UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA
DEPARTAMENTO DE ESTATÍSTICA

Regressão Linear Robusta: O Método de TELBS e uma Aplicação a Dados de E-Commerce

Autor: Gabriel Fumagalli Fontoura
Orientador: Professor Dr. Guilherme Pumi

Porto Alegre, 27 de Novembro de 2015.

Universidade Federal do Rio Grande do Sul
Instituto de Matemática
Departamento de Estatística

Regressão Linear Robusta: O Método de TELBS e uma Aplicação a Dados de E-Commerce

Autor: Gabriel Fumagalli Fontoura

Trabalho de Conclusão de Curso
apresentado para obtenção
do grau de Bacharel em Estatística.

Banca Examinadora:
Professor Dr. Guilherme Pumi
Professor Dr. Marcio Valk

Porto Alegre, 27 de Novembro de 2015.

Dedico este trabalho a minha família, que sempre me apoiou e fez de tudo para que eu conseguisse chegar até aqui, sem eles nada disso seria possível.

“Se eu vi mais longe, foi por estar sobre ombros de gigantes.”

Isaac Newton (1643 – 1727)

Agradecimentos

Agradeço a minha mãe, por sempre me incentivar e me apoiar em todas as decisões que tomei na minha vida e, em especial, por ter sugerido para que eu olhasse o currículo do curso de Bacharelado em Estatística, visto que eu tinha dúvidas em qual curso escolher para prestar o vestibular. Sem a sugestão dela, talvez, eu jamais teria escolhido o curso de Estatística.

Ao meu pai, por todo esforço que ele faz e sempre fez para que eu pudesse ter todas as condições necessárias para estudar, nunca deixando de me incentivar de todas as maneiras possíveis e sempre me colocando em primeiro lugar em suas prioridades.

A minha irmã por todo apoio e incentivo que ela me dá, pois sempre está presente na minha vida, mesmo quando separados por tantos quilômetros de distância, sempre preocupada comigo e disposta a me ajudar no que for preciso.

Aos meus colegas de faculdade, que tornaram essa caminhada um pouco mais fácil. Em especial ao meu amigo Jonathan Farinela, pois iniciamos o curso juntos e nos tornamos amigos desde então, escolhíamos sempre as mesmas cadeiras para sermos colegas e agora estamos nos formando juntos.

A todos os meus amigos que de alguma forma contribuíram para que eu conseguisse superar todos os obstáculos encontrados nessa caminhada. Em especial para o Leonardo Rocha, que sugeriu a ideia inicial para este trabalho e me auxiliou no que foi necessário para a realização deste.

Ao meu Professor e orientador, Guilherme Pumi, que me conduziu durante todo o percurso deste trabalho, sempre amigável, disponível e demonstrando um enorme conhecimento quando solicitado, um exemplo de profissional a ser seguido.

Resumo

Este trabalho apresenta dois métodos de estimação para regressão linear robusta, o M-Estimador e modelo de TELBS. Estes são comparados com o método dos mínimos quadráticos ordinários, motivados por um exemplo simulado e uma análise de dados reais de *e-commerce*.

Para melhor compreensão da eficiência do modelo de regressão linear robusta de TELBS, é apresentada uma simulação e duas análises de bancos de dados reais, feitas pelos autores do modelo, que comparam esse método com o dos mínimos quadráticos, M-Estimador e MM-Estimador.

Também são brevemente descritos alguns diagnósticos de medidas influentes comumente abordados na literatura: medida de alavancagem, distância de Cook, DFBETAS, DFFITS e COVRATIO.

As análises e o exemplo simulado deste trabalho são feitos com o software R (Versão 3.2.0) e a sintaxe está disponível em anexo.

Sumário

1. Introdução

Modelos de regressão linear são muito utilizados em diversas áreas de estudo, porém esses podem apresentar alguns problemas em determinadas condições, que são muito comuns de se observar em dados reais. Então, uma das possíveis soluções, é a utilização de métodos robustos de estimação de regressão linear, capazes de amenizar ou, até mesmo, corrigir esses problemas.

Este trabalho pretende apresentar um novo método de estimação de regressão linear robusta, conhecido como TELBS, proposto por Tabatabai et al. (2012), que foi desenvolvido para ser eficaz, tanto em simulações, quanto em dados reais, e de simples aplicação. Esta nova forma de estimação será comparada com duas comumente utilizadas em regressão linear: o método dos mínimos quadráticos ordinários para regressão linear simples e o método com um M-Estimador para regressão linear robusta.

Uma das principais causas que geram esses problemas, ou imprecisões, nos modelos de regressão linear estimados por certos métodos, são os valores extremos (*outliers*). Então, serão brevemente descritas algumas das principais medidas de identificação desses valores, assim como uma nova maneira de medição, proposta por Tabatabai et al. (2012), com o objetivo de se esclarecer a influência dos *outliers* no processo de estimação.

Para a comparação dos métodos de estimação será analisada uma amostra de dados reais de uma empresa de vendas *online* e um exemplo feito por simulação. Para essas análises foi utilizado o software R (Versão 3.2.0) com as respectivas sintaxes em anexo. Também serão apresentadas, uma simulação e duas análises de dados reais, feitas por Tabatabai et al. (2012), que comparam o método de estimação de TELBS com os métodos dos mínimos quadráticos, M-Estimador e MM-Estimador.

2. Regressão linear robusta

A análise de regressão estuda o relacionamento entre uma variável resposta – dependente – e uma ou mais variáveis explicativas – independentes – através de um modelo matemático. Um dos modelos matemáticos mais populares e amplamente utilizados em todos os campos de estudo é o de regressão linear. Esse modelo ajusta uma equação para explicar a relação linear entre essas variáveis, com o objetivo de fazer predições, seleção de variáveis, estimação e inferência de parâmetros.

Um método comumente utilizado para estimar os parâmetros dos coeficientes de regressão é o método dos mínimos quadráticos, porém com a presença de *outliers* ou quando a variável resposta não segue uma distribuição normal, essas estimativas podem não ser mais confiáveis. A reta ajustada pelo método dos mínimos quadráticos ordinários é fortemente influenciada pelos *outliers*, pois o método minimiza a soma dos resíduos ao quadrado. Para demonstrar essa influência, utilizando o software R (Versão 3.2.0), será utilizado o seguinte exemplo.

Exemplo 1: os valores de x são fixados por uma sequencia de 0 a 15 e duas amostras da variável y, de tamanho $n = 15$, são simuladas com a seguinte equação

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0,1),$$

onde os valores dos coeficientes da regressão são iguais a $\beta_0 = 3$ e $\beta_1 = 0{,}8$. Para comparar resultados, na primeira amostra simulada de y foi adicionado um *outlier* – representado no gráfico por um triângulo –, enquanto que na segunda amostra não. O gráfico abaixo mostra as retas ajustadas pelo método dos mínimos quadráticos ordinários para as duas simulações.

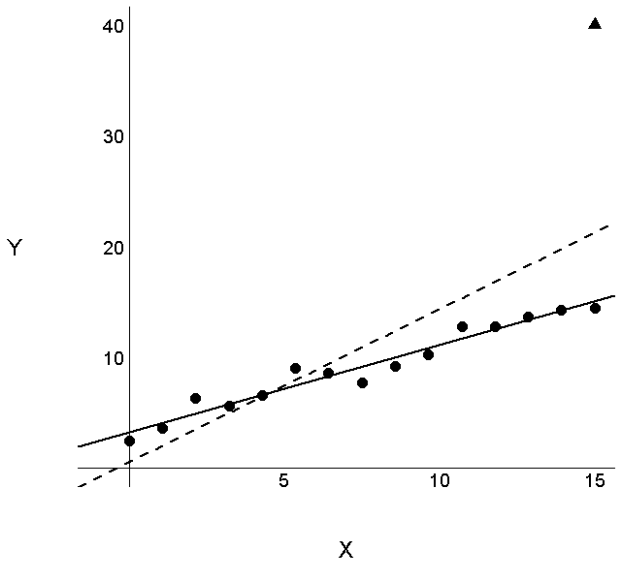

Figura 1: Gráfico de dispersão e retas ajustadas pelo método dos mínimos quadráticos ordinários.

A reta ajustada pelo método dos mínimos quadráticos ordinários para a simulação com o *outlier* – reta tracejada – sofre forte influência em sua inclinação em comparação à reta da simulação sem o *outlier* – reta contínua –, consequentemente também influenciando a estimativa do intercepto da equação. As estimativas pontuais para os coeficientes de regressão pelo método dos mínimos quadráticos ordinários encontram-se na tabela abaixo:

Tabela 1 – Estimativas dos coeficientes de regressão pelo método dos mínimos quadráticos ordinários.

Parâmetros	Verdadeiro valor	Estimativas	
		Sem *outlier* (EP)	Com *outlier* (EP)
β_0	3	3,2709 (0,4295)	0,5023 (2.9717)
β_1	0,8	0,7842 (0,0487)	1,3805 (0.3372)

Abreviaturas: EP (Erro Padrão).

As estimativas pontuais para a simulação sem o *outlier* são próximas dos verdadeiros valores dos coeficientes de regressão, enquanto que as estimativas da simulação com o *outlier* não.

A proporção da variância da variável dependente (y) explicada pela variável independe (x) também sofre grande influência nos dois modelos. Essa proporção pode ser encontrada através do coeficiente de determinação ajustado de cada modelo (R_a^2), iguais a 0,9485 e 0,5296 nos modelos sem e com o *outlier*, respectivamente.

Então, para ajustar uma equação que produza boas estimativas dos coeficientes de regressão de observações que possuem *outliers*, ou quando a variável resposta não segue uma distribuição normal – e essa variável se relaciona de forma linear com as demais –, utiliza-se o método de regressão linear robusta, cujo objetivo é reduzir a influência de pontos discrepantes que afetam a qualidade da estimação dos parâmetros do modelo de regressão. Outro caso interessante é o de observações que seguem distribuições que tenham caudas mais longas, ou pesadas, que a da distribuição normal, pois tendem a gerar valores discrepantes.

Além de ser sensível à influência de *outliers*, um método de estimação robusto deve produzir as mesmas estimativas que o método dos mínimos quadráticos ordinários produziria na ausência desses ou quando a variável resposta é normalmente distribuída.

2.1. Diagnósticos de medidas influentes

De acordo com Barnett e Lewis (1984), um *outlier* é uma observação inconsistente com os dados. Tanto a variável resposta como as variáveis explicativas da regressão podem conter *outliers*, mas esses não necessariamente são pontos influentes nas estimativas do modelo. Um ponto influente é aquele que afeta as estimativas dos coeficientes de regressão do

modelo ajustado, pois este possui valores incomuns para variável x, y ou até mesmo para ambas. Já um ponto de alavanca não afeta as estimativas dos coeficientes de regressão, mas influenciará no coeficiente de determinação (R^2) e o erro padrão dos coeficientes de regressão. A Figura 2 ilustra um exemplo de cada um desses pontos.

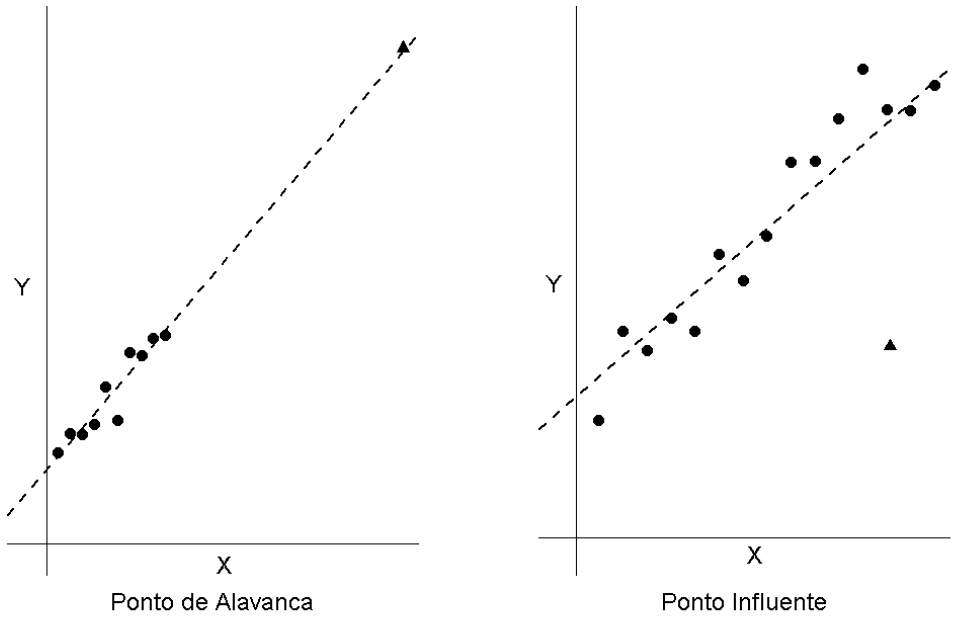

Figura 2: Exemplo de ponto de alavanca e ponto influente.

Note que se fossem ajustados modelos de regressão linear para ambos os exemplos, a reta ajustada passaria pelo ponto de alavanca, mas não pelo ponto influente. Para identificar e medir pontos de influencia existem diversas formas apresentadas na literatura, algumas dessas serão brevemente descritas nesta seção.

2.1.1. Medida de alavancagem

A medida de alavancagem é definida por

$$h_{ii} = x_i'(X'X)^{-1}x_i \,,$$

onde x_i' é a i-ésima linha da matriz de design X

$$X = \begin{bmatrix} x_{10} & x_{11} & \cdots & x_{1k} \\ x_{20} & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n0} & x_{n1} & \cdots & x_{nk} \end{bmatrix} = (X_0 X_1 \ldots X_k),$$

h_{ii} é o valor do i-ésimo elemento da diagonal principal da matriz chapéu, definida por

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

n é o tamanho da amostra e $p = \sum_{i=1}^{h} h_{ii}$, ou seja, o número total de parâmetros do modelo.

Essa estatística mede a importância da i-ésima observação na determinação do ajuste do modelo. Verificam-se então os valores de $h_{ii} > 2p/n$, pois estes podem representar pontos de alavanca – *outliers* na variável explicativa – para amostras não pequenas.

2.1.2. Distância de Cook

A distância de Cook – introduzida por R. Dennis Cook (1977) – é definida por

$$D_i = \frac{e_i^2}{pQMRes}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right], \quad i = 1,2,\dots,n,$$

onde $e_i = y_i - \hat{y}_i$ são os resíduos e QMRes é o erro quadrático médio do modelo de regressão. Essa estatística mede o afastamento do vetor de estimativas dos coeficientes de regressão provocado pela retirada da i-ésima observação, em que valores de $D_i > 1$ podem ser considerados influentes. Outra forma, sugerida na literatura, para encontrar valores influentes através da distância de Cook, é a de comparar D_i a com a mediana de uma distribuição $F(p, n-p)$.

