Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Estatística Robusta Aplicada aos Títulos do Tesouro Direto

por
Rhayssa Maia Costa Pinto

Brasília, 2015

RHAYSSA MAIA COSTA PINTO

Estatística Robusta Aplicada aos Títulos do Tesouro Direto

Relatório Final apresentado, como parte dos requisitos para a obtenção do grau de Bacharel em Estatística.

Orientador: Prof. Dr. Raul Yukihiro Matsushita

Brasília

2015

Agradecimentos

Agradeço primeiramente à Deus pelo dom da vida e pelas oportunidades maravilhosas que tem me proporcionado, e por mais essa conquista.

Aos meus pais, Carlos Alberto e Ivanilda, que são a base da minha vida, por terem me apoiado em todas as etapas da minha vida acadêmica, pelo amor e carinho e pela compreensão, por terem confiado em mim. Ao meu noivo Guilherme Ambrózio que desde o início da minha graduação sempre me apoiou e nunca deixou que eu desistisse dos meus sonhos.

Aos meus tios e tias e minha avó, mesmo distantes sempre me apoiaram.

Aos meus primos e primas, que confiaram em mim.

Aos professores do Departamento de Estatística, pela ensino e por todo auxílio na minha formação acadêmica.

Ao meu orientador Raul Yukihiro Matsushita, pela paciência, prestatividade e por me oferecer uma ótima orientação.

Aos profissionais na área de Estatística em que tive oportunidade de trabalhar, pelo conhecimento e prática.

As minhas queridas amigas, Ariane Santos, Isabella Cintra, Waleska Souza e Inaê Mota pela amizade, companheirismo, pela parceria, por sempre me apoiarem, e me proporcionarem momentos especiais. Ao Anderson Silva e aos meus amigos de graduação que sempre estiveram ao meu lado.

Aos meus amigos que sempre me motivaram, que estiveram ao meu lado e torceram por mim.

Resumo

A estatística robusta teve início no fim do século XIX, com o intuito de estudar interferências que *outliers* ocasionavam aos métodos clássicos. Os métodos robustos produzem estimativas mais resistentes à ações desses valores anômalos.

A média e o desvio padrão da amostra são estimadores clássicos de modelos de locação e escala, respectivamente, neste trabalho iremos perceber que estes estimadores são pouco confiáveis quando há presença de valores atípicos, e será visto alternativas de estimação robusta para esses modelos.

O conhecimento das ferramentas das medidas de robustez é importante para a obtenção de estimadores, como por exemplo a função de influência e ponto de ruptura.

Um outro objeto de estudo deste trabalho será propor um modelo de regressão em que disponha de boas estimativas. Para o ajustamento do modelo de regressão será utilizado o método de mínimos quadrados ordinários (MQO) e o método robusto M-estimador, este consiste nos métodos de mínimos quadrados reponderados interativos.

A aplicação dos referidos métodos será nos títulos do tesouro direto, em que há condições preestabelecidas para o seu resgate. O intuito será estudar o risco financeiro, ou seja a variação desse preço em um determinado tempo.

Palavras-chave: Estatística robusta, M-estimador, Medidas de robustez, estimador de escala, estimador de locação.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

A história da estatística robusta tem início no fim do século XIX, com Simon Newcomb. Porém, os grandes avanços nesse estudo ocorreram na década de 1960 e início dos anos da década 1970 com os trabalhos de John Tukey, Peter Huber e Frank Hampel.

A estatística robusta surgiu para estudar a interferência que os valores atípicos, os chamados *outliers*(observações inconsistentes em um conjunto de dados), ocasionam nas estimativas dos métodos clássicos.

Algumas técnicas da estatística clássica rejeitam automaticamente esses valores anômalos, porém esse procedimento não é correto, pois esses valores podem ser portadores de informações vitais em uma determinada população. Por isso os métodos robustos foram propostos para a produção de estatísticas mais resistentes à ação desses valores atípicos.

Para a obtenção de estimadores robustos deve-se ter conhecimento das ferramentas básicas de medidas de robustez que são: a robustez qualitativa, a robustez quantitativa e a função de influência. A robustez qualitativa expressa que uma pequena alteração nos dados deve causar efeitos pequenos.

A robustez quantitativa é embasada no conceito de ponto de ruptura, este aspecto revela a magnitude de pertubação para que o modelo entre em colapso. A função de influência mede se o estimador responde suavemente a pequenas violações nos dados.

Essas ferramentas auxiliam na compreensão da eficiência dos estimadores robustos, em relação aos estimadores da estatística clássica. O objeto de estudo do trabalho consiste em propor um modelo de regressão que disponham de boas estimativas.

Para o ajustamento de modelos de regressão linear na estatística clássica, o estimador mais utilizado é o método de mínimos quadrados ordinários (MQO). Este método dispõe de pressupostos que precisam se observados na análise estatística, como ausência de valores influentes e pontos de alavanca.

E por isso, inevitavelmente a violação, a regressão robusta torna-se uma técnica alternativa interessante. Um dos estimadores mais utilizados na regressão linear robusta é o M-estimador, este estimador generaliza o conceito de estimação de máxima verossimi-

lhança.

A aplicação dos referidos métodos robustos será nos títulos do tesouro direto, que se classificam como renda fixa, ou seja, existem condições preestabelecidas para o seu resgate.

Diante disso, tem-se a curiosidade de conhecer os métodos robustos e observar o seus efeitos em um conjunto de dados. Uma das justificativas para um estudo mais aprofundado em estatística robusta é a utilização desses métodos em diversas áreas como por exemplo em finanças.

Em finanças o objetivo é analisar o risco financeiro, este é medido pelas variações no tempo. O log-retorno é muito utilizado para realizar essa medição, pois são livres de escalas. Há uma carência de textos em português sobre os métodos robustos, e este trabalho tem como finalidade ser um objeto de consulta para posteriores aplicações.

1.1 Objetivos

1.1.1 Objetivo geral

Mostrar incentivos para a utilização de métodos robustos para estimação de modelos de regressão e apresentar a eficiência dos métodos robustos na elaboração de estatísticas mais resistentes a valores atípicos.

1.1.2 Objetivos específicos

- Comparar o M-estimador e o estimador de mínimos quadrados da estatística clássica.
- Compreender a importância dos conceitos de curva de sensibilidade, função de influência e suas medidas de robustez qualitativas e quantitativas.
- Explorar aspectos computacionais do programa R 3.1.0 para o modelo de regressão linear e identificar observações atípicas.
- Utilizar os métodos robustos aplicando-os em um conjunto de dados.

Capítulo 2

Estimação Robusta

2.1 Teoria de estimação robusta

Segundo Huber (1972) a teoria da estimação originou-se com o problema da variabilidade estatística devido a erros de medição. Esses erros podem ser produzidos na coleta de dados ou no registro incorreto desses dados.

A teoria da estimação clássica surgiu no século XIX e hoje é muito aplicada em várias áreas. Essa teoria é sensível a dados que contenham valores atípicos, os chamados *outliers*, pois os estimadores da estatística clássica mostram-se sensíveis à presença desses valores. A necessidade de minimizar o impacto que esses valores atípicos provocam nas estimativas, acarretou no surgimento de métodos mais robustos, ou seja menos sensíveis.

A teoria da estatística robusta lida com desvios das suposições sobre o modelo e refere-se à construção de procedimentos estatísticos que se aplicam a distribuições de caudas pesadas ou contaminados com valores extremos provenientes de erros.

Os procedimentos de estimação e teste robustos têm atraído a atenção de muitos pesquisadores historicamente, e existem três principais linhas de investigação que formalizam a teoria da robustez.

A primeira linha de pesquisa iniciou em 1964 com a teoria *minimax* de Huber, em que o problema estatístico é visto como um jogo entre a Natureza (que escolhe uma distribuição em torno do modelo) e o estatístico (que escolhe um procedimento estatístico). O estatístico atinge a robustez pela construção de um procedimento *minimax* que minimiza a perda segundo um critério, como por exemplo, a variância assintótica.

A segunda linha se baseia no conceito de pontos de influência concentrando-se em características de robustez global, ou seja , o impacto de valores discrepantes em um procedimento estatístico .

A terceira linha de pesquisa, introduzido por Hampel (1968), considera a robustez local (ou robustez infinitesimal), em que se avalia o impacto da distribuição de desvios moderados a partir de modelos ideais em um procedimento estatístico .

2.2 Modelo de locação

Segundo Maronna (2006) o modelo de locação é descrito assumindo que o resultado de cada observação da amostra x_i depende de um parâmetro de locação desconhecido μ a ser estimado, e um erro aleatório que denotaremos por u_i. Supondo que os erros sejam aditivos, o modelo pode ser escrito como:

$$x_i = \mu + u_i, \qquad \forall i = 1, ..., n. \tag{2.1}$$

Se as observações da amostra são independentes e identicamente distribuídas (iid), pode-se assumir que:

- $u_1, ..., u_n$ tem mesma função de distribuição $F_0(u)$.
- $u_1, ..., u_n$ são independentes.

Então conclui-se que $x_1, ..., x_n$ são variáveis aleatórias independentes com função de distribuição $F(x)$ simétrica em torno de μ. Buscamos estimativas de $\hat{\mu}$=μ com probabilidade alta de ocorrência. Uma maneira de mensurar a aproximação é usando o erro quadrático médio (EQM).

$$EQM(\hat{\mu}) = E(\hat{\mu} - \mu)^2. \tag{2.2}$$

Segundo Bustos (1981) na estimação robusta o ponto inicial é supor que a função de distribuição seja parcialmente conhecida. Neste caso, esta função de distribuição pertence a uma vizinhança de uma distribuição hipotética que pode ser totalmente conhecida. A função pode ser escrita com base em uma mistura da forma:

$$F = (1 - \epsilon) * F_0 + \epsilon * H, \tag{2.3}$$

em que $0 < \epsilon < 1$, F_0 é uma distribuição conhecida e H é uma distribuição simétrica. Existem vários estimadores possíveis para o parâmetro de locação. Iremos definir que T_n estimador de μ. É um requisito que o estimador T_n seja estimador equivariante, ou seja, isso significa que não varia sobre transformações:

$$T_n(ax_1 + b, ..., ax_n + b) = aT_n(x_1, ..., x_n) + b. \tag{2.4}$$

Para todo a e b pertencente aos $\Re$. Partindo de conclusões expostas em diversos trabalhos tem-se chegado à conclusão de que uma alternativa robusta muito conveniente quando se quer evitar as consequências catastróficas de usar a média amostral como estimador de locação, na presença de *outliers*, seria utilizar uma classe de estimadores similares ao de máxima verossimilhança. Essa classe é a dos M-estimadores.

2.2.1 M-estimador de locação

Definição 1. *Seja $\rho : \Re \to \Re_+$ uma função não negativa. Chama-se M-estimador T_n de μ definido por $\rho(x)$ como:*

$$\sum_{i=1}^{n} \rho(x_i - T_n) \leq \sum_{i=1}^{n} \rho(x_i - \mu) \qquad \forall \mu \to \Re. \tag{2.5}$$

Em principio, $\rho(x)$ poderia ser qualquer função. Porém, se algumas propriedades sobre $\rho(x)$ não forem requeridas, um T_n que satisfaça (2.5) poderá existir apenas para um conjunto de observações que dificilmente apareceram na prática. Daí, em geral pede-se que ρ seja simétrica, convexa e $\rho(x) \to \infty$ quando $|x| \to \infty$.

