UNIVERSIDADE DE BRASÍLIA
DEPARTAMENTO DE ESTATÍSTICA

LAÍZA MENDES JAIME DE REZENDE E BRITO

INTRODUÇÃO À REGRESSÃO LINEAR ROBUSTA

Brasília
2023

LAÍZA MENDES JAIME DE REZENDE E BRITO

INTRODUÇÃO À REGRESSÃO LINEAR ROBUSTA

Relatório Final apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.
Orientadora: Profa. Terezinha Késsia de Assis Ribeiro.

Brasília
2023

Resumo

Os modelos de regressão linear são vastamente utilizados para modelagem de dados reais em diversas áreas do conhecimento. A inferência para os parâmetros destes modelos é baseada no clássico método de mínimos quadrados. Entretanto, este método é conhecido por sua sensibilidade a observações discrepantes, o que pode conduzir a conclusões equivocadas sobre as características de interesse. Nessa perspectiva, os métodos de estimação robustos surgem como alternativa para lidar com esses dados atípicos, atribuindo um peso menor para esses pontos no procedimento de estimação. Portanto, neste trabalho, realizamos uma revisão sobre regressão linear e sobre o método inferencial de mínimos quadrados. Para realizar inferência robusta, realizamos um estudo introdutório sobre a classe de M-estimadores e como esta pode ser aplicada sob o contexto de regressão linear. Por fim, para ilustrar a aplicabilidade dos métodos de estimação estudados apresentamos duas aplicações a dados reais.

Palavras-chaves: Regressão linear, Mínimos quadrados, M-estimadores, Robustez.

Lista de Tabelas

Lista de Figuras

Sumário

1 Introdução

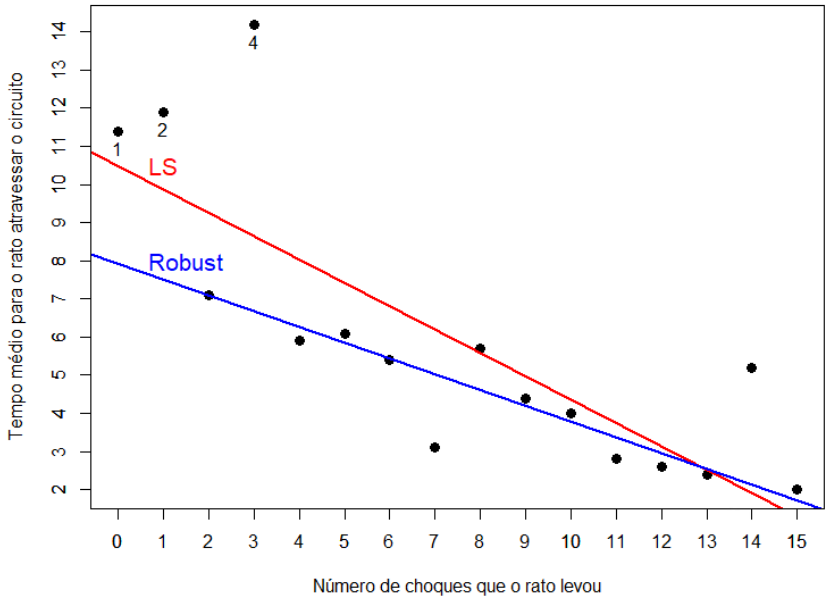
A regressão linear é uma abordagem vastamente utilizada para auxiliar no entendimento de como as variáveis se relacionam em diversos contextos. Por exemplo, pode-se querer entender como as variáveis idade e escolaridade de indivíduos de uma população afetam a respectiva renda mensal média. No caso mais simples, há interesse no comportamento médio de uma variável y, chamada de resposta, em função de uma única variável x, explicativa, usualmente denominada de covariável.

Para uso desta técnica estatística é necessário estimar os coeficientes da regressão que são parâmetros desconhecidos. Considerando uma amostra da populacão, utiliza-se o método de mínimos quadrados (LSM, *Least Squares Method*). Esse método baseia-se na minimização da soma de quadrados dos desvios da regressão (DRAPER; SMITH, 1998; WEISBERG, 2005). O estimador de mínimos quadrados (LSE, *Least Squares Estimator*) possui ótimas propriedades, como não-viciosidade, consistência e normalidade sob dados comportados. Entretanto, caso a amostra contenha observações atípicas, isto é, dados com comportamento distinto da maioria, o LSE possui viés e pode conduzir a conclusões errôneas sobre as relações reais existentes entre as variáveis em estudo. Ademais, estimadores com esse comportamento são chamados de estimadores sensíveis ou não robustos.

Sob essa perspectiva, na literatura, existem diversas propostas para lidar com observações discrepantes sob o contexto de regressão linear. Uma solução natural é utilizar um método de estimação robusto ao invés de mínimos quadrados. Em outras palavras, mantém-se a abordagem simples, que é a regressão linear para estudar o fenômeno aleatório de interesse, no entanto, troca-se o procedimento inferencial para atribuir valores aos coeficientes da regressão. Nesse sentido, diversos estimadores robustos foram propostos sob o contexto de regressão linear, entre os mais utilizados estão os M-estimadores (HUBER, 1973; HUBER, 1981), S-estimadores (ROUSSEEUW; YOHAI, 1984), e MM-estimadores (YOHAI, 1987).

Para ilustrar a ideia de estimação robusta, considere um conjunto de dados discutido previamente em Maronna et al. (2019, p. 87). A Figura 1 é referente aos dados sobre o número de choques que ratos levaram em um experimento no decorrer de tentativas de atravessar uma caixa, tentativas essas que tiveram o tempo cronometrado. No gráfico de dispersão da Figura 1 tem-se o número de choques e o tempo médio de cada rato para atravessar, juntamente com a reta de regressão linear ajustada pelo LSM (referenciado como "LS") e por um método de estimação robusto (referenciado como "*Robust*"). Como observado por Maronna et al. (2019), nota-se que existem três observações atípicas que influenciam consideravelmente o ajuste da reta de regressão via o LSM. A reta "LS" possui uma inclinação que não conduz um bom ajuste aos dados, enquanto que, a reta "*Robust*" apresenta um melhor ajuste a maior parte dos dados.

Figura 1: Gráfico de dispersão do tempo médio para travessia versus número de choques levados pelo rato juntamente com retas de regressão ajustadas via LSM e robusto.



O método de estimação robusto mais conhecido e amplamente utilizado em modelos de regressão gerais é o de M-estimação proposto inicialmente por Huber (1964). Este autor propôs um novo método de estimação baseado na minimização de uma generalização do logaritmo da função de verossimilhança para modelos de localização. Os M-estimadores, além de serem robustos a observações atípicas, possuem ótimas propriedades, como consistência e normalidade assintótica.

Diante do exposto, o presente Relatório Final, que está organizado em quatro capítulos, tem por objetivo principal apresentar um estudo sobre alguns estimadores robustos sob modelos de regressão linear, bem como observar as vantagens do uso desses tipos de estimadores em comparação ao LSE, que não é robusto. No Capítulo 2 é apresentada uma fundamentação teórica sobre o modelo de regressão linear e os métodos de estimação dos parâmetros que serão utilizados. Na Seção 2.1 é feita uma revisão sobre modelos de regressão linear e na Seção 2.2 uma revisão do método de mínimos quadrados. Por conseguinte, na Seção 2.3, são definidos os métodos de estimação robustos que serão utilizados, na qual será introduzida a classe dos M-estimadores sob modelos de localização (Subseção 2.3.1), sob modelos de escala (Subseção 2.3.2), sob modelos de localização e escala (Subseção 2.3.3), e uma generalização sob modelos paramétricos gerais (Subseção 2.3.4). Além disso, será definido o conceito de função de influência (Subseção 2.3.5) que é uma medida importante utilizada para classificação de estimadores robustos. Para finalizar a fundamentação teórica deste trabalho, na Seção 2.4, os estimadores robustos serão apresentados sob o contexto de regressão linear. No Capítulo 3, o LSE é comparado com dois estimadores robustos sob regressão linear por meio de duas aplicações a dados reais. Por fim, no Capítulo 4 são apresentadas as considerações finais deste trabalho.

2 Fundamentação Teórica

Para uma melhor compreensão do que foi estudado ao longo do desenvolvimento deste trabalho, a seguir será feita uma revisão sobre alguns pontos importantes que devem ser entendidos previamente. Dessa forma, busca-se apresentar o processo de modelagem usual em regressão linear, a fim de definir o modelo de regressão linear múltiplo e os diferentes métodos de estimação, clássico e robusto, para os parâmetros de interesse.

2.1 Modelo de regressão linear múltiplo

2.1.1 Definição

Como apontado na introdução desse relatório, a regressão linear é utilizada para compreender a relação entre variáveis e, principalmente, a forma com que uma ou mais variáveis explicativas $x_1, x_2, ..., x_k$ influenciam na média da variável resposta y. Dessa maneira, um modelo de regressão linear trata-se de uma representação matemática que descreve a relação estatística entre duas ou mais variáveis, na qual supõe-se uma distribuição de probabilidade gaussiana para a resposta y dados os valores fixados para $x_1, x_2, ..., x_k$ (MICHAEL et al., 2004), entre outras suposições adotadas que serão abordadas mais adiante.