2.1.3. DFBETAS

A medida de influência *DFBETAS* mede a influencia da i-ésima observação sobre o valor estimado do j-ésimo $\hat{\beta}$. A estatística é dada por

$$DFBETAS_{j(i)} = \frac{\hat{\beta}_{(j)} - \hat{\beta}_{j(i)}}{\sqrt{QMRes_i C_{jj}}},$$

onde C_{jj} é o (j)-ésimo elemento da diagonal da matriz $(X'X)^{-1}$ e $\hat{\beta}_{j(i)}$ é a estimativa do parâmetro β_j com a retirada da observação i. Consideramos que podem ser pontos influentes aqueles em que $\left|DFBETAS_{j(i)}\right| > 2/\sqrt{n}$ para amostras grandes ou $\left|DFBETAS_{j(i)}\right| > 1$ para amostras pequenas e médias.

2.1.4. DFFITS

A medida de influência *DFFITS* mede a influência provocada no valor ajustado pela retirada da i-ésima observação. A estatística é definida pela seguinte equação

$$DFFITS_i = \frac{\hat{y}_{i-}\hat{y}_{i(i)}}{\sqrt{QMRes_i h_{jj}}},$$

onde $\hat{y}_{i(i)}$ é a resta estimada para a variável y obtida com a retirada da observação i.Valores de $|DFFITS_i| > 2\sqrt{p/n}$ podem ser considerados como pontos influentes para amostras grandes ou $|DFFITS_i| > 1$ para amostras pequenas e médias. As medidas de *DFBETAS* e *DFFITS* foram introduzidas por Belsley, Kuh e Welsch (1980).

2.1.5. COVRATIO

A medida conhecida por *COVRATIO*, definida pela equação

$$COVRATIO_i = \frac{\det[QMRes_i(X_i{}'X_i)^{-1}]}{\det[QMRes_i(X'X)^{-1}]},$$

é utilizada para se obter informações sobre a precisão geral das estimativas, a qual não é fornecida pelas medidas anteriormente citadas.

Essa estatística mede o efeito da retirada da i-ésima observação no determinante da matriz de covariância das estimativas. Valores de $COVRATIO_i > 1 + 3p/n$ ou valores de $COVRATIO_i < 1 - 3p/n$ podem ser considerados como pontos influentes. O limite inferior é apropriado apenas quando $n > 3p$ (Belsley, Kuh e Welsch, 1980).

As estatísticas para medida de distancia de Cook, *DFFITS*, *DFBETAS*, *COVRATIO* podem ser calculadas pelo software R através do pacote "stats".

2.2. M-Estimadores de regressão robusta

Existem diversos métodos de estimação de regressão linear robusta, este trabalho pretende apresentar alguns dos principais e mais comumente utilizados métodos de estimação da classe de M-Estimadores, de forma a disponibilizar alternativas para estimação de regressão linear robusta e, também, para comparar com o método de TELBS – descrito no capítulo seguinte – e o método dos mínimos quadráticos ordinários.

A classe de M-Estimadores foi introduzida primeiramente por Huber (1964) e posteriormente discutida por diversos outros autores, como: Andrews (1972), Bunke e Bunke (1986), Hampel (1986), Lecoutre e Tassi (1987), Robusseeuw e Leroy (1987), Staudte e Sheather (1990), Rieder (1994), Jureckova e Sen (1996), Antoch (1998), Dodge e Jureckova (2000),

Jureckova e Picek (2006), entre outros. A classe de M-estimadores foi estendida para todas as distribuições de probabilidade e generaliza o método da máxima verossimilhança, produzindo estimadores consistentes e assintoticamente normais (Heritier, 2009). A letra "M" vem da função de Máxima verossimilhança e essa classe busca minimizar a soma de certa função ρ dos erros aleatórios, definida pela seguinte equação

$$\hat{\beta} = \underset{\beta \in \Omega}{\arg\min} \left\{ \sum_{i=1}^{n} \rho(e_i) \right\}$$

A função ρ está relacionada com a função de probabilidade, apropriadamente escolhida, para a distribuição dos erros. O M-estimador não necessariamente é invariante para mudança de escala, para obter a versão invariante para mudança de escala deste estimador é preciso resolver a equação

$$\hat{\beta} = \underset{\beta \in \Omega}{\arg\min} \left\{ \sum_{i=1}^{n} \rho\left(\frac{e_i}{s}\right) \right\}, \tag{2}$$

onde s é uma estimativa robusta de escala. Uma dessas estimativas, muito comum em aplicações, é o desvio absoluto da mediana

$$s = Mediana(|e_i - Mediana(e_i)|)/0{,}6745.$$

A constante 0,6745 faz s um estimador aproximadamente não viciado de σ para amostras grandes e distribuição normal dos erros.

Para minimizar a equação (2) deve-se igualar à zero a primeira derivada parcial da função ρ em função de β_j, resultando em uma condição mínima necessária e um sistema de $p = k + 1$ equações

$$\sum_{i=1}^{n} x_{ij} \psi\left(\frac{e_i}{s}\right) = 0, \quad j = 0{,}1, \ldots, k, \tag{3}$$

onde $\psi = \rho'$, x_{ij} é a i-ésima observação no j-ésimo regressor e $x_{i0} = 1$. Geralmente a função ψ é não linear e a equação (3) precisa ser resolvida por métodos iterativos. Um método iterativo comumente usado e sugerido na literatura é o IRLS (Iteratively Reweighted Least Squares em inglês), abordagem geralmente atribuída a Beaton e Tukey (1974).

Para resolver a equação (3) através do método iterativo de mínimos quadráticos reponderados (IRLS), devemos supor uma estimativa inicial para $\hat{\beta}_0$ e que s é uma estimativa de escala, e então, escrever as $p = k + 1$ equações de (3)

$$\sum_{i=1}^{n} x_{ij}\psi\left(\frac{e_i}{s}\right) = \sum_{i=1}^{n} \frac{x_{ij}[\psi(e_i/s)/e_i]e_i}{s} = 0, \quad j = 0,1,\dots,k$$

da seguinte forma

$$\sum_{i=1}^{n} x_{ij} w_{i0}(e_i) = \sum_{i=1}^{n} x_{ij} w_{i0}(y_i - x_i'\beta) = 0, \quad j = 0,1,\dots,k,$$

onde

$$w_{i0} = \begin{cases} \dfrac{\psi\left[\left(y_i - x_i'\hat{\beta}_0\right)/s\right]}{\left(y_i - x_i'\hat{\beta}_0\right)/s}, & se\ \ y_i \neq x_i'\beta_0 \\ 1, & se\ \ y_i = x_i'\beta_0 \end{cases} \qquad (4)$$

Em notação matricial temos a equação na seguinte forma

$$X'W_0X\beta = X'W_0y, \qquad (5)$$

onde W_0 é uma matriz diagonal de pesos, de tamanho $n \times n$, com os elementos das diagonais $w_{10}, w_{20}, \dots, w_{n0}$ dados pela equação (5). A equação (4) é comumente reconhecida como a equação normal ponderada de mínimos quadráticos ordinários, consequentemente o estimador no primeiro passo para o parâmetro é

$$\hat{\beta}_1 = (X'W_0X)^{-1}X'W_0y. \qquad (6)$$

Os pesos de $\hat{\beta}_0$ da equação (4) podem então ser substituídos pelos valores estimados de $\hat{\beta}_1$ encontrados através da equação (6). Geralmente apenas algumas iterações são necessárias para alcançar a convergência. Com o pacote "MASS" do software R consegue-se estimar um modelo de regressão linear robusta utilizando um M-Estimador através do método iterativo IRLS.

Voltando ao Exemplo 1, a Figura 3 mostra o gráfico de dispersão de x e y e as retas ajustadas para as duas simulações, com o M-Estimador de Huber, através do método iterativo de mínimos quadráticos reponderados (IRLS).

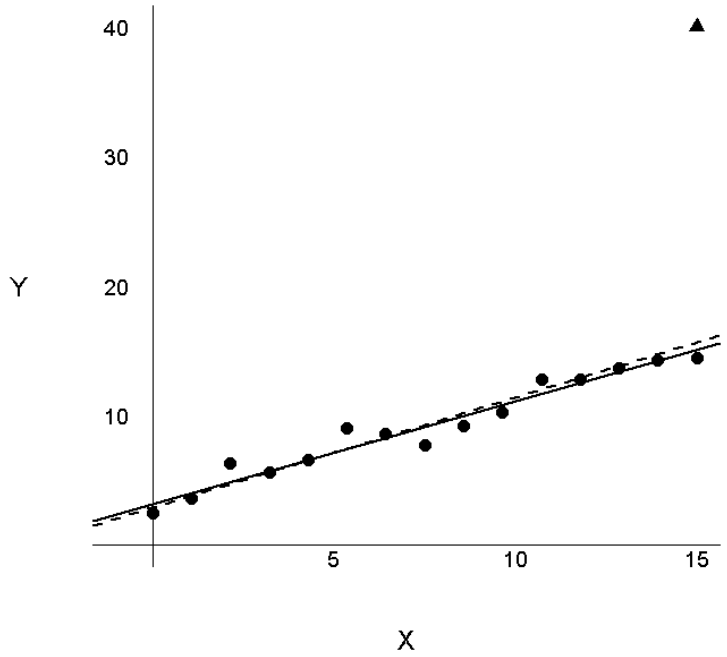

Figura 3: Gráfico de dispersão e retas ajustadas com o M-Estimador de Huber através do método iterativo IRLS.

A reta ajustada com o M-Estimador de Huber para a simulação com *outlier* – reta tracejada – não sofre mais uma forte influência em sua inclinação como quando estimada pelo método dos mínimos quadráticos ordinários. Note que, quase não é possível distinguir as retas ajustadas, na Figura 3, o que indica um bom ajuste para as duas simulações do exemplo. As estimativas pontuais para os coeficientes de regressão com o M-Estimador através do método iterativo IRLS encontram-se na tabela abaixo.

Tabela 2 – Estimativas dos coeficientes de regressão com o M-Estimador através do método iterativo IRLS.

Parâmetros	Verdadeiro valor	Estimativas	
		Sem *outlier* (EP)	Com *outlier* (EP)
β_0	3	3,1710 (0.5170)	2,8835 (0.5274)
β_1	0,8	0,7923 (0.0587)	0,8483 (0.0598)

Abreviaturas: EP (Erro Padrão).

Tanto as estimativas pontuais para a simulação sem *outlier*, quanto às estimativas da simulação com *outlier*, estão próximas dos verdadeiros valores dos coeficientes de regressão. A estimação com o M-Estimador de Huber – para este exemplo – também se mostrou mais próxima dos verdadeiros valores dos parâmetros para a simulação sem o *outlier* do que a estimação pelo método dos mínimos quadráticos ordinários, inclusive.

Cada função robusta necessita que se especifiquem certas constantes para a função ψ. Alguns dos critérios mais comumente utilizados encontram-se na tabela abaixo com os respectivos valores das constantes de sintonização.

Tabela 3 – Critérios para as funções robustas.

Critério	$\rho(z)$	$\psi(z)$	$w(z)$	Domínio
Mínimos quadráticos ordinários	$\frac{1}{2}z^2$	z	1,0	$\lvert z\rvert < \infty$
Função t de Hubert	$\frac{1}{2}z^2$	z	1,0	$\lvert z\rvert \leq t$
$t = 2$	$\lvert z\rvert t - \frac{1}{2}t^2$	$t = sign(z)$	$\frac{t}{\lvert z\rvert}$	$\lvert z\rvert > t$
Função E_a de Ramsay $a = 0{,}3$	$a^{-2}[1 - exp(-a\lvert z\rvert) \cdot (1 + a\lvert z\rvert)$	$ze^{(-a\lvert z\rvert)}$	$e^{(-a\lvert z\rvert)}$	$\lvert z\rvert < \infty$
Função de onda de Andrews	$a[1 - \cos(z/a)]$	$sen(z/a)$	$\frac{sen(z/a)}{z/a}$	$\lvert z\rvert \leq a\pi$
$a = 1{,}339$	$2a$	0	0	$\lvert z\rvert \leq a\pi$
Função 17A de Hampel	$\frac{1}{2}z^2$	z	1,0	$\lvert z\rvert \leq a$
$a = 1{,}7$	$a\lvert z\rvert - \frac{1}{2}a^2$	$a\ sen(z)$	$\frac{a}{\lvert z\rvert}$	$a < \lvert z\rvert \leq b$
$b = 3{,}4$	$\frac{a(c\lvert z\rvert - \frac{1}{2}z^2)}{c-b} - (7/6)a^2$	$\frac{a\ sign(z)(c - \lvert z\rvert)}{c-b}$	$\frac{a(c - \lvert z\rvert)}{\lvert z\rvert(c-b)}$	$b < \lvert z\rvert \leq c$
$c = 8{,}5$	$a(b + c - a)$	0	0	$\lvert z\rvert > c$

Fonte: Montgomery, Peck e Vining (2001).

As Figuras 4 e 5 mostram o comportamento das funções ρ e suas funções ψ correspondentes que classificam o processo de regressão robusta. As funções ψ são chamadas de funções influentes, pois controlam os pesos dados a cada resíduo.

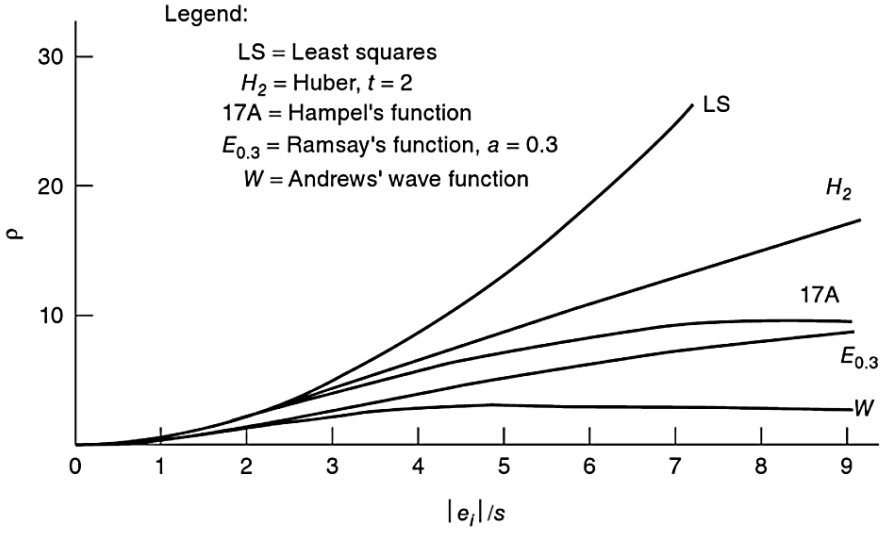

Figura 4: Critérios populares para as funções robustas.
Fonte: Montgomery, Peck e Vining (2001).