Considerando o modelo de locação (2.1), se F for uma função distribuição de x_i a densidade $f = F^{'}$. As observações x_i possui distribuição com densidade $f(x - \mu)$ do tipo contínua, em que μ é o parâmetro de locação. A função de máxima verossimilhança é dada por:

$$L_{(x_1,\ldots,x_n)} = \prod_{i=1}^{n} f(x_i - \mu). \tag{2.6}$$

O estimador de máxima verossimilhança (EMV) de μ é um valor que depende dos valores observados que maximiza $L_{(x_1,\ldots,x_n)}$, ou seja:

$$T_n = \widehat{\mu} = \text{argmax}_{\mu} L_{(x_1,\ldots,x_n)}. \tag{2.7}$$

Sendo f uma função positiva, uma vez que o logaritmo é uma função crescente então o estimador pode ser escrito como:

$$\widehat{\mu} = \text{argmin}_{\mu} \sum_{i=1}^{n} \rho(x_i - \mu), \tag{2.8}$$

em que $\rho = -logf$. Suponha que essa minimização possa ser realizada pela diferenciação (derivação) da equação (2.8) em relação ao parâmetro μ, ou seja achando o $\widehat{\mu}$ apropriado, este satisfaria:

$$\sum_{i=1}^{n} \Psi(x_i - \mu) = 0, \tag{2.9}$$

em que $\Psi = \rho^{'}$. Assim com o M-estimador robusto desejamos determinar a função (2.5) de modo que o resultado do estimador seja protegido contra uma pequena porcentagem de *outliers*. O M-estimador pode ser visto como uma média ponderada, e terá uma função peso, denominada $w(x)$. Usando a definição mais técnica, Huber derivou a seguinte robustez de $\Psi(x)$, $\rho(x)$ e $w(x)$:

$$\Psi(x) = \begin{cases} -k, & x < -k \\ x, & |x| \leq k \\ k, & x > k \end{cases} \tag{2.10}$$

$$\rho(x) = \begin{cases} x^2/2, & |x| < k \\ |x| - \frac{x^2}{2}, & |x| > k \end{cases} \tag{2.11}$$

$$w(x) = \begin{cases} \Psi^{'}(x)/x, & \text{se } x \neq 0 \\ \Psi^{'}(0), & \text{se } x = 0 \end{cases} \tag{2.12}$$

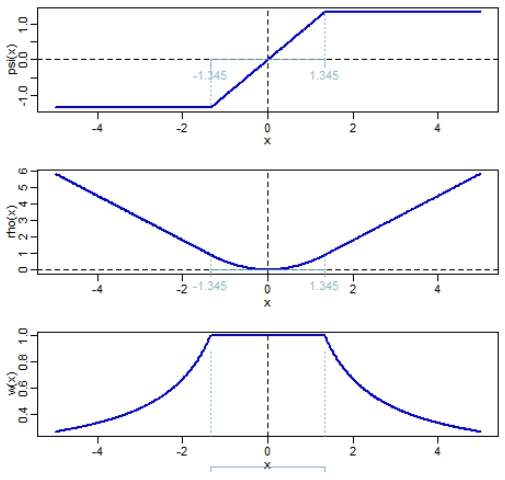

Figura 2.1: Funções $\Psi(x)$, $\rho(x)$ e $w(x)$ de Huber

O gráfico das funções de Huber mostra que $\Psi(x)$, $\rho(x)$ e $w(x)$ possuem comportamento distintos. A função $\Psi(x)$ mantem-se constante, $\rho(x)$ é crescente e $w(x)$ decresce quando atinge determinado valor de k. O valor de k é escolhido de modo a assegurar uma determinada variância assintótica. Quando $k = 1,345$ temos uma perda em eficiência que seria de aproximadamente de 5% em relação a distribuição Normal. Isto é (2.11) e (2.10) são funções associadas a distribuição "Normal"no meio e com a distribuição "exponencial dupla"nas caudas.

Outras funções são comumente utilizadas, quando a distribuições dos dados tem cauda pesada, é melhor utilizarmos estimadores "*redescending*", esses estimadores possuem uma função ρ que tem um aumento mais lento em relação a função Huber. Abaixo apresentaremos uma dessas funções com seus respectivos valores de constantes sugeridos, para a obtenção de 95% de eficiência em relação a distribuição Normal.

- M-estimador de Tukey-bisquare

Tukey sugeriu a seguinte robustez de $\Psi(x)$, $\rho(x)$ e $w(x)$:

$$\Psi(x) = \begin{cases} x(1-(x/h)^2)^2, & |x| \leq h \\ 0, & |x| > h \end{cases} \tag{2.13}$$

$$\rho(x) = \begin{cases} 1-[1-(x/h)^2]^3, & |x| < h \\ 0, & |x| > h \end{cases} \tag{2.14}$$

$$w(x) = \begin{cases} \Psi'(x)/x, & \text{se } x \neq 0 \\ \Psi'(0), & \text{se } x = 0 \end{cases} \tag{2.15}$$

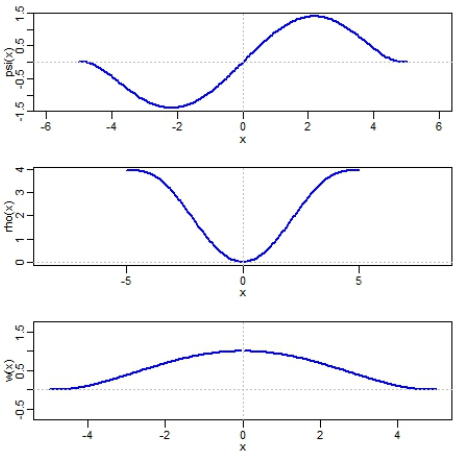

Figura 2.2: Funções $\Psi(x)$, $\rho(x)$ e $w(x)$ de Tukey-bisquare

A função $\Psi(x)$ de Tukey-bisquare tem um comportamento crescente e decrescente, e $\rho(x)$ é uma função crescente a medida que x aumenta e a função $w(x)$ tem um decrescimento gradual quando atinge determinado valor de h. Quando o parâmetro de escala é desconhecido, o melhor valor para a constante é $h = 4.685$, para termos um nível de eficiência de 95% em relação a distribuição Normal.

2.2.2 Outras classes de estimadores

L-estimador

Consideraremos o caso mais simples de uma amostra aleatória de uma distribuição continua em que μ é o parâmetro de locação. Definiremos a estatística de ordem das observações como $x_{(1)} \leq ... \leq x_{(n)}$.

Definição 2. *Sejam $a_1, .., a_n$ números reais $\sum_{i=1}^{n} a_i$ chama-se L-estimador induzido por $a_1, .., a_n$ baseado em $x_1, .., x_n$:*

$$L_n = L_n(x_1, ..., x_n) = \sum_{i=1}^{n} a_i x_i. \tag{2.16}$$

Os diferentes tipos de L-estimadores se distinguem segundo a maneira que se derivam os a_i. Ou seja, o L-estimador é uma combinação linear da estatística de ordem. Os tipos mais tradicionais são:

- Mediana Amostral

Chama-se mediana das observações $x_1, ...x_n$:

$$Mediana = \frac{x_M + x_L}{2}. \tag{2.17}$$

Sendo $M = [(n+1)/2]$, $L = [(n+2)/2]$. Observa-se que a mediana é um L-estimador:

Se n=2k-1

$$\begin{cases} a_k = 1 \\ a_i = 0, \quad \text{i} \neq k \end{cases} \tag{2.18}$$

Se n=2k+1

$$\begin{cases} a_k = 1/2 \\ a_{k+1} = 1/2 \\ a_i = 0, \qquad \text{i} \neq k \qquad e \qquad i \neq k+1 \end{cases} \tag{2.19}$$

- Trimédia

Chama-se estimador trimédia baseado nas observações $(x_1, ...x_n)$:

$$T = T(x_1, ...x_n) = 1/4x_{(q-)} + 1/2\text{MED} + 1/4x_{(q+)}, \tag{2.20}$$

em que $q- = [n/4] + 1$,$q+ = n + [n/4]$ e MED é a mediana das observações definida em (2.17). Então podemos concluir que o estimador trimédia é a média ponderada do 1^o, 2^o e 3^o quartil com seus respectivos pesos.

R-estimador

A família dos R-estimadores é baseada nos testes não paramétricos para amostras emparelhadas baseados em postos.

Definição 3. *Seja* $J : (0,1) \to \Re$ *uma função não decrescente tal que* $J(1-t) = -J(t)$, $a_n : 1, 2..., 2n \to \Re$ *uma função definida por* $a_n(k) = J(\frac{k}{2n+1})$. *E* $S_n = \frac{1}{n}\sum_{i=1}^{n} a_n(R_i)$. *Chama-se R-estimador de* μ *definido por* J *baseado em* $x_1, ..., x_n$ *a um* $\hat{\mu}_{R,J}$ *tal que:*

$$T_n(\hat{\mu}_{R,J}; x_1, ...x_n) = 0 \tag{2.21}$$

Pode-se mostrar que o posto de R_i de $x_i - m$ em $\{x_1 - m, ..., x_n - m\}$ é o mesmo que o posto de x_i em $\{x_1 - 2m, ..., x_n - 2m\}$, em que m é a mediana das observações. Por isto $\hat{\mu}_{R,J}$ está definido por $T_n^*(\hat{\mu}_{R,J}; x_1, ...x_n) = 0$, em que $T_n^* : \Re x \Re^+ \to \Re$ é uma função definida por:

$$T_n^*(m; x_1, ...x_n) = \frac{1}{n}\sum_{i=1}^{n} a_n(R_i^*) \tag{2.22}$$

Onde R_i^* é o posto de x_i no conjunto $\{x_1, ..., x_n, 2m - x_1, ..., 2m - x_n\}$. Usando a nomenclatura usual "a_n" é chamado de escores e "J" a função geratriz de escores. As funções J podem ser definidas baseadas da distribuição F, por meio de sua densidade, através de:

$$J(t) = \frac{f'(F^{-1}(t))}{f(F^{-1}(t))} \qquad 0 < t < 1 \tag{2.23}$$

Como exemplo temos:

- R-estimador baseado em "escores mediana", neste caso a densidade será uma exponencial dupla, em que:

$$J(t) = \begin{cases} 1 & \text{se} \quad 0 < t < 1/2 \\ -1 & \text{se} \quad 1/2 < t < 1 \end{cases} \tag{2.24}$$

Os M-estimadores, caracterizam-se por ser uma generalização dos estimadores de máxima verossimilhança (EMV), e por ter uma melhor eficiência que os demais estimadores apresentados. Portanto, daremos mais ênfase nessa família de estimadores. A estimação de locação robusta é feita por métodos interativos e para melhores detalhes sobre o algoritmo ver Apêndice B.

2.3 Modelo de escala

Neste caso, iremos considerar que as observações x_i satisfazem o seguinte modelo multiplicativo, o erro aleatório u_i:

$$x_i = \sigma u_i, \qquad \forall i = 1, ..., n. \tag{2.25}$$

Com as mesmas suposições feitas para o modelo de locação, em que a densidade dos erros definimos como f_0, neste caso o parâmetro desconhecido é o σ, este mensura a dispersão das observações, e deve ser estritamente positivo. Então considerando o modelo (2.25), a distribuição de x_i constitui uma forma que compõe uma "família escala"em que a densidade será:

$$\frac{1}{\sigma} f_0 \frac{x}{\sigma}. \tag{2.26}$$

Existem vários estimadores possíveis para o parâmetro de escala. Então iremos definir que S_n é o estimador de σ. Como para o estimador de locação é requisito que o estimador S_n seja estimador equivariantes, ou seja, isso significa que não varia sobre transformações:

$$S_n(ax_1 + b, ..., ax_n + b) = |a|\, S_n(x_1, ..., x_n). \tag{2.27}$$

Para todo a e b pertencente aos $\Re$. Segundo Huber(1991) problemas de escala puro são raros, na prática o parâmetro de escala na distribuição ocorre normalmente como um parâmetro "*nuisance*"no modelo de locação, ou seja, este parâmetro não é de interesse imediato, mas faz parte na análise do parâmetro que é de interesse. Esses problemas de escala pura são complexos. Porém tem como vantagem que podem ser convertidos em problemas de locação, neste caso iremos usar o logaritmo. E temos a desvantagem que as distribuições resultantes destas transformações, não há uma escala natural (correspondente ao centro de simetria).

