Universidade Federal da Paraíba
Centro de Ciências Exatas e da Natureza
Departamento de Estatística

Saul de Azevêdo Souza

Comparação dos Estimadores Robustos e de Mínimos Quadrados Ordinários

João Pessoa, 20 de fevereiro de 2015

Saul de Azevêdo Souza

Avaliação dos Estimadores Robustos e de Mínimos Quadrados Ordinários

Monografia apresentada ao Curso de Graduação em Estatística da Universidade Federal da Paraíba, conforme exigência acadêmica parcial para obtenção do Grau de Bacharel. Área de Concentração: Estatística Aplicada.

Aprovada em 20 de fevereiro de 2015.

BANCA EXAMINADORA

Prof. Dra. TATIENE CORREIA DE SOUZA - Orientadora
UFPB

Prof. Dr. LUIZ MEDEIROS DE ARAÚJO LIMA FILHO
UFPB

Prof. Dra. MARIA LÍDIA COCO TERRA
UFPB

Este trabalho é carinhosamente dedicado
aos meus pais, Sílvio e Maria, e ao meu irmão Sílvio Jr.

Agradecimentos

Agradeço primeiramente à DEUS, pelo seu cuidado para comigo, guiando e conduzindo-me para o melhor caminho. Por me proteger, ajudar-me a tomar as decisões e enfrentar os obstáculos que a vida me propôs.

Aos meus pais, Sílvio e Maria, pelo apoio, cuidado e dedicação, aos quais dedico todas as minhas conquistas.

Ao meu irmão Sílvio Jr., aos meus avós Lourival e Herenilde, aos meus tios, Sátiro e Sérgio, a minha tia Marluce, e aos meus primos, Sátiro, Samantha, Sarah, Aline e Sílvia, pelo apoio, incentivo, amizade e momentos de descontração.

Aos meus colegas de curso, Jodavid, Andreza, Alisson, Henrique, Marina, Maizza, Aldine e Michelle, pela amizade, companheirismo, descontração, incentivo e paciência, tornando aquele local de estudo mais agradável.

A professora Tatiene, pela orientação constante, amizade, confiança, ensinamentos, paciência e respeito, jamais me levando ao estresse mesmo quando estive em profundo desespero.

Ao professor Hemílio, pela confiança, amizade, conselhos e oportunidade nos projetos. Por engrandecer meus conhecimentos ao longo desses anos e ser paciente quando eu tinha dúvidas.

Aos Professores, Ana Flávia, Hemílio, Sydnei, Ronei, Tatiene, João Agnaldo e Andrea, pelos conselhos, momentos de descontração, por serem excelentes professores e amigos.

A todos os Professores do DE-UPPB, por contribuírem para minha formação acadêmica.

A todos os funcionários do Departamento de Estatística.

Ao CNPq, pelo apoio financeiro.

A vitória pertence aquele que acredita nela por mais tempo.

(Pearl Harbor)

Resumo

A regressão robusta é uma técnica que permite obter estimativas mais seguras quando os dados apresentam *outliers* ou pontos de alavanca. Segundo Draper & Smith (1998) o método dos mínimos quadrados ordinários atribui pesos iguais a cada observação na obtenção dos parâmetros ao contrário da regressão robusta que atribui a cada observação uma ponderação diferente, essencialmente observações que produzem grandes resíduos recebem menores pesos pelo método de estimação robusto, ocasionando uma menor influência dessas observações atípicas nas estimativas dos parâmetros. Os estimadores robustos *least median of square*, denotado por LMS, e *least trimmed square*, denotado por LTS, propostos por Rousseeuw (1984), assumem que se até 50% dos dados apresentarem pontos de alavanca ou *outliers* é possível ainda se ter boas estimativas. O objetivo dessa monografia é comparar os estimadores robustos e de mínimos quadrados ordinários. Para a avaliação desses estimadores foram realizadas simulações de Monte Carlo considerando cenários sob homoscedasticidade e heteroscedasticidade, cenários balanceado e não-balanceado. Além das simulações de Monte Carlo foi realizada uma aplicação com dados extraídos de Greene (1997, Tabela 12.1, p.541), disponível no pacote *sandwich* do *software* estatístico `R` (Kleiber & Zeileis, 2008). Para a identificação dos pontos de influência e de alavanca foram consideradas respectivamente medidas de influência como distância de Cook e matriz de alavancagem, além da análise gráfica que permite visualizar a distribuição dos dados. A suposição de homoscedasticidade foi verificada através do teste de Koenker (1978) e foi observada a presença de heteroscedasticidade, sendo assim, o estimador usual da matriz de covariâncias deve ser substituído por outros estimadores consistentes. Por fim, verificamos que o estimador LTS apresentou o melhor ajuste de reta, considerando o cenário sob heteroscedasticidade e com pontos de alavanca.

Palavras-chave: Estimador de mínimos quadrados ordinários, Matriz de covariâncias, *Outliers*, Pontos de alavanca, Estimadores robustos.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A regressão linear é uma ferramenta utilizada quando se deseja analisar os impactos de variáveis independentes denotadas por X sobre a variável resposta denotada por Y. A princípio o estimador de mínimos quadrados ordinários, denotado por MQO, é o método de estimação mais utilizado quando pretendemos estimar os parâmetros desconhecidos do modelo de regressão, contudo as estimativas do MQO são facilmente influenciadas pela presença de pontos de alavanca, sendo necessário o uso de técnicas mais robustas para se obter melhores estimativas. A regressão robusta é uma técnica que permite obter estimativas mais seguras quando os dados apresentam *outliers* ou pontos de alavanca. Os *outliers* são observações extremas com relação a variável resposta Y, já os pontos de alavanca são relacionados a variável independente X. A identificação dessas observações pode ser feita por meio de técnica gráfica como *boxplot* ou através de medidas de influência como distância de Cook e valores da matriz de alavancagem. Na busca por outros estimadores mais robustos, Rousseeuw (1984) propôs um estimador que era capaz de suportar até 50% dos dados contaminados por observações extremas, denotado por *least median of squares* (LMS), esse estimador foi construído com base na mediana dos resíduos do MQO, porém o LMS tinha a desvantagem de apresentar uma taxa de convergência pequena de $n^{-1/3}$ sendo ineficiente com erros normalmente distribuídos, para resolver esse problema Rousseeuw (1984) propôs outro estimador com as mesmas qualidades do LMS, denotado por *least trimmed squares* (LTS), com taxa de convergência igual $n^{-1/2}$, ou seja, o LTS converge mais rápido que o LMS, contudo ainda sendo pouco eficiente com erros normalmente distribuídos. Esses métodos foram introduzidos por Rousseeuw como uma alternativa ao MQO quando este não produz estimativas tão seguras devido a influência de observações

extremas.

Segundo Draper & Smith (1998) o método dos mínimos quadrados ordinários atribui pesos iguais a cada observação na obtenção dos parâmetros, já a regressão robusta atribui a cada observação uma ponderação desigual, essencialmente observações que produzem grandes resíduos recebem menores pesos pelo método de estimação robusto, ocasionando uma menor influência dessas observações atípicas nas estimativas dos parâmetros. Portanto a regressão robusta é uma alternativa para o método dos mínimos quadrados ordinários quando a distribuição dos erros não é normal ou quando existe *outliers* que afetem a equação da reta (Ryan, 2009). Quando consideramos a suposição de homoscedasticidade, variância constante dos erros, segundo o Teorema de Gauss-Markov, o MQO é o estimador de variância mínima, ou seja, apresenta as melhores estimativas entre todos os possíveis estimadores do parâmetro desconhecido, entretanto quando essa suposição é violada o estimador usual da matriz de covariâncias do MQO torna-se viesado e inconsistente, tornando pouco confiáveis estimativas intervalares e testes de hipóteses que utilizam tais valores (Souza, 2003). Para resolver esse problema em 1980, Halbert White, um econometrista, propôs um estimador consistente para a matriz de covariâncias denotado por HC0, porém este estimador pode apresentar alguns problemas quando o tamanho da amostra é pequeno, então para contorna esse problema outros estimadores foram desenvolvidos com base no estimador de White e posteriores, a saber: HC1 (Hinkley, 1977), HC2 (Horn et. al, 1975), HC3 (Davidson & MacKinnon, 1993), HC4 (Cribari-Neto & Zarkos, 2004) e HC5 (Cribari-Neto et. al, 2007).

Capítulo 2

Objetivos

Esta monografia tem como objetivo comparar os estimadores robustos e de mínimos quadrados ordinários quando estes são utilizados para realizar estimativas em situações que os dados apresentam pontos de alavanca, sob homoscedasticidade (variância constante dos erros) e heteroscedasticidade. A comparação desses estimadores foi realizada por meio das simulações de Monte Carlo considerando quatro tipos de cenário, a saber, o primeiro cenário, cenário 1, sob homoscedasticidade e sem pontos de alavanca (balanceado); o segundo cenário, cenário 2, sob homoscedasticidade e com pontos de alavanca (não-balanceado). O terceiro e o quarto cenário foram gerados sob heteroscedasticidade, sendo o cenário 3 balanceado, e por fim, o cenário 4, não-balanceado. Para cada um desses cenários foram ajustados três modelos da forma $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$, $i = 1, ..., n$, em que o erro, ε, foi obtido a partir de três distribuições diferentes, a saber, $\mathcal{N}(0, 1)$, χ^2_2 e t_3 e as variáveis independentes foram obtidas a partir da distribuição $\mathrm{U}(0, 1)$. Então por meio do viés e erro quadrático médio foram feitas todas as comparações entre os estimadores robustos e de mínimos quadrados ordinários.

Capítulo 3

Referencial teórico

Neste capítulo será apresentado as técnicas de regressão linear e o estimador de mínimos quadrados ordinários, apresentando suas características e definições, justificando a necessidade de se utilizar os estimadores robustos em certas situações como presença de *outliers* e pontos de alavanca. Por fim, será apresentado as consequências de se considerar o estimador usual da matriz de covariâncias quando não se verifica a suposição de homoscedasticidade do modelo, necessitando assim de estimadores consistentes para a matriz de covariâncias, a saber, o estimadores HC0 (White, 1980), HC1 (Hinkley, 1977), HC2 (Horn et. al, 1975), HC3 (Davidson & MacKinnon, 1993), HC4 (Cribari-Neto & Zarkos, 2004) e HC5 (Cribari-Neto et. al, 2007).

3.1 Modelo de regressão linear

A regressão linear múltipla é uma técnica estatística que permite associar variáveis independentes denotadas por X a uma variável resposta denotada por Y, afim de explicar a relação entre elas. Sua forma matricial é dada da seguinte forma

$$Y = X\beta + \varepsilon,$$

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, X = \begin{pmatrix} 1 & X_{1,1} & X_{1,2} & \cdots & X_{1,p} \\ 1 & X_{2,1} & X_{2,2} & \cdots & X_{2,p} \\ 1 & X_{3,1} & X_{3,2} & \cdots & X_{3,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & X_{n,2} & \cdots & X_{n,p} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

em que Y é um vetor de dimensão $n \times 1$, X representa a matriz de regressores de dimensão $n \times p+1$, β é um vetor $n \times 1$ que contém as estimativas dos parâmetros e ε é um vetor $n \times 1$ dos erros, sendo os erros uma sequência aleatória, não correlacionada e independente.

Para aplicarmos as técnicas de regressão linear múltipla é necessário verificar se algumas suposições são atendidas, a saber:

- $E(\varepsilon_i) = 0$, $i = 1, ..., n$, ou seja, o vetor de erros apresenta média igual a zero.
- $Var(\varepsilon_i) = \sigma^2$, $\sigma^2 > 0$, $i = 1, ..., n$, ou seja, a variância dos erros é constante e igual a σ^2.
- $Cov(\varepsilon_i, \varepsilon_u) = 0$, $\forall i \neq u$, ou seja, os erros são independentes.
- Os erros $\varepsilon_1, ..., \varepsilon_n$ apresentam distribuição normal $\mathcal{N}(0, \sigma^2)$.

Para testarmos a suposição de normalidade dos erros utilizamos o teste proposto por Shapiro & Wilks (1965). A estatística de teste para a normalidade é definida por

$$\mathrm{W} = \frac{(\sum_{i=1}^{n} a_i y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

em que as constantes $(a_1, ..., a_n)$ são calculadas a partir da resolução

$$(a_1, ..., a_n) = \frac{m^\top V^{-1}}{(m^\top V^{-1} V^{-1} m)^{1/2}},$$

$m = (m_1, ..., m_n)^\top$ denota o vetor dos valores esperados da estatística de ordem da amostra e $V = (v_{ij})$ corresponde a matriz de covariâncias de ordem $n \times n$. Rejeitamos a hipótese nula de normalidade dos resíduos se $\mathrm{W}_{calculado} > \mathrm{W}_{tabelado}$, em que $\mathrm{W}_{tabelado}$ é o quantil de nível $1 - \alpha$ da distribuição de W sob a hipótese nula.

O teste de linearidade proposto por Ramsey (1969) consiste em incluir variáveis independentes ao quadrado e ao cubo no modelo de regressão múltipla $y = X\beta + Z\gamma + \varepsilon$, em que Z é o vetor das novas variáveis e γ o vetor dos novos parâmetros. A partir de um testes F é possível verificar a significância das novas variáveis, assim rejeita-se a hipótese nula de linearidade, $\gamma = 0$, caso as variáveis sejam significativas em conjunto.

O teste de homoscedasticidade proposto por Koenker & Bassett (1978) é baseado nos quadrados dos resíduos, $\hat{v}_i^2$ (Gujarati, 2006), dado por

$$\hat{v}_i^2 = \alpha_1 + \alpha_2(\hat{Y}_i)^2 + o_i,$$

em que $\hat{Y}_i$ são os valores estimados por meio de $\hat{Y}_i = \beta_0 + ... + \beta_{p-1}X_{i,p-1}$, α são os coeficientes do modelo e o_i os resíduos. Assim por meio dos teste F ou t podemos verificar a hipótese nula de homoscedasticidade, ou seja, $\alpha_2 = 0$.

Quando utilizamos a regressão estamos interessados em estimar o vetor de parâmetros, β. O estimador de mínimos quadrados ordinários é o método de estimação mais utilizado na regressão e tem como objetivo minimizar a soma dos quadrados dos erros $\Sigma_{i=1}^{n}\varepsilon_i^2$. Levando essa ideia para os termos matriciais, temos que

$$\Sigma_{i=1}^{n}\varepsilon^\top\varepsilon = \Sigma_{i=1}^{n}(Y - X\beta)^\top(Y - X\beta),$$

aplicando a primeira derivada em relação a β e igualando a equação a zero, temos que o estimador dos parâmetros desconhecidos é dado por

$$\hat{\beta} = (X^\top X)^{-1}X^\top Y.$$

Um bom estimador apresenta algumas propriedades, a saber: não-viesado, consistente e eficiente. Assim um estimador é dito não-viesado se $E(\hat{\beta}) = \beta$, ou seja, o viés denotado por $B(\hat{\beta}) = 0$, é possível também se ter um estimador assintoticamente não-viesado quando $\lim_{n\to\infty}E(\hat{\beta}) = \beta$.

Um estimador é dito ser consistente se a medida que o tamanho da amostra cresce a variância do estimador tende a zero, ou seja, $\lim_{n\to\infty}Var(\hat{\beta}) = 0$. Por fim, um estimador é dito ser eficiente se ele for classificado como não-viesado e atingir o limite inferior da desigualdade de Cramer-Rao para todos os possíveis valores do parâmetro. Então o estimador $\hat{\beta}$ via mínimos quadrados ordinários é não-viesado pois

$$\begin{aligned}
E(\hat{\beta}) &= E((X^\top X)^{-1}X^\top Y) \\
&= E((X^\top X)^{-1}X^\top(X\beta + \varepsilon) \\
&= E((X^\top X)^{-1}X^\top X\beta + (X^\top X)^{-1}X^\top\varepsilon) \\
&= \beta + (X^\top X)^{-1}X^\top E(\varepsilon) \\
&= \beta,
\end{aligned}$$

ou seja, precisamos garantir apenas que a $E(\varepsilon_i) = 0$, $i = 1, ..., n$, para que a $E(\hat{\beta}) = \beta$.