A função para os Mínimos quadráticos ordinários – curva LS (Least squares em inglês) da Figura 4 – é ilimitada o que tende a torna-la não robusta

para observações que provenham de distribuições com caudas mais pesadas que a da distribuição normal. A função t de Hubert (1964) – curva H_2 da Figura 4 – possui uma função ψ monótona e acaba não dando tanto peso para grandes resíduos quanto a função de mínimos quadráticos ordinários. A função E_a de Ramsay (1977) – curva $E_{0,3}$ da Figura 4 – decai lentamente conforme os valores dos resíduos crescem, o que a faz ser assintoticamente igual à zero para valores grandes do $|z|$. Tanto a função 17A de Hampel – curva 17A da Figura 4 – quanto à função de onda de Andrews (1972) – curva W da Figura 4 –, possuem forte decaimento conforme os resíduos crescem, fazendo a função ψ ser igual à zero para valores suficientemente grandes de $|z|$.

A Figura 5 mostra as curvas de influência dos estimadores.

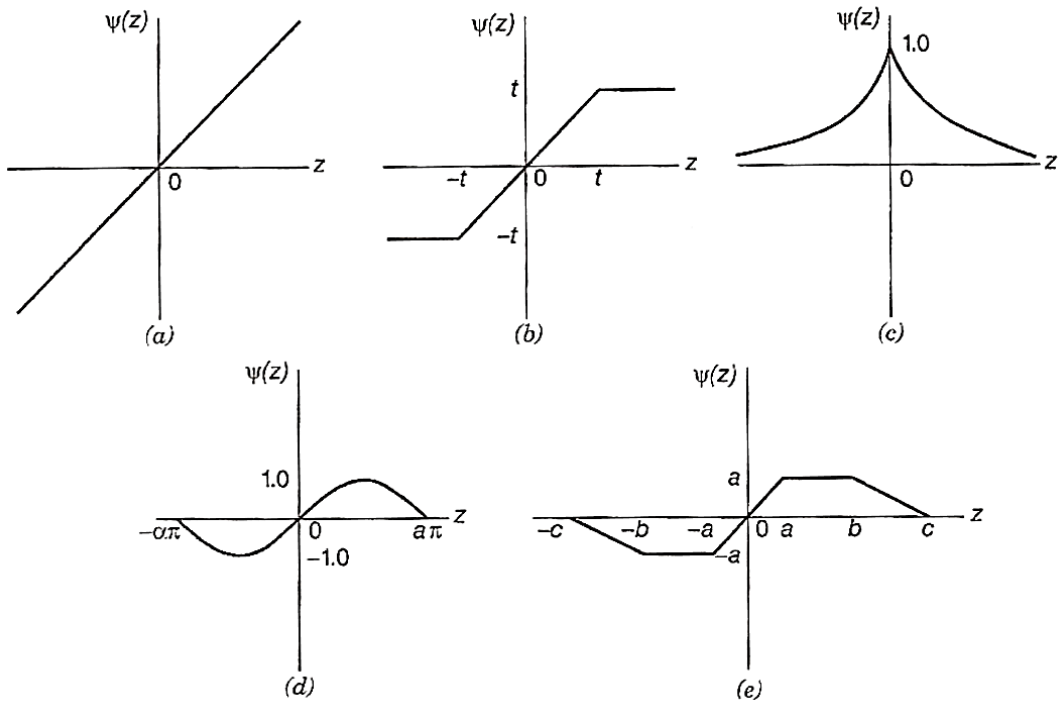

Figura 5: Funções de influência robusta: (a) mínimos quadráticos ordinários; (b) função t de Hubert; (c) função E_a de Ramsay; (d) função de onda de Andrews; (e) função 17A de Hampel.
Fonte: Montgomery, Peck e Vining (2001).

Depois de definido o modelo de regressão final com suas respectivas estimativas do parâmetro $\hat{\beta}$, é importante determinar a matriz de covariância desse parâmetro, para que se possam construir intervalos de confiança, proceder com testes de hipóteses e fazer outras inferências. Assintoticamente o parâmetro $\hat{\beta}$ se aproxima da distribuição normal com média β e a seguinte matriz de covariância (Hubert 1973)

$$\sigma^2 \frac{E[\psi^2(\varepsilon/\sigma)]}{\left\{E\{\psi'(\varepsilon/\sigma)\}\right\}^2}(X'X)^{-1},$$

que pode ser aproximada, em amostras finitas, por

$$\frac{ns^2}{n-p}\frac{\sum_{i=1}^{n}\psi^2(e_i/s)}{[\sum_{i=1}^{n}\psi'(e_i/s)]^2}(X'X)^{-1}.$$

Também é possível estimar a matriz de covariância através dos mínimos quadráticos ordinários ponderados

$$\frac{\sum_{i=1}^{n} w_i(e_i)^2}{n-p}(X'WX)^{-1},$$

onde W é uma matriz diagonal de pesos – de tamanho $n \times n$ – e os w_i's são os elementos das diagonais dessa matriz.

Outros dois métodos são sugeridos por Welsch (1975) e outro por Hill (1979), no entanto, não há um consenso sobre qual a melhor maneira de estimação. De qualquer forma, ambos os métodos permitem que se façam inferências a respeito de $\hat{\beta}$ usando a teoria normal.

2.3. Modelo de regressão linear robusta de TELBS

O modelo de regressão linear robusta de TELBS (Tabatabai, Eby, Li, Bae e Singh) propõe um método simples e eficaz de estimação a ser aplicado em diversos campos de pesquisa, inclusive para biomedicina. Percebe-se que nesta área existem diversos casos em que é necessário o uso de regressão linear robusta, pois é comum que os dados apresentem valores discrepantes.

Primeiramente, para comparação com o método de TELBS, considere o modelo de regressão linear padrão, com n observações, definido pela seguinte equação

$$y_i = x_i'\beta + \varepsilon_i,$$

onde $i = 1, 2, \dots, n$, $\beta' = (\beta_0, \beta_1, \dots, \beta_k)$ é o vetor de parâmetros e os ε_i's são os erros aleatórios. Em experimentos planejados – como no exemplo da seção anterior – os x_{ij}'s são fixos, mas quando observados, são variáveis aleatórias. A variável independente pode ser fixa, aleatória ou mista. Pelo método dos Mínimos Quadráticos Ordinários é possível estimar o vetor de parâmetros β da seguinte forma:

$$\hat{\beta}_L = \underset{\beta \in R^{k+1}}{\arg\min}\left\{\sum_{i=1}^{n}\varepsilon_i^2\right\},$$

onde

$$x_i' = (x_{i0}, x_{i1}, \dots, x_{ik}); 1 \leq i \leq n,$$

com $x_{i0} \equiv 1$ e k é o número de variáveis independentes do modelo.

Também é importante definir o conceito de ponto de ruptura para melhor entendimento do estimador de TELBS. O ponto de ruptura de um estimador mede qual seria a maior porcentagem de contaminação que um estimador poderia suportar e ainda assim fornecer informação confiável sobre o parâmetro de interesse. Quanto mais próximo de 0,5 for o ponto de ruptura, mais resistente é o estimador à presença de *outliers*.

O método de TELBS estima o vetor de parâmetros β da seguinte maneira

$$\hat{\beta} = \underset{\beta \in R^{k+1}}{\arg\min} \left\{ \sum_{i=1}^{n} \frac{\rho_\omega(t_i)}{L_i} \right\}, \tag{7}$$

onde

$$\rho_\omega(x) = 1 - \operatorname{sech}(\omega x),$$

e ω é um número real positivo definido pelo usuário do modelo. A função $\operatorname{sech}(\cdot)$ é a função secante hiperbólica e os t_i's são definidos por

$$t_i = \frac{(y_i - x_i'\beta)(1 - h_{ii})}{\sigma} \tag{8}$$

onde σ é o desvio padrão do erro e os h_{ii}'s são os elementos da diagonal da matriz chapéu, definida em (1).

Para $j = 1, 2, \dots, k$, seja

$$M_j = Med\left(|x_{1j}|, |x_{2j}|, \dots, |x_{nj}|\right),$$

e para $i = 1, 2, \dots, n$, defina

$$L_i = \sum_{j=1}^{k} Max\left(M_j, |x_{ij}|\right).$$

Para casos em que σ é desconhecido, pode-se usar um dos seguintes estimadores propostos por Rousseeuw e Croux (1993):

$$\hat{\sigma} = 1{,}1926 \underset{\{i:1\leq i\leq n\}}{\text{Mediana}} \left(\underset{\{j:1\leq j\leq n\}}{\text{Mediana}} |r_i - r_j| \right) \tag{9}$$

ou

$$\hat{\sigma} = 2{,}2219 \{ |r_i - r_j|; i < j, i, j = 1, \dots, n \}_{(p)}, \tag{10}$$

onde $r_i = y_i - x_i^t \hat{\beta}$, $p = \binom{[n/2]+1}{2}$ e $\{\cdot\}_{(p)}$ é a p-ésima estatística de ordem.

Os estimadores de σ acima referidos possuem pontos de ruptura muito elevados e, sob o pressuposto de normalidade, são mais eficientes do que o desvio absoluto da mediana.

A função $\rho_\omega : \mathbf{R} \to \mathbf{R}$ é uma função diferenciável que satisfaz as seguintes propriedades:

i. ρ_ω é não negativo, limitado, simétrico com $\rho_\omega(0) = 0$,
ii. $\forall a, b \in \mathbf{R}, |a| > |b| \Rightarrow \rho_\omega(a) \geq \rho_\omega(b)$,
iii. $\lim_{x\to\infty} \rho_\omega(x) = \lim_{x\to-\infty} \rho_\omega(x) = 1$,
iv. $\forall k > 0, \lim_{x\to\infty} \rho_\omega(kx)/\rho_\omega(x) = 1$,
v. $\lim_{|x|\to\infty} d\,\rho_\omega(x)/\,dx = 0$.

Derivando (7) em relação à β_j, para $j = 1, \dots, k$, e igualando-se a zero, obtemos o seguinte sistema de equações

$$\sum_{i=1}^{n} \frac{\psi_\omega(t_i)}{L_i} \frac{\partial t_i}{\partial \beta_j} = 0, \tag{11}$$

onde $\psi_\omega = \omega Sech(\omega x) Tanh(\omega x)$ é a derivada de ρ_ω.

Definimos os pesos w_i como

$$w_i = \frac{\psi_\omega(t_i)(1 - h_{ii})}{\sigma(y_i - x_i'\beta)L_i}, \tag{12}$$

assim a equação (11) pode ser escrita como

$$\sum_{i=1}^{n} w_i (y_i - x_i'\beta) x_i = 0.$$

A matriz de pesos Wé uma matriz diagonal, cujos elementos da diagonal principal são $w_1, w_2, \dots, w_n$, e o estimador do vetor de parâmetros β é dado por

$$\hat{\beta}(X, y) = (X'WX)^{-1}X'Wy.$$

O estimador $\hat{\beta}(X, y)$ de TELBS possui as seguinte propriedades:

1. Equivariância por regressão, $\forall \alpha \in \mathbf{R}^{k+1}, \hat{\beta}(X, y + X\alpha) = \hat{\beta}(X, y) + \alpha$.
2. Equivariância escalar, $\forall \gamma \in \mathbf{R}, \hat{\beta}(X, \gamma y) = \gamma\hat{\beta}(X, y)$.
3. Equivariância afim, $\forall M, \hat{\beta}(XM, y) = M^{-1}\hat{\beta}(X, y)$,

onde M é uma matriz invertível de tamanho $k + 1$. Assintoticamente a estimativa $\hat{\beta}$ possui distribuição normal com média igual a β e matriz de covariância definida por:

$$V = \frac{\sigma^2 E\left(\psi_\omega^2(t)\right)}{\left[E\left(\psi_\omega'(t)\right)\right]^2} E((X'X)^{-1}),$$

onde $E[\psi_\omega'(t)]$ e $E[\psi_\omega^2(t)]$ são esperanças tomadas em relação à densidade normal padrão.

Note que

$$\psi_\omega'(t) = \omega^2[Sech^3(\omega t) - Sech(\omega t)Tanh^2(\omega t)].$$

Sob a hipótese de normalidade para a distribuição, podemos definir a eficiência assintótica como

$$Aeff = \frac{(E[\psi_\omega'(t)])^2}{E[\psi_\omega^2(t)]}. \qquad (13)$$

A constante ω pode ser encontrada através de (13). Para o modelo de TELBS, os valores numéricos para ω nos níveis de eficiência assintótica de 0,70, 0,75, 0,80, 0, 85, 0,90 e 0,95 são de aproximadamente 0,901, 0,8118, 0,721, 0,628, 0,525 e 0,405, respectivamente. A definição para a escolha da constante ω fica a critério de quem estiver analisando os dados, mas os autores do modelo recomendam que se utilize uma eficiência de 0,85, que corresponde ao valor de $\omega = 0{,}628$.

As curvas contínuas e tracejadas, na Figura 6, indicam os gráficos das funções ρ_ω e ψ_ω, respectivamente, para valores da constante $\omega = 0{,}901$ e $\omega = 0{,}405$. As Figuras 7 e 8 indicam os gráficos tridimensionais de $\rho_\omega(x)$ e $\psi_\omega(x)$ como funções de ω e x.