2.3.1 M-estimador de escala

O estimador de máxima verossimilhança (EMV) de σ é definido através da função de densidade (2.26) como: (Maronna,2006)

$$S_n = \hat{\sigma} = \text{argmax}_\sigma \frac{1}{\sigma^n} \prod_{i=1}^{n} f_0(\frac{x_i}{\sigma}). \tag{2.28}$$

Tomando o logaritmo e diferenciando com relação a σ, temos que é a solução de:

$$\frac{1}{n} \sum_{i=1}^{n} \rho_{escala}(\frac{x_i}{\sigma}) = \delta. \tag{2.29}$$

Observe que, para (2.29) tenha uma solução então $0 < \delta < \rho_{escala}(\infty)$. Em que ρ_{escala} é uma função diferenciável (derivável) e não decrescente. Então devemos ajustar as propriedades de estimativas de escala em relação ao parâmetro de locação.

Huber(1964) propôs que a função ρ_{escala} seria:

$$\rho_{escala}(x) = \begin{cases} x^2 - g, & |x| \leq k \\ k^2 - g, & |x| > k \end{cases} \tag{2.30}$$

Para alguns valores na constante k, com g determinado de tal modo que (2.29) seja igual a 0.

Tukey, definiu que a função ρ_{escala} seria da forma:

$$\rho_{escala}(x) = \min\left\{1 - (1 - x^2)^3, 1\right\}, \tag{2.31}$$

em que $\delta = 0.5$.

O desvio mediano absoluto (MAD) tem sido uma boa estimativa de escala, pois tem a característica de ser mais robusto. Em que é construído pela função ρ_{escala} :

$$\rho_{escala} = \text{sign}(|x_i| - 1) \tag{2.32}$$

Portanto definimos o desvio mediano absoluto (MAD), como:

$$MAD = Med(|x - Med(x)|). \tag{2.33}$$

Ou seja, o desvio mediano absoluto (MAD) determina todos os desvios absolutos entre cada observação e a mediana, e a mediana dessas distância produz uma boa estimativa de escala.

A estimação de escala robusta é feita por métodos interativos e para melhores detalhes sobre o algoritmo ver Apêndice B.

2.4 Medidas da robustez

Enquanto a abordagem da estatística clássica visa estimativas que tenham propriedades desejáveis em um determinado modelo, o objetivo dos métodos robustos é desenvolver estimativas que não são sensíveis a pequenas violações (ou desvios de suposições).

Existem várias abordagens de mensuração da robustez. O trabalho de Huber (1996) apresenta a definição de robustez sob os seguintes aspectos: robustez qualitativa, robustez quantitativa baseada no conceito de ponto de ruptura e robustez infinitesimal.

- Robustez Qualitativa

Baseado no princípio da continuidade da distribuição assumida, ou seja, pequenas pertubações na distribuição devem causar pequenos efeitos nos métodos estatísticos utilizados.

Assim se X_1,...X_n são variáveis iid (independentes e identicamente distribuídas), com distribuição comum F e o estimador T_n é baseado na distribuição dessas variáveis, $T_n = T_n(X_1, ...X_n)$, então pode ser interpretado da seguinte forma: Uma pequena alteração em $F = \ell(X)$ deve resultar arbitrariamente em uma pequena mudança em $F = \ell(T_n)$. (HUBER,1996)

- Robustez Infinitesimal

Baseado no conceito de função de influência. Esta função mostra que uma pequena alteração nas observações pode resultar em uma pequena mudança na estimativa do parâmetro θ. Este conceito será abordado posteriormente.

- Robustez Quantitativa

Baseado no conceito de ruptura, que é uma medida global de robustez. A grosso modo o ponto de ruptura mostra a quantidade máxima de *outliers* que o estimador pode suportar antes de produzir erros nas estimativas.

2.4.1 Função de influência

Mostra como o estimador responde a uma pequena contaminação em vários pontos nos dados, ou seja, mede a sensibilidade do estimador quando ocorre pequenas mudanças no processo que gera dados originários da distribuição F.

Segundo Staudte e Sheather (1990) a distribuição $F_{x_0,\epsilon} = (1-\epsilon)F + \epsilon\Delta_{x_0}$ é um modelo apropriado quando há contaminações nos dados. Em que Δ_{x_0} é uma função de distribuição no qual o valor de x_0 ocorre com probabilidade $P(x = x_0) = 1$.

Staudte e Sheather (1990) mostram que a influência relativa, que possui $\hat{\theta}(F)$ com probabilidade ϵ, é definida por :

$$\frac{\hat{\theta}(F_{x_0,\epsilon}) - \hat{\theta}(F)}{\epsilon}. \tag{2.34}$$

E a função de influência de $\hat{\theta}$ em F quando a amostra contém uma pequena proporção de *outliers*. É defina como :

$$IF = \lim_{\epsilon \downarrow 0} \frac{\hat{\theta}(F_{x_0,\epsilon}) - \hat{\theta}(F)}{\epsilon}. \tag{2.35}$$

Então:

$$IF = \frac{\partial}{\partial \epsilon} \hat{\theta}((1-\epsilon)F + \epsilon \Delta_0) \downarrow \epsilon \tag{2.36}$$

Desde que seus limites existam e pertença ao conjunto de números reais. A função de influência pode ser considerada como uma "versão limite"da curva de sensibilidade, cuja definição é:

Definição 4. *Seja $\hat{\theta_{n+1}}$ e $\hat{\theta_n}$ estimadores do parâmetro θ baseados em tamanhos de amostras $n+1$ e n respectivamente (ambos são estimadores definidos pela mesma regra, o que diferencia é o tamanho da amostra) e seja $x_1, ..., x_n$ números reais. Chama-se curva de sensibilidade a função $SC_n : \Re \to \Re$ definida por:*

$$SC_n(x_0) = ((n+1)(\widehat{\theta}_{n+1}(x_1, ..., x_{n+1}), x_0)) - \widehat{\theta}_n(x_1, ..., x_n). \tag{2.37}$$

Neste caso $\epsilon = 1/(n+1)$. É de se esperar que os $x_{i's}$ são iid com função de distribuição F, então $SC_n(x_0) \approx IF(x_0, F)$ para grandes amostras. Pode-se notar que para cada x_0, $SC_n(x_0)$ é uma variável aleatória.

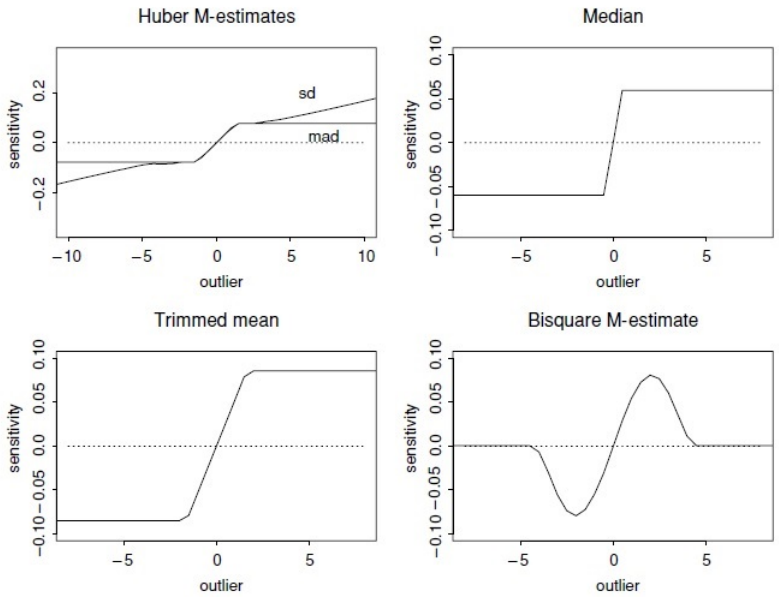

Figura 2.3: Curva de sensibilidade estimadores de locação. Fonte: Maronna, 2006

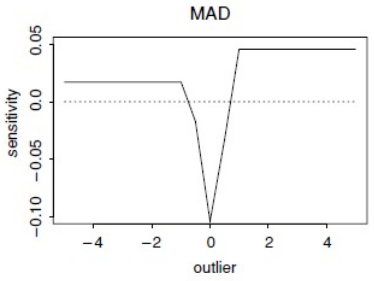

Figura 2.4: Curva de sensibilidade MAD. Fonte: Maronna, 2006

A Figura (2.3) mostra a curva de sensibilidade para alguns estimadores robustos. O M-estimador de Huber para k=1.345 e utilizando MAD e o desvio padrão como estimadores de escala, observa-se que há uma diferença entre essas curvas, quando utiliza-se o desvio padrão e a quantidade de *outlier* é alta, a sensibilidade deste estimador aumenta. A curva de sensibilidade da mediana e da trimédia ficam constantes independentes da quantidade de *outliers*. Para o estimador de Tukey-bisquare foi utilizado o MAD como estimativa de escala, quando há uma quantidade razoável de *outliers* a curva de sensibilidade deste estimador torna-se constante . A Figura (2.4) mostra a curva de sensibilidade para o estimador de escala MAD, a medida que aumenta a quantidade de *outliers* a curva é crescente e depois de atingir uma determinada quantidade de *outliers* a curva de sensibilidade torna-se constante.

2.4.2 Ponto de ruptura

O ponto de ruptura (BP) do estimador do parâmetro θ é a máxima quantidade de proporção de valores atípicos que os dados podem conter de tal modo que $\widehat{\theta}$ ainda fornecerá informações sobre o parâmetro θ.

O parâmetro θ varia ao longo de um conjunto Θ. A fim de que o estimador $\widehat{\theta}$ resulte alguma informação sobre θ, a proporção de *outliers* não deve ser capaz de conduzir $\widehat{\theta}$ para o infinito ou para o limite quando o conjunto Θ é não vazio. Por exemplo, para o parâmetro de escala, temos que o conjunto está delimitado por $[0, \infty]$, ou seja a estimativa também estará limitada para este intervalo.

Definição 5. *A contaminação assintótica do ponto de ruptura da estimativa de* $\widehat{\theta}$ *em* F*, é denotada por* $\epsilon^*(\theta, F)$*, é o maior* $\epsilon^* \in (0,1)$ *de tal modo que* $\epsilon^* < \epsilon$ *,* $\hat{\theta}((1-\epsilon)F + \epsilon G)$*. Como a função* G *é contínua e limitada, e também limitada a partir dos limites de* Θ*.*

O ponto de ruptura para cada tipo de estimador, tem que ser tratado separadamente. Pode-se observar que é fácil encontrar estimativas com alto ponto de ruptura. Por exemplo, quando a estimava é 0, o $\epsilon^* = 1$. No entanto, para estimativas razoáveis, é intuitivamente claro que deve haver mais observações típicas do que observações atípicas, portanto neste caso o $\epsilon^* \leq 1/2$.