A matriz de covariâncias é definida como $cov(\hat{\beta}) = E((\hat{\beta} - \beta)(\hat{\beta} - \beta)^\top)$, e sabendo que $\hat{\beta} - \beta = (X^\top X)^{-1}X^\top\varepsilon$, temos que

$$\begin{aligned} cov(\hat{\beta}) &= E((X^\top X)^{-1}X^\top \varepsilon\varepsilon^\top X(X^\top X)^{-1}) \\ &= (X^\top X)^{-1}X^\top E(\varepsilon\varepsilon^\top)X(X^\top X)^{-1}, \end{aligned}$$

podemos escrever a $E(\varepsilon\varepsilon^\top)$ em termos matriciais da seguinte forma

$$E(\varepsilon\varepsilon^\top) = \begin{pmatrix} E(\varepsilon_1^2) & E(\varepsilon_1\varepsilon_2) & \cdots & E(\varepsilon_1\varepsilon_n) \\ E(\varepsilon_2\varepsilon_1) & E(\varepsilon_2^2) & \cdots & E(\varepsilon_2\varepsilon_n) \\ E(\varepsilon_3\varepsilon_1) & E(\varepsilon_3\varepsilon_2) & \cdots & E(\varepsilon_3\varepsilon_n) \\ \vdots & \vdots & \ddots & \vdots \\ E(\varepsilon_n\varepsilon_1) & E(\varepsilon_n\varepsilon_2) & \cdots & E(\varepsilon_n^2) \end{pmatrix} = \sigma^2 I,$$

como os erros são independentes temos que $E(\varepsilon_i\varepsilon_u) = 0$, com $i = 1, ..., n$ e $u = 1, ..., n$ $\forall i \neq u$. Então se consideramos a suposição de homoscedasticidade e substituirmos o resultado $\sigma^2 I$ na equação da matriz de covariâncias teremos

$$\begin{aligned} cov(\hat{\beta}) &= (X^\top X)^{-1}X^\top E(\varepsilon\varepsilon^\top)X(X^\top X)^{-1} \\ &= \sigma^2(X^\top X)^{-1}X^\top X(X^\top X)^{-1} \\ &= \sigma^2(X^\top X)^{-1}, \end{aligned}$$

como não dispomos do valor da variância dos erros é necessário estimá-la, sabendo que $\hat{\varepsilon} = Y - X\hat{\beta}$, temos

$$\hat{\sigma}^2 = \frac{\hat{\varepsilon}^\top \hat{\varepsilon}}{n-p},$$

em que n é o número de observações, p é o número de parâmetros do modelo, $\hat{\varepsilon}$ é o vetor dos resíduos e $\hat{\sigma}^2$ é a variância estimada dos resíduos.

O Teorema de Gauss-Markov, garante que sob homoscedasticidade, o estimador da matriz de covariâncias de $\hat{\beta}$ possui variância mínima, ou seja, o estimador $\hat{\beta}$ é eficiente e não-viesado.

TEOREMA (Gauss-Markov). Considere um modelo linear $Y = X\beta + \varepsilon$, com as seguintes suposições

$$E(\varepsilon_i) = 0,\ Var(\varepsilon_i) = \sigma^2 \text{ e } E(\varepsilon_i\varepsilon_u) = 0,$$

em que $i = 1, ..., n$, $u = 1, ..., n$, $\forall i \neq u$. Seja $\beta^* = CY$, um estimador qualquer em que C é uma matriz de constantes $n \times n$ e $E(\beta^*) = \beta$. Então $\hat{\beta}$ via MQO é mais preciso que β^*, se $\hat{\beta} \neq \beta^*$, ou seja,

$$cov(\beta^*) = cov(\hat{\beta}) + A,$$

em que A é uma matriz positiva-definida.

Portanto o Teorema de Gauss-Markov sugere que sob a suposição de homoscedasticidade o melhor estimador é o estimador MQO. Entretanto nosso interesse está em verificar o quão eficiente é o estimador MQO em relação aos estimadores LMS e LTS, quando os dados apresentam pontos de alavanca sob homoscedasticidade e heteroscedasticidade.

Considerando a hipótese de heteroscedasticidade temos que o estimador da matriz de covariâncias não é consistente nem não-viesado. Para tentar solucionar esse problema Halbert White (1980), apresentou um estimador para a matriz de covariâncias quando a suposição de homoscedasticidade é violada. O estimador de White é expresso como

$$\text{HC0}= (X^\top X)^{-1} X^\top \hat{\Phi}_0 X (X^\top X)^{-1},$$

em que $\hat{\Phi}_0 = \text{diag}\,\{\hat{\varepsilon}_1^2, ..., \hat{\varepsilon}_n^2\}$. Segundo Cribari-Neto & Soares (2003), resultados de simulações apontam que o estimador HC0 de White pode ser muito viesado em amostras finitas e que a presença de pontos de alavancagem tem grande influência sobre o desempenho de estimadores consistentes e testes associados. Para resolver esse novo problema outros pesquisadores desenvolveram novos estimadores consistentes baseados nas ideias de White.

Segundo MacKinnon & White (1985), uma maneira simples de modificar o estimador HC0 é usar uma correção de graus de liberdade similar ao que é convencionalmente usado para obter estimativas imparciais de σ^2. Este estimador modificado foi sugerido por Hinkley (1977) e denotado por HC1, dado por

$$\text{HC1}= \frac{n}{(n-p)} (X^\top X)^{-1} X^\top \hat{\Phi}_0 X (X^\top X)^{-1},$$

em que n é o número de observações e p o número de parâmetros.

Outra maneira de compensar o fato de que o estimador HC0 pode subestimar a variância dos resíduos do MQO foi proposta por Horn et. al (1975), denotado por HC2, dado por

$$\text{HC2}= (X^\top X)^{-1} X^\top \hat{\Phi}_2 X (X^\top X)^{-1},$$

em que

$$\hat{\Phi}_2 = diag\left\{\hat{\varepsilon}_1^2/(1-h_i), ..., \hat{\varepsilon}_n^2/(1-h_n)\right\},$$

e h_i é o i-ésimo elemento da matriz de alavancagem $X(X^\top X)^{-1}X^\top$.

Davidson & Mackinnon (1993) propuseram um estimador baseado na técnica de *Jackknife*. Essa técnica consiste em recalcular n vezes as estimativas de MQO para o vetor de β, cada vez retirando uma das observações, e então usar a variabilidade das estimativas obtidas como estimativas da variância do estimador MQO original. O estimador HC3 é dado por

$$\text{HC3}= (X^\top X)^{-1} X^\top \hat{\Phi}_3 X (X^\top X)^{-1},$$

em que

$$\hat{\Phi}_3 = diag\left\{\hat{\varepsilon}_1^2/(1-h_i)^2, ..., \hat{\varepsilon}_n^2/(1-h_n)^2\right\}$$

e h_i são os elementos da matriz de alavancagem.

Cribari-Neto & Zarkos (2004) propõe uma modificação do estimador HC3, denotado por HC4, dado por

$$\text{HC4}= (X^\top X)^{-1} X^\top \hat{\Phi}_4 X (X^\top X)^{-1},$$

em que

$$\hat{\Phi}_4 = diag\left\{\hat{\varepsilon}_1^2/(1-h_1)^{\delta_1}, ..., \hat{\varepsilon}_n^2/(1-h_n)^{\delta_n}\right\},$$

$\delta_i = \min\left\{4, nh_i/\sum_{j=1}^n h_i\right\}$ e $\sum_{i=1}^n h_i = tr(H) = p$. O expoente que controla o grau de desconto para a observação i é dado pela razão entre o valor de h_i e a média dos h_i. Quanto maior a alavancagem da i-ésima observação, mais inflado é $\hat{\varepsilon}^2$ pelo fator de desconto.

Um novo estimador para matriz de covariâncias sob heteroscedasticidade foi proposto por Cribari-Neto et. al (2007), denotado por HC5. O estimador HC5 é definido como

$$\text{HC5}= (X^\top X)^{-1} X^\top \hat{\Phi}_5 X (X^\top X)^{-1},$$

em que

$$\hat{\Phi} = diag\left\{\hat{\varepsilon}_1^2/\sqrt{(1-h_1)^{\alpha_1}}, ..., \hat{\varepsilon}_n^2/\sqrt{(1-h_n)^{\alpha_n}}\right\}$$

e

$$\alpha_i = \min\left\{\frac{h_i}{\bar{h}}, \max\left\{4, \frac{kh_{\max}}{\bar{h}}\right\}\right\} = \min\left\{\frac{nh_i}{p}, \max\frac{nkh_{\max}}{p}\right\},$$

sendo k uma constante pré definida $0 < k < 1$, $h_{\max} = \max\{h_1, ..., h_n\}$ é o valor máximo de alavancagem e temos que $\bar{h} = n^{-1}\sum_{i=1}^{n} h_i = p/n$. A constante α_i determina o quanto o i-ésimo quadrado dos resíduos deve ser inflado afim de explicar a i-ésima observação de alavancagem.

Capítulo 4

Modelo de Regressão Robusto

Neste capítulo será apresentado os estimadores robustos propostos por Rousseeuw (1984), a saber, os estimadores LMS e LTS, destacando suas características e definições, além do algorítimo computacional para obter as estimativas robustas. Por fim, será apresentado uma motivação para o uso dos estimadores robustos, retratando a influência dos pontos de alavanca para o ajuste da reta de regressão.

4.1 Estimadores Robustos

O estimador de mínimos quadrados ordinários não apresenta nenhuma resistência a observações discrepantes produzindo assim estimativas pouco confiáveis, contudo os estimadores robustos propostos por Rousseeuw (1984) apresentam a noção de ponto de ruptura introduzida por Hampel (1971) como sendo a maior proporção de dados contaminados por observações extremas que um estimador pode suporta para produzir estimativas satisfatórias, ou seja, os estimadores robusto propostos por Rousseeuw assumem que se até 50% dos dados apresentarem pontos de alavanca ou *outliers* ainda é possível se ter boas estimativas enquanto que o estimador de mínimos quadrados ordinários apresenta ponto de ruptura igual a zero.

Rousseeuw (1984) apresentou uma nova forma de realizar estimativas mais robustas a partir da menor mediana dos quadrados dos resíduos, já que a mediana é uma medida de tendência central com grande resistência a valores extremos garantindo uma certa robustez ao estimador, denotado por *least median of squares* (LMS). A estimativa LMS de $\hat{\beta}$ é dada por:

$$\hat{\beta} = \min[mediana(\hat{\varepsilon}_i^2)],$$

em que o resíduo $\hat{\varepsilon}_i = y_i - \hat{\beta}_0 - ... - \hat{\beta}_{p-1}x_{i,p-1}$, é obtido a partir das estimativas do MQO e p é o número de parâmetros no modelo. Esse estimador possui ponto de ruptura de 50% e possui um grau de convergência pequeno igual $n^{-1/3}$, sendo assim ineficiente com erros normalmente distribuídos já que apresentam uma pequena taxa de convergência.

Bulhões (2013), detalha o algorítimo para o cálculo dos estimadores LMS e LTS. Assim a computação das estimativas de LMS é dada pela execução das seguintes etapas.

- Determine todos $\binom{n}{p}$ subconjuntos de tamanho p de $\{1, ..., n\}$.
- Para cada um dos subconjuntos $\{i_1, ..., i_p\}$, compute as estimativas $\hat{\beta}$ de MQO.
- Calcule os resíduos associados a cada uma das estimativas obtidas no passo anterior e observe o valor do resíduo mediano correspondente a cada subconjunto.
- Identifique qual subconjunto $\{i_1, ..., i_p\}$ gerou menor valor de resíduo mediano. A estimativa $\hat{\beta}$ de LMS é fornecida por esse subconjunto.

Rousseeuw (1984) apresentou um outro estimador com ponto de ruptura de 50% e taxa de convergência igual $n^{-1/2}$ sendo mais eficiente que o estimador LMS, denotado por *least trimmed squares* (LTS), como uma solução para a pequena convergência do estimador LMS. Assim a estimativa LTS de $\hat{\beta}$ é dado por:

$$\hat{\beta} = \min(\Sigma_{i=1}^{h}(\varepsilon_{1:n}^2)),$$

em que os resíduos obtidos através das estimativas do MQO $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$, são ordenados e o valor ótimo de h é o maior inteiro contido na quantidade $(n + p + 1)/2$ (Souza, 2011).

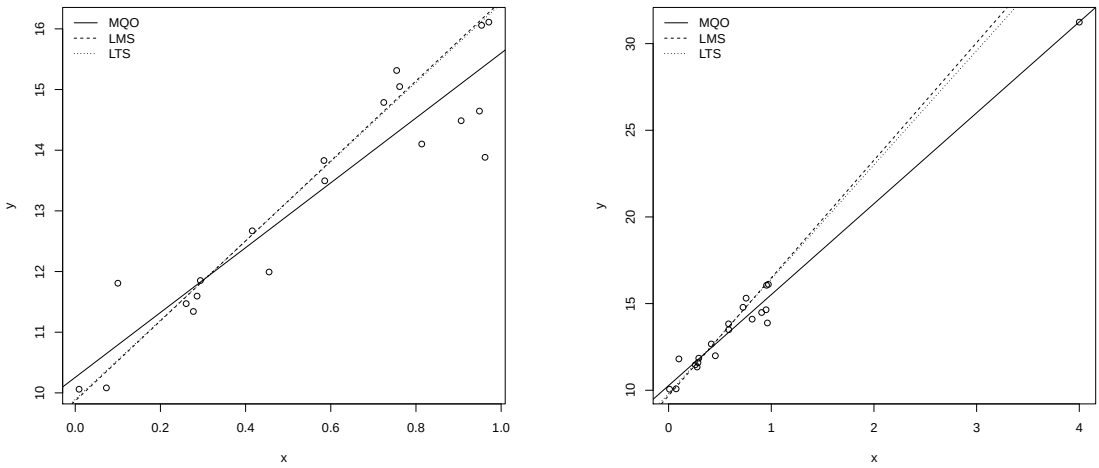
As estimativas do LTS é dada pela execução das seguintes etapas:

- Determine todos $\binom{n}{p}$ subconjuntos de tamanho p de $\{1, ..., n\}$.
- Para cada um dos subconjuntos $\{i_1, ..., i_p\}$, compute as estimativas $\hat{\beta}$ de MQO.
- Calcule os resíduos associados a cada uma das estimativas obtidas no passo anterior, ordene as componentes de cada vetor e observe o valor da soma do quadrado dos h primeiros resíduos correspondentes a cada subconjunto.
- Identifique qual subconjunto $\{i_1, ..., i_p\}$ gerou menor valor daquela soma de resíduos. A estimativa $\hat{\beta}$ de LTS é fornecida por este subconjunto.

4.2 Comparação dos estimadores robustos e de mínimos quadrados ordinários

Souza (2011) apresenta uma figura que ajuda a entender o que acontece com as retas de regressão: MQO, LMS e LTS, quando a variável independente apresenta pontos de alavanca. Assim a Figura (4.1) apresenta os ajustes do modelo de regressão $y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$, $i = 1, ..., 20$ obtidos via MQO, LMS e LTS. Dois cenários foram considerados, um balanceado (sem pontos de alavanca) e outro não-balanceado (com pontos de alavanca). Os valores da covariável correspondem a valores igualmente espaçados entre 0 e 1, esse é o que chamamos de cenário 1 (dados sem pontos de alavanca). No segundo cenário, cenário 2, foi substituído a última observação da covariável por 4 a fim de introduzir um ponto de alavanca nos dados, ou seja, para que o último elemento da diagonal da matriz $X(X^\top X)^{-1}X^\top$ ultrapasse $3p/n = 0.30$, o valor limite comumente utilizado na identificação de pontos de alavanca.

Nos dois cenários tomamos $\beta_0 = 10$ e $\beta_1 = 5$. Foi feito o ajuste do modelo considerando os estimadores MQO, LMS e LTS como apresenta a Figura (4.1). Observamos que no cenário 1, os resíduos obtidos através do LMS, LTS e MQO são semelhantes, diferenciando-se um pouco nas últimas observações. No cenário 2, há um ponto de alta alavancagem ($h_{max} \simeq 0.8584$), podemos observar que a reta de regressão do MQO é arrastada para longe das retas de regressão do LMS e LTS mostrando a influência que os pontos de alavanca podem causa no ajuste da reta de regressão quando se utiliza o estimador MQO, ou seja, os estimadores robustos são capazes de suportar a presença de pontos de alavanca e ainda gerar boas estimativas o que não acontece com o estimador de mínimos quadrados ordinários já que este possui ponto de ruptura igual a zero.



(a) cenário 1 (b) cenário 2

Figura 4.1: Ajustes do modelo de regressão $y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$, $i = 1, ..., 20$, considerando os estimadores, MQO, LMS e LTS nos cenários 1 e 2.