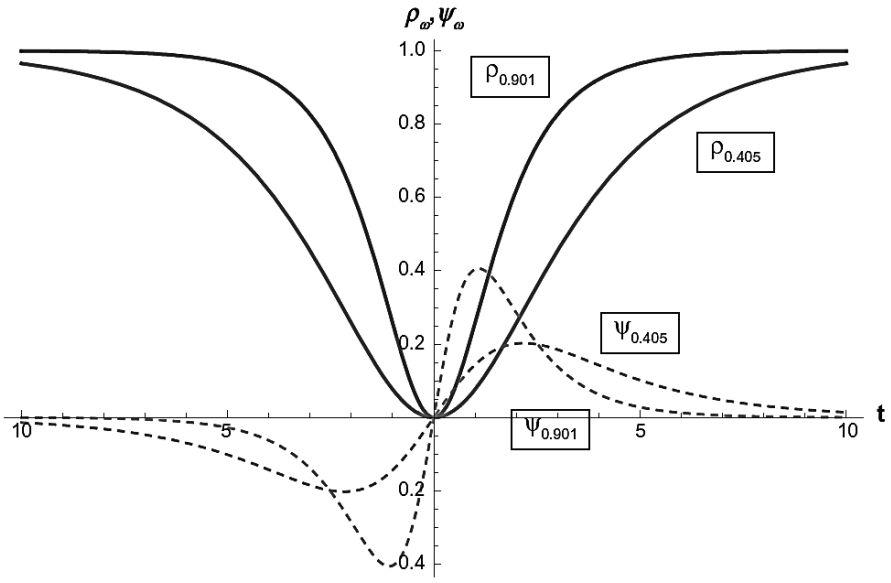

Figura 6: Gráficos das funções ρ_ω e ψ_ω para $\omega = 0{,}901$ e $\omega = 0{,}405$.
Fonte: Tabatabai et al. (2012)

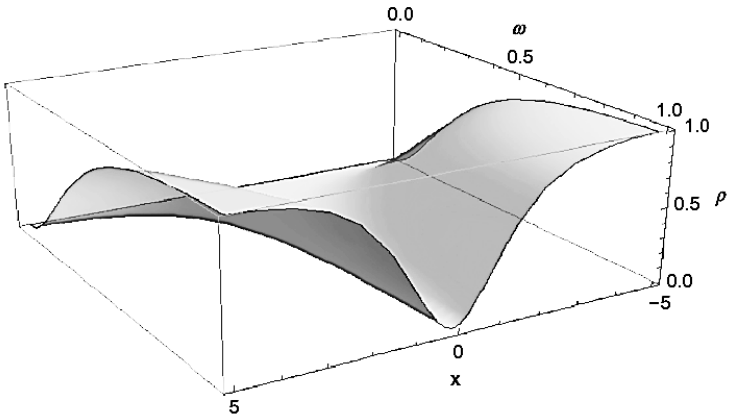

Figura 7: Gráfico tridimensional da função ρ_ω como função de ω e x.
Fonte: Tabatabai et al. (2012)

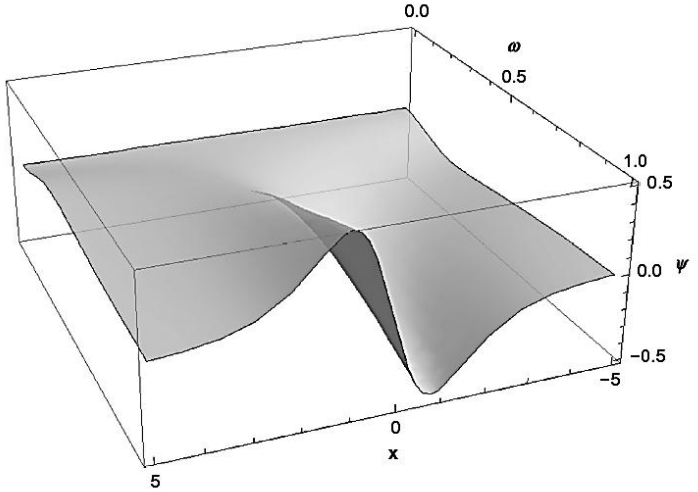

Figura 8: Gráfico tridimensional da função ψ_ω como função de ω e x.
Fonte: Tabatabai et al. (2012)

Uma estimativa da matriz de covariância é dada por

$$\hat{V} = \frac{n^2 \hat{\sigma}^2 \sum_{i=1}^{n} \psi_{\omega}^2(t_i)}{(n-p)(\sum_{i=1}^{n} \psi'_{\omega}(t_i))^2} (X'X)^{-1}.$$

O desvio robusto é definido por

$$D = 2\hat{\sigma}^2 \sum_{i=1}^{n} \frac{1 - Sech(\omega t_i)}{L_i}.$$

O desvio tem um papel importante no ajuste do modelo, em que se desejam valores menores possíveis para ele. Através do critério de informação de Akaike (1974), Ronchetti (1985) definiu o critério de informação robusto equivalente, por

$$AICR = \frac{D}{\hat{\sigma}^2} 2p \frac{E[\psi_{\omega}^2(t)]}{E[\psi'_{\omega}(t)]},$$

onde D é o desvio robusto, $\hat{\sigma}$ é a estimativa do desvio padrão e p é o número de parâmetros do modelo.

O critério de informação bayesiano proposto por Schwarz (1978) tem sua forma robusta definida como

$$BICR = \frac{D}{\hat{\sigma}^2} + 2p \ln(n),$$

onde D é o desvio robusto, $\hat{\sigma}$ é a estimativa do desvio padrão, p é o número de parâmetros do modelo e n o tamanho da amostra.

A versão robusta do coeficiente de determinação, proposta por Rosseeuw e Leroy (1987), é dada por

$$R^2 = 1 - \left(\frac{\underset{\{i:1 \le i \le n\}}{\text{Mediana}}(|r_i|)}{\underset{\{i:1 \le i \le n\}}{\text{Mediana}}\left(\left|y_i - \underset{\{j:1 \le j \le n\}}{\text{Mediana}}(y_i)\right|\right)} \right)^2$$

e para o caso da equação de regressão não possuir intercepto, deve-se utilizar

$$R^2 = 1 - \left(\frac{\underset{\{i:1 \le i \le n\}}{\text{Mediana}}(|r_i|)}{\underset{\{i:1 \le i \le n\}}{\text{Mediana}}(y_i)} \right)^2.$$

A estatística F é sugerida para medir o efeito global das covariáveis na variável resposta

$$F = \frac{2E[\psi'_\omega(t)]\left(\sum_{i=1}^{n}\rho_\omega\,(M_{const}) - \sum_{i=1}^{n}\rho_\omega\,(M_{comp})\right)}{(p-1)E[\psi^2_\omega(t)]}$$

onde M_{const} é o modelo apenas com a constante e M_{comp} é o modelo completo. Para a seleção das variáveis do modelo existem várias técnicas sugeridas na literatura, como o procedimento *Stepwise* pelos métodos *Foreward* ou *Backward*.

Em Marona (2006) os autores definiram que, para cada conjunto $S \subseteq \{x_1, x_2, \dots, x_n\}$ de variáveis explicativas, o erro de predição robusto final como

$$RFPE(S) = \frac{\sum_{i=1}^{n}\rho_\omega\,(t_i)}{n} + \frac{\#(S)\sum_{i=1}^{n}\psi^2_\omega\,(t_i)}{n\sum_{i=1}^{n}\psi'_\omega\,(t_i)},$$

onde $\#(S)$ denota a cardinalidade de S. Tanto para o método de seleção de variáveis *Foreward*, quanto *Backward* deve-se escolher aquele em que a inclusão ou exclusão das variáveis resulte no menor valor de $FRPE$.

Tabatabai et al. (2012) apresentam um novo coeficiente de determinação, baseado no erro de predição robusto final RFPE, definido por

$$R^2_{RFPR} = 1 - \left(\frac{RFPE\left(M_{comp}\right)}{RFPE(M_{const})}\right)^2,$$

onde M_{const} é o modelo apenas com a constante e M_{comp} é o modelo completo. Essa estatística mede a performance global do modelo, em que $0 \leq R^2_{RFPR} \leq 1$.

Para proceder com testes de hipóteses deve-se considerar o espaço paramétrico $\Omega \subseteq \boldsymbol{R}^{k+1}$ e $\{\beta_{j1}, \beta_{j2}, \dots, \beta_{jq}\}$ como um subconjunto de $\{\beta_0, \beta_1, \dots, \beta_k\}$.

Define-se

$$\Omega_0 = \{\beta \in \Omega: \beta_{j1} = \beta_{j2} = \cdots = \beta_{jq} = 0\} \qquad (14)$$

e a função $f(\beta)$ como

$$f(\beta) = \sum_{i=1}^{n}\frac{\rho_\omega(t_i)}{L_i}.$$

Então uma estatística robusta, do tipo razão de verossimilhanças, para testes de hipóteses pode ser obtida com

$$S_n^2 = \frac{2\left(\sup_{\beta\in\Omega_0}\{f(\beta)\} - \sup_{\beta\in\Omega}\{f(\beta)\}\right)}{q}$$

onde q é o número de parâmetros na restrição (14).

Assintoticamente sob a hipótese nula $\{E[\psi'_\omega(t)]\}/\{E[\psi^2_\omega(t)]\}S_n^2$ possui distribuição $\mathcal{X}_q^2$, em que a estatística de teste do tipo de Wald, é definida por

$$W_n^2 = n\left(\hat{\beta}_{j1}, \hat{\beta}_{j2}, \dots, \hat{\beta}_{jq}\right)V_q^{-1}\left(\hat{\beta}_{j1}, \hat{\beta}_{j2}, \dots, \hat{\beta}_{jq}\right)',$$

onde $(1/n)V_q$ é a matriz de covariância assintótica do vetor $\{\hat{\beta}_{j1}, \hat{\beta}_{j2}, \dots, \hat{\beta}_{jq}\}$. Sob a hipótese nula a estatística $W_n^2 \sim \mathcal{X}_q^2$.

Para os diagnósticos de *outliers*, medidas influentes e pontos de alavancagem, Tabatabai et al. (2012) sugerem as seguintes estatísticas, que devem ser usadas em conjunto com análises gráficas.

Resíduos robustos studentizados, usando as estimativas robustas de TELBS dos parâmetros, são dados por

$$SR_i(TELBS) = \frac{t_i}{\hat{\sigma}\sqrt{1-h_{ii}}},$$

onde $\hat{\sigma}$ é dado nas equações (9) e (10), propostas por Rousseeuw e Croux (1993) e t_i é dado pela equação (8).

A distância robusta de Cook, usando as estimativas robustas de TELBS dos parâmetros, é encontrada por

$$CD_i(TELBS) = \frac{h_{ii}t_i^2}{p(1-h_{ii})^4}.$$

A medida de influencia definida por

$$S_h(i) = \frac{h_{ii} - \underset{\{i:1\le i\le n\}}{\text{Mediana}}(h_{ii})}{\hat{\sigma}_h},$$

onde

$$\hat{\sigma}_h = 1{,}1926\underset{\{i:1\le i\le n\}}{\text{Mediana}}\left(\underset{\{j:1\le j\le n\}}{\text{Mediana}}\left(|r_i - r_j|\right)\right).$$

Essa estatística parece ser muito boa para identificação de pontos de alavancagem (Tabatabai et al. 2012). Valores grandes de $|S_h(i)|$ podem indicar a presença de observações influentes.

Voltando ao Exemplo 1, o gráfico abaixo mostra as retas ajustadas para as duas simulações com o estimador de TELBS para regressão linear robusta, com eficiência assintótica de 95%, ou seja, com o valor de $\omega = 0{,}405$.

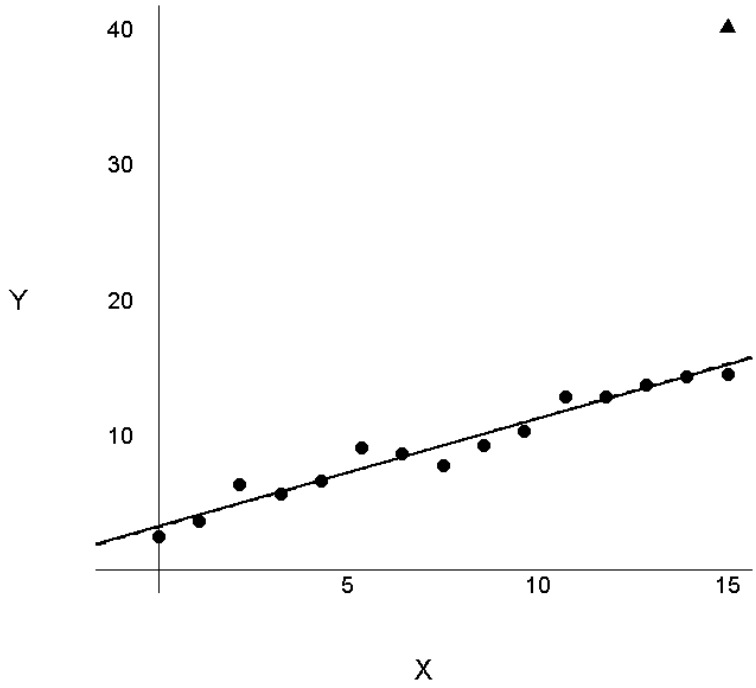

Figura 9: Gráfico de dispersão e retas ajustadas com o estimador de TELBS para regressão linear robusta para $\omega = 0{,}405$.

A reta ajustada com o estimador de TELBS, com eficiência assintótica de 95%, para a simulação com *outlier* – reta tracejada – não sofre mais uma forte influência em sua inclinação como quando estimada pelo método dos mínimos quadráticos ordinários. Note que, não é sequer possível distinguir as retas ajustadas na Figura 9, o que indica que os ajustes para as duas simulações do exemplo são muito próximos. As estimativas pontuais para os coeficientes de regressão através do modelo regressão linear robusto de TELBS encontram-se na tabela abaixo.

Tabela 4 – Estimativas dos coeficientes de regressão com o estimador de TELBS para regressão linear robusta para $\omega = 0{,}405$.

Parâmetros	Verdadeiro valor	Estimativas	
		Sem *outlier* (EP)	Com *outlier* (EP)
β_0	3	3,2344 (0,3946)	3,1968 (0,3978)
β_1	0,8	0,7934 (0,0448)	0,8042 (0,0451)

Abreviaturas: EP (Erro Padrão).

Tanto as estimativas pontuais, para a simulação sem *outlier*, quanto às estimativas da simulação com *outlier*, estão próximas dos verdadeiros valores dos coeficientes de regressão. O método de TELBS foi o que obteve a melhor estimativa do parâmetro β_1 na simulação com *outlier*, comparado com o M-Estimador de Huber e os mínimos quadráticos ordinários, utilizados anteriormente.

O erro padrão das estimativas sem *outlier* é muito próximo para os três métodos abordados, o que indica que não há um grande aumento na variância quando se utiliza o método de TELBS ou o M-Estimador de Huber. Já o erro padrão das estimativas com *outlier* é menor quando se utiliza um M-Estimador de Huber ou o método de TELBS, visto que o método dos mínimos quadráticos ordinários sofre grande influência de pontos extremos.