2.5 Detecção de *outliers*

Existem dois tipos de métodos de detecção de *outliers* que são por testes formais ou por testes informais. A maioria dos testes formais (ou métodos de discordância) precisam de estatística do teste para obtenção de resultados nos testes de hipóteses. Essa estatística do teste normalmente é baseada assumindo que há uma distribuição adequada aos dados, e neste caso é testado se o valor extremo é um *outlier* da distribuição.

A escolha destes testes depende principalmente do tipo de *outliers* e do tipo da distribuição dos dados. Os testes formais são poderosos quando há suposição de uma distribuição especifica, como Normal, Gama ou Exponencial.

Um método formal de detecção de *outliers* é do Escore Z. A ideia básica desse método é que os dados seguirem uma distribuição Normal $N(\mu, \sigma^2)$, então Z segue uma distribuição Normal $N(0, 1)$. Em que:

$$Z_i = \frac{X_i - \bar{X}}{sd}. \tag{2.38}$$

Onde $\bar{X}$ é a média da amostra e sd é o desvio padrão da amostra. Neste método, será considerado *outliers* caso o valor de Z_i for superior a 3 em valor absoluto.

Outra limitação desses valores atípicos são problemas de mascaramento, ou seja, os valores atípicos menos extremos não influenciam os métodos por haver mais valores extremos, ou vice e versa.

Uma outra abordagem para a detecção desses valores atípicos é o boxplot padrão, este baseia-se no intervalo interquatílico. A regra comumente utilizada é:

$$x_i < q_1 - k(q_3 - q_1) \quad \text{e} \qquad x_i > q_3 + k(q_3 - q_1),$$

em que $k = 1.5$ é o mais utilizado, a menos que indicado por outra forma.

Capítulo 3

Regressão

3.1 Regressão linear simples

O modelo de regressão auxilia a compreender como determinadas variáveis influenciam na predição de outra variável. A Análise de Regressão define um conjunto de técnicas estatísticas que possibilitam encontrar uma relação, que pode ser funcional ou de associação (variam em conjunto), entre as variáveis, e também predizer o valor das variáveis dependentes (ou de resposta) com base em um conjunto de variáveis independentes (ou de explicação).

A forma do modelo de regressão linear simples para uma amostra de tamanho n e com uma variável explicativa é escrito da seguinte foma:

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

em que:

Y_i é a valor da variável resposta na i-ésima observação;

β_0 e β_1 são os parâmetros do modelo, que serão estimados;

X_i é o valor da variável explicativa na i-ésima observação;

ϵ_i é o uma variável aleatória representa o erro gerado pelo modelo;

3.1.1 Método de mínimos quadrados ordinários

As estimativas serão obtidas a partir de uma amostra, em que temos Y como a variável resposta e X como variável explicativa, obtemos então o par (X, Y), em que (x_1, y_1) representa a primeira observação no conjunto de dados, (x_2, y_2) representa a segunda observação e (x_i, y_i) é a i-ésima observação no conjunto de dados. Para a obtenção dessas estimativas deve-se supor no modelo (3.1) que os erros aleatórios são independentes, homoscedasticidade, ou seja a variância constante, e esses erros seguem uma distribuição Normal $(0, \sigma^2)$

O objetivos é estimar β_0 e β_1 minimizando a diferença entre o valor real de Y e o valor

predito, que nesse caso será representado por $\widehat{Y}$. Essa diferença é dada por:

$$e_i = y_i - \widehat{y_i}, \tag{3.2}$$

em que e_i é chamado de resíduo da i-ésima observação e $\widehat{y_i} = b_0 + b_1 x_i$. Considerando b_0 e b_1 como estimadores de β_0 e β_1 respectivamente, o método de mínimos quadrados consiste em minimizar a soma de quadrado dos resíduos, que será definido pela função L.

$$L = \sum_{i=1}^{n} e_i^2. \tag{3.3}$$

Portanto, temos que:

$$L = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1)^2. \tag{3.4}$$

Para minimizar L, derivamos a função L em relação aos parâmetros β_0 e β_1, e obtemos:

$$\frac{\partial L(\beta_0, \beta_1)}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2. \tag{3.5}$$

$$\frac{\partial L(\beta_0, \beta_1)}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 x_i. \tag{3.6}$$

Igualando a zero as derivadas parciais e substituindo β_0 e β_1 por b_0 e b_1 que indicam os os parâmetros , teremos o seguinte sistema de equações normais:

$$-2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 = 0. \tag{3.7}$$

$$-2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 x_i = 0. \tag{3.8}$$

Obtemos a seguinte solução desse sistema:

$$b_0 = \bar{y} - b_1 \bar{x}. \tag{3.9}$$

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})}. \tag{3.10}$$

Para melhores detalhes sobre as propriedades desse método ver Apêndice A.

3.1.2 Diagnóstico

O diagnóstico do modelo de regressão linear se dá essencialmente pela análise dos resíduos, ou seja, pela verificação das suposições do modelo por métodos inferenciais.

- Normalidade dos Resíduos

Um dos requisitos para que o ajuste do modelo seja confiável, é a distribuição dos resíduos ser uma Normal $(0, \sigma^2)$. Neste caso, os testes não paramétricos verificam essa suposição. Um teste utilizado é o teste de Shapiro-Wilk em que as hipóteses são:

$$\begin{cases} H_0\text{: Os resíduos seguem uma distribuição Normal } (0,\sigma^2) \\ H_1\text{: Os resíduos não seguem uma distribuição Normal } (0,\sigma^2) \end{cases}$$

Em que a estatística de teste será:

$$W = \frac{1}{D}\left[\sum_{i=1}^{k} a_i (x^{(n-i+1)} - x^{(i)})\right]^2, \tag{3.11}$$

em que k é aproximadamente $n/2$, $x^{(i)}$ é um estatística de ordem i e $D = \sum_{i=1}^{n} (x_i - \bar{x})$. A regra de decisão será feita sob um nível de significância α, se a estatística W for menor que o quantil α do teste de normalidade de Shapiro-Wilk então rejeitamos a hipótese H_0.

- Homoscedasticidade dos Resíduos

A variância homogênea dos resíduos é outro requisito para que o modelo seja adequado, para a verificação, pode-se visualizar o gráfico de valores preditos "*versus*"resíduos. Para o diagnóstico de homoscedasticidade, busca-se verificar se os resíduos encontram-se distribuídos em torno de zero.

- Independência dos Resíduos

Pode ser verificada visualmente pelo gráfico de valores preditos "*versus*"resíduos, se ocorrer tendência então esses erros não são independentes, ou seja, a distribuição dos resíduos deve ser de forma aleatória.

- *Outliers*

O gráfico de valores preditos "*versus*"resíduos também auxilia na detecção de valores atípicos. Se o *outlier* for influente nas estimações dos parâmetros, ele irá interferir nos valores ajustados. Porém uma observação atípica, não é consequentemente um ponto influente.

- Pontos Influentes

É dito ponto influente se a a eliminação deste causa mudança nos valores dos parâmetros estimados. Há várias técnicas para a identificação desses valores influentes. Uma técnica utilizada é a distância de Cook, esta mede a influência da observação sobre todos os valores ajustados. A distância de Cook é definida como:

$$D_i = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_{i(j)})^2}{p * QME}, \tag{3.12}$$

em que p é o número de parâmetros do modelo, o QME é o quadrado médio dos resíduos. Quando $D_i > 1$, o valor do resíduo é um valor influente nas estimativas do modelo.

3.1.3 Método robusto de regressão

As estimativas de mínimos quadrados podem não ter um comportamento eficiente quando a distribuição dos erros não é normal, particularmente quando esses erros atribuem pesos na cauda. Neste caso uma solução seria a retirada dessas observações, porém temos um outro recurso que seria a utilização do M-estimador, pois esse não se mostra tão vulnerável a valores atípicos.

A regressão robusta é feita por mínimos quadrados reponderados interativos, devido ao fato desses estimadores não terem uma expressão analítica explicita. As ponderações dependem dos resíduos, estes dependem dos coeficientes estimados, por este fato é utilizada uma solução interativa.

Substituindo o método dos mínimos quadrados (3.4) pelo critério do M-estimador (Huber,1964) a estimação dos $\beta_i s$ será:

$$b_n = \operatorname{argmin} \sum_{i=1}^{n} \rho(\frac{y_i - x_i\beta}{\hat{\sigma}}). \tag{3.13}$$

Neste caso denotaremos $r_i = y_i - x_i\beta$. A derivada de $\rho(r_i/\hat{\sigma})$ será denotada por $\psi(r_i/\hat{\sigma}) = \rho(r_i/\hat{\sigma})$. A função de influência (IF) do estimador b_n nesta situação depende de dois argumentos, x e y. Do mesmo modo, a medição do ponto de ruptura de b_n não deve ser considerado apenas com respeito as alterações nos valores de y, mas também no que diz respeito aos valores de x.

Em particular $\rho(t) = \frac{1}{2}t^2$, em que consideraremos $t = (r_i/\hat{\sigma})$,mostra a estimativa pelo método dos mínimos quadrados.

Rousseeuw e Yohai (1984) mostram que as estimativas pelo métodos dos mínimos quadrados apresentam ponto de ruptura (mede qual seria a porcentagem de contaminação que um estimador pode suportar e ainda assim fornecer informação confiável sobre o parâmetro considerado) mensurado a $\frac{1}{n}$, e este tende a zero quando o tamanho da amostra é muito grande. Portanto uma observação atípica pode causar grande impacto na estimação pelo método de mínimos quadrados.

Propriedades da função $\rho(.)$:

- Sempre não negativa $\rho(t) \geq 0$.
- Igual a zero quando $\rho(0) = 0$.
- Simétrica $\rho(t) = \rho(-t)$.
- Monótona em $|t|$,$\rho(t) = \rho(t')$ para $|t| = |t'|$.

Uma das funções ρ normalmente usadas é a função ψ de Huber (HUBER,1981), em que $\psi(t) = \rho'(t) = max\{k, min(-k, t)\}$. Huber(1981) recomenda usar k=1.345 na prática. Esta escolha produz uma eficiência relativa de aproximadamente 95% em relação a distribuição Normal.

Tukey propôs que :

$$\psi(t) = \rho'(t) = \begin{cases} t(1-(t/h)^2)^2, & |t| \leq h \\ 0, & |t| > h \end{cases} \tag{3.14}$$

Para o modelo de regressão a constante mais utilizada é $h = 4.685$ na qual também tem uma eficiência relativa de aproximadamente 95% em relação a distribuição Normal.

No diagnóstico com estimativas robustas observamos apenas os valores influentes na estimação, pois esses métodos não são baseados em suposições. A estimação de regressão robusta é feita por métodos interativos e para melhores detalhes sobre o algoritmo ver Apêndice B.

Capítulo 4

Aplicação

4.1 Apresentação dos dados

A aplicação dos referidos métodos será nos títulos de tesouro Direto, os dados situam-se no site do Tesouro Nacional, (http://www.tesouro.fazenda.gov.br) no link Balanços e Estatísticas. A atualização do preço de venda ocorre diariamente, iremos descrever as características de duas modalidades de título do tesouro Nacional.

O intuito é analisar os riscos financeiros através dos log-retornos do ano de 2014 das letras do tesouro nacional (LTN) e letras financeiras do tesouro (LFT). A análise robusta será realizada no software estatístico R, em que utilizar-se-á os pacotes *robustbase*, $MASS$, *qualityTools*, *stats*, *smoothmest* e $Rfit$.

4.1.1 Letra do tesouro nacional - LTN

É um título prefixado, ou seja a rentabilidade é definida no momento da compra. Esta rentabilidade é calculada pela diferença entre o preço da compra do título e seu valor nominal no vencimento, este valor será sempre R$ 1000,00.