Capítulo 5

Avaliação numérica

As simulações de Monte Carlo foram feitas na plataforma computacional do `R` , por ser um *software* livre que permiti realizar cálculos e elaborar gráficos (Kleiber & Zeileis, 2008). O experimento de Monte Carlo foi baseado no modelo de regressão linear

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i,\ i = 1, ..., n,$$

as covariáveis x_1 e x_2 foram obtidas a partir da distribuição uniforme de tamanho n no intervalo $[0, 1]$. O vetor de erros ε_i foi obtido de três formas diferentes:

- $\varepsilon_i \sim \mathcal{N}(0, 1)$,
- $\varepsilon_i \sim t_{(3)}$,
- $\varepsilon_i \sim \chi^2_{(2)}$.

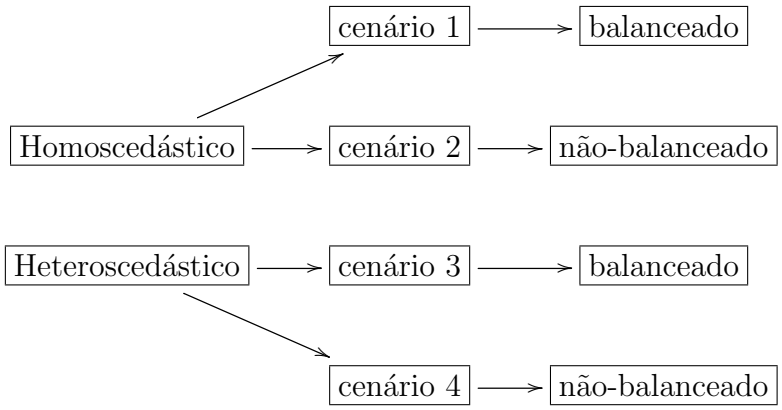
Para as simulações de Monte Carlo foram considerados quatro cenários, a saber:

- cenário 1: refere-se a um conjunto de dados sem pontos de alavanca (cenário balanceado) e sob homoscedasticidade, considerando os erros obtidos a partir das distribuições $\mathcal{N}(0, 1)$, χ^2_2 e t_3 com os tamanhos amostrais $n = 50, 100, 150$.
- cenário 2: refere-se a um conjunto de dados com a substituição das duas últimas observações por pontos de alavanca (cenário não-balanceado) e sob homoscedasticidade, considerando os erros obtidos a partir das distribuições $\mathcal{N}(0, 1)$, χ^2_2 e t_3 com os tamanhos amostrais $n = 50, 100, 150$.

- cenário 3: refere-se a um conjunto de dados sem pontos de alavanca (cenário balanceado) e sob heteroscedasticidade, considerando os erros obtidos a partir das distribuições $\mathcal{N}(0,1)$, χ^2_2 e t_3 com os tamanhos amostrais $n = 50, 100, 150$.

- cenário 4: refere-se a um conjunto de dados com a substituição das duas últimas observações por pontos de alavanca (cenário não-balanceado) e sob heteroscedasticidade, considerando os erros obtidos a partir das distribuições $\mathcal{N}(0,1)$, χ^2_2 e t_3 com os tamanhos amostrais $n = 50, 100, 150$.

O esquema a seguir apresenta os cenários utilizados nas simulações de Monte Carlo.



Para o cenário 2 e o cenário 4 as duas últimas observações da covariável x_2 foram substituídas por pontos que ultrapassaram o limite comumente utilizado de $3p/n$ para classificar uma observação como ponto de alavanca. As medidas de alavancagem de cada observação podem ser obtidas através da matriz de alavancagem definida como

$$H = X(X^\top X)^{-1}X^\top,$$

em que X é a matriz de regressores de dimensão $n \times p + 1$ e os elementos da matriz de alavancagem serão definidos por h_i, assim o elemento h_i que ultrapassar o valor de $3p/n$ pode ser interpretado como ponto de alavanca.

Para analisar o cenário 3 e o cenário 4 foi considerada uma medida de grau de heteroscedasticidade que aumenta a medida que se aumenta o tamanho amostral, definida como:

$$\lambda = \max(\sigma_i^2)/\min(\sigma_i^2),$$

em que o valor de λ é a razão entre a máxima e a mínima variância e depende apenas da função cedástica

$$\sigma_i = \sqrt{exp(c_1 + c_2 x_1)}.$$

em que $c > 0$. Sob o cenário homoscedástico foi considerado $c_1 = c_2 = 0$, ou seja, $\sigma_i = 1$, já em relação ao cenário heteroscedástico foram considerados 2 casos, a saber:

- $c_1 = c_2 = 7.900$, gerando $\lambda \simeq 50$,
- $c_1 = c_2 = 9.309$, gerando $\lambda \simeq 100$.

Foram realizadas 10.000 réplicas de Monte Carlo considerando que cada tamanho amostral foi duplicado e triplicado a partir do $n = 50$, assim é possível analisar os impactos das estimativas à medida que o tamanho da amostra aumenta, garantindo também que a medida do grau de heteroscedasticidade permaneça constante com os diferentes tamanhos amostrais.

Nos estudos de simulação foram considerados que os verdadeiros valores dos parâmetros são $\beta_0 = 1$, $\beta_1 = 1$ e $\beta_2 = 0$. As estimativas dos parâmetros foram obtidas através dos estimadores MQO, LMS e LTS, para cada tamanho amostral, com objetivo de analisar o erro quadrático médio (EQM) e o viés a medida que se aumenta o tamanho da amostra nos cenários considerados.

5.1 Resultados da simulação

Através das simulações de Monte Carlo é possível avaliar o desempenho dos estimadores MQO, LMS e LTS, considerando as suposições de homoscedasticidade e heteroscedasticidade. Os principais resultados encontram-se resumidos a seguir.

Primeiro, no cenário homoscedástico apresentado nas Tabelas 5.1, 5.2 e 5.3, representando respectivamente os erros obtidos a partir das distribuições $\mathcal{N}(0,1)$, $\chi^2_{(2)}$ e $t_{(3)}$, foi observado que o estimador MQO apresentou o maior número de estimativas com menor viés e EQM. Contudo em alguns momentos os estimadores LMS e LTS apresentaram os menores vieses, por exemplo, na Tabela 5.1 (cenário balanceado) com $\varepsilon \sim \mathcal{N}(0,1)$, observamos que o MQO apresentou menor viés a medida que se aumentou o tamanho amostral, entretanto no cenário não-balanceado o MQO com $n = 150$ apresentou um

único viés menor que o dos outros estimadores. Na Tabela 5.2 com $\varepsilon \sim \chi^2_{(2)}$, o estimador LTS apresentou em vários momentos um viés menor que o produzido via estimativas de MQO, por exemplo, no cenário balanceado para as estimativas de β_0 com $n = 50, 100, 150$ o LTS obteve o menor viés. Por fim, na Tabela 5.3 (cenário balanceado) com $\varepsilon \sim t_{(3)}$, o MQO apresentou um menor viés a medida que se aumentou o tamanho da amostra, exceto com $n = 100$ para a estimativa de β_2, já para a estimativa de β_2 com $n = 50$ o MQO subestimou o verdadeiro valor do parâmetro em relação aos estimadores robustos. No cenário não-balanceado com $n = 150$ os estimadores robustos apresentaram menor viés, ou seja, o MQO apresentou menor viés com menores tamanhos amostrais além de subestimar o verdadeiro valor de β_0 e β_2.

Segundo, na Tabela 5.4 cenário heteroscedástico e $\varepsilon \sim \mathcal{N}(0, 1)$ com $\lambda \simeq 50$, ao contrário do estimador MQO que apresentou menor viés para $n = 150$ o estimador LMS apresentou menor viés com menor tamanho amostral ($n = 50$), já com $n = 100$ o MQO subestimou as estimativas de β_0 e superestimou as estimativas β_2 em relação aos estimadores robustos. Para $\lambda \simeq 100$ verificamos que o LTS apresentou o maior número de estimativas com menor viés, exceto para as estimativas de β_1 com $n = 100$ além de β_0 e β_1 com $n = 150$ pelo MQO, em que as estimativas dos parâmetros foram subestimadas e superestimadas em relação as estimativas robustas, respectivamente. Considerando a Tabela 5.5 com $\lambda \simeq 50$, o MQO apresentou o menor viés com $n = 100, 150$, exceto na estimativa de β_2 para $n = 150$ pelo LMS, já o LTS apresentou menor viés para $n = 50$. No caso com $\lambda \simeq 100$, o MQO continuou apresentando o menor viés com $n = 100, 150$, exceto em um momento para a estimativa de β_2 com $n = 150$ pelo LMS, entretanto o LMS apresentou o menor viés para $n = 50$. Esses resultados são justificados pela pouca eficiência dos estimadores robusto com erros normalmente distribuídos.

Terceiro, no cenário heteroscedástico Tabelas 5.6 e 5.7 com erros obtidos a partir da distribuição $\chi^2_{(2)}$, o estimador MQO não gerou nenhuma estimativa com menor viés em relação aos estimadores robustos. Na Tabela 5.6 (cenário balanceado) com $\lambda \simeq 50$, o LTS apresentou menor viés para $n = 150$, exceto para a estimativa do β_1 pelo LMS. Considerando $\lambda \simeq 100$ o LMS apresentou o menor viés, exceto em três momentos, com $n = 50, 100, 150$ para as estimativas de β_0 pelo LTS. Na Tabela 5.7 (cenário não-balanceado) com $\lambda \simeq 50$ e $\lambda \simeq 100$, o estimador LMS se manteve com menor viés, exceto em três momentos para as estimativas de β_0 com $n = 50, 100, 150$ pelo LTS.

Quarto, no cenário heteroscedástico apresentado nas Tabelas 5.8 e 5.9 com erros provenientes da distribuição $t_{(3)}$, os estimadores robustos apresentaram o maior número de estimativas com menor viés. Por exemplo, na Tabela 5.8 (cenário balanceado) com $\lambda \simeq 50$, o LTS apresentou menor viés com $n = 50, 150$, exceto em um momento na estimativa de β_0 com $n = 50$ pelo MQO. Considerando $\lambda \simeq 100$ o MQO não apresentou nenhum viés menor que os estimadores robustos, sendo que o LTS apresentou o menor viés nos tamanhos amostrais $n = 100, 150$. Na Tabela 5.9 (cenário não-balanceado) com $\lambda \simeq 50$, o MQO não apresentou nenhuma viés menor que os outros estimadores robustos, sendo que o LTS apresentou menor viés a medida que se aumentou o tamanho amostral ($n = 100, 150$), exceto na estimativa de β_2 com $n = 100$ pelo LMS, entretanto quando se considera o caso com $\lambda \simeq 100$ o LMS apresentou menor viés com $n = 50, 100, 150$, exceto em alguns tamanhos amostrais com $n = 50, 100$ pelos estimadores MQO e LTS.

Considerando os cenários homoscedásticos, os resultados mostraram que quando os erros são gerados a partir das distribuições normal padrão e *t-student* com 3 graus de liberdade o MQO apresentou o menor EQM, já em relação a distribuição $\chi^2_{(2)}$ os estimadores robustos geraram o menor EQM com maiores tamanhos amostrais, contudo nos cenários heteroscedásticos os valores do EQM obtidos a partir das estimativas do MQO foram muito maiores que o EQM dos estimadores robustos, por exemplo, na Tabela 5.4 (cenário balanceado) com $\lambda \simeq 100$ o EQM de $\hat{\beta}_{0MQO}$ foi quase quatro vezes maior que o EQM de $\hat{\beta}_{0LTS}$, ou seja, os estimadores robustos apresentaram melhores estimativas sob o cenário heteroscedástico, entretanto sob o cenário homoscedástico seus valores de EQM não foram muito diferentes do MQO.

Por fim, quando considerado o cenário homoscedástico balanceado, foi visto que os estimadores robustos apresentaram melhores estimativas nos menores tamanhos amostrais, por outro lado, considerando o caso não-balanceado os estimadores robustos apresentaram menor viés a medida que se aumentou o tamanho das amostras. No cenário heteroscedástico balanceado, os estimadores robustos não apresentaram um ótimo desempenho quando os erros são gerados a partir da distribuição normal padrão, contudo as estimativas robustas passam a apresentar menor viés quando se aumenta o grau de heteroscedasticidade. Considerando as outras distribuições os estimadores robustos apresentaram melhores estimativas em relação ao estimador MQO, além de apresentarem um erro quadrático médio muito inferior ao do estimador MQO, concluindo assim que o uso

dos estimadores robustos tem maior impacto quando se considera os cenários heteroscedásticos, sendo o estimador LTS o que apresentou o maior número de estimativas com menor viés nos tamanhos amostrais superiores, por outro lado, o estimador MQO apresentou menor viés e EQM quando considerado um cenário homoscedástico balanceado.

Tabela 5.1: Viés e EQM obtidos pelos estimadores MQO, LMS e LTS com base no modelo $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(0,1)$, $i = 1, ..., n$, no cenário 1 (balanceado) e cenário 2 (não-balanceado), sob homoscedasticidade.

Cenário	Parâmetros	$n = 50$		$n = 100$		$n = 150$	
	Estimados	Viés	EQM	Viés	EQM	Viés	EQM
1	β_{0MQO}	−0.0025	**0.1594**	**−0.0026**	**0.0814**	**0.0022**	**0.0545**
	β_{0LMS}	−0.0106	0.8875	−0.0085	0.5092	0.0063	0.3804
	β_{0LTS}	**0.0018**	0.8656	−0.0134	0.5536	0.0027	0.4174
	β_{1MQO}	0.0076	**0.2420**	**0.0001**	**0.1213**	**−0.0030**	**0.0809**
	β_{1LMS}	**0.0033**	1.3183	0.0182	0.7641	−0.0096	0.5645
	β_{1LTS}	−0.0107	1.2992	0.0178	0.8228	−0.0101	0.6273
	β_{2MQO}	**−0.0025**	**0.2159**	−0.0015	**0.1084**	**0.0014**	**0.0725**
	β_{2LMS}	0.0140	1.1799	−0.0075	0.6775	−0.0040	0.4907
	β_{2LTS}	0.0077	1.1625	**0.0001**	0.7220	0.0046	0.5444
Cenário	Parâmetros	$n = 50$		$n = 100$		$n = 150$	
	Estimados	Viés	EQM	Viés	EQM	Viés	EQM
2	β_{0MQO}	−0.0036	**0.1074**	**−0.0011**	**0.0577**	0.0021	**0.0404**
	β_{0LMS}	−0.0113	0.7518	−0.0030	0.4541	0.0044	0.3467
	β_{0LTS}	**−0.0107**	0.7368	−0.0160	0.4917	**−0.0018**	0.3778
	β_{1MQO}	0.0077	**0.2424**	**0.0003**	**0.1213**	**−0.0031**	**0.0808**
	β_{1LMS}	0.0053	1.2918	0.0154	0.7630	−0.0115	0.5607
	β_{1LTS}	**0.0037**	1.2975	0.0211	0.8156	−0.0073	0.6347
	β_{2MQO}	**−0.0003**	**0.0222**	−0.0038	**0.0198**	0.0014	**0.0178**
	β_{2LMS}	0.0134	0.7517	−0.0112	0.4713	**−0.0008**	0.3653
	β_{2LTS}	0.0126	0.6870	**−0.0036**	0.4997	0.0075	0.3992

Tabela 5.2: Viés e EQM obtidos pelos estimadores MQO, LMS e LTS com base no modelo $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$, $\varepsilon_i \sim \chi^2_{(2)}$, $i = 1, ..., n$, no cenário 1 (balanceado) e cenário 2 (não-balanceado), sob homoscedasticidade.

Cenário	Parâmetros	$n = 50$		$n = 100$		$n = 150$	
	Estimados	Viés	EQM	Viés	EQM	Viés	EQM
1	β_{0MQO}	1.9921	4.6223	2.0075	4.3616	1.9977	4.2059
	β_{0LMS}	0.9096	**1.5903**	0.8188	0.9183	0.7699	0.7103
	β_{0LTS}	**0.9060**	1.5959	**0.7709**	**0.9006**	**0.7074**	**0.6627**
	β_{1MQO}	**0.0080**	**0.9683**	**−0.0138**	0.4925	**−0.0010**	0.3246
	β_{1LMS}	−0.0485	0.9894	−0.0373	**0.3200**	−0.0218	**0.1589**
	β_{1LTS}	−0.0623	0.9973	−0.0367	0.3916	−0.0272	0.2216
	β_{2MQO}	**−0.0020**	**0.8675**	**0.0001**	0.4305	0.0071	0.2849
	β_{2LMS}	0.0148	0.9068	−0.0058	**0.2868**	0.0051	**0.1433**
	β_{2LTS}	0.0133	0.9189	−0.0070	0.3624	**0.0042**	0.1992
Cenário	Parâmetros	$n = 50$		$n = 100$		$n = 150$	
	Estimados	Viés	EQM	Viés	EQM	Viés	EQM
2	β_{0MQO}	1.9948	4.4273	2.0062	4.2581	1.9998	4.1580
	β_{0LMS}	0.8483	1.3248	0.7808	0.8191	0.7507	0.6651
	β_{0LTS}	**0.8351**	**1.3169**	**0.7243**	**0.7900**	**0.6761**	**0.5985**
	β_{1MQO}	**0.0086**	0.9683	**−0.0139**	0.4920	**−0.0012**	0.3246
	β_{1LMS}	−0.0431	**0.9243**	−0.0276	**0.2999**	−0.0202	**0.1525**
	β_{1LTS}	−0.0579	0.9684	−0.0275	0.3829	−0.0256	0.2217
	β_{2MQO}	**−0.0057**	**0.0852**	**0.0022**	**0.0808**	**0.0030**	**0.0736**
	β_{2LMS}	0.0759	0.5208	0.0357	0.1982	0.0337	0.1031
	β_{2LTS}	0.0875	0.5030	0.0468	0.2447	0.0498	0.1450

Tabela 5.3: Viés e EQM obtidos pelos estimadores MQO, LMS e LTS com base no modelo $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$, $\varepsilon_i \sim t_{(3)}$, $i = 1, ..., n$, no cenário 1 (balanceado) e cenário 2 (não-balanceado), sob homoscedasticidade.