Para comparação do método de estimação de TELBS com outros métodos existentes, também foi realizada um simulação por Tabatabai et al. (2012) com os softwares R e Mathematica. Nessa simulação foi avaliado seu desempenho em comparação com os estimadores de mínimos quadráticos, M-Estimador e MM-Estimador (ver Yohai, 1987) com uma eficiência assintótica de 95%. Para isso, foram geradas 1000 amostras de tamanho $n = 20$, $\varepsilon_i \sim iid\ N(0,1)$, $x_i \sim N(0,1)$ e y_i definido pela equação

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

onde os valores dos coeficientes da regressão são iguais a $\beta_0 = 1$ e $\beta_1 = 3$. Também foram simuladas outras 1000 amostras de tamanho $n = 100$, $\varepsilon_i \sim iid\ N(0,1)$, x_{1i},$x_{2i} \sim N(0,1)$ e y_i definido pela equação

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i,$$

onde os valores dos coeficientes da regressão são iguais a $\beta_0 = 1$, $\beta_1 = 3$ e $\beta_2 = 0{,}5$. Para avaliar a robustez dos estimadores, foram aleatoriamente escolhidas 5%, 25% e 40% das observações simuladas, que foram contaminadas através do aumento do seu tamanho por um fator de 1000, em que primeiro contaminou-se a variável x, então y e por último em ambas. Para a comparação dos resultados foram utilizadas as estimativas de

$$Bias = \left| \frac{\sum_{l=1}^{m} (\hat{\beta}_l)}{m} - \beta \right|$$

e

$$EQM = \frac{\sum_{l=1}^{m} \left(\hat{\beta}_l - \beta\right)^2}{m},$$

onde m é o número de replicações da simulação, EQM é o erro quadrático médio e Bias é o vício.

Através dos resultados encontrados nessa simulação, os autores concluíram que, o método de estimação dos mínimos quadráticos tem um desempenho ruim para todos os níveis de contaminação e para ambos os tamanhos de amostras. O M-Estimador possui um baixo desempenho em todos os casos, com exceção no nível de 5% de contaminação de, apenas, a variável y. O MM-Estimador possui melhor desempenho do que os métodos de mínimos quadráticos e M-Estimador. No geral, o método de estimação não parece funcionar tão bem quando os níveis de contaminação são de 25% e 40% e, sobretudo, quando a razão entre o número de parâmetros e o tamanho da amostra é grande. O estimador de TELBS superou o desempenho dos demais métodos de estimação para todos os níveis de contaminação e em ambos os tamanhos de amostra.

Tabela 5: Vício e erro quadrático médio para amostras de tamanho $n = 20$ e $n = 100$, com níveis de 5%, 25% e 40% das observações simuladas contaminadas na variável y.

	5%		25%		40%	
	β_0	β_1	β_0	β_1	β_0	β_1
$n = 20$						
LS						
Vício	43,6	154,4	249,7	749,8	414,8	1185,6
EQM	28576	72966	156694	156694	300353	1643825
M						
Vício	0,0201	0,0777	15,6	79,4	263,9	832,7
EQM	0,0745	0,0786	13866	95715	187651	1098011
MM						
Vício	0,0134	0,0109	0,0020	0,0115	0,0076	0,0016
EQM	0,0728	0,0884	0,0920	0,1245	0,1026	0,1094
TELBS						
Vício	0,0068	0,0002	0,0113	0,0160	0,0015	0,0081
EQM	0,0648	0,0862	0,0891	0,1135	0,0939	0,1319
$n = 100$						
LS						
Vício	45,3	145,3	252,3	751,2	405,8	1199,1
EQM	6507	30716	81952	602531	189135	1485956
M						
Vício	0,0251	0,0699	0,2283	0,7080	232,7	707,8
EQM	0,0140	0,0171	0,1238	0,6627	80514	631891
MM						
Vício	0,0002	0,0014	0,0043	0,0019	0,0017	0,0045
EQM	0,0113	0,0116	0,0132	0,0134	0,0174	0,0185
TELBS						
Vício	0,0034	0,0000	0,0006	0,0017	0,0006	0,0056
EQM	0,0123	0,0140	0,0145	0,0179	0,0164	0,0192

Abreviaturas: LS (Mínimos quadráticos); EQM (Erro quadrático médio); TELBS (Tabatabai, Eby, Li, Bae e Singh).
Fonte: Tabatabai et al. (2012).

A Tabela 5 traz um resumo da simulação, com o vício e erro quadrático médio para os dois tamanhos de amostras simulados, para os níveis de contaminação de 5%, 25% e 40% na variável y. Os resultados referentes às contaminações das variáveis x, y e x conjuntas e dos modelos com três parâmetros, para as variáveis x_1, x_2 e y , encontram-se no Anexo 5.5.

Em, Tabatabai et al. (2012), os autores também compararam o método de TELBS com os demais métodos de estimação anteriormente citados em duas análise de dados reais. O primeiro conjunto de dados analisados trata de um estudo do fluxo sanguíneo cerebral, em que tanto o estimador de TELBS quanto o MM-Estimador fizeram estimativas igualmente precisas. A Figura 10 mostra o gráfico de dispersão e a reta ajustada pelo método de TELBS, das variáveis *Bolus tracking* e *Spin labeling.*

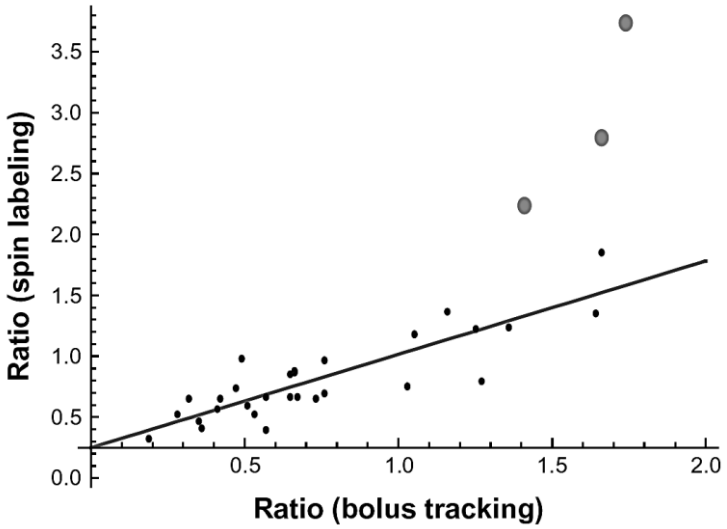

Figura 10: Gráfico de dispersão e reta ajustada pelo método de TELBS.
Fonte: Tabatabai et al. (2012)

A segunda análise foi feita utilizando dados de um experimento com fígados de ratos, em que nenhum dos estimadores – mínimos quadráticos, M-Estimador, MM-Estimador e TELBS – fez boas estimativas sobre os parâmetros de regressão. Entretanto, para ambos os conjuntos de dados reais, a medida $S_h(i)$, usando as estimativas robustas de TELBS dos parâmetros, mostrou-se muito eficiente na identificação de observações influentes (Tabatabai et al. 2012).

3. Análise de dados reais

Para a análise dos dados reais foi utilizada uma amostra de diferentes informações provenientes de uma empresa de vendas online. Essa empresa atua no mercado digital, na área de investimentos, a mais de quatro anos, com a venda de *e-books* (livros eletrônicos) e vídeo-aulas para ensinar futuros investidores. Ela trabalha através de um site que vende quatro diferentes tipos de produtos – produzidos pelo próprio dono da empresa – e também vende, sob o ganho de comissão, outros diferentes produtos de sites parceiros.

O dono da empresa não autorizou a divulgação do nome da empresa bem como da denominação de seus produtos de forma que, no que segue, será utilizada uma denominação fantasia para os produtos vendidos pela empresa. Dos quatro diferentes produtos feitos pelo proprietário da empresa, tem-se um maior interesse em avaliar um produto específico, aqui denominado de "Produto M", pois este é o único que possui estratégias de propaganda e marketing que potencializam sua venda, consequentemente influenciando na sua receita e no faturamento total da empresa.

As estratégias de marketing, assim como as vendas, também são feitas de forma virtual, através de *facebook ads* – anúncios na rede social *facebook* – e variam de diversas formas, conforme o desejo do próprio dono da empresa. Foram coletados dados diários, no período de 26/06/2015 a 10/08/2015 – com um total de $n = 46$ observações – de sete variáveis quantitativas, que são:

1. Faturamento Total (FT): representa o faturamento total – em reais – em certo dia i. Esse faturamento compreende tanto os quatro produtos do proprietário da empresa, quanto os demais produtos vendidos de sites parceiros.
2. Produto M (PM): representa a receita total – em reais – gerada pelo Produto M – aquele que recebe investimentos em propaganda e marketing *online* através da rede social *facebook* – em certo dia i.
3. Sessões (SS): representam o número total de acessos ao site em certo dia i. A sessão independe do meio ao qual se chegou ao site ou de quantas páginas foram acessadas dentro do site.
4. *Web Clicks* (WC): representa o número total de acessos ao site, provenientes exclusivamente da rede social *facebook*, em certo dia i.
5. Alcance (REACH): representa o número total de visualizações das campanhas de marketing do Produto M anunciadas no *facebook*, em certo dia i.
6. *Checkouts* (CKO): representa o número total de Produtos M que foram comprados, em certo dia i, exclusivamente através da rede social *facebook*.

7. *Amount Spent* (AS): representa o total gasto – em dólares – com as campanhas de marketing feitas através da rede social *facebook*, em certo dia i.

O objetivo principal da análise destes dados é o de estudar as relações entre o faturamento total da empresa com as demais variáveis e, principalmente, o relacionamento do faturamento total com a variável que representa a receita total gerada pelo Produto M, pois se pressupõem que este gera um grande impacto naquele, visto que é o único produto que recebe investimento em publicidade e propaganda.

3.1. Análise descritiva

Para melhor compreensão dos dados, será apresentada uma breve análise descritiva das principais variáveis de interesse – Faturamento Total e Produto M – e de algumas das demais variáveis que possam fornecer informações relevantes para o estudo.

A variável resposta – faturamento total – não apresenta distribuição normal, aparentemente, conforme podemos observar no gráfico da variável padronizada abaixo, o que sugere que as estimativas dos parâmetros para o modelo de regressão linear, utilizando o método dos mínimos quadráticos ordinários, podem não ser precisas.

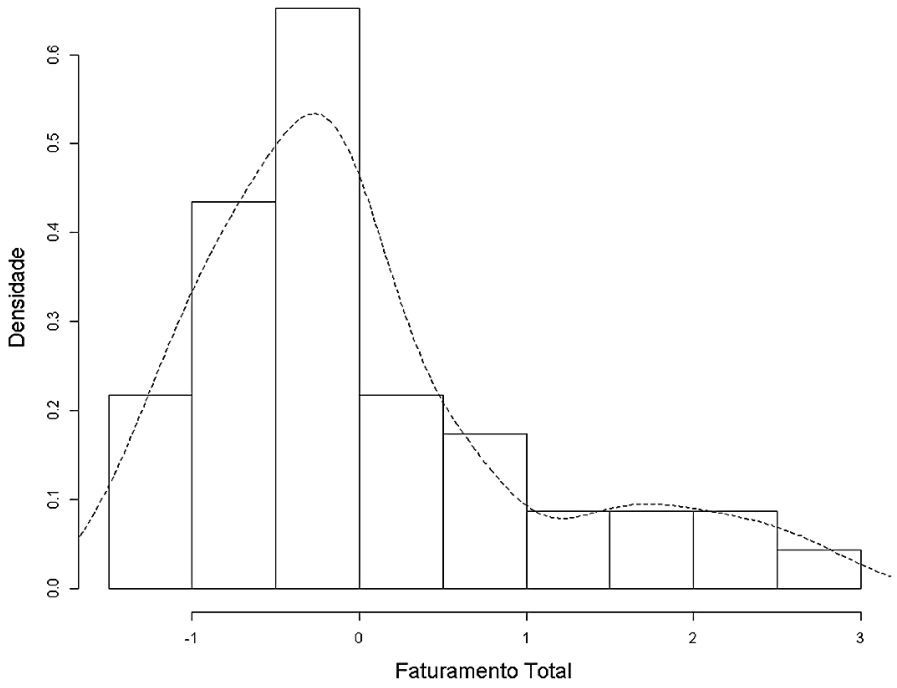

Figura 11: Histograma da variável Faturamento Total.

A Tabela 6 fornece informações das médias, medianas, desvios padrão, valores máximos e mínimos de cada uma das variáveis. Percebe-se que esses valores são relativamente próximos nas variáveis do Faturamento Total e da

receita gerada pelo Produto M, o que pode sugerir que estas tenham uma forte relação entre si.

Tabela 6: Análise descritiva da amostra.

Variável	Mínimo	Média	Máximo	Mediana	Desvio Padrão
Faturamento Total (R$)	115,13	971,82	2.698,32	852,77	623,01
Produto M (R$)	0,00	791,21	2.471,20	737,29	563,1
Sessões	2662	5089,5	8761	4930,5	1437,69
Web Clicks	103	852,72	2012	693,5	525,23
Reach	8243	40957,83	91803	33714	23699
Checkouts	0	10,67	35	9	7,75
Amount Spent ($)	10,74	105,64	263,15	73,09	74,68

Outro indício de que as estimativas dos parâmetros do modelo de regressão, feitas através do método dos mínimos quadráticos ordinários, podem não ser precisas, é a presença de pontos extremos (*outliers*), tanto na variável resposta – Faturamento Total –, quanto na variável explicativa – Produto M –, os quais podem ser vistos na Figura 12.