Essa diferença é conhecida como depreciação do título. Este título possui fluidez de pagamento simples, o investidor faz a compra do título e recebe seu rendimento uma única vez, na data de vencimento deste, junto com o valor nominal.

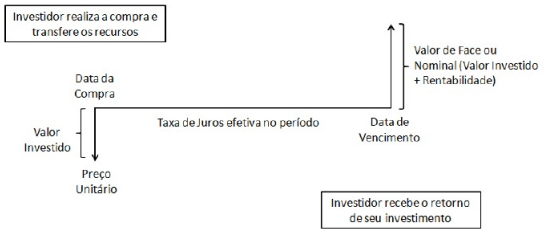

Figura 4.1: Fluxo de pagamento LTN - Fonte Tesouro Direto

4.1.2 Letra financeira do tesouro - LFT

A letra financeira do Tesouro é um título pós-fixado, em que sua rentabilidade segue a variação da taxa de juros básica da economia, mas comumente chamada de taxa Selic. O valor de resgate é dado pela variação desta taxa diária registrada entre a data de liquidação da compra e a data de vencimento do título, que pode ser acrescida de ágio ou deságio no momento da compra.

Este deságio é uma taxa acrescida à variação da taxa Selic para medir a rentabilidade de acordo com uma menor demanda pelas LFT. Na ocorrência deste deságio, o investidor recebe a Selic mais o valor do deságio.

Já o ágio da LFT é uma taxa deduzida à oscilação da taxa Selic para medir a rentabilidade do título de acordo com uma maior demanda pelas letras financeiras do tesouro. Neste caso o investidor recebe a Selic menos o ágio.

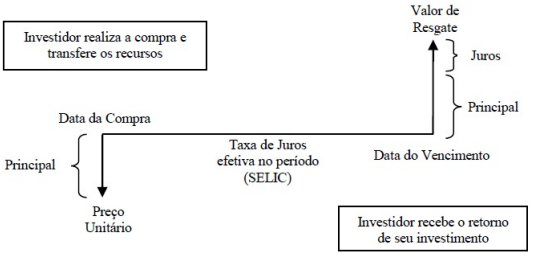

Figura 4.2: Fluxo de pagamento LFT - Fonte Tesouro Direto

Um dos objetivos em finanças é avaliar os riscos financeiros, e este risco é medido em termos de variações de preço. A variação de preços entre os instantes $t-1$ e t é dada por:

$$\Delta P = P_t - P_{t-1}. \tag{4.1}$$

Costuma-se usar os log-retornos, que iremos definir como $r_t = log\frac{P_t}{P_{t-1}}$ pois estes são livres de escala e possuem propriedades estatísticas mais significativas, em relação aos preços propriamente ditos.

4.2 Análise descritiva dos dados

Os dados foram extraídos do site do tesouro direto, porém os log-retornos do ano 2014 das letras do tesouro nacional e das letras financeiras do tesouro foram calculados no software R.

- Letra do Tesouro nacional - LTN

O gráfico do log-retornos:

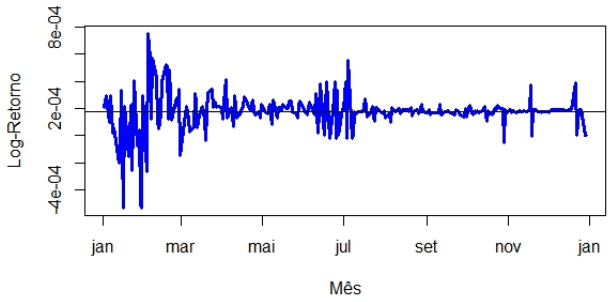

Figura 4.3: Gráfico dos log-retornos dos valores de venda de 2014 das letras do tesouro nacional

Observa-se que no meses de janeiro e fevereiro ocorreram os menores retornos possíveis, mas ao final de fevereiro observa-se uma alta desse retorno. Para uma análise mais profunda, verificaremos algumas estatísticas descritivas.

Tabela 4.1: Estatísticas descritivas dos log-retornos - LTN

Estatística	Valor
Mínimo	-0.000533
1º Quartil	0.000153
Média	0.000174
Mediana	0.000181
3º Quartil	0.000211
Máximo	0.000754
Assimetria	-1.074107
Curtose	7.513282

Temos que o valor mínimo do log-retorno é -0.000533, neste caso o investidor não estava possuindo rentabilidade, e o valor máximo do log-retorno é 0.00754, neste caso o retorno a rentabilidade era satisfatória.

Para ver os possíveis valores discrepantes desses dados, iremos observar o box-plot da distribuição dessas observações.

Observa-se que há muitos *outliers* nessas observações, pois há muitos pontos fora do limite determinado. As medidas de assimetria e curtose nos mostra como será distribuição desses log-retornos no ano de 2014.

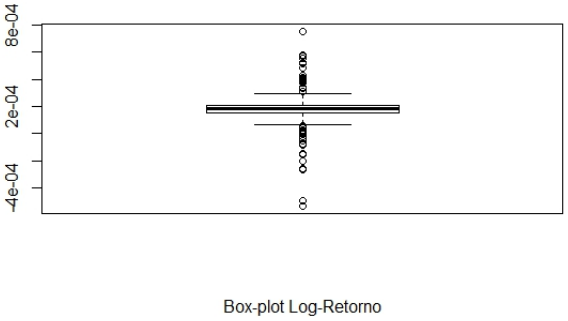

Figura 4.4: Blox-plot log-retornos - LFT

Pode-se observar que a assimetria é negativa, ou seja, as caudas da distribuição distinguem-se. Já a curtose como o valor é positivo e alto, mostra que a distribuição desses log-retornos tem o pico mais agudo. Portanto a distribuição dos log-retornos será:

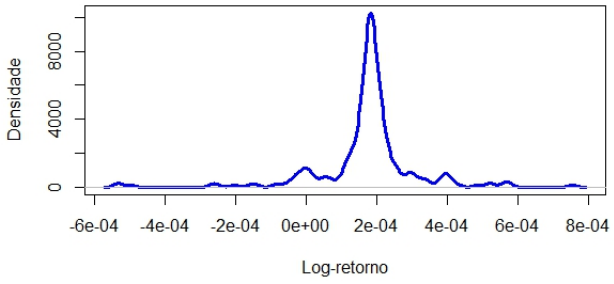

Figura 4.5: Densidade dos log-retornos da LTN em 2014

Como dito anteriormente, pelo gráfico que há uma diferença entre as caudas da distribuição e o pico desta é agudo, ou seja esses dados têm uma distribuição com cauda pesada, isto é comumente visto em dados financeiros.

- Letra Financeira do Tesouro - LFT

O gráfico do log-retornos:

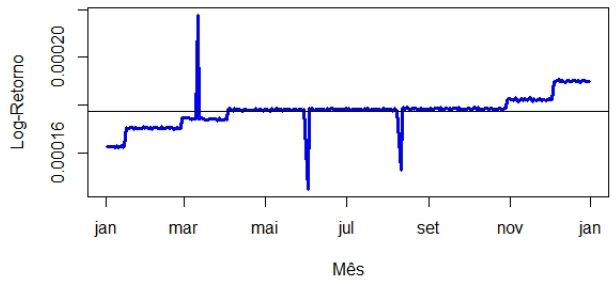

Figura 4.6: Gráfico dos log-retornos dos valores de venda de 2014 das letras financeiras do tesouro

Diferente dos dados da Letra do tesouro Nacional, observa-se que há poucas variações nos log-retornos da letras financeiras do tesouro.

No mês de março ocorreu a maior rentabilidade possível para o investidor, já no mês de junho ocorreu o menor retorno possível desta letra no ano de 2014. Verificaremos algumas estatísticas descritivas:

Tabela 4.2: Estatísticas descritivas dos log-retornos - LFT

Estatística	Valor
Mínimo	0.000145
1º Quartil	0.000175
Média	0.000178
Mediana	0.000178
3º Quartil	0.000179
Máximo	0.000218
Assimetria	0.12038
Curtose	7.876604

O valor mínimo do log-retorno é 0.00145 e o valor máximo do log-retorno é 0.000218, neste caso o retorno a rentabilidade para o investidor foi satisfatória. Para visualizar possíveis *outliers* nas observações, iremos observar o box-plot da distribuição desses dados.

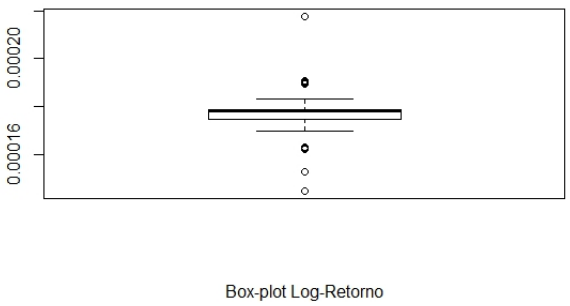

Figura 4.7: Blox-plot log-retornos - LFT

No boxplot a mediana não se encontra ao centro do caixa, isso nos dá indícios de que a distribuição desses dados não é simétrica, observa-se que a mediana está próxima ao 1^o quartil da distribuição, mostrando que a assimetria da distribuição é positiva, e verificamos isso através das estatísticas, em que a assimetria é 0.12038.

A curtose dessa distribuição é semelhante ao da LTN, pois possui valor positivo e alto, mostrando que a distribuição desses log-retornos tem o pico mais agudo. Então a distribuição dos log-retornos da Letra Financeira do tesouro será:

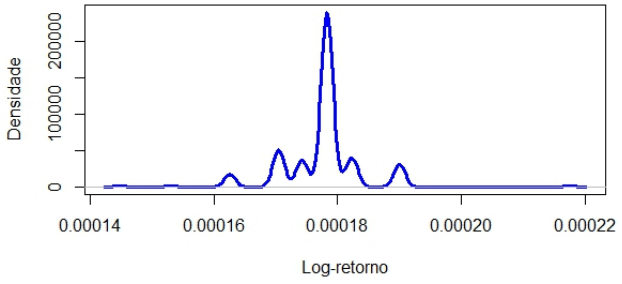

Figura 4.8: Densidade dos log-retornos da LTN em 2014

Portanto, a distribuição dos log-retornos das observações da LFT é distribuição leptocúrtica, com cauda pesada, e média igual a 0.000178.

4.3 Estimadores de locação

As tabelas abaixo irão mostrar o comportamento dos estimadores de locação nos dados do log-retornos da LTN e da LFT , primeiramente iremos comparar os estimadores de locação quando diminuímos o tamanho da amostra e a média geral dos log-retornos de 2014.

Consideraremos $n = 40$ o período 01/11/2014 a 30/12/2014, quando $n = 28$ compreenderá o período 20/11/2014 a 30/12/2014, para $n = 15$ o período será 30/11/2014 a 20/12/2014 e para $n = 7$ o intervalo será entre os dias 19/12/2014 a 29/12/2014. E a média(geral) calculada dos log-retornos compreende o período de 02/01/2014 a 31/12/2014.