Cenário	Parâmetros	$n = 50$		$n = 100$		$n = 150$	
	Estimados	Viés	EQM	Viés	EQM	Viés	EQM
1	β_{0MQO}	−0.0101	**0.4837**	**0.0067**	**0.2498**	**0.0026**	**0.1616**
	β_{0LMS}	−0.0058	0.8860	0.0104	0.4915	0.0042	0.3403
	β_{0LTS}	**−0.0037**	0.8471	0.0078	0.4889	0.0049	0.3512
	β_{1MQO}	0.0148	**0.7160**	**−0.0087**	**0.3650**	**−0.0067**	**0.2361**
	β_{1LMS}	0.0100	1.3611	−0.0174	0.7262	−0.0069	0.4963
	β_{1LTS}	**0.0077**	1.3096	−0.0167	0.7305	−0.0073	0.5269
	β_{2MQO}	**−0.0052**	**0.6637**	−0.0051	**0.3368**	**0.0030**	**0.2141**
	β_{2LMS}	0.0092	1.1657	−0.0110	0.6158	−0.0052	0.4339
	β_{2LTS}	0.0103	1.1179	**−0.0003**	0.6387	−0.0049	0.4566
Cenário	Parâmetros	$n = 50$		$n = 100$		$n = 150$	
	Estimados	Viés	EQM	Viés	EQM	Viés	EQM
2	β_{0MQO}	−0.0125	**0.3301**	0.0060	**0.1705**	0.0030	**0.1169**
	β_{0LMS}	**0.0005**	0.7319	**0.0045**	0.4244	**−0.0004**	0.3041
	β_{0LTS}	−0.0058	0.6982	0.0102	0.4293	0.0033	0.3113
	β_{1MQO}	0.0150	**0.7165**	**−0.0084**	**0.3647**	-0.0068	**0.2357**
	β_{1LMS}	**0.0065**	1.2715	−0.0091	0.7029	**−0.0036**	0.4959
	β_{1LTS}	0.0151	1.2480	−0.0171	0.7175	−0.0059	0.5186
	β_{2MQO}	−0.0007	**0.0699**	**−0.0035**	**0.0596**	0.0021	**0.0536**
	β_{2LMS}	**0.0007**	0.6879	−0.0047	0.4179	−0.0020	0.3160
	β_{2LTS}	0.0018	0.6197	−0.0072	0.4315	**−0.0016**	0.3232

Tabela 5.4: Viés e EQM obtidos pelos estimadores MQO, LMS e LTS com base no modelo $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(0,1)$, $i = 1, ..., n$, no cenário 3 (balanceado), sob heteroscedasticidade.

Cenário 3	Parâmetros	$n = 50$		$n = 100$		$n = 150$	
	Estimados	Viés	EQM	Viés	EQM	Viés	EQM
$\lambda \simeq 50$	β_{0MQO}	0.1030	242.7359	−0.0970	121.4345	**−0.0108**	80.0703
	β_{0LMS}	**−0.0642**	76.9977	0.0586	52.4878	0.0962	44.2336
	β_{0LTS}	−0.1155	**51.6435**	**0.0483**	**27.2510**	0.0842	**19.0919**
	β_{1MQO}	0.4315	783.3296	**−0.0882**	378.4522	**0.0024**	257.9440
	β_{1LMS}	**0.0385**	536.8560	−0.2690	401.2726	−0.2413	329.8416
	β_{1LTS}	−0.0442	**430.2427**	−0.1728	**276.9176**	−0.3331	**204.0680**
	β_{2MQO}	−0.4978	518.5016	0.1230	267.4833	**0.0467**	173.2797
	β_{2LMS}	**0.1125**	99.6297	**−0.0654**	61.7771	−0.0818	51.3341
	β_{2LTS}	0.1825	**64.5015**	−0.0669	**31.3867**	−0.0506	**21.6044**
Cenário 3	Parâmetros	$n = 50$		$n = 100$		$n = 150$	
	Estimados	Viés	EQM	Viés	EQM	Viés	EQM
, $\lambda \simeq 100$	β_{0MQO}	0.4690	3971.2062	−0.3991	1989.5060	**−0.0951**	1310.0102
	β_{0LMS}	−0.2025	703.5927	0.2589	574.6174	0.1089	500.5373
	β_{0LTS}	**−0.1933**	**450.9921**	**0.0803**	**234.6791**	0.1276	**168.6021**
	β_{1MQO}	1.7538	13207.9850	**−0.3101**	6396.2460	**0.0563**	4355.1002
	β_{1LMS}	−0.2656	5374.2943	−0.9563	4256.8365	−0.1842	3545.1876
	β_{1LTS}	**0.0875**	**4137.8363**	−0.4972	**2543.1170**	−0.4114	**1843.8107**
	β_{2MQO}	−2.1037	8445.0310	0.5641	4359.5426	0.2143	2818.5023
	β_{2LMS}	0.5143	858.5592	−0.2891	603.3591	−0.1296	518.7054
	β_{2LTS}	**0.3311**	**514.2304**	**−0.0931**	**243.9557**	**−0.0806**	**173.9792**

Tabela 5.5: Viés e EQM obtidos pelos estimadores MQO, LMS e LTS com base no modelo $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(0,1)$, $i = 1, ..., n$, no cenário 4 (não-balanceado), sob heteroscedasticidade.

Cenário 4	Parâmetros	$n = 50$		$n = 100$		$n = 150$	
	Estimados	Viés	EQM	Viés	EQM	Viés	EQM
$\lambda \simeq 50$	β_{0MQO}	-0.1097	86.5515	**0.0024**	47.0113	**−0.0123**	34.0755
	β_{0LMS}	-0.0410	48.7228	0.0203	41.5855	0.0531	38.1517
	β_{0LTS}	**−0.0299**	**29.6938**	0.0438	**19.2162**	0.0549	**15.1997**
	β_{1MQO}	0.4534	791.5821	**−0.0900**	379.7102	**0.0005**	258.2831
	β_{1LMS}	-0.1194	511.6064	-0.2183	392.4202	-0.2554	332.5391
	β_{1LTS}	**−0.0983**	**396.9508**	-0.1594	**255.9039**	-0.2734	**193.5378**
	β_{2MQO}	-0.0814	**21.0354**	**−0.0572**	**20.2611**	0.0461	20.0394
	β_{2LMS}	0.1054	52.4411	-0.0778	40.8795	**0.0012**	36.1656
	β_{2LTS}	**0.0697**	34.8091	-0.0829	22.1174	-0.0063	**17.1117**
Cenário 4	Parâmetros	$n = 50$		$n = 100$		$n = 150$	
	Estimados	Viés	EQM	Viés	EQM	Viés	EQM
$\lambda \simeq 100$	β_{0MQO}	-0.4309	1415.9155	**−0.0063**	763.7023	**−0.0791**	550.6751
	β_{0LMS}	**−0.0253**	480.5549	0.4055	473.8190	0.0878	450.2132
	β_{0LTS}	-0.0609	**264.0626**	0.1753	**173.0782**	0.2157	**137.1892**
	β_{1MQO}	1.8463	13353.4790	**−0.3217**	6420.2402	**0.0481**	4362.4504
	β_{1LMS}	**−0.3430**	5145.4757	-1.0419	4154.4986	-0.2527	3606.0675
	β_{1LTS}	-0.4472	**3728.9144**	-0.4647	**2314.0941**	-0.5824	**1743.1379**
	β_{2MQO}	-0.3428	**281.5485**	**−0.1550**	279.7515	0.1718	284.3209
	β_{2LMS}	**0.1501**	458.0559	-0.2676	394.1828	**−0.1537**	366.6043
	β_{2LTS}	0.2465	286.3586	-0.2817	**184.7148**	-0.1714	**145.5670**

Tabela 5.6: Viés e EQM obtidos pelos estimadores MQO, LMS e LTS com base no modelo $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$, $\varepsilon_i \sim \chi^2_{(2)}$, $i = 1, ..., n$, no cenário 3 (balanceado), sob heteroscedasticidade.

Cenário 3	Parâmetros	$n = 50$		$n = 100$		$n = 150$	
	Estimados	Viés	EQM	Viés	EQM	Viés	EQM
$\lambda \simeq 50$	β_{0MQO}	-48.3694	3298.5970	-48.0640	2797.7183	-48.1783	2637.6196
	β_{0LMS}	2.9449	103.8993	5.6892	83.0483	7.0529	84.7555
	β_{0LTS}	**1.9377**	**71.8054**	**3.0856**	**38.0938**	**3.5169**	**29.9329**
	β_{1MQO}	202.3658	43985.9154	201.7028	42254.7602	202.0823	41872.2938
	β_{1LMS}	**24.4983**	1113.3366	**17.5528**	**579.8892**	**14.1177**	**379.1203**
	β_{1LTS}	25.8650	**1062.1201**	20.2548	594.8290	17.9506	435.8559
	β_{2MQO}	3.8303	2083.0876	3.8267	1082.8684	3.8683	705.8771
	β_{2LMS}	1.5671	125.4610	**0.6689**	55.7226	0.4759	34.6933
	β_{2LTS}	**1.4226**	**100.9894**	0.6852	**39.3517**	**0.4569**	**23.8852**
Cenário 3	Parâmetros	$n = 50$		$n = 100$		$n = 150$	
	Estimados	Viés	EQM	Viés	EQM	Viés	EQM
$\lambda \simeq 100$	β_{0MQO}	-225.7814	66662.6395	-224.6243	58449.4755	-224.9817	55807.8292
	β_{0LMS}	4.1848	949.4205	12.9291	738.4108	17.7877	750.8508
	β_{0LTS}	**0.4561**	**582.2402**	**4.6355**	**259.9589**	**6.1709**	**192.5418**
	β_{1MQO}	840.0878	756938.4406	837.5018	727911.5459	838.8433	721111.8858
	β_{1LMS}	**86.0794**	12671.0096	**64.6942**	7337.6283	**53.2525**	**5052.4063**
	β_{1LTS}	90.8737	**12109.0116**	72.6343	**7019.7255**	65.2713	5316.7373
	β_{2MQO}	14.0786	33921.8654	13.9807	17636.5743	14.0928	11480.7913
	β_{2LMS}	**4.9005**	1134.2855	**2.4260**	557.2109	**1.4882**	367.7193
	β_{2LTS}	5.3765	**817.0711**	2.6449	**323.8611**	1.7700	**203.7822**

Tabela 5.7: Viés e EQM obtidos pelos estimadores MQO, LMS e LTS com base no modelo $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$, $\varepsilon_i \sim \chi^2_{(2)}$, $i = 1, ..., n$, no cenário 4 (não-balanceado), sob heteroscedasticidade.

Cenário 4	Parâmetros	$n = 50$		$n = 100$		$n = 150$	
	Estimados	Viés	EQM	Viés	EQM	Viés	EQM
$\lambda \simeq 50$	β_{0MQO}	-42.8013	2177.4190	-43.4694	2082.7776	-44.0460	2075.4868
	β_{0LMS}	2.6796	74.1705	5.4658	73.5785	7.0226	81.2224
	β_{0LTS}	**1.2845**	**43.5640**	**2.5082**	**28.1162**	**3.0704**	**23.7697**
	β_{1MQO}	202.7484	44170.6770	201.7249	42269.8431	202.0091	41844.6969
	β_{1LMS}	**24.0163**	1078.5979	**17.3892**	**579.5943**	**13.7228**	**365.4355**
	β_{1LTS}	25.9028	**1050.9644**	20.3769	595.3411	18.0839	439.8238
	β_{2MQO}	-5.4276	109.4182	-4.3156	101.8198	-3.6143	94.2604
	β_{2LMS}	**2.2188**	72.8568	**1.3529**	38.0177	**0.9646**	26.1416
	β_{2LTS}	2.4273	**61.0832**	1.6064	**30.3431**	1.2023	**19.4755**
Cenário 4	Parâmetros	$n = 50$		$n = 100$		$n = 150$	
	Estimados	Viés	EQM	Viés	EQM	Viés	EQM
$\lambda \simeq 100$	β_{0MQO}	-201.3683	46171.4600	-204.5340	44967.2995	-206.9934	45035.3917
	β_{0LMS}	3.5646	699.9527	12.5961	660.5279	17.6299	708.4723
	β_{0LTS}	**1.0219**	**356.1358**	**3.0465**	**198.2889**	**4.9520**	**151.2040**
	β_{1MQO}	842.0474	760769.9109	837.7556	728450.6519	838.6327	720797.0838
	β_{1LMS}	**83.6840**	12327.5125	**63.0823**	7111.9114	**51.8630**	**4857.6212**
	β_{1LTS}	90.1109	**11798.2946**	73.0790	**7058.9700**	65.4994	5340.3679
	β_{2MQO}	-26.0290	1750.1546	-21.3535	1605.8769	-18.2947	1485.6636
	β_{2LMS}	**7.4229**	655.0594	**4.6602**	380.8857	**3.2229**	267.3034
	β_{2LTS}	7.9310	**505.4370**	5.2319	**264.4007**	3.8499	**173.9398**

Tabela 5.8: Viés e EQM obtidos pelos estimadores MQO, LMS e LTS com base no modelo $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i$, $\varepsilon_i \sim t_{(3)}$, $i = 1, ..., n$, no cenário 3 (balanceado), sob heteroscedasticidade.

Cenário 3	Parâmetros	$n = 50$		$n = 100$		$n = 150$	
	Estimados	Viés	EQM	Viés	EQM	Viés	EQM
$\lambda \simeq 50$	β_{0MQO}	**−0.0884**	714.2408	0.2183	410.5405	−0.0475	245.3492
	β_{0LMS}	−0.2016	104.3000	**0.0568**	71.0408	0.0149	59.3358
	β_{0LTS}	−0.1580	**71.0014**	0.0702	**33.9157**	**−0.0091**	**24.1527**
	β_{1MQO}	0.2916	2209.5673	**−0.2920**	1246.9174	−0.2685	780.7232
	β_{1LMS}	0.3480	615.9287	−0.3100	429.8290	−0.1681	351.3087
	β_{1LTS}	**0.1974**	**489.5319**	−0.3261	**285.9005**	**0.0125**	**208.7375**
	β_{2MQO}	−0.1624	1559.1139	−0.2657	863.6414	0.2918	517.5385
	β_{2LMS}	0.2013	140.3297	**0.0192**	93.6644	0.0318	76.2346
	β_{2LTS}	**0.1515**	**91.0544**	−0.0293	**42.6112**	**0.004**	**29.2993**
Cenário 3	Parâmetros	$n = 50$		$n = 100$		$n = 150$	
	Estimados	Viés	EQM	Viés	EQM	Viés	EQM
$\lambda \simeq 100$	β_{0MQO}	−0.2779	11672.9662	0.8306	6742.9250	−0.2047	4018.1343
	β_{0LMS}	**−0.1217**	991.8979	0.3178	738.5978	−0.1346	641.8092
	β_{0LTS}	−0.1485	**603.6136**	**0.1656**	**295.6942**	**0.0562**	**208.3122**
	β_{1MQO}	1.0766	37276.9012	−1.1007	21035.9024	−1.0623	13183.8187
	β_{1LMS}	**−0.1072**	6313.9607	−1.0763	4592.6456	−0.3565	3769.9462
	β_{1LTS}	0.1721	**4686.6692**	**−0.9805**	**2689.5960**	**−0.3089**	**1922.1297**
	β_{2MQO}	−0.5762	25372.1823	−1.0264	14091.0066	1.1544	8431.8643
	β_{2LMS}	0.3111	1267.9446	0.0333	900.8144	0.2341	764.9551
	β_{2LTS}	**0.0600**	**728.6870**	**0.0173**	**341.3841**	**0.0118**	**235.4093**

Tabela 5.9: Viés e EQM obtidos pelos estimadores MQO, LMS e LTS com base no modelo $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$, $\varepsilon_i \sim t_{(3)}$, $i = 1, ..., n$, no cenário 4 (não-balanceado), sob heteroscedasticidade.