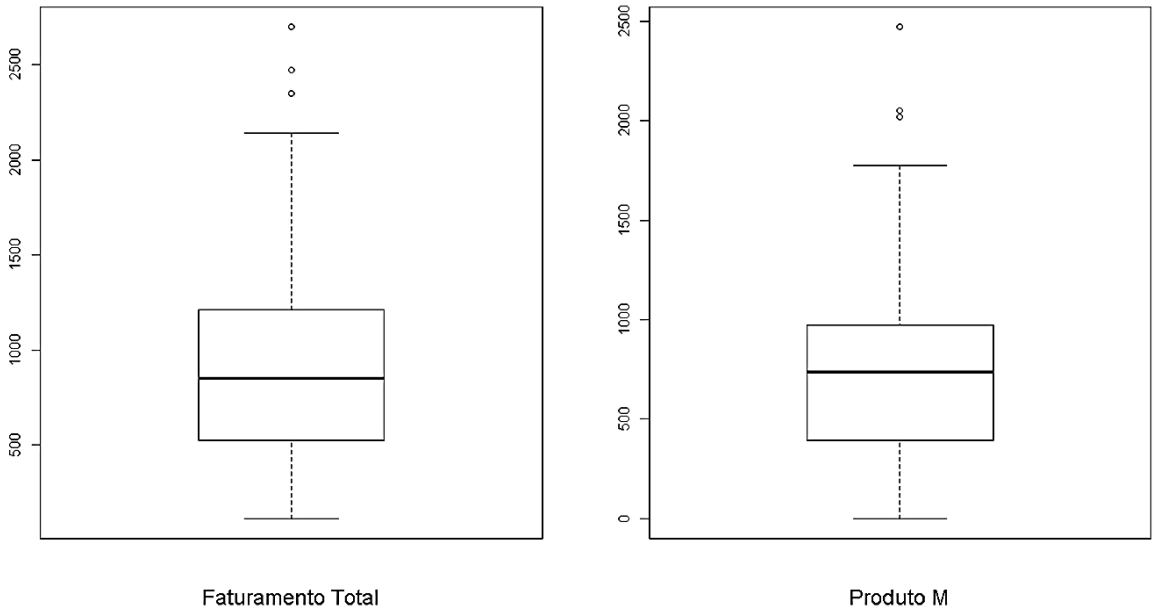

Figura 12: *Boxplots* das variáveis Faturamento total e Produto M.

3.2. Análise de regressão

A análise de regressão foi feita com as variáveis Faturamento Total – variável resposta – e Produto M – variável explicativa –, pois acredita-se que estas se relacionem de forma linear e são as de maior interesse para o dono da empresa. Para comparação, foi estimado um modelo de regressão linear, através de três métodos distintos: Mínimos quadráticos ordinários para regressão linear simples, M-Estimador e TELBS para regressão linear robusta.

Como estas duas variáveis são séries temporais, pois são coletadas diariamente, primeiramente é necessário se preocupar com o problema de regressão espúria, que ocorre quando as séries temporais, das variáveis que se pretende regredir, não são estacionárias. Para que a regressão seja legítima – não espúria – as variáveis devem cointegrar (ver Bueno, 2008), para isso foi aplicado o Teste de raiz unitária de Phillips-Perron (ver Phillips et al. 1988), conforme a tabela abaixo.

Tabela 7 – Teste de raiz unitária de Phillips-Perron. H_0: a série apresenta raiz unitária.

	Faturamento Total	Produto M
p-valor	0,01	0,01

Conforme a Tabela 7, pelo teste de raiz unitária de Phillips-Perron, com um p-valor = 0,01, as séries podem ser consideradas estacionárias e a regressão entre as variáveis não apresentará resultado espúrio.

O gráfico de dispersão abaixo mostra as retas ajustadas, pelos métodos de TELBS – com eficiência assintótica de 95% –, M-Estimador de Huber e mínimos quadráticos ordinários, para as variáveis Faturamento total e Produto M.

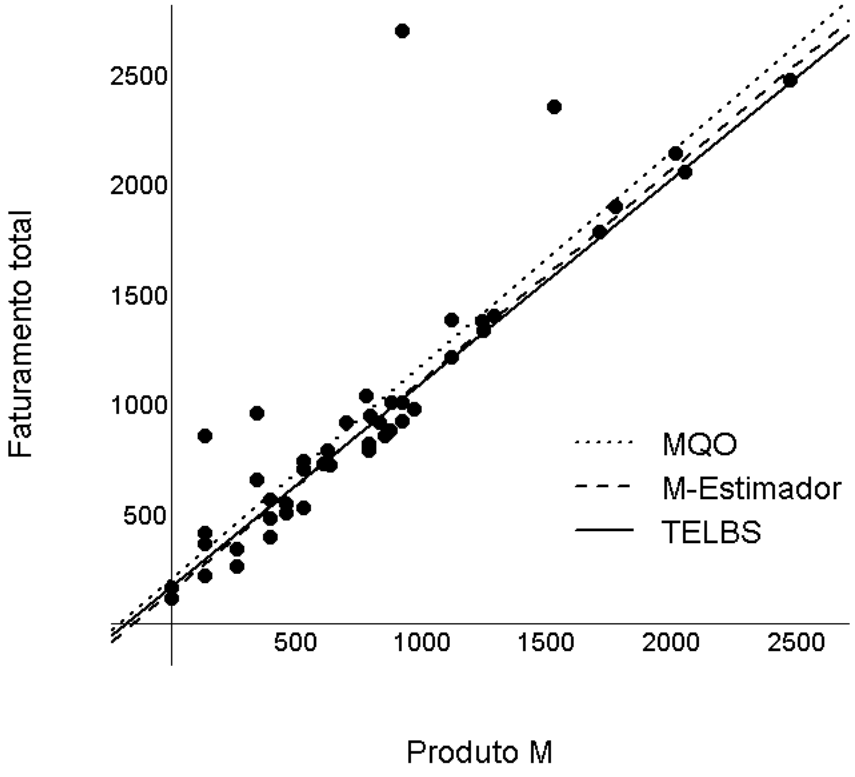

Figura 13: Gráfico de dispersão e retas ajustadas com os estimadores de TELBS, M-Estimador de Huber e mínimos quadráticos ordinários.

Na Figura 13, pode-se observar que as curvas ajustadas estão próximas na área do gráfico de dispersão em que se encontram a maior parte das observações, mas começam a se separar na área das observações com maiores valores. Note que, na área do gráfico em que se encontram a maior parte das observações, as retas ajustadas pelos métodos de TELBS – reta contínua – e M-Estimador de Huber – reta tracejada – estão mais próximas do centro do conjunto de pontos, enquanto que, a reta ajustada pelo método dos mínimos quadráticos ordinários – reta pontilhada – é deslocada em direção aos pontos discrepantes, se distanciando deste centro. Mas a reta ajustada com o M-Estimador de Huber sofre maior influência em sua inclinação, em comparação com a reta ajustada pelo método de TELBS.

As estimativas pontuais para os coeficientes de regressão linear, obtidas através dos estimadores de TELBS – com eficiência assintótica de 95% –, M-Estimador de Huber e mínimos quadráticos ordinários, encontram-se na tabela abaixo.

Tabela 8: Estimativas para os parâmetros do modelo de regressão linear pelos métodos de TELBS, M-Estimador de Huber e mínimos quadráticos ordinários.

Estimador	Estimativas	
	β_0	β_1
MQO	202,9702	0,9717
M-Estimador H.	145,2149	0,9597
TELBS	165,5284	0,9280

Abreviaturas: MQO (Mínimos quadráticos ordinários), H. (Huber) e TELBS (Tabatabai, Eby, Li, Bae e Singh).

As estimativas para o parâmetro β_1 estão relativamente próximas, com a maior diferença entre o método de TELBS e dos mínimos quadráticos ordinários. O intervalo de confiança, a 95%, para o parâmetro β_1, estimado pelo método de TELBS, é [0,8717 ; 0,9844]. Já as estimativas do parâmetro β_0 estão mais distantes, se considerar o que efetivamente representam as variáveis para a empresa, essa diferença é de, aproximadamente R$57,00, entre o método dos mínimos quadráticos ordinários e o M-Estimador de Huber, valor que pode ser considerável para o dono do site, visto que se está medindo uma variação diária do faturamento total. O intervalo de confiança, a 95%, para o parâmetro β_0, estimado pelo método de TELBS, é [110,99 ; 220,06].

O coeficiente de determinação, baseado no erro de predição robusto final RFPE, que mede a performance global do modelo, é de $R^2_{RFPR} = 0{,}9623$, o que indica que o ajuste, utilizando apenas a variável que representa a receita total gerada pelo Produto M, explica aproximadamente 96% da variação total do faturamento da empresa. Esse valor pode ser considerado muito alto, tornando desnecessário o acréscimo de mais variáveis no modelo, visto que essas explicariam apenas um pouco mais da variação total do faturamento.

A equação final, que explica a relação linear entre as variáveis faturamento total e a receita total gerada pelo Produto M, pelo método de estimação robusto de TELBS, é

$$y_i = 165{,}5284 + 0{,}9280x_i,$$

onde, y_i representa o faturamento total da empresa, em certo dia i, e x_i representa a receita total gerada pelo Produto M, em certo dia i. O faturamento total, em certo dia i, pode ser então previsto a partir da venda do produto M, através da equação 165 + 0,93 × a receita gerada pelo produto M neste mesmo dia i.

4. Conclusão

A análise de regressão linear robusta é um método de simples implementação e de grande eficiência para dados que possuem pontos discrepantes, que são comumente encontrados em observações reais, quando comparada ao método dos mínimos quadráticos ordinários.

No exemplo simulado neste trabalho, as estimativas dos parâmetros de regressão, feitas através do método dos mínimos quadráticos ordinários, foram fortemente distorcidas quando se colocou um *outlier*. Por outro lado, as estimativas feitas pelos métodos do M-Estimador de Huber e TELBS, foram muito próximas dos verdadeiros valores dos parâmetros. O método de estimação de regressão linear robusta de TELBS obteve a melhor estimativa do parâmetro β_1 para o exemplo simulado, em comparação aos outros dois métodos.

Na análise dos dados reais de *e-commerce*, as estimativas dos parâmetros de regressão, feitas através do método dos mínimos quadráticos ordinários, para as variáveis faturamento total e produto M, foram influenciadas pelos pontos discrepantes, tendo sua reta ajustada deslocada na direção desses pontos, enquanto que as retas ajustadas pelos métodos de TELBS e M-Estimador de Huber sofreram pouca influência dos valores extremos. As estimativas pontuais do parâmetro β_0 foram relativamente diferentes entre o método dos mínimos quadráticos ordinários e o M-Estimador de Huber, se considerar o que representam efetivamente essas variáveis para a empresa, essa diferença é de, aproximadamente, R$57,00, valor que pode ser considerável para o dono do site, visto que se está medindo uma variação diária do faturamento total da empresa. A equação que explica a relação linear entre as variáveis faturamento total e a receita total gerada pelo Produto M, permite concluir que o faturamento total, em certo dia i, pode ser previsto a partir da venda do produto M, através da equação $165 + 0{,}93 \times$ a receita gerada pelo produto M neste mesmo dia i.

Na simulação feita por Tabatabai et al. (2012), os autores concluíram que o método de estimação dos mínimos quadráticos tem um desempenho ruim para todos os níveis de contaminação e para ambos os tamanhos de amostras. O M-Estimador possui um baixo desempenho em todos os casos, com exceção no nível de 5% de contaminação apenas da variável y. O MM-Estimador possui melhor desempenho do que os métodos de mínimos quadráticos e M-Estimador. No geral, o método de estimação não parece funcionar tão bem quando os níveis de contaminação são de 25% e 40% e, sobretudo, quando a razão entre o número de parâmetros e o tamanho da amostra é grande. O estimador de TELBS superou o desempenho dos demais métodos de estimação para todos os níveis de contaminação e em ambos os tamanhos de amostra.

Na análise de dados reais, feita por Tabatabai et al. (2012), os autores concluíram que tanto o estimador de TELBS quanto o MM-Estimador fizeram estimativas igualmente precisas para o estudo do fluxo sanguíneo cerebral. No

experimento com fígados de ratos nenhum dos estimadores – mínimos quadráticos, M-Estimador, MM-Estimador e TELBS – fez boas estimativas sobre os parâmetros de regressão. Entretanto, para ambos os conjuntos de dados reais, a medida $S_h(i)$, usando as estimativas robustas de TELBS dos parâmetros, mostrou-se muito eficiente na identificação de observações influentes.

A regressão linear robusta mostra-se uma eficaz alternativa para a análise de dados que possuem pontos discrepantes, tanto simulados, quanto reais. Estimadores robustos como o de TELBS e a classe de M-Estimadores, são sensíveis à influência de *outliers* e de fácil implementação, produzindo as mesmas estimativas que o método dos mínimos quadráticos ordinários produziria na ausência de valores extremos ou quando a variável resposta é normalmente distribuída.

Como trabalhos futuros, seria interessante analisar o relacionamento da variável *Amount Spent* com as demais, para se compreender as possíveis influências do investimento diário em propaganda. Outra análise importante, é a relação dos diferentes tipos de campanhas publicitárias utilizadas, para se encontrar as de melhor eficiência e que possam produzir os melhores resultados para a empresa.