Tabela 4.3: Estimação de locação - letras do tesouro nacional

Dias(n)	Bisquare	Huber	Trimédia	Mediana	R-estimador	Média	Média(Geral)
40	0.0001848	0.0001849	0.0001853	0.0001849	0.0001842	0.0001759	0.0001736
28	0.0001868	0.0001876	0.0001862	0.0001876	0.0001856	0.0001727	0.0001736
15	0.0001920	0.0001917	0.0001903	0.0001917	0.0001896	0.0001881	0.0001736
7	0.0001921	0.0001913	0.0001919	0.0001913	0.0001913	0.0001925	0.0001736

Quando n=40 dias as estimativas robustas estão se comportando de maneira semelhante, e a estimativa clássica (média das observações) está se diferenciando. Observa-se no gráfico (4.3) que há valores discrepantes neste intervalo, o que ocasiona essa diferença significativa. Para n=7, nota-se que há uma desconformidade entre os estimadores robustos e a média geral do ano de 2014, este fato nos mostra que os *outliers* estão interferindo na média geral.

Para as letras financeiras do tesouro as observações serão $n = 39$ que compreenderá o período 01/03/2014 a 01/05/2014, quando $n = 26$ consideraremos o período 01/03/2014 a 10/04/2014, para $n = 11$ o período será 01/03/2014 a 20/03/2014 e para $n = 5$ o intervalo será entre os dias 07/03/2014 a 17/03/2014. E a média(geral) calculada dos log-retornos compreende o período de 02/01/2014 a 31/12/2014.

Tabela 4.4: Estimação de locação - letra financeira do tesouro

Dias(n)	Bisquare	Huber	Trimédia	Mediana	R-estimador	Média	Média(Geral)
39	0.0001706	0.0001748	0.0001755	0.0001748	0.0001762	0.0001771	0.0001776
26	0.0001742	0.0001744	0.0001744	0.0001744	0.0001744	0.0001766	0.0001776
11	0.0001742	0.0001742	0.0001742	0.0001742	0.0001742	0.0001781	0.0001776
5	0.0001743	0.0001743	0.0001743	0.0001744	0.0001744	0.0001830	0.0001776

Tendo em vista as estimativas de locação para a letra financeira do tesouro, observa-se que para n=252, ou seja, todas as observações de 2014, a média não se mostra significantemente desigual quando o tamanho da amostra se reduz até n=11. Já quando n=5 no gráfico (4.6) observa-se que há *outliers*, e percebemos que há uma diferença expressiva entre as estimativas robustas e a estimativa clássica.

4.4 Estimadores de escala

Assim como analisamos os estimadores de locação para os log-retornos da letra do tesouro nacional e da letra financeira do tesouro, iremos observar o comportamento dos estimadores robustos e estimador clássico de escala. Considerando os períodos mencionados no estimador de locação.

Tabela 4.5: Estimação de escala - letra do tesouro nacional

Dias(n)	Tukey-bisquare	Huber	MAD	Desvio-padrão	Desvio-padrão(Geral)
42	$1.00 * 10^{-5}$	$1.01 * 10^{-5}$	$6.75 * 10^{-6}$	$7.31 * 10^{-5}$	0.00014
28	$7.21 * 10^{-6}$	$7.36 * 10^{-5}$	$4.86 * 10^{-6}$	$7.36 * 10^{-5}$	0.00014
15	$1.12 * 10^{-6}$	$4.77 * 10^{-6}$	$7.57 * 10^{-7}$	$6.29 * 10^{-6}$	0.00014
7	0.00014	0.00018	$9.56 * 10^{-5}$	0.00015	0.00014

Percebe-se que pelo desvio-padrão da série, há uma variação maior entre esses dados, já os estimadores robustos mostram que essa variação é pequena, neste caso, como dito anteriormente, nessas observações há muitos *outliers*, o que ocasiona uma estimativa tendenciosa. Para a amostra de tamanho n=7 , observa-se que há uma diferença significativa no estimador robusto MAD para os demais estimadores robustos. O estimador MAD é determinado pelo desvio absoluto entre a observação e a mediana, neste caso a valores próximo da mediana, acarretando pequenos desvios.

Tabela 4.6: Estimação de escala - letra financeira do tesouro

Dias(n)	Tukey-bisquare	Huber	MAD	Desvio-padrão	Desvio-padrão(Geral)
39	$2.43 * 10^{-6}$	$2.03 * 10^{-6}$	$1.37 * 10^{-6}$	$6.94 * 10^{-6}$	$6.53 * 10^{-6}$
26	$5.87 * 10^{-6}$	$5.22 * 10^{-6}$	$3.96 * 10^{-7}$	$8.51 * 10^{-6}$	$6.53 * 10^{-6}$
11	$4.13 * 10^{-7}$	$4.38 * 10^{-7}$	$2.79 * 10^{-7}$	$1.31 * 10^{-5}$	$6.53 * 10^{-6}$
5	$3.38 * 10^{-7}$	$6.64 * 10^{-7}$	$2.28 * 10^{-7}$	$1.93 * 10^{-5}$	$6.53 * 10^{-6}$

Pela tabela 4.6, podemos analisar que não há uma grande diferença entre os estimadores de escala. Mas pode-se notar que quando a amostra é pequena, neste caso n=5, o estimador clássico mostra-se mais expressivo que os demais estimadores robustos. Isso pode ser pelo mesmo fato do estimador de locação, pois existem *outliers* neste intervalo, ocasionando assim uma maior variação.

4.5 Estimação do modelo de regressão linear simples

A modelagem dos log-retornos da LTN e da LFT será feita através de estimadores clássicos e estimadores robustos, em ambos os casos apenas com apenas uma variável explicativa. O comportamento dessas letras do tesouro se distinguem, pois existem diferentes indexadores. Como o objetivo é analisar esses log-retornos de maneira análoga, por este motivo a variável explicativa será o tempo.

4.5.1 Método de mínimos quadrados ordinários

- Letra do tesouro nacional

O modelo inicialmente sugerido será da forma:

$$\hat{y}_i = b_0 + b_1 x_i. \tag{4.2}$$

Em que $\hat{y}_i$ serão os valores ajustados dos log-retornos das letras do tesouro e x_i é a variável resposta que neste caso será o tempo. Para o modelo de regressão utilizar uma amostra, e esta compreenderá entre 01/11/2014 e 31/12/2014. Portanto as estimativas dos parâmetros serão:

Tabela 4.7: Estimação mínimos quadrados ordinários (MQO) - LTN

Coeficientes	Estimativas	Erro Padrão	Valor t	P-valor
Intercepto	$5.396{*}10^{-3}$	$1.100 * 10^{-2}$	0.49	0.6266
Dias	$-3.180 * 10^{-7}$	$6.706 * 10^{-7}$	-0.47	0.6381

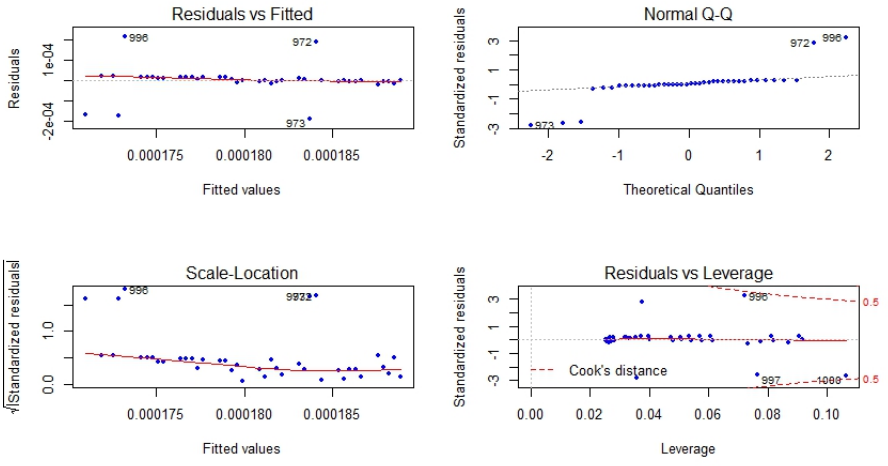

Figura 4.9: Gráficos de diagnóstico do modelo de regressão (MQO) - LTN

Os resultados da Tabela (4.7) foram obtidos considerando o nível de confiança de 95%. Pelo gráfico QxQ verifica-se que os os resíduos não seguem um distribuição Normal. Para uma melhor

análise foi feito o teste de Shapiro - Wilk, em que a estatística do teste foi 0.6411 e o seu p-valor associado foi de $1.194 * 10^{-8}$. Com esses resultados a hipótese nula é rejeitada para um nível de significância de 5%.

Pelos gráficos de diagnósticos (4.9) as observações #996, #972 e #973 são valores discrepantes, para verificarmos qual a influência que esses valores causam na estimação dos parâmetros do modelo, iremos retirá-los e fazer novamente a regressão.

Tabela 4.8: Estimação mínimos quadrados ordinários (MQO) / sem *outliers* - LTN

Coeficientes	Estimativas	Erro Padrão	Valor t	P-valor
Intercepto	-0.0038	0.0012	-3.13	0.00362
Dias	$2.424 * 10^{-7}$	7.379	3.28	0.00242

Os resultados da Tabela (4.8) foram obtidos considerando o nível de significância de 5%. Nota-se uma diferença significativa entre parâmetros comparado ao modelo (4.7), ou seja, os *outliers* estão influenciando na estimação desses parâmetros. Ao contrário do modelo (4.7), os parâmetros estão sendo significativos ao nível de 5% de confiança. Abaixo iremos observar o gráfico QxQ dos resíduos, e podemos observar que estes estão próximos da normalidade em relação ao modelo da Tabela (4.7):

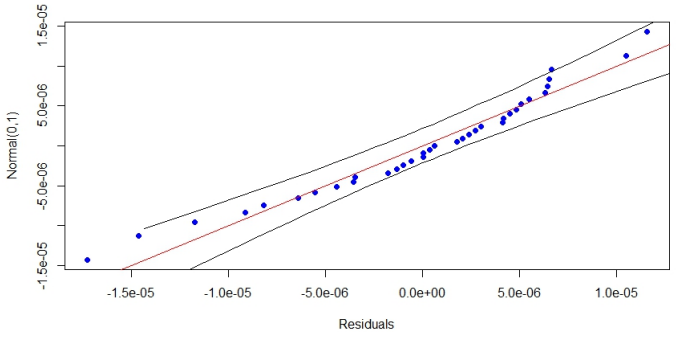

Figura 4.10: Gráfico QxQ dos resíduos do modelo de regressão / sem *outliers* - LTN

4.5.2 Métodos robustos

A fim de diminuir o impacto dessas observações na estimação, iremos utilizar o critério de Huber e Tukey-bisquare para ajustar o modelo (4.2).

A estimativas dos parâmetros considerando $k = 1.345$ para o M-estimador de Huber e $h = 4.865$ para o M-estimador de Tukey-bisquare :

Tabela 4.9: M-estimador (Huber) - LTN

Coeficientes	Estimativas	Erro Padrão	Valor t
Intercepto	-0.0030	0.0014	-2.1794
Dias	$1.956469 * 10^{-7}$	$1.1 * 10^{-11}$	2.3126

Tabela 4.10: M-estimador (Tukey-bisquare) - LTN

Coeficientes	Estimativas	Erro Padrão	Valor t
Intercepto	-0.0038	0.0011	-3.3227
Dias	$2.425087 * 10^{-7}$	$1 * 11 * 10^{-11}$	3.4851

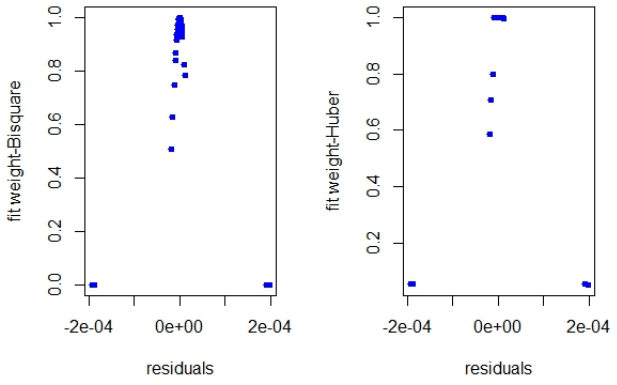

Figura 4.11: Gráficos dos pesos dos modelos robustos - LTN

As estimativas dos parâmetros β_0 e β_1 são próximas para os dois métodos robustos. Então o diagnóstico para esses métodos serão semelhantes, por este fato, analisaremos o M- estimador de Huber.