Cenário 4	Parâmetros	$n = 50$		$n = 100$		$n = 150$	
	Estimados	Viés	EQM	Viés	EQM	Viés	EQM
$\lambda \simeq 50$	β_{0MQO}	-0.1473	247.5596	0.1446	156.5685	0.0355	103.9515
	β_{0LMS}	$\mathbf{-0.0392}$	71.8700	0.0564	58.0368	0.0163	52.2073
	β_{0LTS}	-0.0555	**42.2765**	$\mathbf{-0.0051}$	**25.4704**	$\mathbf{-0.0111}$	**19.3987**
	β_{1MQO}	0.3002	2232.3419	-0.2807	1248.8067	-0.2784	780.9837
	β_{1LMS}	0.2129	585.3762	-0.2487	430.0633	-0.1396	360.1354
	β_{1LTS}	$\mathbf{-0.0165}$	**449.9509**	$\mathbf{-0.2431}$	**272.0059**	**0.0265**	**197.3769**
	β_{2MQO}	-0.0426	65.6936	-0.1150	65.5657	0.1267	60.7799
	β_{2LMS}	**0.0209**	77.9782	$\mathbf{-0.0030}$	61.1240	0.0294	54.1255
	β_{2LTS}	0.0811	**49.9047**	0.0831	**30.8313**	**0.0012**	**23.1377**
Cenário 4	Parâmetros	$n = 50$		$n = 100$		$n = 150$	
	Estimados	Viés	EQM	Viés	EQM	Viés	EQM
$\lambda \simeq 100$	β_{0MQO}	-0.4818	4034.9074	0.5237	2542.9012	0.1351	1682.7222
	β_{0LMS}	$\mathbf{-0.1486}$	686.3036	0.2806	629.3401	**0.0434**	563.2685
	β_{0LTS}	-0.2962	**362.9387**	$\mathbf{-0.0338}$	**223.9443**	0.1856	**169.9140**
	β_{1MQO}	1.1079	37682.9383	-1.0582	21072.2161	-1.1012	13193.0496
	β_{1LMS}	**0.0754**	5961.3657	-1.3451	4576.7652	$\mathbf{-0.3458}$	3835.2393
	β_{1LTS}	0.3848	**4223.9856**	$\mathbf{-0.7805}$	**2513.4278**	-0.4240	**1828.2686**
	β_{2MQO}	$\mathbf{-0.1595}$	873.7174	-0.4072	910.9466	0.4827	860.8979
	β_{2LMS}	0.3268	693.9104	**0.2979**	603.1401	**0.0734**	537.8401
	β_{2LTS}	0.2572	**421.7471**	0.3491	**261.0256**	-0.1746	**193.4563**

Capítulo 6

Aplicação

Neste capítulo apresentaremos uma aplicação do uso da regressão robusta quando os dados apresentam *outliers*, pontos de alavanca e encontram-se sob o cenário heteroscedástico, características essas que nas simulações de Monte Carlo mostraram o melhor desempenho dos estimadores robustos em ajustar a reta de regressão. Para esta aplicação, os dados considerados foram extraídos de Greene (1997, Tabela 12.1, p.541) e sua fonte original é o Departamento de Comércio dos Estados Unidos. Adicionalmente, este conjunto de dados está disponível no pacote *sandwich* do *software* estatístico `R` (Kleiber & Zeileis, 2008). A variável resposta refere-se aos gastos per capita em escolas públicas e a variável independente representa a renda per capita por Estado, ambas definidas em dólares, sendo está reescalonada por 10^{-4}. O Estado de Wisconsin foi excluído da amostra por apresentar valores faltantes, totalizando assim uma amostra de tamanho $n = 50$.

Os dados usados para essa aplicação encontram-se dispostos na Tabela 6.1. Através de uma análise descritiva podemos verificar que o menor valor do gasto per capita é 259.00 US$ e o maior é 821.00 US$, correspondendo aos Estados do Mississippi e Alaska, respectivamente, com um gasto per capita médio de 373.26 US$ e desvio padrão 94.55 US$. O menor valor da renda per capita por Estado é 5736 US$ e o maior é 10851 US$, correspondendo também aos Estados do Mississippi e Alaska, respectivamente, com uma renda média por Estado de 7608.56 US$ e desvio padrão 1050.64 US$. Para a identificação dos pontos de influência e alavanca foram consideradas medidas de influência como Distância de Cook e matriz de alavancagem, além da análise gráfica que também permite visualizar a distribuição dos dados. Foi visto através do teste de linearidade de Ramsey (1969), que a variável renda não foi suficiente para a adequação do modelo,

necessitando-se assim da variável renda ao quadrado, denotada por x^2. A suposição de homoscedasticidade foi verificada pelo teste de Koenker & Bassett (1978), resultando em um modelo heteroscedástico, ou seja, o estimador usual da matriz de covariâncias terá que ser substituído por outros estimadores consistentes.

Tabela 6.1: Dados sobre o gasto per capita em escolas públicas e a renda per capita por Estado em 1979 nos Estados Unidos.

Estado	Gasto	Renda	Estado	Gasto	Renda	Estado	Gasto	Renda
Alabama	275	6247	Maine	327	6333	Oregon	397	7839
Alaska	821	10851	Maryland	427	8306	Pennsylvania	412	7733
Arizona	339	7374	Massachusetts	427	8063	Rhode Island	342	7526
Arkansas	275	6183	Michigan	466	8442	South Carolina	315	6242
California	387	8850	Minnesota	477	7847	South Dakota	321	6841
Colorado	452	8001	Mississippi	259	5736	Tennessee	268	6489
Connecticut	531	8914	Missouri	274	7342	Texas	315	7697
Delaware	424	8604	Montana	433	7051	Utah	417	6622
Florida	316	7505	Nebraska	294	7391	Vermont	353	6541
Georgia	265	6700	Nevada	359	9032	Virginia	356	7624
Hawaii	403	8380	New Hampshire	279	7277	Washington	415	8450
Idaho	304	6813	New Jersey	423	8818	Washington DC	428	10022
Illinois	437	8745	New Mexico	388	6505	West Virginia	320	6456
Indiana	345	7696	New York	447	8267	Wisconsin	*	7597
Iowa	431	7873	North Carolina	335	6607	Wyoming	500	9096
Kansas	355	8001	North Dakota	311	7478			
Kentucky	260	6615	Ohio	322	7812			
Louisiana	316	6640	Oklahoma	320	6951			

* Dado faltante

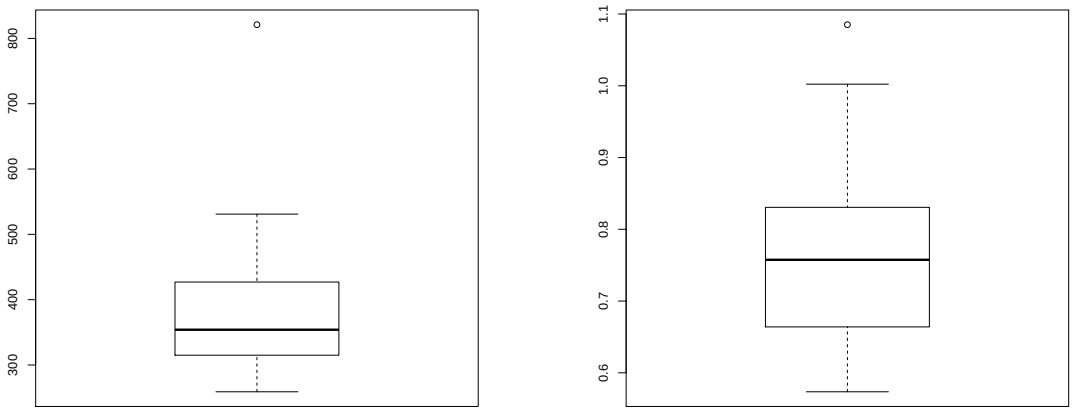
Inicialmente o modelo adotado apresentou a forma

$$y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i,\ i = 1, ..., 50,$$

em que y representa o gasto per capita em escolas públicas e x representa a renda per capita por Estado nos Estados Unidos. Os parâmetros do modelo foram estimados via MQO e suas estimativas pontuais são $\hat{\beta}_0 = -151.3$ e $\hat{\beta}_1 = 689.4$.

Através da Figura 6.1 podemos observa a presença de *outliers* na variável resposta e pontos de alavanca na variável independente, sendo estes pontos que excedem os limites

superiores do *boxplot*, é possível verificar também que a variável gasto per capita apresenta uma certa assimetria a direita já a variável renda per capita apresenta assimetria a esquerda.



(a) Gasto per capita (b) Renda per capita

Figura 6.1: *Boxplot* da variável gasto per capita em escolas públicas e renda per capita por Estado em 1979 nos Estados Unidos.

A Distância de Cook permite determinar o grau de influência da i-ésima observação nos dados sobre a estimativa de β, quando a i-ésima observação é excluída (Cook, 1977). Na Figura 6.2 podemos visualizar que a observação 2 (Estado do Alaska) ultrapassa o valor 0.70 correspondente ao quantil aproximado da distribuição $F_{(0.5,2,48)}$, identificando o Estado do Alaska como possível ponto de influência sobre as estimativas do parâmetro, sendo o resíduo correspondente a essa observação 4.11 e a distância de Cook é 2.31.

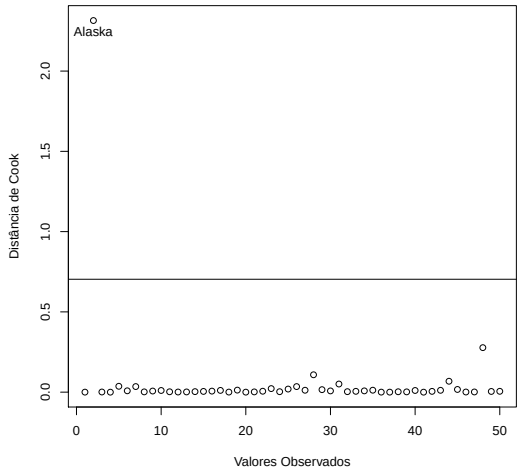
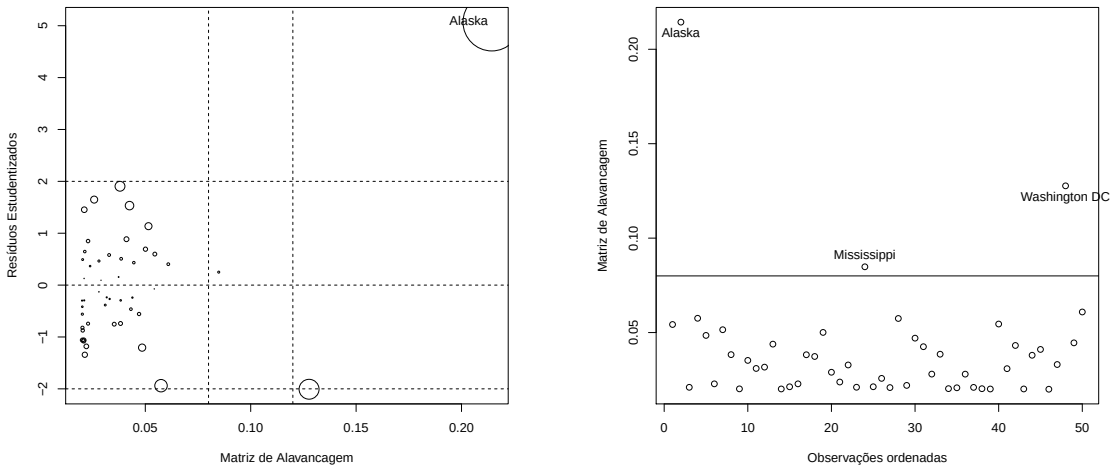


Figura 6.2: Distância de Cook do modelo $y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$, $i = 1, ..., 50$.

A Figura 6.3 permite identificar um ponto de influência e três pontos de alavanca. Por exemplo, o gráfico (a) é o cruzamento dos resíduos estudentizados e alavancagem, em que as áreas dos círculos são proporcionais a distância de Cook (Fox, 2002), é possível identificar o Estado do Alaska como ponto de influência devido ao tamanho do circulo, já o gráfico (b) apresenta os valores da matriz de alavancagem versus as observações ordenadas. Através da matriz de alavancagem, $H = X(X^\top X)^{-1}X^\top$, é possível identificar os elementos h_i que possuem alta alavancagem, ou seja, os elementos que ultrapassarem o valor $2p/n \simeq 0.08$, destacando-se os Estados do Alaska com $h_i \simeq 0.214$, Washington DC com $h_i \simeq 0.127$ e Mississippi com $h_i \simeq 0.084$. Então para minimizar a influência sobre os parâmetros de regressão quando existem observações potencialmente influentes ou pontos de alavanca é recomendado o uso da regressão robusta, caso seja confirmada a plausibilidade dos valores (Barbieri, 2012).



(a) (b)

Figura 6.3: Gráficos para identificar possíveis pontos de influência e de alavanca no modelo $y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$, $i = 1, ..., 50$.

As técnicas estatísticas anteriormente permitiram observar um ponto de influência referente ao Estado do Alaska na variável gasto per capita e três pontos de alavanca referentes ao Estado do Alaska, Washington DC e Mississippi na variável renda per capita. A Figura 6.4 apresenta os impactos sofridos pela reta de regressão considerando os estimadores MQO, LMS e LTS na presença dessas observações discrepantes, sendo possível observar que os estimadores MQO e LMS sofreram os maiores desvios em relação ao estimador LTS devido a observação do Estado do Alaska que possui maior alavancagem. A Tabela 6.2 apresenta as estimativas pontuais do modelo ajustado via os estimadores MQO, LMS e LTS. Podemos observar que as estimativas de β_0 são negativa para os três estimadores, entretanto o $\hat{\beta}_0$ via MQO é quase três vezes maior que $\hat{\beta}_0$ via LTS.

Tabela 6.2: Estimativas pontuais de β_0 e β_1 considerando o modelo $y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$, $i = 1, ..., 50$, via estimadores MQO, LMS e LTS.

Estimadores	$\hat{\beta}_0$	$\hat{\beta}_1$
MQO	-151	689
LMS	-99	626
LTS	-60	562

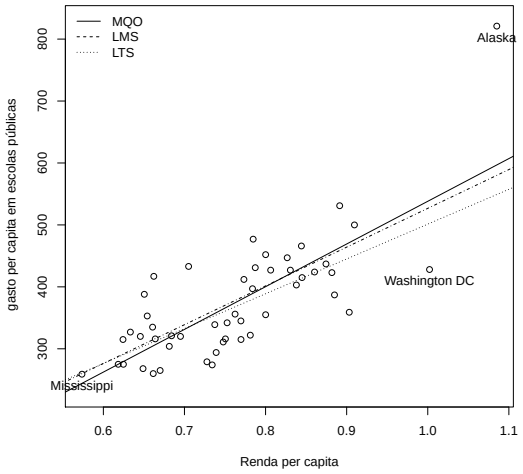
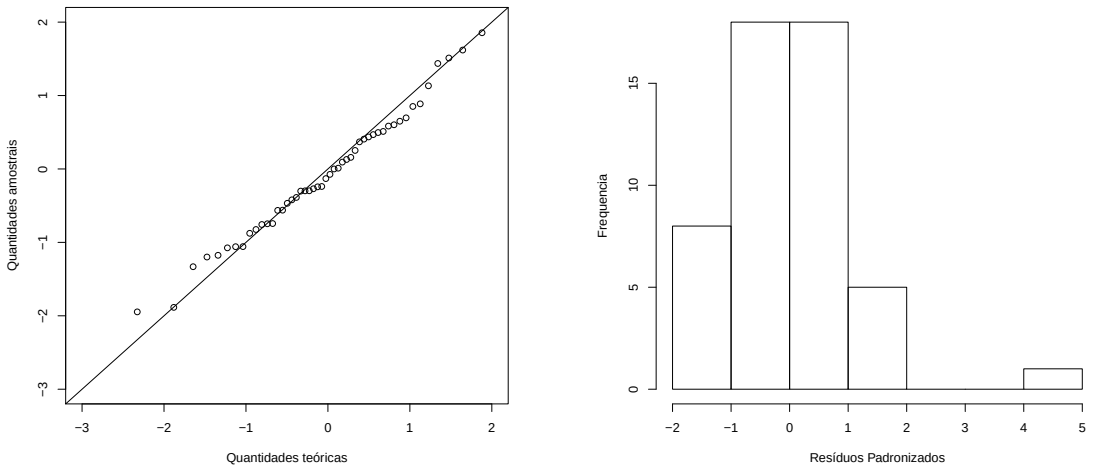


Figura 6.4: Ajustes do modelo de regressão $y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$, $i = 1, ..., 50$, considerando os estimadores MQO, LMS e LTS.