O modelo que descreve o comportamento da variável resposta y é composto por duas partes. A primeira parte é determinística, na qual encontram-se os parâmetros desconhecidos e as variáveis explicativas que contém informações relacionadas com o fenômeno estudado. Enquanto a segunda parte é aleatória, composta por um erro aleatório que contém os fatores desconhecidos que afetam a resposta y, e que se supõe que obedeça a alguma distribuição de probabilidades.

Como o modelo de regressão linear simples é um caso particular do modelo de regressão linear múltiplo, definiremos o modelo partindo deste segundo. Assim, o modelo de regressão linear múltiplo amostral segue o seguinte formato:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \ , \ i \in \{1, 2, ..., n\}, \tag{2.1.1}$$

em que

- y_i é uma variável aleatória observável conhecida como resposta ou dependente;
- $x_{i1}, x_{i2}, ..., x_{ik}$ são os valores fixados das k variáveis não aleatórias para a i-ésima observação, chamadas de explicativas ou independentes, em que se $k = 1$, tem-se um modelo de regressão linear simples, e se $k \geq 2$, então, o modelo é múltiplo;

- $\beta_1, \beta_2, ..., \beta_k$ são os k parâmetros desconhecidos, chamados de coeficientes da regressão linear para cada uma das k variáveis explicativas, e β_0 é o parâmetro chamado de intercepto da reta de regressão;
- ε_i é uma variável aleatória não observável denominada como erro da regressão.

Os modelos de regressão podem ser classificados como simples, contendo apenas uma variável explicativa, ou como múltiplos, considerando mais de uma variável explicativa. Além disso, podem ser lineares nos parâmetros ou não-lineares nos parâmetros. Com isso, por exemplo, se um modelo é definido como $y_i = \beta_0 + \sqrt{\beta_1} x_i + \varepsilon_i$, ele é não é linear, entretanto, um modelo indicado como $y_i = \beta_0 + \beta_1 x_i^3 + \varepsilon_i$ é linear.

2.1.2 Notação matricial do modelo

A forma matricial do modelo facilita algumas manipulações matemáticas, como, por exemplo, a estimação dos parâmetros via o método de mínimos quadrados. Sendo assim, tem-se, em termos matriciais:

$$Y = X\beta + \varepsilon, \tag{2.1.2}$$

em que

- Y é um vetor não aleatório n-dimensional, no qual tem-se as n variáveis respostas;
- X é uma matriz não aleatória com n linhas e $(k+1)$ colunas, em que a primeira coluna é toda composta por 1's, pois está associada ao termo constante β_0, e as demais colunas contém os n valores fixados para cada uma das k variáveis explicativas;
- β é um vetor $(k+1)$-dimensional que contém os $(k+1)$ coeficientes desconhecidos da regressão, no qual o primeiro é o intercepto e os demais são os coeficientes das variáveis explicativas;
- ε é um vetor aleatório n-dimensional de erros da regressão.

Dessa forma, os elementos matriciais descritos acima são representados por

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}, \; X = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & x_{23} & \cdots & x_{2k} \\ 1 & x_{31} & x_{32} & x_{33} & \cdots & x_{3k} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} & \cdots & x_{nk} \end{bmatrix}, \; \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}, \; \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

2.1.3 Suposições adotadas para o modelo de regressão linear

Para que o processo inferencial dos parâmetros desconhecidos β seja possível, a partir de estimadores com ótimas propriedades, será necessário fazer algumas suposições:

S_1. A relação entre as covariáveis $x_1, x_2, \ldots, x_k$ e a variável dependente y deve ser linear, ou seja, a relação entre ambas pode ser descrita corretamente por uma equação linear nos parâmetros, logo, o modelo deve ser linear e estar corretamente especificado;

S_2. A média dos erros aleatórios é igual a zero, ou seja, $\mathrm{E}(\varepsilon_i) = 0,\ \forall\ i \in \{1, 2, ..., n\}$ ou $\mathrm{E}(\varepsilon) = 0$. Ademais, devido a essa suposição, vale que

$$\mathrm{E}(Y) = \mathrm{E}(X\beta + \varepsilon) = \mathrm{E}(X\beta) + \mathrm{E}(\varepsilon) = X\beta + 0 = X\beta.$$

Portanto, ao analisar o comportamento da variável Y, implicitamente, modela-se a média da variável resposta;

S_3. Os erros aleatórios associados às observações diferentes não podem ser correlacionados entre si, ou seja, $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \mathrm{E}(\varepsilon_i\varepsilon_j) - \mathrm{E}(\varepsilon_i)\mathrm{E}(\varepsilon_j) = 0,\ \forall\ i \neq j$;

S_4. A variância dos erros aleatórios deve ser constante (a mesma) para todos os erros, ou seja, $\mathrm{Var}(\varepsilon_i) = \mathrm{E}(\varepsilon_i^2) - [\mathrm{E}(\varepsilon_i)]^2 = \sigma^2,\ \forall\ i \in \{1, 2, ..., n\}$. Dessa forma, supõe-se que o modelo é homoscedástico. Matricialmente, a matriz de covariâncias de Y é obtida de

$$\mathrm{Var}(Y) = \mathrm{Var}(X\beta + \varepsilon) = \mathrm{Var}(\varepsilon) = \sigma^2 I_n,$$

em que I_n é a matriz identidade de ordem n, ou seja, é uma matriz com a diagonal principal composta por uns e os demais valores iguais a zero;

S_5. A matrix X deve ser de posto completo, ou seja, as colunas de X devem ser linearmente independentes. Para tanto, basta ter que $r(X) = k + 1$, em que $r(X)$ denota o posto da matrix X;

S_6. Os erros da regressão, considerados independentes e identicamente distribuídos($i.i.d.$), devem seguir uma distribuição normal, logo, $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$. Ou ainda, matricialmente, para uma distribuição normal multivariada de dimensão n, tem-se $\varepsilon \overset{i.i.d.}{\sim} N_n(0, \sigma^2 I_n)$, sendo I_n uma matriz identidade de dimensão $n \times n$. Com isso, partindo de que $Y \sim N_n(X\beta, \sigma^2 I_n)$, obtém-se $y_i \sim N(\mu_i, \sigma^2),\ \forall\ i \in \{1, 2, \ldots, n\}$, ou seja, apesar de também ter distribuição normal, as variáveis y_i têm suas médias variando de acordo com as observações. Ainda, sob normalidade e S_3, tem-se independência entre $y_1, y_2, \ldots, y_n$. Vale salientar que esta última suposição sobre os erros ε é adicional, não sendo necessariamente obrigatória para estimar pontualmente β.

2.2 Método de estimação por mínimos quadrados

"Many methods have been suggested for obtaining estimates of parameters in a model. The method discussed here is called ordinary least squares, or ols, in which parameter estimates are chosen to minimize a quantity called the residual sum of squares." (WEISBERG, 2005, p. 21).

O LSM é uma das técnicas mais utilizadas para ajustar modelos de regressão linear. Diante disso, como evidenciado na citação acima de Weisberg (2005, p. 21), o objetivo desse método é encontrar os valores dos parâmetros que conjuntamente minimizam a soma dos quadrados dos desvios, os quais são as diferenças entre os valores observados de y_i e os valores esperados pelo modelo de regressão linear. Com isso, o modelo que melhor se ajustará aos dados será o que apresentar os menores desvios possíveis.

Nesse sentido, o LSM consiste em obter o valor de $\beta = (\beta_0 \ \beta_1 \cdots \beta_k)^\top$ que minimiza a soma

$$S(\beta) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - E(y_i))^2,$$

em que $\mathrm{E}(y_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$, com $S(\beta)$ denotando a soma dos quadrados dos desvios. É notório que $S(\beta)$ não depende da variância σ^2, logo, o estimador para σ^2 não será obtido de forma direta por meio deste método.

Reescrevendo $S(\beta)$ sob notação matricial, tem-se

$$S(\beta) = \varepsilon^\top \varepsilon = (Y - X\beta)^\top (Y - X\beta) = Y^\top Y - 2Y^\top X\beta + \beta^\top X^\top X\beta.$$

Para encontrar o candidato a mínimo global da função $S(\beta)$, considere

$$\frac{\partial}{\partial \beta} \left[S(\beta)\right] = -2X^\top Y + 2X^\top X\beta.$$

Igualando a expressão anterior da derivada ao vetor de zeros para encontrar o candidato a ponto de mínimo global, $\widehat{\beta}$, e isolando-se $\widehat{\beta}$, tem-se que

$$X^\top X\widehat{\beta} = X^\top Y \quad \Longrightarrow \quad \widehat{\beta} = (X^\top X)^{-1} X^\top Y. \tag{2.2.1}$$

Este resultado é válido devido a suposição S_5, que garante posto completo da matrix X, e portanto, garante que $(X^\top X)^{-1}$ sempre exista desde que $r(X^\top X) = r(X) = k+1$ ($X^\top X$) é de posto completo.

Diante do exposto, a expressão (2.2.1) define $\widehat{\beta}$ como estimador de β via o LSM. Além disso, garante-se que $\widehat{\beta}$ é de fato um ponto de mínimo global pois tem-se que $\partial^2[S(\beta)]/\partial\beta\partial\beta^\top = 2X^\top X$ que é uma matriz positiva pois $r(X) = r(X^\top X)$, e isto implicará que $X^\top X$ é positiva definida. Logo, $\widehat{\beta}$ é o valor de β que minimiza $S(\beta)$, e este é denominado por estimador de mínimos quadrados para β. Com isso, pode-se encontrar o valor $\widehat{\beta}$ substituindo X e Y definidos em (2.1.2). A fórmula que resume os resultados para os parâmetros estimados pelo LSM, no caso em que se tem apenas β_0 e β_1, ou seja, sob regressão linear simples, é

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\,\overline{x}\,\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \quad \text{e} \quad \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1\overline{x},$$

em que $\widehat{\beta}_0$ e $\widehat{\beta}_1$ são o intercepto e o coeficiente angular, respectivamente, da reta de regressão ajustada (CHARNET et al., 1999).