Os *outliers* tiveram grande influência na estimativa do MQO, pois a sua eliminação causa variação desproporcional na estimativa do β_0. Devido ao fato dos métodos robustos serem interativos, ocorrem seis interações no M-estimador de Huber e cinco interações no M-estimador de Tukey-bisquare para alcançar um modelo adequado.

No gráfico de pesos observa-se que as observações próximas a zero tem o maior peso, e há poucas observações que não possuem peso na estimação.

Pelos gráficos de diagnósticos, não há muitos valores influentes para a estimação robusta como observa-se nos gráficos de diagnósticos na estimação clássica.

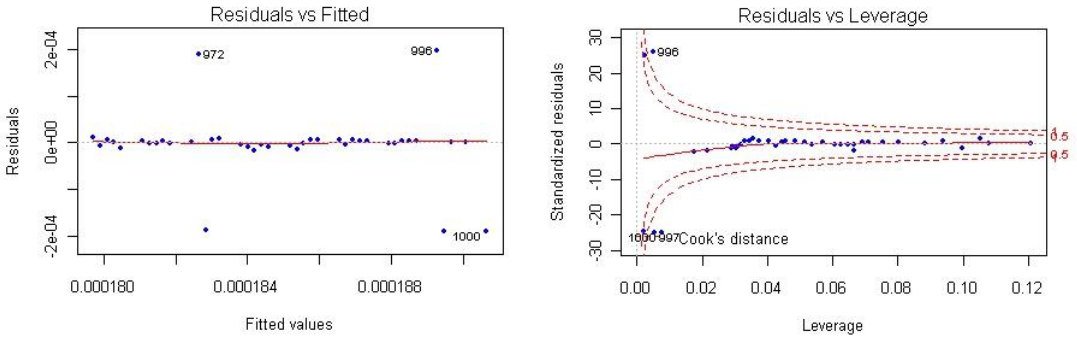

Figura 4.12: Gráficos de diagnóstico de resíduos do modelo robusto - LTN

Portanto o melhor modelo de regressão para esses dados é o robusto, pois as observações atípicas recebem o menor peso, melhorando de forma significativa o ajuste do modelo. Então investidor terá um log-retorno de $1.956469 * 10^{-7}$ a cada dia que seu valor inicial de compra estiver investido.

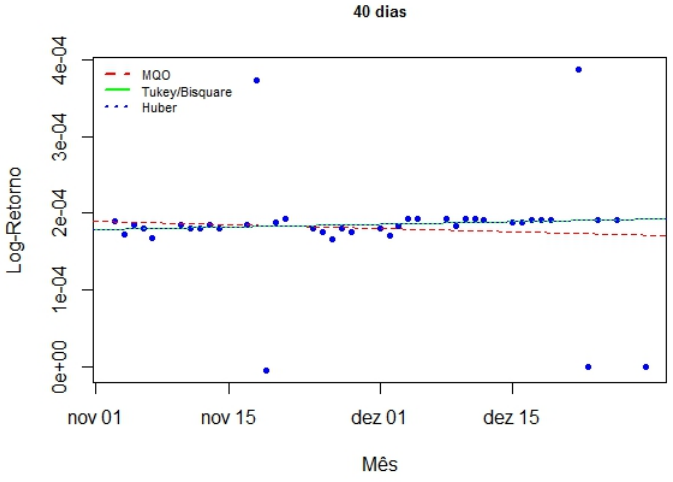

Figura 4.13: Gráfico dos métodos de regressão - LTN

Pelo gráfico as observações finais estão influenciando na reta de regressão, deixando a reta mais inclinada. Já os métodos robustos dão menos pesos para esses *outliers*, mostrando ter um melhor resultado na resistência desses valores e acompanhando a tendência dos valores restantes.

- Letra financeira do tesouro

Considerando o modelo (4.2) de regressão dos log-retornos da LFT. Para esse modelo utilizaremos uma amostra em que é situado o período 01/03/2014 e 01/05/2014. As estimativas dos parâmetros serão:

Tabela 4.11: Estimação mínimos quadrados ordinários (MQO) - LFT

Coeficientes	Estimativas	Erro Padrão	Valor t	p-valor
Intercepto	-4.503*10^{-5}	$1.102 * 10^{-}3$	-0.041	0.968
Dias	$1.375 * 10^{-8}$	$6.817 * 10^{-8}$	0.202	0.8413

Esses resultados foram obtidos considerando o nível de confiança de 95%. Os parâmetros não são significativos, para saber se os valores atípicos estão influenciando será feita a análise de resíduos.

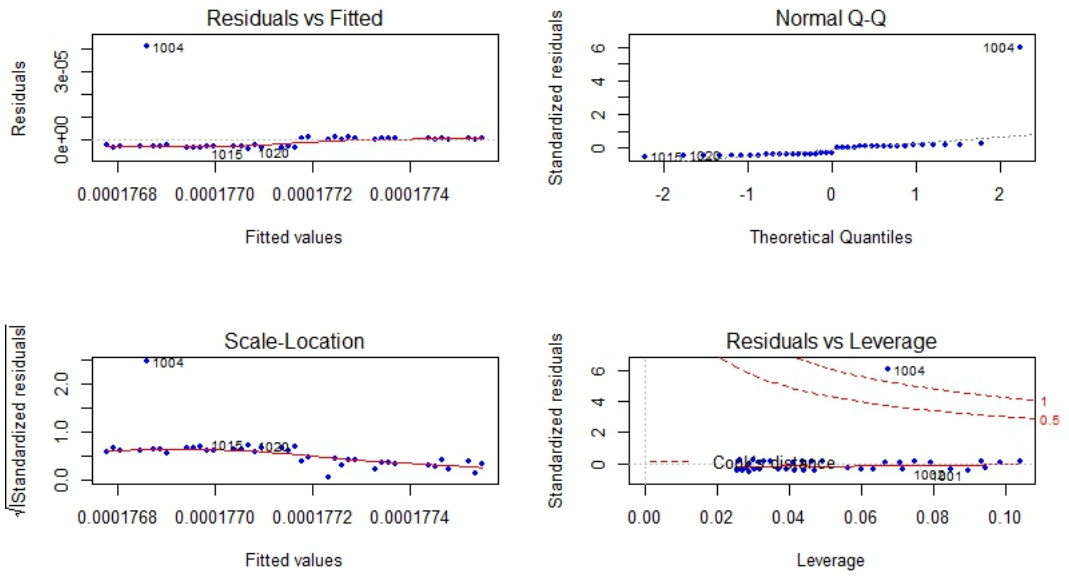

Figura 4.14: Gráficos de dignóstico do modelo de regressão (MQO) - LFT

O p-valor do teste de Shapiro-Wilk foi de aproximadamente $5.981 * 10^{-12}$. Dessa forma, para um nível de significância de 5% existem evidências suficientes para afirmar que os resíduos do modelo não seguem uma distribuição Normal. Isto pode ser visto no gráfico QxQ, nas caudas não se observa-se uma linearidade.

Pelo gráfico de resíduos "versus"valores preditos há valores atípicos, e alguns estão influenciando nas estimativas dos parâmetros, neste caso retiraremos esses valores e analisaremos se há diferença significativa na estimação dos parâmetros do modelo.

Tabela 4.12: Estimação mínimos quadrados ordinários (MQO) / sem *outliers* - LFT

Coeficientes	Estimativas	Erro Padrão	Valor t	p-valor
Intercepto	$-1.426 * 10^{-3}$	$1.676 * 10^{-4}$	-8.508	$6.161 * 10^{-10}$
Dias	$9.915 * 10^{-8}$	$1.037 * 10^{-8}$	9.559	$3.451 * 10^{-11}$

Os resultados da Tabela (4.12) foram obtidos considerando o nível de confiança de 95%. Os *outliers* estão influenciando na estimação desses parâmetros, pois há uma diferença entre os parâmetros do modelo de regressão. Neste novo modelo sem os valores atípicos, a covariável é significante ao nível de 5% de significância. No gráfico QxQ de resíduos, observa-se que estes estão próximos da normalidade em relação ao modelo da Tabela (4.11):

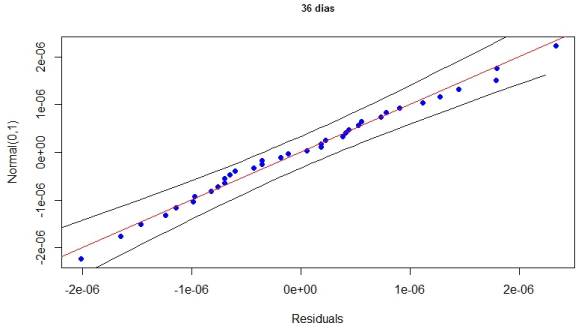

Figura 4.15: Gráfico QxQ dos resíduos do modelo de regressão / sem *outliers* - LFT

Como feito nos log-retornos das LTN iremos aplicar a estimação robusta no modelo de regressão, com o objetivo de diminuir o impacto desses valores atípicos na estimação desses parâmetros. As estimativas dos parâmetros será considerando k = 1.345 para o M-estimador de Huber e h = 4.865 para o M-estimador de Tukey-bisquare.

Tabela 4.13: M-estimador (Huber)- LFT

Coeficientes	Estimativas	Erro Padrão	Valor t
Intercepto	$-1.380 * 10^{-3}$	0.0002	-6.8609
Dias	$9.632 * 10^{-8}$	1.12^{-11}	7.736

Tabela 4.14: M-estimador (Tukey-bisquare) - LFT

Coeficientes	Estimativas	Erro Padrão	Valor t
Intercepto	$-1.428 * 10^{-3}$	0.0002	-7.3265
Dias	$9.928 * 10^{-8}$	1.1^{-11}	8.2292

Não há uma diferença significativa entre as estimativas dos parâmetros e os valores preditos entre a estimativa clássica sem os valores discrepantes e a estimativa robusta. Ocorreram cinco interações para os dois métodos robustos citados, para a obtenção de um modelo apropriado na presença desses valores atípicos.

No gráfico dos pesos dos estimadores robustos, verifica-se que os maiores pesos localizam-se nos resíduos mais próximo de zero. O estimador de Bisquare mostra que o peso adequado para a estimação está entre 0.8 e 1, na qual verificamos uma maior concentração de resíduos.

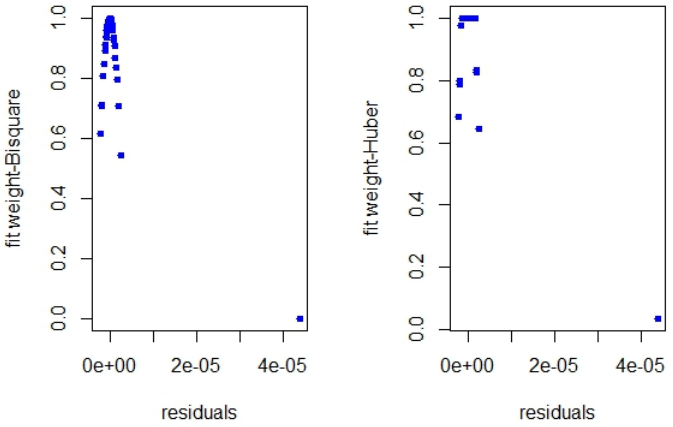

Figura 4.16: Gráficos dos pesos dos modelos robustos - LFT

Como a estimação robusta está tendo um comportamento semelhante para os estimadores robustos mencionados, então daremos ênfase na análise de resíduos do estimador de Huber.