Para verificarmos a suposição de normalidade dos resíduos, utilizamos o teste de normalidade de Shapiro & Wilks (1965) que é útil quando os dados apresentam poucas observações, cuja hipótese nula é que os resíduos seguem distribuição normal padrão, através desse teste obtivemos um p-valor$< 5 \times 10^{-3}$, sendo p-valor a probabilidade de rejeitar a hipótese nula quando ela é verdadeira, rejeitando assim a hipótese que os resíduos seguem distribuição normal padrão ao nível de 5% de significância. Na Figura 6.5, o gráfico (a) apresenta os resíduos padronizados versus os quantis da distribuição normal padrão, denotado por QQplot, é possível verificar que as observações estão se distanciando da reta em vários momentos, já o gráfico (b) apresenta a distribuição de frequência dos resíduos padronizados. Então é possível confirmar junto com os resultado obtidos pelo teste de normalidade, QQPlot e o histograma que os resíduos não seguem distribuição normal padrão.



(a) QQPlot (b) Histograma

Figura 6.5: Gráfico para verificar suposição de normalidade dos resíduos do modelo $y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$, $i = 1, ..., 50$.

Para verificar a suposição de homoscedasticidade do modelo, foi utilizado o teste de Koenker & Bassett (1978), pois este funciona bem quando os resíduos não seguem distribuição normal padrão. Com um p-valor$< 6 \times 10^{-4}$, rejeitamos a hipótese nula de homoscedasticidade ao nível de significância de 5%, ou seja, sendo necessário o uso dos estimadores consistentes da matriz de covariâncias já que o estimador usual da matriz de covariâncias não é mais confiável. A Figura 6.6 apresenta os resíduos padronizados versus os valores ajustados, assim os resíduos que estiverem fora do limite $[-2, 2]$ são considerados muito grandes.

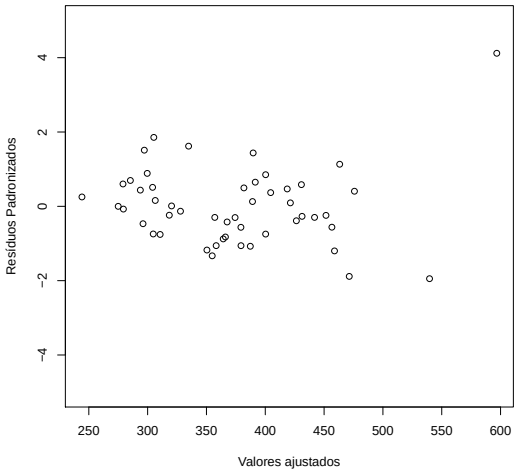


Figura 6.6: Gráfico para verificar suposição de homoscedasticidade dos resíduos do modelo $y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$, $i = 1, ..., 50$.

Para verificar a relação entre a variável resposta e a variável independente foi utilizado o teste de linearidade de Ramsey (1969), que consiste em incluir uma variável x^2 no modelo e verificar se ela é significativa em conjunto através de um teste F. Com p-valor$< 3 \times 10^{-3}$ podemos rejeitar a hipótese nula de linearidade do modelo ao nível de 5%, ou seja, concluímos que o modelo é da forma

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}^2 + \varepsilon_i,$$

A partir da Figura 6.7 podemos observar que os pontos encontram-se bastante dispersos, sendo necessário a inclusão da variável x^2 por causa da possível relação quadrática. O R^2, coeficiente de determinação, avalia o poder explicativo do modelo, ou seja, quantos por cento da variabilidade da variável resposta é explicada pelas variáveis independentes. Com a inclusão da nova variável ao modelo o poder explicativo passou de $R^2 \simeq 0.58$ para $R^2 \simeq 0.65$.

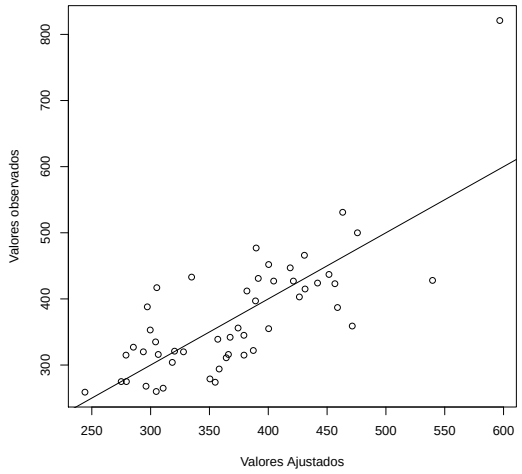


Figura 6.7: Gráfico para verificar a relação linear entre a variável resposta e a variável independente.

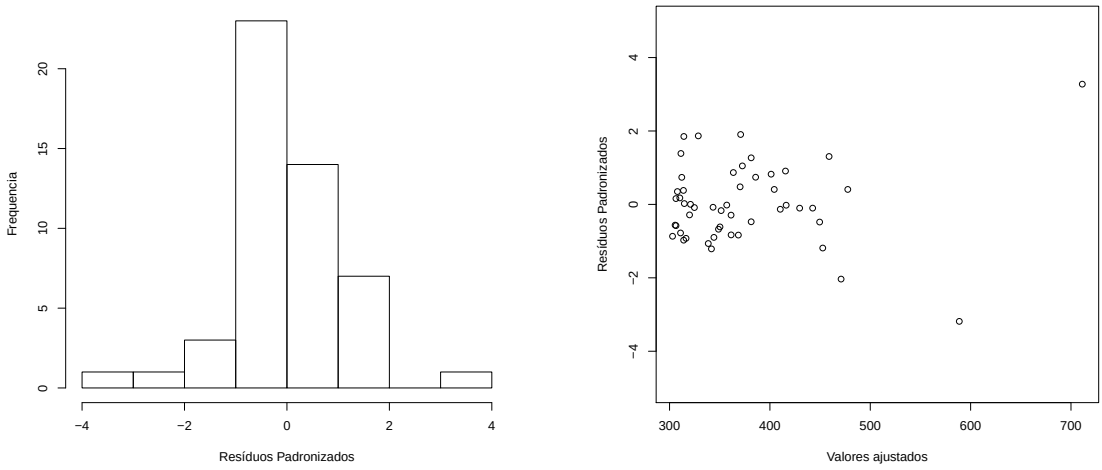
A Tabela 6.3 apresenta as estimativas pontuais do modelo $y_i = \beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}^2+\varepsilon_i$, $i = 1, ..., 50$, via os estimadores MQO, LMS e LTS. Podemos observar que as estimativas via LTS para β_0, β_1 e β_2 são quase o dobro das estimativas dos mínimos quadrados ordinários, porém apesar disso todos os estimadores apresentaram o mesmo sinal nas estimações.

Tabela 6.3: Estimativas pontuais de β_0, β_1 e β_2 considerando o modelo $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}^2 + \varepsilon_i$, $i = 1, ..., 50$, via os estimadores MQO, LMS e LTS.

Estimadores	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
MQO	832	-1834	1587
LMS	1323	-3250	2558
LTS	1808	-4365	3196

O teste de normalidade de Shapiro & Wilks (1965), aplicado nos resíduos padronizados do modelo $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2}^2 + \hat{\varepsilon}_i$ com $i = 1, ..., 50$, gerou um p-valor< 0.1721 garantindo a suposição de normalidade do modelo ao nível de 5% de significância. O teste homoscedasticidade de Koenker & Bassett (1978) gerou um p-valor $< 3 \times 10^{-4}$, ou seja, estamos diante de um cenário heteroscedástico com 5% de significância. A Figura 6.8 apresenta o comportamento dos resíduos padronizados, sendo possível observa no gráfico (b) que os resíduos são tendenciosos, já que as observações não encontram-se espalhadas

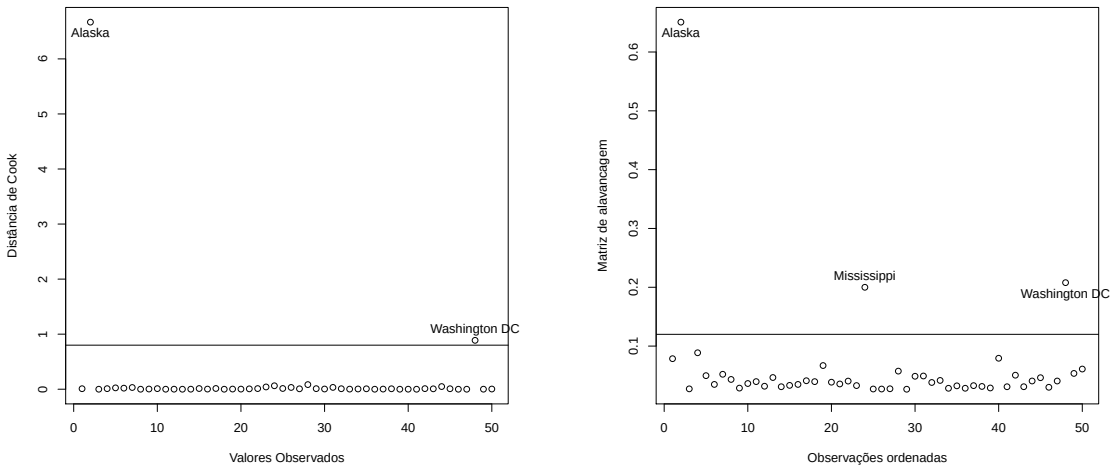
aleatoriamente entre os limites $[-2, 2]$ configurando um cenário heteroscedástico.



(a) (b)

Figura 6.8: Histograma (a) e gráfico para verificar a suposição de homoscedasticidade dos resíduos (b), considerando os resíduos padronizados do modelo $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2}^2 + \hat{\varepsilon}_i$, $i = 1, ..., 50$.

Na Figura 6.9 (a), referente a Distância de Cook, é possível observa a presença de dois pontos de influência na variável resposta, correspondente aos Estados do Alaska com distância de CooK 6.66 e Washington DC com distância de Cook 8.87×10^{-1}, pois ultrapassam o quantil da distribuição $F_{(0.5,3,47)} \simeq 0.8$, diferente do modelo $y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$, $i = 1, ..., 50$, que apresentava apenas um ponto de influência correspondente ao Estado do Alaska. A Figura 6.9 (b) apresenta três pontos de alavanca referentes aos Estados do Alaska com $h_i \simeq 0.650$, Washington DC com $h_i \simeq 0.207$ e Mississippi $h_i \simeq 0.200$, pois ultrapassam a quantidade $2p/n \simeq 0.12$. Por fim, na Figura 6.10 é possível também verifica que os Estados do Alaska e Washington DC se destacam dos demais por apresentarem círculos com uma grande área em relação aos demais Estados.



(a) (b)

Figura 6.9: Distância de Cook (a) e matriz de alavancagem (b) correspondente ao modelo $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}^2 + \varepsilon_i$, $i = 1, ..., 50$.

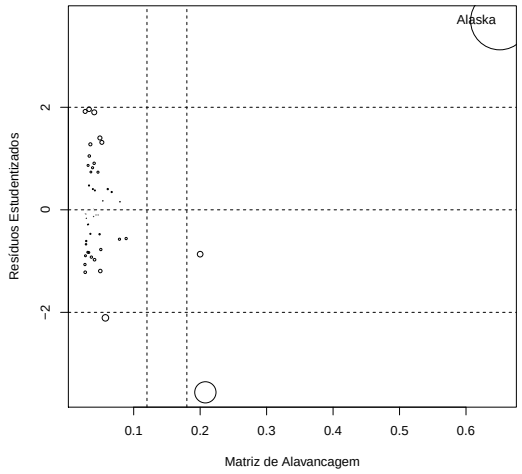


Figura 6.10: Gráfico para verificar possíveis pontos de influência no modelo $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}^2 + \varepsilon_i$.

Portanto no modelo $y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$ vimos que o ajuste de regressão via LTS sofreu o menor desvio devido a alta alavancagem do Estado do Alaska. Por fim, com a inclusão da variável x^2 o modelo $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}^2 + \varepsilon_i$ apresentou melhor capacidade explicativa e a diferença nas estimativas pontuais, pelos estimadores robustos e de mínimos quadrados ordinários, pode ser devido a influência dos *outliers* e pontos de

alavanca, concluindo que a regressão robusta é uma alternativa viável já que as estimativas de mínimos quadrados ordinários é bastante afetada por essas observações.

Capítulo 7

Considerações Finais

7.1 Conclusões

Através das simulações de Monte Carlo foi visto que no cenário homoscedástico balanceado o estimador MQO apresentou o menor viés com maiores tamanhos amostrais, já no caso não-balanceado esse resultado se inverte, ou seja, os estimadores robustos geraram o menor viés com maiores tamanhos amostrais. No cenário heteroscedástico balanceado e não-balanceado os estimadores robustos foram muito superiores ao estimador MQO, pois apresentaram menor viés, exceto no cenário com erros obtidos a partir da distribuição normal padrão, entretanto quando se aumenta o grau de heteroscedasticidade os estimadores robustos passam a apresentar menor viés nos tamanhos amostrais superiores. Por fim, foi visto que o estimador LTS apresentou o maior número de estimativas com menor viés além de gerar o menor erro quadrático médio nos cenários heteroscedásticos.

Considerando a aplicação dos dados extraídos de Greene (1997, Tabela 12.1, p.541), no primeiro modelo $y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$, $i = 1, ..., 50$, através do teste de Shapiro & Wilks (1965) foi visto que os erros não eram normalmente distribuídos e através das medidas de influência, como distância de Cook e matriz de alavancagem, foi possível identificar um ponto de influência referente ao Estado do Alaska e três pontos de alavanca referente aos Estados do Alaska, Mississippi e Washington DC, pois suas observações excederam os limites estabelecidos pelo quantil da distribuição $F_{(0.5,p,n-p)}$ e $2p/n$, respectivamente. Através do teste de Koenker & Bassett (1978) foi possível identificar um cenário sob heteroscedasticidade e foi visto que a reta de regressão do modelo via estimador LTS sofreu o menor desvio devido as observações discrepantes. Em relação as estimativas pontuais

vimos que o estimador MQO gerou um $\hat{\beta}_0$ quase três vezes menor que a estimativa de β_0 via LTS. Com o teste de linearidade de Ramsey (1969) foi observado que o modelo precisava da variável x^2 devido a possível relação quadrática, gerando um novo modelo denotado por $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}^2 + \varepsilon_i,\ i = 1, ...50$, aumentando o coeficiente de determinação $R^2 \simeq 0.58$ para $R^2 \simeq 0.65$. Por fim, através da distância de Cook foi possível observar que o novo modelo apresentou dois pontos de influência, a saber, os Estados do Alaska e Washington DC. Observou-se também que as estimativas do LTS para β_0,β_1 e β_2 foram quase o dobro das estimativas via MQO. Portanto sob o cenário heteroscedástico e com pontos de alavanca, considerando que os erros não são normalmente distribuídos, os estimadores robustos podem se tornar uma alternativa viável ao método de estimação de mínimos quadrados ordinários.

7.2 Trabalhos Futuros

Serão foco das nossas pesquisas futuras:

- Avaliar o desempenho dos estimadores na presença de *outliers*.
- Avaliar o desempenho dos estimadores na presença de *outliers* e pontos de alavanca.
- Avaliar os estimadores via teste de hipótese e intervalos de confiança.

Referências Bibliográficas

[1] Barbiere, N.B. (2012). Estimação robusta para o modelo de regressão logística. Monografia, Instituto de Matemática, Universidade Federal do Rio Grande do Sul, Brasil.

[2] Bulhões, R.S. (2013). Contribuição a análise de *outliers* em modelos de equação estruturais. Dissertação de mestrado, Instituto de Matemática e Estatística, Universidade Federal de São Paulo, Brasil.

[3] Cribari-Neto, F.; Soares, A.C.N. (2003). Inferência em modelos heteroscedásticos. *RBE*, 57(2), 319–335.

[4] Cribari-Neto, F.; Souza, T.C; Vasconcellos, K.L.P. (2007). Inference under heteroskedasticity and laverage data. *Comunications in Statistics–Theory and Methods*, 36, 1877–1888.