Das suposições de S_1 a S_6, segue que $\widehat{\beta}$ possui as seguintes propriedades (CHARNET et al., 1999):

P_1. $\widehat{\beta}$ é linear em Y, pois $\widehat{\beta} = AY$ com $A = (X^\top X)^{-1}X^\top$;

P_2. $\widehat{\beta}$ é um estimador não viciado para β, tendo em vista que

$$\mathrm{E}(\widehat{\beta}) = \mathrm{E}\left[(X^\top X)^{-1}X^\top Y\right] = (X^\top X)^{-1}X^\top \mathrm{E}(Y) = (X^\top X)^{-1}X^\top X\beta = I_{k+1}\beta = \beta;$$

P_3. Dentre todos os estimadores lineares e não viciados, $\widehat{\beta}$ é o estimador de β que possui a menor variância, pelo o Teorema de Gauss-Markov, e essa variância é expressa por

$$\begin{aligned}\mathrm{Var}(\widehat{\beta}) &= \mathrm{Var}\left[(X^\top X)^{-1}X^\top Y\right] \\ &= (X^\top X)^{-1}X^\top\ \mathrm{Var}(Y)\ \left[(X^\top X)^{-1}X^\top\right]^\top \\ &= (X^\top X)^{-1}X^\top\ \sigma^2 I_n\ X(X^\top X)^{-1} \\ &= \sigma^2(X^\top X)^{-1}\ ;\end{aligned}$$

P_4. Sob a suposição S_6, segue que $\widehat{\beta} \sim N_{k+1}(\beta, \sigma^2(X^\top X)^{-1})$, em que $N_{k+1}(\mu, \Sigma)$ denota a distribuição normal $(k+1)$-variada de média μ e matriz de covariâncias Σ. Neste caso, $\widehat{\beta}$ é o estimador não viciado de variância uniformemente mínima (ENVVUM) de β;

P_5. $\widehat{\beta}_j \sim N(\beta_j,\ \sigma^2\ a_{(j+1)(j+1)}\ )$, em que $a_{(j+1)(j+1)}$ é o $(j+1)$-ésimo elemento da diagonal principal da matriz $(X^\top X)^{-1}$, para $j = 0, 1, 2, \ldots, k$;

P_6. $\mathrm{Cov}(\widehat{\beta}_j,\ \widehat{\beta}_l) = \sigma^2\ a_{(j+1)(l+1)}$, em que $j, l = 0, 1, 2, \ldots, k$, e $a_{(j+1)(l+1)}$ é o elemento $(j+1, l+1)$ da matriz $(X^\top X)^{-1}$. Além disso, segue da propriedade P_4 que $\widehat{\beta}_j$ e $\widehat{\beta}_l$ são independentes se $a_{(j+1)(l+1)} = 0$.

P_7. O estimador $\widehat{\beta}$ obtido pelo LSM coincide com o estimador $\widehat{\beta}_{ML}$ obtido pelo método de máxima verossimilhança (esse método será discutido na Subseção 2.3). Entretanto, a variância σ^2 estimada pelo LSM denotada por $\widehat{\sigma}^2$ não coincide com $\widehat{\sigma}^2_{ML}$ obtido pelo método da máxima verossimilhança. Na verdade, tem-se que

$$\widehat{\sigma}^2 = \frac{n}{n-k-1}\widehat{\sigma}^2_{ML}$$

em que

$$\widehat{\sigma}^2_{ML} = \frac{(Y - X\widehat{\beta})^\top (Y - X\widehat{\beta})}{n}.$$

Além da validação das suposições necessárias para o modelo ajustado, é possível avaliar se os coeficientes de regressão são diferentes de zero. Para tal, são feitos testes de hipóteses e intervalos de confiança para os coeficientes β_j, $j = 0, 1, 2, \ldots, k$, do modelo. Se não há evidências contra $\beta_j = 0$, isso implicaria que a j-ésima variável explicativa não tem relação com a variável resposta. De forma geral, considere as hipóteses

$$\begin{cases} H_0 : \beta_j = b_j; \\ H_1 : \beta_j \neq b_j. \end{cases}$$

Para testar a hipótese nula H_0 contra a hipótese de teste H_1, utiliza-se a seguinte estatística de teste, que, sob H_0, segue uma distribuição *t-student* com $(n-k-1)$ graus de liberdade,

$$t_0 = \frac{\widehat{\beta}_j - b_j}{\sqrt{\widehat{\sigma}^2\ a_{(j+1)(j+1)}}} \overset{H_0}{\sim}\ t_{(n-k-1)}, \tag{2.2.2}$$

em que $\widehat{\beta}_j$ é a estimativa de β_j pelo LSM, $\sqrt{\widehat{\sigma}^2\ a_{(j+1)(j+1)}}$ é o erro padrão estimado de $\widehat{\beta}_j$, e b_j é uma constante real. Se $b_j = 0$, estaremos testando a importância da j-ésima variável explicativa, $\forall\ j \in \{1, 2, ..., k\}$. Nesse sentido, rejeita-se H_0 ao nível de significância de $\alpha \cdot 100\%$, com $0 < \alpha < 1$, se $t_0 \leq -t_c$ ou $t_0 \geq t_c$, com t_c obtido tal que $P(t_{n-k-1} \geq t_c) = \frac{\alpha}{2}$. Ainda, em termos do p-valor, rejeita-se H_0 se p-valor $= 2P(t_{n-k-1} \geq |t_0|) \leq \alpha$.

Pode-se obter um estimador intervalar o qual conterá o valor verdadeiro do coeficiente β_j, com uma certa probabilidade fixada, $\alpha \cdot 100\%$. Este intervalo de confiança pode ser obtido pelo cálculo

$$\widehat{\beta}_j \pm t_c \cdot \sqrt{\widehat{\sigma}^2\ a_{(j+1)(j+1)}}\ , \tag{2.2.3}$$

em que t_c é o valor crítico da distribuição *t-student* com $(n-k-1)$ graus de liberdade para o nível de confiança $(1-\alpha) \cdot 100\%$ desejado.

Outro conceito importante para avaliação da qualidade do ajuste de um modelo de regressão é o resíduo. Um resíduo é uma medida calculada sob um modelo de regressão ajustado, e sua principal função é indicar se o modelo suposto para os dados está bem ajustado. Sendo assim, sob um modelo bem ajustado aos dados, espera-se que os resíduos assumam valores próximos de zero. Ainda, deseja-se que o resíduo tenha comportamento próximo ao erro aleatório, pois é sobre esse erro que são feitas diversas suposições, entretanto o erro não é observável.

O resíduo mais simples é o resíduo ordinário, denotado por $\widehat{\varepsilon_i}$, em que estes são as diferenças entre os valores observados da variável dependente e os valores ajustados pela reta de regressão, ou seja, $\widehat{\varepsilon_i} = y_i - \widehat{y_i}$. Matricialmente, denota-se por

$$\widehat{\varepsilon} = Y - \widehat{Y} = Y - X\widehat{\beta} = Y - X(X^\top X)^{-1}X^\top Y = (I_n - H)Y, \tag{2.2.4}$$

em que $H = X(X^\top X)^{-1}X^\top$, denota a matriz de projeção gerada pelo espaço de colunas de X, isto é, pelas combinações lineares das colunas de X, obtendo $HX = X$.

Por conseguinte, tem-se que

$$\begin{aligned}
\mathrm{E}(\widehat{\varepsilon}) &= (I_n - H)\mathrm{E}(Y) \\
&= (I_n - X(X^\top X)^{-1}X^\top)X\beta \\
&= (X - X(X^\top X)^{-1}X^\top X)\beta \\
&= 0.
\end{aligned}$$

Além disso, segue que

$$\begin{aligned}
\mathrm{Var}(\widehat{\varepsilon}) &= (I_n - H)\mathrm{Var}(Y)(I_n - H)^\top \\
&= (I_n - H)\sigma^2 I_n(I_n - H) \\
&= \sigma^2(I_n - H)(I_n - H) \\
&= \sigma^2(I_n - H).
\end{aligned}$$

Isso porque têm-se que $(I_n - H)(I_n - H) = (I_n - H)$, já que H é uma matriz idempotente, o que resulta dizer que a matriz $(I_n - H)$ também é idempotente. Assim, $\mathrm{Var}(\widehat{\varepsilon_i}) = \sigma^2(1 - h_{ii})$ com $i = 1, \ldots, n$ e h_{ii} sendo o i-ésimo elemento da diagonal principal de H. Nesse contexto, os elementos h_{ii} são conhecidos como medidas de alavancagem da regressão ajustada (MICHAEL et al., 2004). Ainda, pode-se mostrar que esses resíduos ordinários seguem distribuição normal, partindo de que $Y \sim N_n(X\beta, \sigma^2 I_n)$. Nota-se, de (2.2.4), que $\widehat{\varepsilon}$ é uma combinação linear de Y. Como qualquer combinação linear dos elementos de Y segue distribuição normal univariada, então $\widehat{\varepsilon_i} \sim N\left(0, \sigma^2(1 - h_{ii})\right),\ i = 1, \ldots, n$.