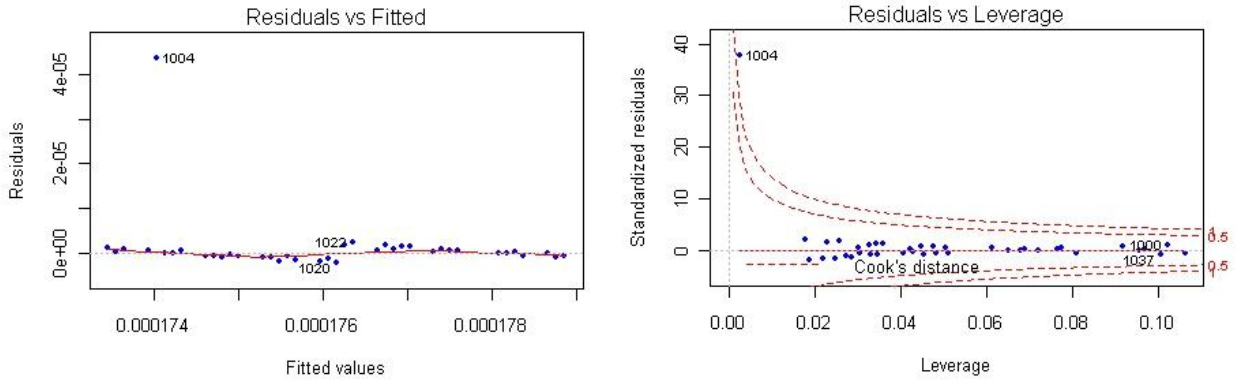

Figura 4.17: Gráficos de diagnóstico de resíduos do modelo robusto - LFT

Pelos gráficos de diagnóstico do modelo de regressão robusta, ocorre ainda valores influentes, ou seja, os métodos robustos parecem ser sensíveis a pontos pontos influentes como mostram os gráficos de resíduos.

A observação #1004 é um *outlier*, porém na estimação robusta ele não mostra ser tão influente como no regressão clássica.

A estimação robusta é mais eficiente neste caso, pois a estimação clássica mostrou-se vulnerável a valores atípicos, então o investidor terá um log-retorno de $9.632 * 10^{-8}$ a cada dia que seu "principal"estiver investido.

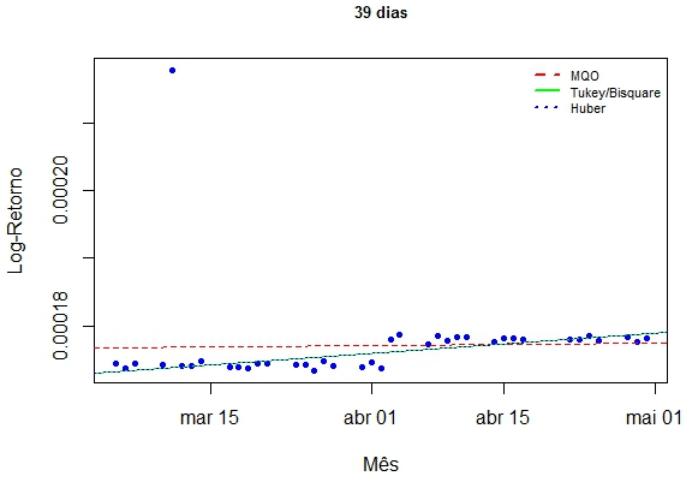

Figura 4.18: Gráfico dos métodos de regressão - LFT

Através do gráfico é possível ter uma visualização de como o método de mínimos quadrados é influenciado por valores discrepantes, este fato acontece pois este método atribui peso unitário a todas observações ocasionando assim essa desproporção.

O M-estimador mostra-se eficaz nestes casos, pois observações atípicas recebem o menor peso. A estimação pelo MQO sem esses valores anômalos tem o mesmo comportamento que os estimadores robustos, porém em dados financeiro temos que levar em conta todas os observações, concluindo que o melhor modelo que se adéqua aos dados é o modelo estimado pelos métodos robustos.

Capítulo 5

Considerações Finais

Os métodos robustos foram aplicados em duas modalidades distintas de títulos de tesouro, a letra do tesouro nacional e a letra financeira do tesouro. Porém a abordagem na estimação foi semelhante para ambos títulos.

O principal objetivo deste estudo foi observar a diferença que os *outliers* ocasionam nas estimativas clássicas, uma vez que dependendo dos casos, esses valores não podem ser retirados do estudo, como por exemplo em finanças.

Na estimação do modelo locação, as estimativas clássicas tiveram um comportamento diferente quando houve uma maior sinalização dos valores atípicos, já para o modelo de escala ocorreu uma sútil desconformidade entre esses estimadores.

No modelo de regressão utilizando o método M-estimador robusto mostrou-se adequado para os log-retornos dos títulos do tesouro, pois as estimativas dos métodos de mínimos quadrados são sensíveis a presença dos valores atípicos, na qual no diagnóstico do modelo, observou-se a influência que esses valores.

Assim, este estudo mostrou que o M-estimador robusto, generalização do método do estimador de máxima verossimilhança, é eficaz quando há presença de *outliers*, pois reduz a sensibilidade das estimativas.

Apêndice A

Propriedades do Método de Mínimos Quadrados

Teorema de Gauss Markov: Sob os pressupostos do Modelo de Regressão linear, os estimadores obtidos pelo Método de Mínimos Quadrados (MQO) b_0 e b_1 possuem menor variância e são os melhores estimadores lineares não tendenciosos.

Um estimador é dito linear quando:

$$\hat{\theta} = \sum_{i=1}^{n} k_i y_i. \tag{A.1}$$

onde $k_i \in \Re$.

Portanto os estimadores b_0 e b_1 :

1. são combinações lineares dos $y_i's$.
2. são não tendenciosos e não viesados.
3. são consistentes.
4. têm variância mínima.
5. são suficientes.

Apêndice B

Cálculo dos estimadores Robustos

Reservamos este apêndice para mostrar em algoritmo o cálculo dos estimadores robustos.

B.1 Cálculo estimador locação

- M-estimador

1. Ordenar as observações $x_1, ..., x_n$.
2. Calcular o estimador inicial $T_0 = MED(x_1, ..., x_n)$.
3. Para cada observação calcular os pesos $W_i = \frac{\psi(x_i - T(m))}{x_i - T(m)}$.
4. Definir $T(m+1) = \frac{\sum_{i=1}^{n} W_i * x_i}{\sum_{i=1}^{n} W_i}$.
5. Se $|T(m+1) - T(m)| \geq \epsilon\,|T(m+1)|$ com ϵ pré-fixado, assumindo $\epsilon = 0.001$, fazer $T_n = T(m+1)$ e parar.
6. Se $|T(m+1) - T(m)| \geq \epsilon\,|T(m+1)|$ e $m+1 < 20$ (em que 20 é o número máximo de interações permitidas), então fazer m=m+1 e voltar ao passo 3.
7. Se $|T(m+1) - T(m)| \geq \epsilon\,|T(m+1)|$ e $m+1 = 20$, o algoritmo não convergiu, mas terá que ser feito $T_n = T(m+1)$.

- L-estimador

1. Ordenar as observações $x_1, ..., x_n$.
2. Calcular os a_i.
3. Fazer $L_n = \sum_{i=1}^{n} a_i * x_i$, neste caso as observações já estão ordenadas de modo que $x_1 \leq ... \leq x_n$.

- R-estimador

1. Ordenar as observações $x_1, ..., x_n$.

2. Calcular o estimador inicial $T_0 = MED(x_1...x_n)$.
3. Seja $Z_i = x_i$ e $Z_{n+i} = 2 * T(m) - x_i$ para $i = 1, ..., n$.
4. Ordenar as observações $Z_1, ..., Z_{2n}$, de modo que $Z_1 \leq ... \leq Z_{2n}$.
5. Para cada $i = 1, ..., n$ calcular R_i^* que é o posto de x_i em $(Z_1, ..., Z_{2n})$.
6. Para cada $i = 1, ..., n$ calcular $W_i = \frac{a_n(R_i^*)}{x_i - T(m)}$.
7. Definir $T(m+1) = \frac{\sum_{i=1}^{n} W_i^* * x_i}{\sum_{i=1}^{n} W_i^*}$.
8. Se $|T(m+1) - T(m)| \geq \epsilon\,|T(m+1)|$ com ϵ pré-fixado, assumindo $\epsilon = 0.001$, fazer $T_n = T(m+1)$ e parar.
9. Se $|T(m+1) - T(m)| \geq \epsilon\,|T(m+1)|$ e $m + 1 < 20$ (em que 20 é o número máximo de interações permitidas), então fazer m=m+1 e voltar ao passo 3.
10. Se $|T(m+1) - T(m)| \geq \epsilon\,|T(m+1)|$ e $m + 1 = 20$, o algoritmo não convergiu, mas terá que ser feito $T_n = T(m+1)$.

B.2 Cálculo M-estimador escala

1. Calcular o estimador inicial $S_0 = MAD(x_1, ..., x_n)$.
2. Para cada observação calcular o peso $w_{k,i} = W(\frac{x_i}{\hat{\sigma_k}})$, em que W é a função peso.
3. Definir $\hat{\sigma_{k+1}} = \sqrt{\frac{1}{n*\delta} * \sum_{i=1}^{n} w_{k,i} * x_i^2}$.
4. Parar quando $|\hat{\sigma_{k+1}}/\hat{\sigma_k} - 1| \geq \epsilon$.

B.3 Cálculo M-estimador regressão

1. Selecionar estimativas iniciais de $b_{i's}$, como por exemplo as estimativas por mínimos quadrados.
2. A cada interação calcula-se os resíduos e os pesos associados $w_i = W(e_i)$.
3. Calcular as novas estimativas de mínimos quadrados ponderados.
4. Repetir os passos 1 e 2 até o algoritmo convergir.

Referências Bibliográficas

[1] Bustos Oscar, *Estimação Robusta no Modelo de Posição*, Rio de Janeiro, 1981.

[2] Casella, G. e Berger, R. L., *Statistical Inference*, Ceange Learning, Tradução da 2ª edição Norte Americana, Brasil, 2014.

[3] Entenda o mercado de Títulos Públicos, Brasília, 2015, Disponível em: http://www.tesouro.fazenda.gov.br. Acesso em: 30 mai. 2015.

[4] Hampel, R.F *Contribuitions to theory of robust estimation*, PhD Thesis, University of California, Berkeley, 1968.

[5] Huber, P.J., *Robust estimation of location parameter*, Ann. Math. Statist, 1964.

[6] Huber, P.J., *Robust Statistics: A Review*, Ann. Math. Statist, 1972.

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

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

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

[10] Ronchetti,E. *The historical development of robust statistics*, 2006.

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

[12] Seo, Songwon *A review and comparison of methods for detecting outliers in univariate datasets*, Kyunghe e University, 2002.

[13] Staudte, G.R., Sheather,J.S *Robust Estimation e Testing*, John Wiley & Sons, Inc., Canadá, 1990.

[14] Yu.C, Yao.W e Bai.X *Robust Linear Regression: A Review and comparison*, Kansas States University, Manhattan, Kansas, USA, 2014