Influence Curve and Its Role in Robust Estimation.* The Annals of Statistics, 69, 383-393, 1974.

[9] Huber, P.J. *Robust Estimation of a Location Parameter*, University of California, Berkeley, 1964.

[10] Huber, P.J. *Robust Statistics*, John Wiley & Sons, Inc., New York, 1981.

[11] Huber, P.J. *Robust Statistical Procedures*, 2nd, SIAM., Philadelphia, 1996.

[12] Jurecková, J., Picek, J. *Robust Statistical Methods with R*, Chapman & Hall/CRC, Taylor & Francis Group, New York, 2006.

[13] Maronna, R. A., Martin, R. D., Yohai, V. J., *Rubust Statistics*, John Wiley & Sons, Inc., New York, 2006.

[14] Rousseeuw, P.J., Leroy, A.M., *Robust Regression and Outlier Detection*, John Wiley & Sons, Inc., New York, 1987.

[15] YAHOO. **Yahoo Finanças**. São Paulo, SP, 2014. Disponível em: <https://br.financas.yahoo.com/actives?e=SA>. Acesso em: 7 set. 2014.

[16] Tukey, J. W., *A Survey of Sampling from Contaminated Distributions.* In Contributions to Probbility ond Statistics (ed. Olkin). Stanford Univ. Press. 1960