Ademais, analisando a covariância entre resíduos ordinários de diferentes observações, tem-se

$$\mathrm{Cov}(\widehat{\varepsilon_i}, \widehat{\varepsilon_j}) = -\sigma^2 h_{ij} \ ,$$

com $i \neq j$ e $h_{ij} = x_i{}^\top (X^\top X)^{-1} x_j$. Então, segue também que, para a correlação, tem-se

$$\mathrm{Corr}(\widehat{\varepsilon_i}, \widehat{\varepsilon_j}) = -\frac{\sigma^2 h_{ij}}{\sqrt{\sigma^2(1-h_{ii})\sigma^2(1-h_{jj})}} = -\frac{h_{ij}}{\sqrt{(1-h_{ii})(1-h_{jj})}}.$$

Logo, observa-se que os resíduos ordinários $\widehat{\varepsilon_i}$ possuem média zero e são normalmente distribuídos, assim como os erros aleatórios. Entretanto, estes resíduos possuem variância não constante e são correlacionados. Como a ideia é que estes resíduos tivessem as mesmas características dos erros aleatórios ε_i, definidas nas suposições da Subseção 2.1.3, é necessário fazer algumas modificações.

Para lidar com a variância não constante, busca-se uma padronização de $\widehat{\varepsilon_i}$. Como $\mathrm{Var}(\widehat{\varepsilon_i}) = \sigma^2(1-h_{ii})$ com σ^2 desconhecido, inicialmente, considera-se o resíduo padronizado (ou resíduo interno studentizado) dado por

$$t_i = \frac{\widehat{\varepsilon_i}}{\sqrt{\widehat{\sigma}^2(1-h_{ii})}}.$$

Contudo, $\widehat{\varepsilon_i}$ não é independente de $\widehat{\sigma}^2$ e, consequentemente, t_i não segue distribuição *t-student* com $(n-(k+1))$ graus de liberdade. Esse problema pode ser contornado substituindo $\widehat{\sigma}^2$ por $\widehat{\sigma}^2_{(i)}$, que é a variância estimada correspondente ao modelo excluindo a i-ésima observação. Como apresentado em Paula (2023), tem-se que

$$\widehat{\sigma}^2_{(i)} = \widehat{\sigma}^2 \left(\frac{n-k-1-t_i^2}{n-k-2} \right).$$

Então, para obter $\widehat{\sigma}^2_{(i)}$, não será necessário ajustar n modelos de regressão.

Portanto, o resíduo final, chamado de resíduo studentizado (ou resíduo externo studentizado), é definido como

$$t_i^* = \frac{\widehat{\varepsilon_i}}{\sqrt{\widehat{\sigma}^2_{(i)}(1-h_{ii})}}.$$

Ademais, t_i^* pode ser reescrito da seguinte forma

$$t_i^* = t_i \sqrt{\left(\frac{n-k-2}{n-k-1-t_i^2} \right)} \ ,$$

do qual segue que $t_i^* \sim t_{n-k-2}$, em que, para n grande, $t_i^* \overset{n\to\infty}{\sim} N(0,1)$, de acordo com o exposto por Paula (2023). Logo, se para n grande, os resíduos studentizados seguirem distribuição normal padrão, isso significa que há indicações de que a suposição de normalidade está adequada. Além disso, para se avaliar a normalidade dos erros, como na suposição S_6, pode-se construir o gráfico normal de probabilidade dos resíduos studentizados, com envelope simulado. Para mais detalhes sobre a construção de envelopes simulados, ver Paula (2023). Para avaliar as suposições S_2 (média dos erros igual a zero) e S_4 (homoscedasticidade), pode-se construir o gráfico dos resíduos studentizados contra os valores ajustados da resposta. Se estas suposições estiverem adequadas, os resíduos devem apresentar média zero e variância constante.

2.3 Métodos de estimação robustos

O estimador obtido pelo LSM possui ótimas propriedades em situações em que todas as suposições supracitadas são garantidas. No entanto, em alguns conjuntos de dados, pode-se ter *outliers* ou observações discrepantes que podem influenciar e tendenciar o modelo ajustado. Ainda, pode-se ter erros considerados "não gaussianos", ou seja, erros que violam a suposição clássica de normalidade. Com isso, os estimadores obtidos pelo LSM não são ideais, pois podem resultar em estimativas imprecisas e enviesadas dos parâmetros do modelo de regressão linear.

Nesses casos, existem outros métodos de estimação, os quais serão estudados ao longo deste trabalho, que são os métodos de estimação robustos. Esses métodos auxiliam na redução da influência de pontos atípicos nas estimativas dos parâmetros. Assim, essas estimativas conduzirão em um modelo que melhor se ajusta a maior parte dos dados, dando pouco peso aos *outliers* no ajuste do modelo. Portanto, os métodos robustos podem ser mais flexíveis, no sentido que a maior parte dos dados pode ser modelada com maior precisão, e as inferências ou estimativas dos coeficientes da regressão linear são robustas e confiáveis.

Sob essa perspectiva, o principal método de estimação que será estudado, em comparação ao LSM, é baseado na classe dos M-estimadores, que parte de uma generalização do processo de estimação por máxima verossimilhança. O M-estimador baseia-se na ideia de atribuição de pesos diferentes no processo de estimação para cada observação, de acordo com a distância dessas quanto aos valores esperados sob o modelo postulado. Assim, espera-se que os pontos mais distantes do valor esperado tenham um peso menor do que os pontos mais próximos do previsto, o que reduz a influência dos *outliers* sob o ajuste da reta de regressão linear e resulta em um ajuste melhor dessa reta a maior parte dos dados (MARONNA et al., 2019).

2.3.1 M-estimadores sob modelos de localização

Nesta subseção, introduziremos o método de M-estimação sob modelos de localização para n variáveis aleatórias independentes e identicamente distribuídas. Este método será generalizado posteriormente no presente trabalho, na Seção 2.4, em um contexto de regressão linear.

Considere $y_1, y_2, \ldots, y_n$, n realizações independentes da variável aleatória y que possui função de distribuição $F(y;\theta)$, em que $\theta \in \mathbb{R}$ é um parâmetro desconhecido. Note que $y_1, y_2, \ldots, y_n$ são variáveis aleatórias independentes e identicamente distribuídas (*i.i.d.*).

Sob a configuração acima, suponha que as n realizações de y podem ser descritas através da relação, denominada de modelo de localização,

$$y_i = \theta + \varepsilon_i,\ i = 1, 2, \ldots, n, \tag{2.3.1}$$

em que $\theta \in \mathbb{R}$ é um parâmetro de localização do modelo, e $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ são as n variáveis aleatórias independentes (erros) com função de distribuição comum $F_0(y)$ que satisfaz

$$F(y;\theta) = F_0(y - \theta).$$

Considerando o modelo de localização definido em (2.3.1), tem-se a função densidade de probabilidade conjunta de $y_1, y_2, \ldots, y_n$, ou seja,

$$\mathrm{L}(\theta) = \prod_{i=1}^{n} f_0(y_i - \theta),$$

em que $f_0(y) = \frac{\partial}{\partial y} F_0(y)$, é a função de densidade comum de $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$. Dizemos que $\mathrm{L}(\theta)$ é a função de verossimilhança de θ sob o modelo definido em (2.3.1).

Consequentemente, considerando o modelo de localização supracitado, o logaritmo da função de verossimilhança de θ é

$$\ell(\theta) = \sum_{i=1}^{n} \log(f_0(y_i - \theta)),$$

em que $\log(y)$ é o logaritmo natural, ou seja, $\log(y) = \log_{\mathrm{e}}(y) = \ln(y)$. Por definição, o estimador de máxima verossimilhança (MLE, *Maximum Likelihood Estimator*), denotado por $\widehat{\theta}_{ML}$, é obtido por meio de $\widehat{\theta}_{ML} = \underset{\theta \in \mathbb{R}}{\operatorname{argmax}}\ [\mathrm{L}(\theta)]$, com $\theta \in \mathbb{R}$. Como $f_0(y)$ é sempre positiva e a função logaritmo é crescente, pode-se obter $\widehat{\theta}_{ML}$ a partir de $\widehat{\theta}_{ML} = \underset{\theta \in \mathbb{R}}{\operatorname{argmax}}\ [\ell(\theta)]$, ou ainda, $\widehat{\theta}_{ML} = \underset{\theta \in \mathbb{R}}{\operatorname{argmin}}[-\ell(\theta)]$.

Dessa forma, a ideia de Huber (1964) foi substituir $-\log(f_0(y_i-\theta))$ por uma função real $\rho(y_i-\theta)$, que deverá satisfazer algumas condições. Vale salientar que

$$\sum_{i=1}^{n}[-\log(f_0(y_i-\theta))]=\sum_{i=1}^{n}\rho(y_i-\theta), \tag{2.3.2}$$

com $\rho(y)=-\log(f_0(y))$.