5. Anexos

5.1. Sintaxe R – Exemplo 1

```
#definindo os parametros e o tamanho da amostra
n <- 15
b0 <- 3
b1 <- 0.8
d <- 1
set.seed(123) # definindo semente para que se recriem os exemplos.
e <- rnorm(n, 0, d) # criando o erro aleatório
X <- seq(0, 15, length=n) # criando a var. X
Y <- b0+b1*X+e  # criando a var. Y
m1 <-  lm(Y~X) # modelo linear sem outlier
## inserindo os outliers em Y
out <- 40   #definindo o valor do outlier
Y2 <- c(Y[1:14],out)
m2 <-  lm(Y2~X)  # modelo linear com outlier
#criando o gráfico das retas ajustadas pelo MQO para ambos os modelos
windows()
plot(X, Y,pch=19,cex=1.5,ylab="",xlab="X",cex.lab=1.5,xlim=c(-
1,15),ylim=c(0,40),axes=F)
abline(m1,lwd=2)
abline(m2,lwd=2,lty=2)
abline(h = 0,v = 0)
mtext("Y", side=2, line=3, cex=1.5,las=2)
text(5,-1,cex=1.2,"5")
text(10,-1,cex=1.2,"10")
text(15,-1,cex=1.2,"15")
text(-.5,10,cex=1.2,"10")
text(-.5,20,cex=1.2,"20")
text(-.5,30,cex=1.2,"30")
text(-.5,40,cex=1.2,"40")
points(15,40,pch=17,cex=1.5)
#estimativas dos parametros
m1
m2
summary(m1)
summary(m2)
## regressão robusta com M-Estimador pelo método iterativo IRLS
require(MASS)
m3 <-  rlm(Y~X,psi="psi.huber")
m4 <-  rlm(Y2~X,psi="psi.huber")
windows()
plot(X, Y,pch=16,cex=1.5,ylab="",xlab="X",cex.lab=1.5,xlim=c(-
1,15),ylim=c(0,40),axes=F)
abline(m3,lwd=2)
abline(m4,lwd=2,lty=2)
abline(h = 0,v = 0)
mtext("Y", side=2, line=3, cex=1.5,las=2)
text(5,-1,cex=1.2,"5")
```

```
text(10,-1,cex=1.2,"10")
text(15,-1,cex=1.2,"15")
text(-1,10,cex=1.2,"10")
text(-1,20,cex=1.2,"20")
text(-1,30,cex=1.2,"30")
text(-1,40,cex=1.2,"40")
points(15,40,pch=17,cex=1.5)
m3
m4

## regressão robusta com estimador de TELBS
b0 <- 3
b1 <- 0.8
set.seed(123) # definindo semente para que se recriem os exemplos.
e <- rnorm(15, 0, 1) # criando o erro aleatório
Xis <- seq(0, 15, length=15) # criando a var. X
Y <- b0+b1*Xis+e  # criando a var. Y
m1 <-  lm(Y~Xis) # modelo linear sem outlier
## inserindo os outliers em Y
out <- 40   #definindo o valor do outlier
Y2 <- c(Y[1:14],out)
banco1<-cbind(Y2,Xis)
banco2<-cbind(Y,Xis)
cs=0.405
lmauto2=function(m)
{
data=banco2 #para ajuste com outlier, use banco1, para ajuste sem outlier use banco2
code1=0
sig=1
y=Y #para ajuste com outlier, use Y2, para ajuste sem outlier use Y
x1=Xis
n=dim(data)[1]
int=array(0,c(m,1))
se=array(0,c(m,1))
slope1=array(0,c(m,1))
des=array(1,c(n,1))
mr=array(0,c(n,1))
ma1=array(0,c(n,1))
x=x1
a=abs(x1)
wd=cbind(des,x)
w=solve(t(wd)%*%wd)
diff1=10
diff2=10
j=1
while(abs(diff1)>0.01|abs(diff2)>0.01)
{
j=j+1
lof=function(par)
{
stdev=sig
prob=array(0,c(n,1))
num=array(0,c(n,1))
```

```
sm=wd%*%solve(t(wd)%*%wd)%*%t(wd)
mf=diag(diag(sm))%*%des
for (i in 1:n)
{
prob[i]=par[1]+par[2]*x1[i]
ma1[i]=max(median(a),abs(x1[i]))
}
num=1-mf
losf=sum((1-1/cosh(cs*(1/stdev)*(y-prob)*num))/ma1)
return(losf)
}
result=optim(par=c(0.2,-0.02),fn=lof)
a=result$par[1]
b=result$par[2]
r=y-(a+b*x1) ## r is a vector of residuals
for (k in 1:n)
{
mr[k]=median(abs(r[k]-r))
}
sig=1.1926*median(mr)
code=result$convergence
code1=code1+code
int[j]=result$par[1]
slope1[j]=result$par[2]
se[j]=sig
diff1=int[j]-int[j-1]
diff2=slope1[j]-slope1[j-1]

#print(diff2)
}
out=array(0,c(3,1))
out[1]=int[j]
out[2]=slope1[j]
out[5]=se[j]
out[6]=median(abs(r))
return(out)
}
a=lmauto2(100)[1]
b=lmauto2(100)[2]
sighat=lmauto2(100)[5]
ma=lmauto2(100)[6]

ratcov2=function(cs){
data=banco2
n=dim(data)[1]
y=Y
x1=Xis
x=x1
des=array(1,c(n,1))
mr=array(1,c(n,1))
num0=array(0,c(n,1))
wd=cbind(des,x);
sm=wd%*%solve(t(wd)%*%wd)%*%t(wd)
```

```
mf=diag(diag(sm))%*%des
s1=array(0,c(n,1))
s2=array(0,c(n,1))
p=dim(data)[2]
num=1-mf
for (i in 1:n)
{
num0[i]=(1-(1/n))
}
##compute R-square
for (k in 1: n){
mr[k]=median(abs(y[k]-y))
}
rsq=1-(ma/median(mr))^2
## Estimate the variance covariance matrix
g1=function(t){
g=cs*1/cosh(cs*t)*tanh(cs*t) ##first derivative
return(g)
}
g2=function(t){
g=cs^2*(1/cosh(cs*t))^3-cs^2*(1/cosh(cs*t))*(tanh(cs*t))^2
return(g) ##second derivative
}
for(i in 1:n){
c1=(y[i]-(a+b*x1[i]))*(1-mf[i])/sighat
s1[i]=g1(c1)
s2[i]=g2(c1)
}
wt=solve(t(wd)%*%wd)
sq=(sum(s2))^2
variance=sum(s1^2)/((n-p)*sq)
##covariance matrix
cov=(sighat^2)*(n^2)*variance*wt
##Standard error
sea=sqrt(diag(cov))[1]
seb=sqrt(diag(cov))[2]
##compute t-value and Wald chi-square
t1=a/sea
chsq1=t1^2
t2=b/seb
chsq2=t2^2
##compute p-value
pvalue1=1-pchisq(chsq1,1)
pvalue2=1-pchisq(chsq2,1)
## Output parameter estimate, S.E., t-value, variance-covariance matrix,
## Wald chi-square, and p-value for each parameter
cat("a:", a,"\n");
cat("b:", b,"\n");
cat("SE (a):", sea,"\n");
cat("SE (b):", seb,"\n");
cat("t value (a):", a/sea,"\n");
cat("t value (b):", b/seb,"\n");
cat("Var-Cov matrix:", cov,"\n");
```

```
cat("Wald (a):", chsq1,"\n");
cat("Wald (b):", chsq2,"\n");
cat("p-value (a):", pvalue1,"\n");
cat("p-value (b):", pvalue2, "\n");
cat("Rsquare:", rsq,"\n");
cat("Sigma hat:", sighat,"\n");
}
ratcov2(cs)

windows()
plot(X, Y,pch=16,cex=1.5,ylab="",xlab="X",cex.lab=1.5,xlim=c(-
1,15),ylim=c(0,40),axes=F)
abline(a=3.234367,b=.7933694,lwd=2)
abline(a=3.196826,b=.8042209,lwd=2,lty=2)
abline(h = 0,v = 0)
mtext("Y", side=2, line=3, cex=1.5,las=2)
text(5,-1,cex=1.2,"5")
text(10,-1,cex=1.2,"10")
text(15,-1,cex=1.2,"15")
text(-1,10,cex=1.2,"10")
text(-1,20,cex=1.2,"20")
text(-1,30,cex=1.2,"30")
text(-1,40,cex=1.2,"40")
points(15,40,pch=17,cex=1.5)
```

5.2. Sintaxe R – Pontos de alavanca e influência

```
#definindo os parametros para Y
b0 <- 3
b1 <- 0.8
set.seed(12345) # definindo semente para que se recriem os exemplos.
e1 <- rnorm(10, 0, 1) # criando o erro aleatório
X1 <- seq(1, 10, length=10) # criando a var. X1
Y1 <- b0+b1*X1+e1 #definindo Y
e2 <- rnorm(15, 0, 1) # criando o erro aleatório
X2 <- seq(1, 16, length=15) # criando a var. X2
Y2 <- b0+b1*X2+e2   #definindo Y
X3 <- c(1,2,3,4,5,14,7,8,9,10,11,12,13,14,15)
Y3 <-  c(Y2[1:5],6,Y2[7:15])
m1 <-  lm(Y1~X1)  #modelo 1 de regressao linear
m2 <-  lm(Y3~X3)  #modelo 2 de regressao linear
windows()
#criando os gráficos
par(mfrow=c(1,2))
plot(X1, Y1,pch=19,cex=1.5,ylab="",xlab="",xlim=c(-2,30),ylim=c(-.5,25),axes=F)
text(-2,12.5,cex=1.5,"Y")
text(15,-1,cex=1.5,"X")
points(30,24,pch=17,cex=1.5)
abline(h = 0,v = 0)
abline(m1,lwd=2,lty=2)
mtext("Ponto de Alavanca", side=1, line=1, cex=1.5,las=1)
```

```
plot(X2, Y2,pch=19,cex=1.5,ylab="",xlab="",xlim=c(-1,16),ylim=c(-.5,16),axes=F)
text(-1,8,cex=1.5,"Y")
text(8,-.5,cex=1.5,"X")
points(14,6,pch=17,cex=1.5)
abline(h = 0,v = 0)
abline(m2,lwd=2,lty=2)
mtext("Ponto Influente", side=1, line=1, cex=1.5,las=1)
```

5.3. Sintaxe R – Análise descritiva

```
windows()
#histograma da variável Faturamento Total
hist(FT,prob=T,main=" ",cex=1.2,ylab="Densidade",xlab="Faturamento Total")
lines(density(FT),lty=2)
windows()

#padronizando a variável FT
mi=mean(FT)
se=sqrt(var(FT))
FTpad=(FT-mi)/se
windows()
#histograma da variável Faturamento Total padronizada
hist(FTpad,prob=T, main=" ", cex.lab=1.5, cex=1.5,ylab="Densidade", xlab="Faturamento
Total")
lines(density(FTpad),lty=2)

par(mfrow=c(1,2))
#boxplot das variáveis Faturamento Total e Produto M
boxplot(FT,xlab="Faturamento Total",cex.lab=1.5)
boxplot(PM,xlab="Produto M",cex.lab=1.5)
```

5.4. Sintaxe R – Análise de regressão

```
prodm=data[,2]
ft=data[,1]

PP.test(prodm)
PP.test(ft)

bancoTCC <- ("Dados") #entrar com seus dados
attach(bancoTCC)
require("MASS")
require("stats")
m1 <-  lm(FT~PM)
m2 <-  rlm(FT~PM,method="M")
m1
m2
cs=0.405
lmauto2=function(m)
{
data=("Dados") #entrar com seus dados
```

```
code1=0
sig=1
y=data[,1] #FT
x1=data[,2] #PM
n=dim(data)[1]
int=array(0,c(m,1))
se=array(0,c(m,1))
slope1=array(0,c(m,1))
des=array(1,c(n,1))
mr=array(0,c(n,1))
ma1=array(0,c(n,1))
x=x1
a=abs(x1)
wd=cbind(des,x)
w=solve(t(wd)%*%wd)
diff1=10
diff2=10
j=1
while(abs(diff1)>0.1|abs(diff2)>0.1)
{
j=j+1
lof=function(par)
{
stdev=sig
prob=array(0,c(n,1))
num=array(0,c(n,1))
sm=wd%*%solve(t(wd)%*%wd)%*%t(wd)
mf=diag(diag(sm))%*%des
for (i in 1:n)
{
prob[i]=par[1]+par[2]*x1[i]
ma1[i]=max(median(a),abs(x1[i]))
}
num=1-mf
losf=sum((1-1/cosh(cs*(1/stdev)*(y-prob)*num))/ma1)
return(losf)
}
result=optim(par=c(0.2,-0.02),fn=lof)
a=result$par[1]
b=result$par[2]
r=y-(a+b*x1) ## r is a vector of residuals
for (k in 1:n)
{
mr[k]=median(abs(r[k]-r))
}
sig=1.1926*median(mr)
code=result$convergence
code1=code1+code
int[j]=result$par[1]
slope1[j]=result$par[2]
se[j]=sig
diff1=int[j]-int[j-1]
diff2=slope1[j]-slope1[j-1]
```

```
#print(diff2)
}
out=array(0,c(3,1))
out[1]=int[j]
out[2]=slope1[j]
out[5]=se[j]
out[6]=median(abs(r))
return(out)
}
a=lmauto2(100)[1]
b=lmauto2(100)[2]
sighat=lmauto2(100)[5]
ma=lmauto2(100)[6]
ratcov2=function(cs){
data=read.csv2("C:\\BancoTCC.csv")
n=dim(data)[1]
y=data[,1] #FT
x1=data[,2] #PM
x=x1
des=array(1,c(n,1))
mr=array(1,c(n,1))
num0=array(0,c(n,1))
wd=cbind(des,x);
sm=wd%*%solve(t(wd)%*%wd)%*%t(wd)
mf=diag(diag(sm))%*%des
s1=array(0,c(n,1))
s2=array(0,c(n,1))
p=dim(data)[2]
num=1-mf
for (i in 1:n)
{
num0[i]=(1-(1/n))
}
##compute R-square
for (k in 1: n){
mr[k]=median(abs(y[k]-y))
}
rsq=1-(ma/median(mr))^2
## Estimate the variance covariance matrix
g1=function(t){
g=cs*1/cosh(cs*t)*tanh(cs*t) ##first derivative
return(g)
}
g2=function(t){
g=cs^2*(1/cosh(cs*t))^3-cs^2*(1/cosh(cs*t))*(tanh(cs*t))^2
return(g) ##second derivative
}
for(i in 1:n){
c1=(y[i]-(a+b*x1[i]))*(1-mf[i])/sighat
s1[i]=g1(c1)
s2[i]=g2(c1)
}
wt=solve(t(wd)%*%wd)
```

```
sq=(sum(s2))^2
variance=sum(s1^2)/((n-p)*sq)
##covariance matrix
cov=(sighat^2)*(n^2)*variance*wt
##Standard error
sea=sqrt(diag(cov))[1]
seb=sqrt(diag(cov))[2]
##compute t-value and Wald chi-square
t1=a/sea
chsq1=t1^2
t2=b/seb
chsq2=t2^2
##compute p-value
pvalue1=1-pchisq(chsq1,1)
pvalue2=1-pchisq(chsq2,1)
## Output parameter estimate, S.E., t-value, variance-covariance matrix,
## Wald chi-square, and p-value for each parameter
cat("a:", a,"\n");
cat("b:", b,"\n");
cat("SE (a):", sea,"\n");
cat("SE (b):", seb,"\n");
cat("t value (a):", a/sea,"\n");
cat("t value (b):", b/seb,"\n");
cat("Var-Cov matrix:", cov,"\n");
cat("Wald (a):", chsq1,"\n");
cat("Wald (b):", chsq2,"\n");
cat("p-value (a):", pvalue1,"\n");
cat("p-value (b):", pvalue2, "\n");
cat("Rsquare:", rsq,"\n");
cat("Sigma hat:", sighat,"\n");
}
ratcov2(cs)

#TELBS beta0=165.5284 e beta1=0.9280519
windows()
plot(PM, FT,pch=19,cex=1.5,ylab="Faturamento total",xlab="Produto M",xlim=c(-
120,2500),ylim=c(-75,2700),cex.lab=1.5,axes=F)
abline(m1,lwd=2,lty=3)
abline(m2,lwd=2,lty=2)
abline(a=165.5284,b=0.9280519,lwd=2,lty=1)
abline(h = 0,v = 0)
text(500,-75,cex=1.2,"500")
text(1000,-75,cex=1.2,"1000")
text(1500,-75,cex=1.2,"1500")
text(2000,-75,cex=1.2,"2000")
text(2500,-75,cex=1.2,"2500")
text(-90,500,cex=1.2,"500")
text(-120,1000,cex=1.2,"1000")
text(-120,1500,cex=1.2,"1500")
text(-120,2000,cex=1.2,"2000")
text(-120,2500,cex=1.2,"2500")

#calculando o intervalor de confianca para Beta0 e Beta1
```

```
seB0=27.82414
B0=165.5284
linfB0=B0-(1.96*seB0)
lsupB0=B0+(1.96*seB0)
linfB0
lsupB0

seB1=0.02875658
B1=0.9280519
linfB1=B1-(1.96*seB1)
lsupB1=B1+(1.96*seB1)
linfB1
lsupB1
```

5.5. Tabelas – Simulação de Tabatabai (2012)

Tabela 9: Vício e erro quadrático médio para amostras de tamanho $n = 20$ e $n = 100$, com níveis de 5%, 25% e 40% das observações simuladas contaminadas na variável x.