[5] Cribari-Neto, F.; Zarkos, S.G. (2004). Leverage-adjusted heteroskedastic bootstrap methods. *Journal of Statistical Computation and Simulation*, 74, 215–232.

[6] Cook, R.D. (2007). Detection of influential observation in linear regression, *Technometrics*, 19, 1, 15-18.

[7] Davidson, R.; MacKinnon, J.G. (1993). *Estimation and Inference in Econometrics*. New York: Oxford University Press.

[8] Draper, N.R.; Smith, H. (1966). *Applied Regression Analysis*, Editora John Wiley & Sons, second edition, .

[9] Draper, N.R.; Smith, H. (1998). *Applied Regression Analysis*, Editora Wily-Interscience, Third edition.

[10] Fox, J. (2002).*An R and S-PLUS Companion to Applied Regression*. SAGE

[11] Gujarati, D.N. (2006). *Econométrica Básica*, quarta edição, Elsevier.

[12] Greene, W.H. (1977). *Econometric Analysis*, Third edition. Upper Saddle River: Prentice Hall.

[13] Hampel, F.R. (1971), A general qualitative definition of robustness. *Annals of Mathematical Statistics*, 42, 6, 1887–1896.

[14] Hinkley, S.D. (1977). Jakkniifing in unbalanced situations. *Technometrics*, 19, 285-292.

[15] Horn, S.D.; Horn, R.A.; Duncan, D.B. (1975). Estimating heteroskedastic variances in linear models. *Journal of the American Statistical Association*, 70, 380–385.

[16] Kleiber, C.; Zeileis, A. (2008). *Applied Econometrics with R*. New York: Springer.

[17] Koenker, R.; Bassett, G.Jr. (1978). Regression quantiles. *Econometrica* , 46, 1, 33-50.

[18] MacKinnon, J.G.; White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite-sample properties. *Journal of Ecnonometrics*, 29, 305-325.

[19] Ramsey, J.B. (1969). Tests for specification erros in classical linear leas-square regression analysis. *Journal of Royal Statistical Society*, Serie B (Methodological), 31, 350–371.

[20] Ramirez, F.A.P. (2014). Testes quasi-t em modelos lineares de regressão: Avaliação numérica. Dissertação de mestrado, Departamento de Estatística, Universidade Federal de Pernambuco, Brasil.

[21] Ryan, T.P. (2009). *Modern Regression Methods*, Wiley series, Second edition.

[22] Rousseeuw, P.J. (1984). Least median of squares regression, *Journal of the American Statistical Association*, 79, 388.

[23] Souza, T.C. (2003). Inferência em modelos heteroscedásticos na presença de pontos de alavanca. Dissertação de mestrado, Departamento de Estatística, Universidade Federal de Pernambuco, Brasil.

[24] Souza, T.C. (2011). Ensaios sobre modelos de regressão com dispersão variável. Tese de doutorado, Departamento de Estatística, Universidade Federal de Pernambuco, Brasil.

[25] Shapiro, S.S.; Wilk, M.B. (1965). An analysis of variance test for normality (Complete Samples). *Biometrika Trust*, 52, 3/4, 591–611.