Por exemplo, considere $f_0(y)$ como sendo a função de densidade da distribuição normal padrão, isto é, se

$$f_0(y)=\frac{1}{\sqrt{2\pi}}e^{-y^2/2},\quad y\in\ \mathbb{R}. \tag{2.3.3}$$

Para calcular $\rho(y_i-\theta)$, tem-se

$$\begin{aligned}\rho(y_i-\theta)&=-\log(f_0(y_i-\theta))\\&=-\log\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{(y_i-\theta)^2}{2}}\right)\\&=-\left(\log\left(\frac{1}{\sqrt{2\pi}}\right)-\left(\frac{(y_i-\theta)^2}{2}\right)\right)\\&=\frac{(y_i-\theta)^2}{2}+\log\left(\sqrt{2\pi}\right).\end{aligned}$$

Assim, exceto uma constante, $\rho(y_i-\theta)\propto(y_i-\theta)^2$, logo, para esse caso, minimizar a expressão (2.3.2) com relação a θ equivale a minimizar

$$\sum_{i=1}^{n}(y_i-\theta)^2.$$

Portanto, obtemos que

$$\widehat{\theta}_{ML}=\underset{\theta\ \in\ \Theta}{\operatorname{argmin}}\sum_{i=1}^{n}(y_i-\theta)^2.$$

Note que este método é equivalente ao LSM com $\mathrm{E}(y_i)$ ao invés de θ.

Agora, considere $f_0(y)$ como sendo a função de densidade da distribuição exponencial dupla, isto é, se

$$f_0(y)=\frac{1}{2}e^{-|y|},\quad y\in\ \mathbb{R}. \tag{2.3.4}$$

Analogamente ao exemplo anterior, calcula-se $\rho(y_i-\theta)$ para a exponencial dupla, obtendo

$$\begin{aligned}\rho(y_i-\theta) &= -\log(f_0(y_i-\theta)) \\ &= -\log\left(\frac{1}{2}e^{-|y_i-\theta|}\right) \\ &= -\left(\log\left(\frac{1}{2}\right)-(|y_i-\theta|)\right) \\ &= |y_i-\theta|+\log(2)\,.\end{aligned}$$

Dessa forma, exceto uma constante, $\rho(y_i-\theta) \propto |y_i-\theta|$, então, para minimizar a soma (2.3.2) em relação a θ, deve-se minimizar

$$\sum_{i=1}^{n}|y_i-\theta|.$$

Portanto, para esse caso,

$$\widehat{\theta}_{ML} = \underset{\theta\,\in\,\Theta}{\operatorname{argmin}}\sum_{i=1}^{n}|y_i-\theta|.$$

Note que ao trocar a função de densidade normal pela exponencial dupla, o procedimento de estimação minimizará o valor absoluto dos desvios, ou seja, $|y_i-\theta|$, ao invés do quadrado dos desvios, ou seja, $(y_i-\theta)^2$.

De modo geral, segundo Maronna et al. (2019), um M-estimador de localização para o parâmetro θ, denotado por $\widehat{\theta}_M$, é obtido de

$$\widehat{\theta}_M = \underset{\theta\,\in\,\Theta}{\operatorname{argmin}}\sum_{i=1}^{n}\rho(y_i-\theta), \tag{2.3.5}$$

em que a função $\rho(y)$ deve satisfazer as seguintes condições:

C_1. $\rho(y)$ é uma função não descrescente de $|y|$, isto é, se $|y_i|<|y_j|$, então $\rho(|y_i|)\leq\rho(|y_j|)$;

C_2. $\rho(0)=0$;

C_3. $\rho(y)$ é crescente para $y>0$, desde que $\rho(y)<\rho(\infty)$;

C_4. Se $\rho(y)$ é limitada, então assume-se que $\rho(\infty)=1$.

Vale salientar que as condições de C_1 a C_4 são válidas para qualquer M-estimador de localização.

Se a função $\rho(y)$ for diferenciável, pode-se obter $\widehat{\theta}_M$ derivando (2.3.5) com relação

ao parâmetro θ e igualando a zero, resultando na equação de estimação dada por (MARONNA et al., 2019)

$$\sum_{i=1}^{n} \psi(y_i - \widehat{\theta}_M) = 0, \tag{2.3.6}$$

em que $\psi(y) = \frac{\partial}{\partial y}\rho(y)$ é uma função ímpar, ou seja, $\psi(-y) = -\psi(y)$ com $\psi(y) \geq 0$, para todo $y \geq 0$. Note que a contribuição de cada observação y_i na equação de estimação é dada por $\psi(y_i - \widehat{\theta}_M)$. Logo, se y_i for uma observação discrepante, pretende-se escolher uma função ψ tal que $\psi(y_i - \widehat{\theta}_M)$ assuma um valor pequeno quando $|y_i - \widehat{\theta}_M|$ é um valor alto, lembrando que $\widehat{\theta}_M$ é um estimador de localização. Em outras palavras, queremos que observações de y que sejam discrepantes recebam um peso baixo no processo de estimação de θ. Além disso, M-estimadores obtidos de (2.3.6) são chamados de M-estimadores tipo-ψ, enquanto se forem obtidos de (2.3.5) são denominados de M-estimadores tipo-ρ.

Voltando ao exemplo da distribuição normal padrão, para $\rho(y) \propto y^2$, tem-se que $\psi(y) = \frac{\partial}{\partial y}\rho(y) \propto y$ (MARONNA et al., 2019, pg.24) e, portanto, $\widehat{\theta}_M$ é obtido de

$$\sum_{i=1}^{n} (y_i - \widehat{\theta}_M) = 0 \implies \widehat{\theta}_M = \frac{1}{n}\sum_{i=1}^{n} y_i = \overline{y}.$$

Para o exemplo da $f_0(y)$ como função de densidade da distribuição exponencial dupla (MARONNA et al., 2019, pg.24), ou seja, $\rho(y) \propto |y|$, tem-se que $\psi(y) = \frac{\partial}{\partial y}\rho(y) \propto sign(y)$, desde que

$$\frac{\partial}{\partial y}|y| = \begin{cases} -1, \text{ se } y < 0, \\ \ \ 1, \text{ se } y > 0. \end{cases} \tag{2.3.7}$$

Diante disso, pode-se reescrever $\psi(y)$ como sendo a função sinal definida por

$$\psi(y) = sign(y) = \begin{cases} -1, \text{ se } y < 0 \\ \ \ 0, \text{ se } y = 0 \\ \ \ 1, \text{ se } y > 0, \end{cases}$$

ou ainda,

$$\psi(y) = sign(y) = \mathrm{I}(y > 0) - \mathrm{I}(y < 0),$$

em que a função $\mathrm{I}(y > 0)$ é a função indicadora, ou seja,

$$\mathrm{I}(y > 0) = \begin{cases} 0, \text{ se } y \leq 0, \\ 1, \text{ se } y > 0. \end{cases}$$

Assim, aplicando $\psi(y)$ em (2.3.6), tem-se

$$\sum_{i=1}^{n} sign(y_i - \widehat{\theta}_M) = 0$$

$$\sum_{i=1}^{n} (\mathrm{I}(y_i - \widehat{\theta}_M > 0) - \mathrm{I}(y_i - \widehat{\theta}_M < 0)) = 0$$

$$\#(y_i - \widehat{\theta}_M > 0) = \#(y_i - \widehat{\theta}_M < 0), \tag{2.3.8}$$

em que a função $\#(y)$ denota o número de observações que satisfazem a condição y. Dessa forma, o valor de $\widehat{\theta}_M$ que satisfaz (2.3.8) é $\widehat{\theta}_M = med(y_i)$, em que $med(y_i)$ é a mediana amostral das y_i observações, com $i = 1, \ldots, n$.

Exemplos de funções $\rho(y)$ e $\psi(y)$ mais utilizadas na literatura de robustez são:

1. Função do tipo Huber:

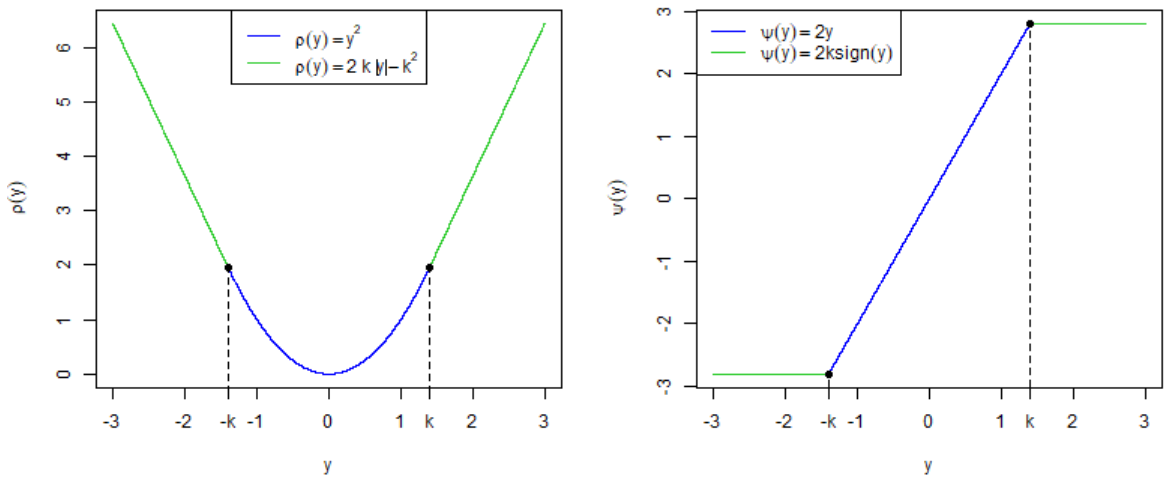
$$\rho(y) = \begin{cases} y^2, \text{ se } |y| \leq k \\ 2k|y| - k^2, \text{ se } |y| > k, \end{cases}$$

em que $k > 0$ é uma constante real conhecida. Usualmente fixa-se $k = 1,4$ para garantir robustez considerável e eficiência assintótica próxima ao MLE (MARONNA et al., 2019, p.28). Note que,

$$\rho'(y) = \psi(y) = \begin{cases} 2y, \text{ se } |y| \leq k, \\ 2k \cdot sign(y), \text{ se } |y| > k. \end{cases}$$

Na Figura 2 apresentamos os gráficos para as funções $\rho(y)$ e $\psi(y)$ do tipo Huber.