	5%		25%		40%	
	β_0	β_1	β_0	β_1	β_0	β_1
$n = 20$						
LS						
Vício	0,0062	3,0	0,0058	3,0	0,0089	3,0
EQM	0,5159	9,0	0,4225	9,0	0,3652	9,0
M						
Vício	0,0081	3	0,0097	3,0	0,0004	3,0
EQM	0,5611	8,8	0,4251	9,0	0,2827	9,0
MM						
Vício	0,0082	0,0008	0,0123	0,2185	0,0289	2,7
EQM	0,0593	0,0739	0,0927	0,7732	0,2965	8,2
TELBS						
Vício	0,0058	0,0031	0,0064	0,0024	0,0043	0,0099
EQM	0,0649	0,0725	0,0723	0,0983	0,0710	0,0995
$n = 100$						
LS						
Vício	0,0033	3,0	0,0010	3,0	0,0025	3,0
EQM	0,1066	9,0	0,0769	9,0	0,0645	9,0
M						
Vício	0,0205	3,0	0,0051	3,0	0,0004	3,0
EQM	0,0178	9,0	0,0769	9,0	0,0500	9,0
MM						
Vício	0,0024	0,0022	0,0003	0,0030	0,0063	2,9
EQM	0,0110	0,0109	0,0144	0,0143	0,0512	8,8
TELBS						
Vício	0,0011	0,0009	0,0042	0,0002	0,0007	0,0038
EQM	0,0111	0,0142	0,0126	0,0141	0,0172	0,0165

Abreviaturas: LS (Mínimos quadráticos); EQM (Erro quadrático médio); TELBS (Tabatabai, Eby, Li, Bae e Singh).
Fonte: Tabatabai et al. (2012)

Tabela 10: Vício e erro quadrático médio para amostras de tamanho $n = 20$ e $n = 100$, com níveis de 5%, 25% e 40% das observações simuladas contaminadas nas variáveis x e y.

	5%		25%		40%	
	β_0	β_1	β_0	β_1	β_0	β_1
$n = 20$						
LS						
Vício	0,2076	1,15	212,1	0,0091	363,2	0,0125
EQM	45,9	382,2	58563	0,6532	152871	0,2283
M						
Vício	0,0647	0,2297	0,5398	0,0174	74,1	0,0398
EQM	8,0	111,9	0,6401	1,07	22616	0,4646
MM						
Vício	0,0092	0,0181	0,0115	0,0215	0,0025	0,0121
EQM	0,0637	0,1116	0,0831	0,1997	0,1082	0,2319
TELBS						
Vício	0,0016	0,0002	0,0173	0,0005	0,0052	0,0089
EQM	0,0553	0,0760	0,0756	0,0821	0,0805	0,1065
$n = 100$						
LS						
Vício	39,9	0,0231	240,4	0,0027	393,5	0,0083
EQM	2096	0,6392	60462	0,0761	159,2	0,0372
M						
Vício	0,0560	0,0202	0,5216	0,0039	35,1	0,0017
EQM	0,0309	1,11	0,3135	0,1665	5043	0,0874
MM						
Vício	0,0020	0,0042	0,0002	0,0026	0,0012	0,0004
EQM	0,0116	0,0231	0,0144	0,0351	0,0167	0,0360
TELBS						
Vício	0,0034	0,0019	0,0021	0,0015	0,0010	0,0001
EQM	0,0125	0,0130	0,0140	0,0144	0,0150	0,0170

Abreviaturas: LS (Mínimos quadráticos); EQM (Erro quadrático médio); TELBS (Tabatabai, Eby, Li, Bae e Singh).
Fonte: Tabatabai et al. (2012)

Tabela 11: Vício e erro quadrático médio para amostras de tamanho $n = 20$ e $n = 100$, com níveis de 5%, 25% e 40% das observações simuladas contaminadas na variável x_1 e x_2.

	5%			25%			40%		
	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
$n = 20$									
LS									
Vício	0,0172	1,5	0,4673	0,0476	3,0	0,5000	0,0445	3,0	0,5000
EQM	0,5778	3,5	1,5548	0,6964	9,0	0,2500	0,6501	9,0	0,2500
M									
Vício	0,0093	1,5	0,5099	0,0168	3,0	0,4999	0,0213	3,0	0,4999
EQM	0,6182	3,6	1,6077	0,7245	9,0	0,2499	0,5762	9,0	0,2499
MM									
Vício	0,0088	0,0208	0,5131	0,0135	0,1752	0,5144	0,0691	2,3	0,4878
EQM	0,3175	0,0832	0,3668	0,3678	0,5797	0,4602	0,5557	7,1	0,2887
TELBS									
Vício	0,0312	0,0020	0,0048	0,0045	0,0021	0,0037	0,0047	0,0093	0,0007
EQM	0,0671	0,0816	0,0839	0,0774	0,0870	0,0907	0,0965	0,1071	0,1196
$n = 100$									
LS									
Vício	0,0212	3,0	0,4999	0,0196	3,0	0,5000	0,0274	3,0	0,5000
EQM	0,3444	9,0	0,2499	0,3274	9,0	0,2500	0,3211	9,0	0,2500
M									
Vício	0,0075	3,0	0,5000	0,0064	3,0	0,5000	0,0218	3,0	0,5000
EQM	0,3553	9,0	0,2500	0,3529	9,0	0,2500	0,2961	9,0	0,2500
MM									
Vício	0,0007	0,0057	0,4986	0,0264	0,0179	0,4972	0,0275	3,0	0,5004
EQM	0,2588	0,0117	0,2686	0,2613	0,0150	0,2802	0,2999	8,6	0,2522
TELBS									
Vício	0,0009	0,0017	0,0033	0,0033	0,0039	0,0007	0,0026	0,0003	0,0053
EQM	0,0121	0,0122	0,0123	0,0151	0,0156	0,0141	0,0160	0,0164	0,0177

Abreviaturas: LS (Mínimos quadráticos); EQM (Erro quadrático médio); TELBS (Tabatabai, Eby, Li, Bae e Singh).

Fonte: Tabatabai et al. (2012)

Tabela 12: Vício e erro quadrático médio para amostras de tamanho $n = 20$ e $n = 100$, com níveis de 5%, 25% e 40% das observações simuladas contaminadas na variável y.

	5%			25%			40%		
	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
$n = 20$									
LS									
Vício	47,5	150,7	7,97	237,2	757,3	15,7	423,9	1231	133813
EQM	27539	7526	3515	174783	763935	136845	370305	1801726	178241
M									
Vício	0,0221	0,0661	0,5003	27,9	136,7	15,1	304,3	942,7	23,6
EQM	0,3266	0,0848	0,3317	25830	146578	36128	271498	1325407	146334
MM									
Vício	0,0035	0,0085	0,5037	0,0157	0,0201	0,5174	0,0036	0,0066	0,5025
EQM	0,3248	0,0726	0,3193	0,3140	0,0938	0,3632	0,3781	0,1139	0,4065
TELBS									
Vício	0,0080	0,0011	0,0138	0,0014	0,0000	0,0172	0,0014	0,0000	0,0172
EQM	0,0688	0,0939	0,0850	0,0800	0,1001	0,1114	0,0800	0,1005	0,1114
$n = 100$									
LS									
Vício	48,2	147,7	0,3145	245,2	746,1	6,78	405,2	1179,1	7,97
EQM	7693	30459	5103	96704	594204	22831	227561	1440607	30060
M									
Vício	0,0489	0,0630	0,5002	0,2612	0,7666	0,5475	236,1	725,7	1,49
EQM	0,2889	0,0158	0,2633	0,5174	2,98	1,73	100135	654402	18714
MM									
Vício	0,0056	0,0026	0,5005	0,0284	0,0014	0,5066	0,0041	0,0033	0,5060
EQM	0,2736	0,0122	0,2628	0,2756	0,0155	0,2709	0,2572	0,0153	0,2731
TELBS									
Vício	0,0019	0,0053	0,0021	0,0031	0,0016	0,0022	0,0019	0,0014	0,0014
EQM	0,0126	0,0134	0,0132	0,0157	0,0157	0,0159	0,0126	0,0150	0,0134

Abreviaturas: LS (Mínimos quadráticos); EQM (Erro quadrático médio); TELBS (Tabatabai, Eby, Li, Bae e Singh).
Fonte: Tabatabai et al. (2012)

Tabela 13: Vício e erro quadrático médio para amostras de tamanho $n = 20$ e $n = 100$, com níveis de 5%, 25% e 40% das observações simuladas contaminadas nas variáveis x_1, x_2 e y.

	5%			25%			40%		
	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
$n = 20$									
LS									
Vício	0,0647	0,1164	0,5577	158,0	0,0587	0,5099	328,3	0,0166	0,5061
EQM	0,5506	5,37	3,87	46225	0,8650	1,09	159787	0,3129	0,5967
M									
Vício	0,0066	0,0507	0,4944	0,5344	0,0284	0,5113	48,2	0,0161	0,5265
EQM	0,6885	2,4124	2,8125	1,4986	1,7021	1,8582	18968	0,6473	0,9399
MM									
Vício	0,0018	0,0014	0,4827	0,0364	0,0182	0,5090	0,0135	0,0108	0,4768
EQM	0,3186	0,1018	0,3426	0,3503	0,2253	0,4639	0,3509	0,2846	O,4742
TELBS									
Vício	0,0033	0,0053	0,0164	0,0139	0,0120	0,0111	0,0217	0,0058	0,0026
EQM	0,0688	0,0745	0,0758	0,0758	0,0889	0,0834	0,0914	0,1086	0,1086
$n = 100$									
LS									
Vício	30,1	0,0031	0,4799	240,4	0,0108	0,4982	391,1	0,0014	0,5022
EQM	1640	0,9444	1,32	73844	0,0782	0,3289	193012	0,0416	0,2949
M									
Vício	0,0465	0,0088	0,5183	0,4865	0,0243	0,4845	71,7	2,66	0,4951
EQM	0,3300	1,60	1,73	0,8035	0,1945	0,4367	12853	7,2	0,2693
MM									
Vício	0,0168	0,0002	0,5069	0,0018	0,0029	0,4954	0,0035	0,0014	0,4979
EQM	0,2606	0,0174	0,2755	0,2522	0,0352	0,2800	0,2711	0,0452	0,2935
TELBS									
Vício	0,0024	0,0075	0,0054	0,0033	0,0036	0,0057	0,0064	0,0001	0,0009
EQM	0,0122	0,0124	0,0119	0,1391	0,0155	0,0144	0,0157	0,0162	0,0169

Abreviaturas: LS (Mínimos quadráticos); EQM (Erro quadrático médio); TELBS (Tabatabai, Eby, Li, Bae e Singh).

Fonte: Tabatabai et al. (2012)

Referências Bibliográficas

Tabatabai, M. A.; Eby, W. M.; Li, H.; Bae, S. e Singh, K. P. **TELBS robust linear regression method**. Open Access Medical Statistics 2: 65-84, 2012.

Montgomery, D. C.; Peck, E. A. e Vining, G. G. **Introduction to Linear Regression Analysis**. Quinta edição. Wiley, 2012.

Neter, J.; Kutner, M.H.; Nachtsheim, C.J. e Li, W. **Applied Linear Statistical Models**. Quinta edição. Irwin: McGraw-Hill, 2005.

Rousseeuw P.J. e Leroy, A.M. **Robust Regression and Outlier Detection**. NY: Wiley Interscience; 1987.

Rousseuw P.J. e Croux C. **Alternatives to the median absolute deviation**. Journal of the American Statistical Association 88:1273-1283, 1993.

Huber, P.J. **Robust regression: asymptotics, conjectures, and Monte Carlo**. Annals of Statistics 1:799–821, 1973

Ronchetti, E. **Robust model selection in regression**. Statistics & Probability Letters 3:21–23, 1985.

Maronna, R.A.; Martin, R.D. e Yohai, V.J. **Robust Statistics: Theory and Methods**. NY: Wiley; 2006.

Hampel, F.R.; Ronchetti, E.M.; Rousseeuw, P.J. e Stahel, W.A. **Robust Statistics: The Approach Based on Influence Functions**. NY: Wiley; 1986.

Rousseeuw, P.J. e Yohai, V.J. **Robust regression by means of S estimators**. In: Franke J, Härdle W, Martin RD, editores. Robust and Non-Linear Time Series Analysis. NY: Springer-Verlag; 1984.

Muthukrishnan, R. e Radha, M. **M-Estimators in Regression Models**. Journal of Mathematics Research 2:23-27, 2010.

Yohai, V.J. **High breakdown point and high efficiency**. Annals of Statistics 15: 642–656, 1987.

Akaike, H. **A new look at the statistical model identification**. IEEE Transactions on Automatic Control 19:716–723, 1974.

Bulhões, R. S. e Lima, C. M. **Comparação de Estimadores de Regressão**. Departamento de Estatística Universidade Federal da Bahia. Salvador, BA. 2007.

Alves, G. I. L. e Lima, V. M. C. **Comparação entre medidas clássicas e robustas para identificação de outliers em regressão**. Universidade Federal da Bahia. Salvador, BA, 2007.

Bueno, R. de L. da S. **Econometria de Séries Temporais**. CENGAGE Learning Edições Ltda, São Paulo, 2008.

Phillips, P. C. B. e Perron, P. **Testing for a Unit Root in Time Series Regression**. Biometrika 75 (2): 335–346, 1988.

R. **The R Project for Statistical Computing**. Disponível em www.r-project.org. Acesso em: 02 novembro 2015.