Apêndice

Neste apêndice apresentamos os comandos utilizados nas simulações de Monte Carlo e na aplicação.

```
######### Regressão Robusta #############
###### Simulação de Monte Carlo #####

#Será necessário reproduzir o cenário 1 varias vezes
#(até um total de 3 simulações)
#trocando o tipo do erro e o número de replicas de x10 e x20 (de 1 até 3 réplicas)

########## Pacotes requeridos ########
library(MASS)
library(car)
#############cenário 1 homoscedástico balanceado
rm(list=ls()) #remover os valores antigos, para comerçar novo cenário
######################  CENÁRIO 1 - EXPERIMENTO BALANCEADO
#é necessário reproduzir esse cenário três vezes para cada
#distribuição. Fazendo os valores das replicas de x1 e x2 variar de 1 até 3.
# Número de Replicas
R=10000
# Gerando as covariaveis
set.seed(4) #semente
x10 = runif(50, min = 0, max = 1) #variável x1
x20 = runif(50, min = 0, max = 1) #variável x2
#replicar a variável x1 e x2:1, 2, e 3 vezes,
#para gerar os tamanhos amostrais diferentes n=50,100,150.
x1=rep(x10,1)
x2=rep(x20,1)

#tamanho da variável x1
len =length(x1)
# funcao cedastica - heteroscedasticidade
#como o cenário é homoscedástico basta igualar alpha_1=alpha_2=0, torna o sigma=1
```

```
alpha_1=0
alpha_2=0
sigma=sqrt(exp(alpha_1+alpha_2*x1))
razao = (max(sigma)/min(sigma))
razao
# Valores para beta (verdadeiro valor do parâmetro)
b0 = 1
b1 = 1
b2 = 0
# Inicializando os vetores
coeflm= matrix(0,R,3)
coeflms= matrix(0,R,3)
coeflts= matrix(0,R,3)
for  (i in 1:R){
 # Obtendo y
 y  = b0+b1*x1+b2*x2+sigma*rnorm(len) #mudar o tipo de erro
 # ajustando o modelo - lm
 ajustelm=lm(y~x1+x2)
 # ajustando o modelo - lms
 ajustelms=lqs(y~x1+x2, method = "lms")
 # ajustando o modelo - lts
 ajustelts=lqs(y~x1+x2, method = "lts")
 # armazenando os coef
 coeflm[i,]=ajustelm$coef
 coeflms[i,]=ajustelms$coef
 coeflts[i,]=ajustelts$coef
  }
############ Vieses
# Obtendo os vieses - lm
viesLMb0 = coeflm[,1]-b0
viesLMb1 = coeflm[,2]-b1
viesLMb2 = coeflm[,3]-b2
#viés médio
mean(viesLMb0)
```

```
mean(viesLMb1)
mean(viesLMb2)
# Obtendo os vieses - lms
viesLMSb0 = coeflms[,1]-b0
viesLMSb1 = coeflms[,2]-b1
viesLMSb2 = coeflms[,3]-b2
#viés médio
mean(viesLMSb0)
mean(viesLMSb1)
mean(viesLMSb2)
# Obtendo os vieses - lts
viesLTSb0 = coeflts[,1]-b0
viesLTSb1 = coeflts[,2]-b1
viesLTSb2 = coeflts[,3]-b2
#viés médio
mean(viesLTSb0)
mean(viesLTSb1)
mean(viesLTSb2)
############ Variancia
# Obtendo variancias - lm
varLMb0 = var(coeflm[,1])
varLMb1 = var(coeflm[,2])
varLMb2 = var(coeflm[,3])
varLMSb0 = var(coeflms[,1])
varLMSb1 = var(coeflms[,2])
varLMSb2 = var(coeflms[,3])
varLTSb0 = var(coeflts[,1])
varLTSb1 = var(coeflts[,2])
varLTSb2 = var(coeflts[,3])
####### EQM
# Obtendo EQM - lm
EQMLMb0 = mean(varLMb0)+ ((mean(viesLMb0))^2)
EQMLMb1 = mean(varLMb1)+ ((mean(viesLMb1))^2)
EQMLMb2 = mean(varLMb2)+ ((mean(viesLMb2))^2)
```

```
# Obtendo EQM - lms
EQMLMSb0 = mean(varLMSb0)+ ((mean(viesLMSb0))^2)
EQMLMSb1 = mean(varLMSb1)+ ((mean(viesLMSb1))^2)
EQMLMSb2 = mean(varLMSb2)+ ((mean(viesLMSb2))^2)
# Obtendo EQM - lts
EQMLTSb0 = mean(varLTSb0)+ ((mean(viesLTSb0))^2)
EQMLTSb1 = mean(varLTSb1)+ ((mean(viesLTSb1))^2)
EQMLTSb2 = mean(varLTSb2)+ ((mean(viesLTSb2))^2)
#TABELAS COM OS VALORES DOS VIES E DO EQM PARA
#criando vetor para armazenar os valores do vies e EQM por método
arg1=c(mean(viesLMb0),mean(viesLMb1),mean(viesLMb2),
EQMLMb0,EQMLMb1,EQMLMb2) #LM
arg2=c(mean(viesLMSb0),mean(viesLMSb1),mean(viesLMSb2),
EQMLMSb0,EQMLMSb1,EQMLMSb2) #LMS
arg3=c(mean(viesLTSb0),mean(viesLTSb1),mean(viesLTSb2),
EQMLTSb0,EQMLTSb1,EQMLTSb2) #LTS
#criando uma matrix linha com o vies e EQM de cada vetor
uu=rbind(arg1,arg2,arg3)
#organizando a saida da tabela
cenario1=matrix(uu,nrow=9,ncol=2)
colnames(cenario1)=c("Vies","EQM")
rownames(cenario1)=c("LMb0","LMSb0",
"LTSb0","LMb1","LMSb1","LTSb1","LMb2","LMSb2","LTSb2")

############################################################

#############cenário 2 homoscedástico não-balanceado
rm(list=ls()) #remover os valores antigos, para comerçar novo cenário
######################  CENÁRIO 2 - EXPERIMENTO NÃO-BALANCEADO
#é necessário reproduzir esse cenário três vezes para cada
#distribuição. Fazendo os valores das replicas de x1 e x2 variar de 1 até 3.
# Número de Replicas
R=10000
# Gerando as covariaveis
```

```
set.seed(4) #semente
x10 = runif(50, min = 0, max = 1) #variável x1
x20 = runif(50, min = 0, max = 1) #variável x2
#replicar a variável x1 e x2: 1, 2, e 3 vezes, para gerar os tamanhos
#amostrais diferentes n=50,100,150.
x1=rep(x10,1)
x2=rep(x20,1)
#pontos de alavanca nas duas últimas observações
n=50
x2[n-1]=4
x2[n]=6
#tamanho da variável x1
len =length(x1)
# funcao cedastica - heteroscedasticidade
#como o cenário é homoscedástico basta igualar alpha_1=alpha_2=0, torna o sigma=1
alpha_1=0
alpha_2=0
sigma=sqrt(exp(alpha_1+alpha_2*x1))
razao = (max(sigma)/min(sigma))
# Valores para beta (verdadeiro valor do parâmetro)
b0 = 1
b1 = 1
b2 = 0
# Inicializando os vetores
coeflm= matrix(0,R,3)
coeflms= matrix(0,R,3)
coeflts= matrix(0,R,3)
for  (i in 1:R){
 # Obtendo y
 y  = b0+b1*x1+b2*x2+sigma*rnorm(len)#mudar o tipo de erro
 # ajustando o modelo - lm
 ajustelm=lm(y~x1+x2)
 # ajustando o modelo - lms
 ajustelms=lqs(y~x1+x2, method = "lms")
```

```
 # ajustando o modelo - lts
 ajustelts=lqs(y~x1+x2, method = "lts")
 # armazenando os coef
 coeflm[i,]=ajustelm$coef
 coeflms[i,]=ajustelms$coef
 coeflts[i,]=ajustelts$coef
  }
############ Vieses
# Obtendo os vieses - lm
viesLMb0 = coeflm[,1]-b0
viesLMb1 = coeflm[,2]-b1
viesLMb2 = coeflm[,3]-b2
#viés médio
mean(viesLMb0)
mean(viesLMb1)
mean(viesLMb2)
# Obtendo os vieses - lms
viesLMSb0 = coeflms[,1]-b0
viesLMSb1 = coeflms[,2]-b1
viesLMSb2 = coeflms[,3]-b2
#viés médio
mean(viesLMSb0)
mean(viesLMSb1)
mean(viesLMSb2)
# Obtendo os vieses - lts
viesLTSb0 = coeflts[,1]-b0
viesLTSb1 = coeflts[,2]-b1
viesLTSb2 = coeflts[,3]-b2
#viés médio
mean(viesLTSb0)
mean(viesLTSb1)
mean(viesLTSb2)
############ Variancia
# Obtendo variancias - lm
```

```
varLMb0 = var(coeflm[,1])
varLMb1 = var(coeflm[,2])
varLMb2 = var(coeflm[,3])
varLMSb0 = var(coeflms[,1])
varLMSb1 = var(coeflms[,2])
varLMSb2 = var(coeflms[,3])
varLTSb0 = var(coeflts[,1])
varLTSb1 = var(coeflts[,2])
varLTSb2 = var(coeflts[,3])
####### EQM
# Obtendo EQM - lm
EQMLMb0 = mean(varLMb0)+ ((mean(viesLMb0))^2)
EQMLMb1 = mean(varLMb1)+ ((mean(viesLMb1))^2)
EQMLMb2 = mean(varLMb2)+ ((mean(viesLMb2))^2)
# Obtendo EQM - lms
EQMLMSb0 = mean(varLMSb0)+ ((mean(viesLMSb0))^2)
EQMLMSb1 = mean(varLMSb1)+ ((mean(viesLMSb1))^2)
EQMLMSb2 = mean(varLMSb2)+ ((mean(viesLMSb2))^2)
# Obtendo EQM - lts
EQMLTSb0 = mean(varLTSb0)+ ((mean(viesLTSb0))^2)
EQMLTSb1 = mean(varLTSb1)+ ((mean(viesLTSb1))^2)
EQMLTSb2 = mean(varLTSb2)+ ((mean(viesLTSb2))^2)
#TABELAS COM OS VALORES DOS VIES E DO EQM PARA
#criando vetor para armazenar os valores do vies e EQM por método
arg1=c(mean(viesLMb0),mean(viesLMb1),mean(viesLMb2),
EQMLMb0,EQMLMb1,EQMLMb2) #LM
arg2=c(mean(viesLMSb0),mean(viesLMSb1),mean(viesLMSb2),
EQMLMSb0,EQMLMSb1,EQMLMSb2) #LMS
arg3=c(mean(viesLTSb0),mean(viesLTSb1),mean(viesLTSb2),
EQMLTSb0,EQMLTSb1,EQMLTSb2) #LTS
#criando uma matrix linha com o vies e EQM de cada vetor
uu=rbind(arg1,arg2,arg3)
#organizando a saida da tabela
cenario2=matrix(uu,nrow=9,ncol=2)
```

```
colnames(cenario2)=c("Vies","EQM")
rownames(cenario2)=c("LMb0","LMSb0",
"LTSb0","LMb1","LMSb1","LTSb1","LMb2","LMSb2","LTSb2")

############################################################

#############cenário 3 heteroscedástico balanceado
rm(list=ls()) #remover os valores antigos, para comerçar novo cenário
######################  CENÁRIO 3 - EXPERIMENTO BALANCEADO
#é necessário reproduzir esse cenário três vezes para cada
#distribuição e grau de heteroscedasticidade.
#Fazendo os valores das replicas de x1 e x2 variar de 1 até 3.
# Número de Replicas
R=10000
# Gerando as covariaveis
set.seed(4) #semente
x10 = runif(50, min = 0, max = 1) #variável x1
x20 = runif(50, min = 0, max = 1) #variável x2
#replicar a variável x1 e x2: 1, 2, e 3 vezes, para gerar os tamanhos
#amostrais diferentes n=50,100,150.
x1=rep(x10,1)
x2=rep(x20,1)
#tamanho da variável x1
len =length(x1)
# funcao cedastica - heteroscedasticidade
# Será considerando 2 graus de heteroscedasticidade - 50 e 100
# Realizar as três simulações para cada grau de heteroscedasticidade
# Razao proxima de 50
alpha_1=7.900
alpha_2=7.900
# Razao proxima 100
#alpha_1=9.309
#alpha_2=9.309
sigma=sqrt(exp(alpha_1+alpha_2*x1))
```

```
razao = (max(sigma)/min(sigma))
# Valores para beta (verdadeiro valor do parâmetro)
b0 = 1
b1 = 1
b2 = 0
# Inicializando os vetores
coeflm= matrix(0,R,3)
coeflms= matrix(0,R,3)
coeflts= matrix(0,R,3)
for  (i in 1:R){
 # Obtendo y
 y  = b0+b1*x1+b2*x2+sigma*rnorm(len)#mudar o tipo de erro
 # ajustando o modelo - lm
 ajustelm=lm(y~x1+x2)
 # ajustando o modelo - lms
 ajustelms=lqs(y~x1+x2, method = "lms")
 # ajustando o modelo - lts
 ajustelts=lqs(y~x1+x2, method = "lts")
 # armazenando os coef
 coeflm[i,]=ajustelm$coef
 coeflms[i,]=ajustelms$coef
 coeflts[i,]=ajustelts$coef
  }
############ Vieses
# Obtendo os vieses - lm
viesLMb0 = coeflm[,1]-b0
viesLMb1 = coeflm[,2]-b1
viesLMb2 = coeflm[,3]-b2
#viés médio
mean(viesLMb0)
mean(viesLMb1)
mean(viesLMb2)
# Obtendo os vieses - lms
viesLMSb0 = coeflms[,1]-b0
```

```
viesLMSb1 = coeflms[,2]-b1
viesLMSb2 = coeflms[,3]-b2
#viés médio
mean(viesLMSb0)
mean(viesLMSb1)
mean(viesLMSb2)
# Obtendo os vieses - lts
viesLTSb0 = coeflts[,1]-b0
viesLTSb1 = coeflts[,2]-b1
viesLTSb2 = coeflts[,3]-b2
#viés médio
mean(viesLTSb0)
mean(viesLTSb1)
mean(viesLTSb2)
############ Variancia
# Obtendo variancias - lm
varLMb0 = var(coeflm[,1])
varLMb1 = var(coeflm[,2])
varLMb2 = var(coeflm[,3])
varLMSb0 = var(coeflms[,1])
varLMSb1 = var(coeflms[,2])
varLMSb2 = var(coeflms[,3])
varLTSb0 = var(coeflts[,1])
varLTSb1 = var(coeflts[,2])
varLTSb2 = var(coeflts[,3])
####### EQM
# Obtendo EQM - lm
EQMLMb0 = mean(varLMb0)+ ((mean(viesLMb0))^2)
EQMLMb1 = mean(varLMb1)+ ((mean(viesLMb1))^2)
EQMLMb2 = mean(varLMb2)+ ((mean(viesLMb2))^2)
# Obtendo EQM - lms
EQMLMSb0 = mean(varLMSb0)+ ((mean(viesLMSb0))^2)
EQMLMSb1 = mean(varLMSb1)+ ((mean(viesLMSb1))^2)
EQMLMSb2 = mean(varLMSb2)+ ((mean(viesLMSb2))^2)
```

```
# Obtendo EQM - lts
EQMLTSb0 = mean(varLTSb0)+ ((mean(viesLTSb0))^2)
EQMLTSb1 = mean(varLTSb1)+ ((mean(viesLTSb1))^2)
EQMLTSb2 = mean(varLTSb2)+ ((mean(viesLTSb2))^2)
#TABELAS COM OS VALORES DOS VIES E DO EQM PARA
#criando vetor para armazenar os valores do vies e EQM por método
arg1=c(mean(viesLMb0),mean(viesLMb1),mean(viesLMb2),
EQMLMb0,EQMLMb1,EQMLMb2) #LM
arg2=c(mean(viesLMSb0),mean(viesLMSb1),mean(viesLMSb2),
EQMLMSb0,EQMLMSb1,EQMLMSb2) #LMS
arg3=c(mean(viesLTSb0),mean(viesLTSb1),mean(viesLTSb2),
EQMLTSb0,EQMLTSb1,EQMLTSb2) #LTS
#criando uma matrix linha com o vies e EQM de cada vetor
uu=rbind(arg1,arg2,arg3)
#organizando a saida da tabela
cenario3=matrix(uu,nrow=9,ncol=2)
colnames(cenario3)=c("Vies","EQM")
rownames(cenario3)=c("LMb0","LMSb0",
"LTSb0","LMb1","LMSb1","LTSb1","LMb2","LMSb2","LTSb2")

############################################################

#############cenário 4 heteroscedástico não-balanceado
rm(list=ls()) #remover os valores antigos, para comerçar novo cenário
######################  CENÁRIO 4 - EXPERIMENTO NÃO-BALANCEADO
#é necessário reproduzir esse cenário três vezes para cada
#distribuição e grau de heteroscedasticidade.
#Fazendo os valores das replicas de x1 e x2 variar de 1 até 3.
# Número de Replicas
R=10000
# Gerando as covariaveis
set.seed(4) #semente
x10 = runif(50, min = 0, max = 1) #variável x1
x20 = runif(50, min = 0, max = 1) #variável x2
```

```
#replicar a variável x1 e x2: 1, 2, e 3 vezes, para gerar os tamanhos
#amostrais diferentes n=50,100,150.
x1=rep(x10,1)
x2=rep(x20,1)
#pontos de alavanca nas duas últimas observações
n=50
x2[n-1]=4
x2[n]=6
#tamanho da variável x1
len =length(x1)
# funcao cedastica - heteroscedasticidade
# Será considerando 2 graus de heteroscedasticidade - 50 e 100
# Realizar as três simulações para cada grau de heteroscedasticidade
# Razao proxima de 50
alpha_1=7.900
alpha_2=7.900
# Razao proxima 100
#alpha_1=9.309
#alpha_2=9.309
sigma=sqrt(exp(alpha_1+alpha_2*x1))
razao = (max(sigma)/min(sigma))
# Valores para beta (verdadeiro valor do parâmetro)
b0 = 1
b1 = 1
b2 = 0
# Inicializando os vetores
coeflm= matrix(0,R,3)
coeflms= matrix(0,R,3)
coeflts= matrix(0,R,3)
for  (i in 1:R){
 # Obtendo y
 y  = b0+b1*x1+b2*x2+sigma*rnorm(len)#mudar o tipo de erro
 # ajustando o modelo - lm
 ajustelm=lm(y~x1+x2)
```

```
 # ajustando o modelo - lms
 ajustelms=lqs(y~x1+x2, method = "lms")
 # ajustando o modelo - lts
 ajustelts=lqs(y~x1+x2, method = "lts")
 # armazenando os coef
 coeflm[i,]=ajustelm$coef
 coeflms[i,]=ajustelms$coef
 coeflts[i,]=ajustelts$coef
  }
############ Vieses
# Obtendo os vieses - lm
viesLMb0 = coeflm[,1]-b0
viesLMb1 = coeflm[,2]-b1
viesLMb2 = coeflm[,3]-b2
#viés médio
mean(viesLMb0)
mean(viesLMb1)
mean(viesLMb2)
# Obtendo os vieses - lms
viesLMSb0 = coeflms[,1]-b0
viesLMSb1 = coeflms[,2]-b1
viesLMSb2 = coeflms[,3]-b2
#viés médio
mean(viesLMSb0)
mean(viesLMSb1)
mean(viesLMSb2)
# Obtendo os vieses - lts
viesLTSb0 = coeflts[,1]-b0
viesLTSb1 = coeflts[,2]-b1
viesLTSb2 = coeflts[,3]-b2
#viés médio
mean(viesLTSb0)
mean(viesLTSb1)
mean(viesLTSb2)
```

```
############ Variancia
# Obtendo variancias - lm
varLMb0 = var(coeflm[,1])
varLMb1 = var(coeflm[,2])
varLMb2 = var(coeflm[,3])
varLMSb0 = var(coeflms[,1])
varLMSb1 = var(coeflms[,2])
varLMSb2 = var(coeflms[,3])
varLTSb0 = var(coeflts[,1])
varLTSb1 = var(coeflts[,2])
varLTSb2 = var(coeflts[,3])
####### EQM
# Obtendo EQM - lm
EQMLMb0 = mean(varLMb0)+ ((mean(viesLMb0))^2)
EQMLMb1 = mean(varLMb1)+ ((mean(viesLMb1))^2)
EQMLMb2 = mean(varLMb2)+ ((mean(viesLMb2))^2)
# Obtendo EQM - lms
EQMLMSb0 = mean(varLMSb0)+ ((mean(viesLMSb0))^2)
EQMLMSb1 = mean(varLMSb1)+ ((mean(viesLMSb1))^2)
EQMLMSb2 = mean(varLMSb2)+ ((mean(viesLMSb2))^2)
# Obtendo EQM - lts
EQMLTSb0 = mean(varLTSb0)+ ((mean(viesLTSb0))^2)
EQMLTSb1 = mean(varLTSb1)+ ((mean(viesLTSb1))^2)
EQMLTSb2 = mean(varLTSb2)+ ((mean(viesLTSb2))^2)
#TABELAS COM OS VALORES DOS VIES E DO EQM PARA
#criando vetor para armazenar os valores do vies e EQM por método
arg1=c(mean(viesLMb0),mean(viesLMb1),mean(viesLMb2),
EQMLMb0,EQMLMb1,EQMLMb2) #LM
arg2=c(mean(viesLMSb0),mean(viesLMSb1),mean(viesLMSb2),
EQMLMSb0,EQMLMSb1,EQMLMSb2) #LMS
arg3=c(mean(viesLTSb0),mean(viesLTSb1),mean(viesLTSb2),
EQMLTSb0,EQMLTSb1,EQMLTSb2) #LTS
#criando uma matrix linha com o vies e EQM de cada vetor
uu=rbind(arg1,arg2,arg3)
```

```
#organizando a saida da tabela
cenario4=matrix(uu,nrow=9,ncol=2)
colnames(cenario4)=c("Vies","EQM")
rownames(cenario4)=c("LMb0","LMSb0",
"LTSb0","LMb1","LMSb1","LTSb1","LMb2","LMSb2","LTSb2")

############################################################

###### Aplicação  ######

#Pacote necessário
library(sandwich)
library(MASS)
library(car)
library(lmtest)
library(psych)

#Banco de Dados
data("PublicSchools")
ps=na.omit(PublicSchools)
#Variáveis
#Expenditure - Gasto
#Income - Renda
#variável Income Reescalonada
ps$Income=ps$Income/10000

#Matriz de alavancagem
ps_hat=hatvalues(ps_lm)
ps_hat

#Identificando possíveis pontos de alavanca
quoc = 2*2/50 #(2*p/n)
ifelse(ps_hat>quoc,1,0)
sum(ifelse(ps_hat>quoc,1,0)) # três possíveis pontos de alavanca
```

```
#Gráfico para identificar os pontos de alavanca
plot(ps_hat, ylab="Matriz de alavancagem", xlab="Observações ordenadas")
abline(h=quoc,col="black") # cota 3p/n
id=which(ps_hat>quoc) #valores que ultrapassam a cota
text(id,ps_hat[id], rownames(ps)[id], pos=c(1,3),xpd=TRUE)

#medidas de influência
influence.measures(ps_lm)

#Distância de Cook
dc=influence.measures(ps_lm)$infmat[,5]#distância de cook

#reta com os ajustes dos estimadores MQO, LMS e LTS
plot(ps$Expenditure~ps$Income, data=ps,ylim=c(230,830), xlab=
"Renda per capita", ylab=" gasto per capita em escolas públicas")
abline(ps_lm)
ps_lm=lm(ps$Expenditure~ps$Income, data=ps)
id=which(apply(influence.measures(ps_lm)$is.inf,1,any))
text(ps[id,2:1],rownames(ps)[id],pos=1,xpd=TRUE)
#ps_noinf=lm(Expenditure~Income, data=ps[-id,])
#abline(ps_noinf,lty=2)
ajusteLMS = lqs(ps$Expenditure~ps$Income, data=ps, method = "lms")
 abline(ajusteLMS,lty=4)
ajusteLTS = lqs(ps$Expenditure~ps$Income, data=ps, method = "lts")
 abline(ajusteLTS,lty=3)
legend("topleft", legend=c("MQO","LMS","LTS"),lty = 1:4,bty="n")

#VERIFICANDO SUPOSIÇÃO DE HOMOSCEDASTICIDADE
#Teste Homoscedasticidade , H0:Homoscedástico x H1:Heteroscedástico
bptest(ps_lm)# Heteroscedástico , rejeita a hipótese nula

#Inclusão da variável renda ao quadrado
Income2 = ps$Income*ps$Income
```

```
X = cbind(ps$Income, Income2)

#Modelo 2
ps_lm2=lm(Expenditure~Income+I(Income^2), data=ps)
 summary(ps_lm2)

#AJUSTANDO NOVO MODELO - LMS e LTS
ajusteLTS = lqs(Expenditure~X, data=ps, method = "lts")
summary(ajusteLTS)
ajusteLMS = lqs(Expenditure~X, data=ps, method = "lms")
summary(ajusteLMS)

#ANÁLISE DESCRITIVA DAS VARIÁVEIS
summary(ps$Income)
summary(ps$Expenditure)

#Medida descritiva
describe(ps) # descreve as variáveis do banco de dados

#BOXPLOT DAS VARIÁVEIS
boxplot(ps$Expenditure)
boxplot(ps$Income)

#MEDIDAS DE DIAGNÓSTICO
#gráfico para identificar pontos de influência

summary(influence.measures(ps_lm)) # resumo das medidas de influência
influencePlot(lm(Expenditure ~ Income, data=ps), xlab="Matriz de Alavancagem",
ylab="Resíduos Estudentizados")

#homocedasticidade tecnica gráfica - modelo1
plot(rstandard(ps_lm)~ps_lm$fitted.values,ylim=c(-5,5), ylab=
"Resíduos Padronizados", xlab="Valores ajustados")
```

```
#######################Ponto de influência
#Ponto Influente dc>F(p,n-p,(0.5)) será influente
plot(dc, xlab="Valores Observados",ylab="Distância de Cook")
abline(h=qf(0.5,2,48))
id=which(dc>qf(0.5,2,48)) #valores que ultrapassam a cota
text(id,dc[id], rownames(ps)[id], pos=c(1,3),xpd=TRUE)

#teste de linearidade; H0: linearidade x H1: não linearidade
## Tecnica Grafica
plot(ps$Expenditure~ps_lm$fitted.values, ylab="Valores observados",xlab=
"Valores Ajustados")
abline(a=0,b=1,col="black")
resettest(ps$Expenditure~ps$Income , power=c(2), type="regressor")

######### Normalidade ######### Modelo 1
## Tecnica Grafica
#Histograma
hist(rstandard(ps_lm), ylab="Frequencia", xlab="Resíduos Padronizados", main="")

#QQplot
qqnorm(rstandard(ps_lm), ylim=c(-3,2),xlim=c(-3,2), main=""
, ylab="Quantidades amostrais", xlab=
"Quantidades teóricas")
abline(a=0,b=1,col="black")
shapiro.test(rstandard(ps_lm))

# Teste Koenker
library(lmtest)
bptest(ps$Expenditure~ps$Income, studentize=TRUE)

#Tipos de resíduo
rstandard(ps_lm) #padronizados
rstudent(ps_lm)  #estudentizados
```

```
#####################################################################
#####################################################################
################## Modelo 2
ps_lm2=lm(Expenditure~Income+I(Income^2), data=ps)
ps_lts2=lqs(Expenditure~Income+I(Income^2), data=ps, method = "lts")
ps_lms2= lqs(Expenditure~Income+I(Income^2), data=ps, method = "lms")

######### Normalidade #########
## Tecnica Grafica
#Histograma
hist(rstandard(ps_lm2), ylab="Frequencia", xlab="Resíduos Padronizados", main="")

#QQplot
qqnorm(rstandard(ps_lm2), ylim=c(-3,2),xlim=c(-3,2), main="", ylab=
"Quantidades amostrais", xlab="Quantidades teóricas")
abline(a=0,b=1,col="black")
shapiro.test(rstandard(ps_lm2))

# Teste K
library(lmtest)
bptest(ps_lm2, studentize=TRUE)
bptest(ps_lm2)

#resíduos
rstandard(ps_lm2) #padronizados
rstudent(ps_lm2)  #estudentizados

#######################Ponto de influência
#Distância de Cook
dc2=influence.measures(ps_lm2)$infmat[,6]#distância de cook

#Ponto Influente dc2>F(p,n-p,(0.5)) será influente
plot(dc2, xlab="Valores Observados",ylab="Distância de Cook")
abline(h=qf(0.5,3,47))
```

```
id=which(dc2>qf(0.5,3,47)) #valores que ultrapassam a cota
text(id,dc2[id], rownames(ps)[id], pos=c(1,3),xpd=TRUE)

#teste de linearidade; H0: linearidade x H1: não linearidade
#tecnica grafica: valores observados x valors estimados
## Tecnica Grafica
plot(ps$Expenditure~ps_lm2$fitted.values, ylab=
"Valores observados",xlab="Valores Ajustados")
abline(a=0,b=1,col="black")
resettest(ps_lm2 , power=c(2), type="regressor")

#homocedasticidade tecnica gráfica
plot(rstandard(ps_lm2)~ps_lm2$fitted.values,ylim=c(-5,5), ylab="Resíduos Padronizados",
 xlab="Valores ajustados")
#abline(h=0,col="black")
#abline(h=2,col="black")
#abline(h=-2,col="black")

#Matriz de alavancagem
ps_hat2=hatvalues(ps_lm2)
ps_hat2
quoc2 = 2*3/50 #(2*p/n)

####### Pontos de alavanca
plot(ps_hat2, ylab="Matriz de alavancagem", xlab="Observações ordenadas")
abline(h=quoc2,col="black") # cota 2p/n
id=which(ps_hat2>quoc2) #valores que ultrapassam a cota
text(id,ps_hat2[id], rownames(ps)[id], pos=c(1,3),xpd=TRUE)
```