Figura 2: Comportamento das funções $\rho(y)$ e $\psi(y)$ do tipo Huber.



Analisando o comportamento de $\psi(y)$ na Figura 2, vemos que, para valores de $-k < y < k$, a contribuição das observações no processo de estimação é igual a $2y$. No entanto, quando $y < -k$ ou $y > k$, essa contribuição se torna constante, sendo $k = 1,4$ para função $\rho(y)$ do tipo Huber.

2. Função do tipo *bisquare* de Tukey:

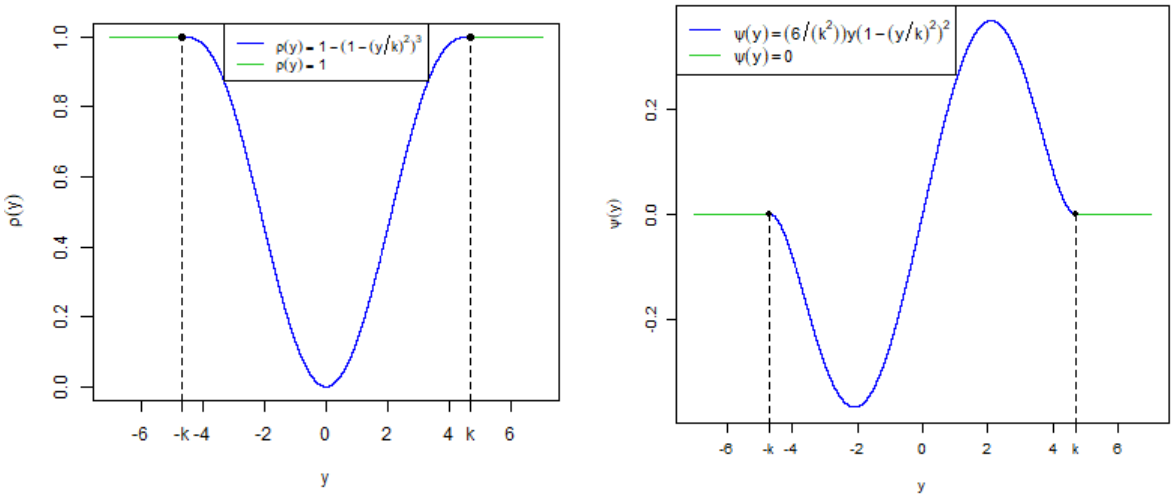
$$\rho(y) = \begin{cases} 1 - [1 - (y/k)^2]^3, & \text{se } |y| \leq k \\ 1, & \text{se } |y| > k, \end{cases}$$

em que $k > 0$ é uma constante real conhecida. Usualmente fixa-se $k = 4,675$ para garantir uma eficiência assintótica de 95% do estimador obtido pela função do tipo *bisquare* (MARONNA et al., 2019, pg.31). Ainda, veja que

$$\rho'(y) = \psi(y) = \begin{cases} (6/k^2)y\,[1 - (y/k)^2]^2, & \text{se } |y| \leq k \\ 0, & \text{se } |y| > k. \end{cases}$$

Na Figura 3 apresentamos os gráficos para as funções $\rho(y)$ e $\psi(y)$ do tipo *bisquare*.

Figura 3: Comportamento das funções $\rho(y)$ e $\psi(y)$ do tipo *bisquare* de Tukey.



Analisando o comportamento de $\psi(y)$ na Figura 3, para valores de $-k < y < k$, a contribuição das observações no processo de estimação aumenta de acordo com que y se distancia de zero, decresce a partir de $y > 2$ ou $y < -2$, e assume valor zero quando $y > k$ ou $y < -k$.

Note que a função $\psi(y)$ do tipo *bisquare* robustifica o procedimento de estimação mais do que a função $\psi(y)$ de Huber, pois a primeira atribui contribuição zero para $|y| > k$, com $k = 4,675$, diferente da segunda em que tem-se contribuição constante igual a $2k \cdot sign(y)$ para $|y| > k$, com $k = 1,4$.

2.3.2 M-estimadores sob modelos de escala

Nesta subseção, trabalharemos um outro tipo de M-estimação, o método sob modelos de escala para n variáveis aleatórias independentes e identicamente distribuídas. Esse método será importante para auxiliar, na Subseção 2.4, no desenvolvimento de estimadores no contexto de regressão.

Nessa abordagem, consideraremos $y_i,\ i = 1, 2, \ldots, n$, como n variáveis aleatórias independentes e identicamente distribuidas ($i.i.d.$), mas agora com função de distribuição $F(y;\sigma)$, em que $\sigma > 0$ é um parâmetro desconhecido.

Sob essa perspectiva, iremos supor um modelo de escala, que descreve as n ocorrências de y sob a seguinte relação

$$y_i = \sigma\varepsilon_i,\ i = 1, 2, \ldots, n, \tag{2.3.9}$$

em que $\sigma > 0$ é um parâmetro de escala do modelo e ε_i são os n erros aleatórios independentes e identicamente distribuídos com função de distribuição $F_0(y)$. Segue-se, então, que $y_1, y_2, \ldots, y_n$ terá função de distribuição comum equivalente a

$$F(y;\sigma) = F_0\left(\frac{y}{\sigma}\right).$$

Isto implica que $f(y;\sigma) = \frac{1}{\sigma} f_0\left(\frac{y}{\sigma}\right)$, em que $f_0(y) = \frac{\partial}{\partial y} F_0(y)$ é a função de densidade comum de $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$.

Considerando o modelo de escala definido em (2.3.9), temos que a função densidade de probabilidade conjunta de $y_1, y_2, \ldots, y_n$ é dada por

$$\mathrm{L}(\sigma) = \prod_{i=1}^{n} \frac{1}{\sigma} f_0\left(\frac{y_i}{\sigma}\right) = \frac{1}{\sigma^n} \prod_{i=1}^{n} f_0\left(\frac{y_i}{\sigma}\right),$$

em que dizemos que $\mathrm{L}(\sigma)$ é a função de verossimilhança de σ sob o modelo definido na expressão (2.3.9).

Consequentemente, seguindo com o procedimento análogo ao adotado para a família de localização, o logaritmo da função de verossimilhança de σ sob o modelo de escala é

$$\ell(\sigma) = \log(\mathrm{L}(\sigma)) = \log\left(\frac{1}{\sigma^n}\right) + \log\left(\prod_{i=1}^{n} f_0\left(\frac{y_i}{\sigma}\right)\right) = -n\log(\sigma) + \sum_{i=1}^{n} \log\left(f_0\left(\frac{y_i}{\sigma}\right)\right).$$

O MLE para σ é obtido de $\widehat{\sigma}_{MV} = \underset{\sigma\,>\,0}{\operatorname{argmax}}\ [\mathrm{L}(\sigma)]$. Podemos obter $\widehat{\sigma}_{ML}$ também por meio de $\widehat{\sigma}_{ML} = \underset{\sigma\,>\,0}{\operatorname{argmax}}\ [\ell(\sigma)]$, ou ainda, $\widehat{\sigma}_{ML} = \underset{\sigma\,>\,0}{\operatorname{argmin}}[-\ell(\sigma)]$. Assim, a fim de maximizar $\ell(\sigma)$, deve-se derivar essa função com relação a σ e igualar a zero. Assim, temos que

$$\frac{\partial}{\partial\sigma}\left[\ell(\sigma)\right]\Big|_{\sigma=\widehat{\sigma}_{ML}} = 0$$

$$-\frac{n}{\widehat{\sigma}_{ML}} + \sum_{i=1}^{n} \frac{1}{f_0\left(\frac{y_i}{\widehat{\sigma}_{ML}}\right)} f_0'\left(\frac{y_i}{\widehat{\sigma}_{ML}}\right) \cdot \left(\frac{-y_i}{\widehat{\sigma}_{ML}^2}\right) = 0$$

$$\left(\frac{1}{\widehat{\sigma}_{ML}}\right) \sum_{i=1}^{n} \left(\frac{-y_i}{\widehat{\sigma}_{ML}}\right) \cdot \frac{f_0'\left(\frac{y_i}{\widehat{\sigma}_{ML}}\right)}{f_0\left(\frac{y_i}{\widehat{\sigma}_{ML}}\right)} = \frac{n}{\widehat{\sigma}_{ML}}$$

$$\left(\frac{1}{n}\right) \sum_{i=1}^{n} \left(\frac{-y_i}{\widehat{\sigma}_{ML}}\right) \cdot \frac{f_0'\left(\frac{y_i}{\widehat{\sigma}_{ML}}\right)}{f_0\left(\frac{y_i}{\widehat{\sigma}_{ML}}\right)} = 1. \tag{2.3.10}$$

Dessa forma, a equação de estimação para $\widehat{\sigma}_{ML}$ (2.3.10) pode ser reescrita por

$$\left(\frac{1}{n}\right) \sum_{i=1}^{n} \rho\left(\frac{y_i}{\widehat{\sigma}_{ML}}\right) = 1, \tag{2.3.11}$$

em que

$$\rho(y) = -y \cdot \frac{f_0'(y)}{f_0(y)} = y \cdot \psi(y)$$

com $\psi(y) = -f_0'(y)/f_0(y)$. Nota-se que a função $\psi(y)$ para o MLE de escala equivale a função $\psi(y)$ definida para o MLE de localização, uma vez que, para o segundo, tinha-se $\rho(y) = -\log(f_0(y))$, e portanto,

$$\psi(y) = \frac{\partial}{\partial y}\rho(y) = -\frac{f_0'(y)}{f_0(y)}.$$

No entanto, a função $\rho(y)$ não coincide para ambos os casos, sendo $\rho(y) = -y \cdot \frac{f_0'(y)}{f_0(y)}$ para o primeiro e $\rho(y) = -\log(f_0(y))$ para o segundo, com $f_0(y)$ sendo uma função densidade padronizada qualquer.

Considerando $f_0(y)$ a função de densidade da distribuição normal padrão definida em (2.3.3), para calcular $\rho\left(\frac{y_i}{\widehat{\sigma}_{ML}}\right)$, faz-se, inicialmente,

$$\rho\left(\frac{y_i}{\widehat{\sigma}_{ML}}\right) = \left(-\frac{y_i}{\widehat{\sigma}_{ML}}\right) \cdot \frac{f_0'\left(\frac{y_i}{\widehat{\sigma}_{ML}}\right)}{f_0\left(\frac{y_i}{\widehat{\sigma}_{ML}}\right)}, \tag{2.3.12}$$

em que

$$f_0'(y) = \frac{1}{\sqrt{2\pi}} \cdot (-y) \cdot \exp\left\{-\frac{y^2}{2}\right\}.$$

Logo, tem-se que

$$\begin{aligned}
\rho\left(\frac{y_i}{\widehat{\sigma}_{ML}}\right) &= \left(-\frac{y_i}{\widehat{\sigma}_{ML}}\right) \cdot \frac{f_0'\left(\frac{y_i}{\widehat{\sigma}_{ML}}\right)}{f_0\left(\frac{y_i}{\widehat{\sigma}_{ML}}\right)} \\
&= \left(-\frac{y_i}{\widehat{\sigma}_{ML}}\right) \cdot \frac{\frac{1}{\sqrt{2\pi}} \cdot \left(-\frac{y_i}{\widehat{\sigma}_{ML}}\right) \cdot \exp\left\{-\left(\frac{y_i}{\widehat{\sigma}_{ML}}\right)^2 / \ 2\right\}}{\frac{1}{\sqrt{2\pi}} \cdot \exp\left\{-\left(\frac{y_i}{\widehat{\sigma}_{ML}}\right)^2 / \ 2\right\}} \\
&= \left(-\frac{y_i}{\widehat{\sigma}_{ML}}\right) \cdot \left(-\frac{y_i}{\widehat{\sigma}_{ML}}\right) \\
&= \left(\frac{y_i}{\widehat{\sigma}_{ML}}\right)^2 .
\end{aligned}$$

Assim, obtemos que $\rho(y) = y^2$ para a função $f_0(y)$ sendo a distribuição normal padrão, com $y \in \mathbb{R}$. Como $\rho(y) = y \cdot \psi(y)$ para modelos de escala, tem-se que $\psi(y) = \rho(y)/y$. Com isso, por meio da expressão (2.3.11), o MLE para σ é obtido de

$$\begin{aligned}
\left(\frac{1}{n}\right) \sum_{i=1}^{n} \left(\frac{y_i}{\widehat{\sigma}_{ML}}\right)^2 &= 1 \\
\left(\frac{1}{n}\right) \left(\frac{1}{\widehat{\sigma}^2_{ML}}\right) \sum_{i=1}^{n} (y_i)^2 &= 1 \\
\left(\frac{1}{n}\right) \sum_{i=1}^{n} y_i^2 &= \widehat{\sigma}^2_{ML} \\
\widehat{\sigma}_{ML} &= \sqrt{\sum_{i=1}^{n} \frac{y_i^2}{n}} \ .
\end{aligned}$$

Portanto, $\widehat{\sigma}_{ML}$ sob a função de densidade da normal padrão será igual à raiz do quadrado médio (RMS, *Root Mean Square*) das observações y_i.

Em contrapartida, para f_0 sendo a função de densidade da distribuição exponencial dupla definida em (2.3.4), analogamente ao caso anterior, podemos calcular $\rho\left(\frac{y_i}{\widehat{\sigma}_{ML}}\right)$ partindo da expressão (2.3.12), considerando

$$f_0'(y) = \frac{1}{2} \cdot \left(-\frac{\partial}{\partial y} |y|\right) \cdot e^{-|y|},$$

em que a derivada do módulo está definida em (2.3.7).

Assim, tem-se que

$$\begin{aligned}
\rho\left(\frac{y_i}{\widehat{\sigma}_{ML}}\right) &= \left(-\frac{y_i}{\widehat{\sigma}_{ML}}\right) \cdot \frac{f_0'\left(\frac{y_i}{\widehat{\sigma}_{ML}}\right)}{f_0\left(\frac{y_i}{\widehat{\sigma}_{ML}}\right)} \\
&= \left(-\frac{y_i}{\widehat{\sigma}_{ML}}\right) \cdot \frac{\frac{1}{2} \cdot \left(-sign\left(\frac{y_i}{\widehat{\sigma}_{ML}}\right)\right) \cdot \exp\left\{-\left|\frac{y_i}{\widehat{\sigma}_{ML}}\right|\right\}}{\frac{1}{2} \cdot \exp\left\{-\left|\frac{y_i}{\widehat{\sigma}_{ML}}\right|\right\}} \\
&= \left(-\frac{y_i}{\widehat{\sigma}_{ML}}\right) \cdot \left(-sign\left(\frac{y_i}{\widehat{\sigma}_{ML}}\right)\right) \\
&= \left|\frac{y_i}{\widehat{\sigma}_{ML}}\right|.
\end{aligned}$$

Logo, tem-se $\rho(y) = |y|$. Como $\rho(y) = y \cdot \psi(y)$ para modelos de escala, tem-se que $\psi(y) = \frac{|y|}{y} = sign(y)$.

Com isso, o MLE para σ pode ser obtido da expressão (2.3.11) substituindo $\rho\left(\frac{y_i}{\widehat{\sigma}_{ML}}\right)$ da seguinte forma

$$\begin{aligned}
\left(\frac{1}{n}\right) \sum_{i=1}^{n} \left|\frac{y_i}{\widehat{\sigma}_{ML}}\right| &= 1 \\
\left(\frac{1}{n}\right) \left(\frac{1}{\widehat{\sigma}_{ML}}\right) \sum_{i=1}^{n} |y_i| &= 1 \\
\left(\frac{1}{n}\right) \sum_{i=1}^{n} |y_i| &= \widehat{\sigma}_{ML} \\
\widehat{\sigma}_{ML} &= \sum_{i=1}^{n} \frac{|y_i|}{n}.
\end{aligned}$$

De modo geral, segundo Maronna et al. (2019), um M-estimador de escala para o parâmetro σ, denotado por $\widehat{\sigma}_M$, é solução da equação

$$\left(\frac{1}{n}\right) \sum_{i=1}^{n} \rho\left(\frac{y_i}{\widehat{\sigma}_M}\right) = \delta, \tag{2.3.13}$$

com $\rho(y)$ uma função que satisfaz as condições C_1 a C_4 e $\delta > 0$ uma constante fixada. Para garantir que esta equação tenha solução, devemos ter que $0 < \delta < \rho(\infty)$. Além disso, se $\rho(y)$ é limitada (C_4), então $\rho(y) = 1$, e portanto, $0 < \delta < 1$.

Entre as funções $\rho(y)$ mais utilizadas, pode-se citar a *step function*, obtida de

$$\rho(y) = \mathrm{I}(|y| > c),$$

em que c é uma constante positiva conhecida e $\delta = 0,5$ fixo. Aplicando em (2.3.13), ficamos com $\widehat{\sigma}_M$ tal que

$$\left(\frac{1}{n}\right)\sum_{i=1}^{n} \mathrm{I}\left(\left|\frac{y_i}{\widehat{\sigma}_M}\right| > c\right) = 0,5 \ .$$

Segue que

$$\begin{aligned}
\sum_{i=1}^{n} \mathrm{I}\left(\left|\frac{y_i}{\widehat{\sigma}_M}\right| > c\right) &= \frac{n}{2} \\
\#\left(\left|\frac{y_i}{\widehat{\sigma}_M}\right| > c\right) &= \frac{n}{2} \\
\#\left(|y_i| > c \cdot \widehat{\sigma}_M\right) &= \frac{n}{2}. \qquad (2.3.14)
\end{aligned}$$

Com isso, o valor $\widehat{\sigma}_M$ que satisfaz (2.3.14) é obtido de

$$\begin{aligned}
med\left(|y_i|\right) &= c \cdot \widehat{\sigma}_M \\
\widehat{\sigma}_M &= \frac{med\left(|y_i|\right)}{c} \ .
\end{aligned}$$

Segundo Maronna et al. (2019), o valor c que garantirá que $\widehat{\sigma}_M$ seja um estimador Fisher-consistente (definição na Subseção 2.3.4) é $c = 0,6745$.

Os M-estimadores de escala não são os estimadores de maior importância nos procedimentos inferenciais robustos. Na verdade, estes possuem papel auxiliar na estimação dos parâmetros sob famílias de localização e escala.

2.3.3 M-estimadores sob modelos de localização e escala

Considere um modelo de localização e escala que descreve n ocorrências independentes e identicamente distribuídas $y_1, y_2, \ldots, y_n$ da variável aleatória y, com função de distribuição $F(y; \theta, \sigma)$, definido por

$$y_i = \theta + \sigma\varepsilon_i, \ i = 1, 2, \ldots, n, \qquad (2.3.15)$$

em que $\theta \in \mathbb{R}$ é um parâmetro de localização, $\sigma > 0$ é o parâmetro de escala e $\varepsilon_1, \ldots, \varepsilon_n$ são variáveis aleatórias independentes (erros) com função de densidade comum $f_0(y)$ que satisfazem

$$F(y; \theta, \sigma) = F_0\left(\frac{y - \theta}{\sigma}\right).$$

Assim, a função densidade de probabilidade de y satisfaz

$$f(y;\theta,\sigma) = \frac{1}{\sigma} f_0\left(\frac{y-\theta}{\sigma}\right).$$

Considerando o modelo de localização e escala definido em (2.3.15), a função de densidade de probabilidade conjunta de $y_1, y_2, \ldots, y_n$ é dada por

$$\mathrm{L}(\theta,\sigma) = \left(\frac{1}{\sigma^n}\right) \prod_{i=1}^{n} f_0\left(\frac{y_i-\theta}{\sigma}\right),$$

em que $\mathrm{L}(\theta,\sigma)$ é a função de verossimilhança de θ e σ sob o modelo definido em (2.3.15).

O logaritmo da função de verossimilhança de θ e σ sob o modelo de localização e escala é

$$\ell(\theta,\sigma) = -n\log(\sigma) + \sum_{i=1}^{n} \log\left(f_0\left(\frac{y_i-\theta}{\sigma}\right)\right) = -\log(\sigma) + \frac{1}{n}\sum_{i=1}^{n} \log\left(f_0\left(\frac{y_i-\theta}{\sigma}\right)\right).$$

Com isso, os estimadores de máxima verossimilhança podem ser obtidos a partir de $\left(\widehat{\theta}_{ML}, \widehat{\sigma}_{ML}\right) = \underset{(\theta,\ \sigma)\ \in\ \Theta}{\operatorname{argmax}}\ [\mathrm{L}(\theta,\sigma)]$, com $(\theta,\ \sigma)\ \in\ \Theta$, em que Θ é o espaço paramétrico. Assim, tem-se que

$$\left(\widehat{\theta}_{ML}, \widehat{\sigma}_{ML}\right) = \underset{(\theta,\ \sigma)\ \in\ \Theta}{\operatorname{argmax}}\ [\mathrm{L}(\theta,\sigma)] = \underset{(\theta,\ \sigma)\ \in\ \Theta}{\operatorname{argmax}}\ \left\{\frac{1}{\sigma^n}\prod_{i=1}^{n} f_0\left(\frac{y_i-\theta}{\sigma}\right)\right\},$$

ou ainda,

$$\begin{aligned}\left(\widehat{\theta}_{ML}, \widehat{\sigma}_{ML}\right) &= \underset{(\theta,\ \sigma)\ \in\ \Theta}{\operatorname{argmin}}\ [-\ell(\theta,\sigma)] \\ &= \underset{(\theta,\ \sigma)\ \in\ \Theta}{\operatorname{argmin}}\ \left\{\log(\sigma) + \frac{1}{n}\sum_{i=1}^{n} -\log\left(f_0\left(\frac{y_i-\theta}{\sigma}\right)\right)\right\}, \qquad (2.3.16)\end{aligned}$$

em que $\rho_0(y) = -\log(f_0(y))$.

Derivando (2.3.16) com relação a θ e σ, respectivamente, obtemos as seguintes equações de estimação associadas aos MLEs de θ e σ

$$\sum_{i=1}^{n} \psi\left(\frac{y_i - \widehat{\theta}_{ML}}{\widehat{\sigma}_{ML}}\right) = 0$$

e

$$\left(\frac{1}{n}\right)\sum_{i=1}^{n}\rho_{scale}\left(\frac{y_i-\widehat{\theta}_{ML}}{\widehat{\sigma}_{ML}}\right)=1,$$

com $\psi(y)=-\rho_0'(y)$, $\rho_{scale}(y)=y\cdot\psi(y)$, em que $\rho_{scale}(y)$ é a função $\rho(y)$ utilizada para o modelo de escala.

De forma geral, os M-estimadores de localização e escala, denotados por $\widehat{\theta}_M$ e $\widehat{\sigma}_M$, são obtidos, respectivadamente, de

$$\sum_{i=1}^{n}\psi\left(\frac{y_i-\widehat{\theta}_M}{\widehat{\sigma}_M}\right)=0$$

e

$$\left(\frac{1}{n}\right)\sum_{i=1}^{n}\rho_{scale}\left(\frac{y_i-\widehat{\theta}_M}{\widehat{\sigma}_M}\right)=\delta,$$

em que $0<\delta<1$, $\rho_{scale}(y)$ é escolhida tal como visto na Subseção 2.3.2 e escolheremos funções $\psi(y)$ como apresentado na Subseção 2.3.1.

2.3.4 M-estimadores sob modelos paramétricos gerais

De forma geral, um M-estimador $\widehat{\theta}_M$ para um parâmetro θ pode ser obtido de

$$\widehat{\theta}_M=\underset{\theta\,\in\,\Theta}{\operatorname{argmin}}\sum_{i=1}^{n}\rho(y_i,\theta),$$

ou como solução da equação

$$\sum_{i=1}^{n}\Psi(y_i,\widehat{\theta}_M)=0, \tag{2.3.17}$$

com $\rho(y_i,\theta)$ e $\Psi(y_i,\theta)$ funções dependentes de y e θ e diferenciáveis em θ. Além disso, para modelos de localização, tem-se que $\Psi(y_i,\widehat{\theta}_M)=\psi(y_i-\widehat{\theta}_M)$, e para modelos de escala tem-se que $\Psi(y_i,\widehat{\theta}_M)=\rho(y_i/\widehat{\theta}_M)-\delta$ (MARONNA et al., 2019). Esta forma de obtenção de M-estimadores é válida tanto para parâmetros de localização e escala, como também, para parâmetros de forma.

Para estabelecer a distribuição de probabilidades aproximada de um M-estimador, considere as definições e conceito a seguir.

Seja $F_n(z)$ a função de distribuição empírica (EDF, *Empirical Distribution Func-*

tion) de $y_1, y_2, \ldots, y_n$ definida por

$$F_n(z) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}(y_i < z).$$

Dessa forma, temos interesse em estimadores $\widehat{\theta}_M$ que dependam dos dados somente através de $F_n(z)$, isto é,

$$\widehat{\theta}_M(y_1, y_2, \ldots, y_n) = \widehat{\theta}_M(F_n),$$

os quais são chamados de M-estimadores funcionais. Também, temos interesse em M-estimadores funcionais que são consistentes, isto é, quando $n \to \infty$, $\widehat{\theta}_M(F_n) \to_p \theta$, em que "$\to_p$" significa convergência em probabilidade, ou seja, o M-estimador $\widehat{\theta}_M(F_n)$ converge em probabilidade para o verdadeiro θ.

Ainda, temos interesse em estimadores Fisher-consistentes. Dizemos que $\widehat{\theta}_M$ é Fisher-consistente se, quando substituímos a EDF F_n pela $F(y;\theta)$ populacional em $\widehat{\theta}_M$, obtemos que

$$\widehat{\theta}_M(F) = \theta, \ \forall\ \theta \in \Theta.$$

Nesse sentido, a Fisher-consistência de $\widehat{\theta}_M$ equivale a

$$\mathrm{E}_F\left(\Psi(y, \widehat{\theta}_M)\right) = 0.$$

Em outras palavras, a Fisher-consistência de $\widehat{\theta}_M$ implica que a equação de estimação (2.3.17) é não viciada. O conceito de Fisher-consistência é vastamente discutido na literatura de estimadores robustos pois existem diversos estimadores robustos obtidos de equações de estimação viciadas.

Supondo que o M-estimador $\widehat{\theta}_M$ é um estimador consistente e Fisher-consistente para θ, e a distribuição suposta para os dados F é adequada, mostra-se que a distribuição aproximada de $\widehat{\theta}_M$ é dada por

$$\widehat{\theta}_M \overset{n\to\infty}{\sim} N\left(\theta, \frac{\upsilon}{n}\right), \text{ com } \upsilon = \frac{\mathrm{E}_F\left(\Psi\left(y,\theta\right)^2\right)}{\left[\mathrm{E}_F\left(\Psi'\left(y,\theta\right)\right)\right]^2}, \tag{2.3.18}$$

em que $\Psi'(y,\theta)$ é a derivada da função $\Psi(y,\theta)$ com relação a θ e (υ/n) é a variância aproximada de $\widehat{\theta}_M$, uma vez que o M-estimador é assintoticamente normal. Se θ for um vetor de parâmetros, esse resultado é generalizável; para mais detalhes, veja Hampel et al. (2011).

2.3.5 Função de Influência

De acordo com Hampel et al. (2011), para a definição da função de influência (FI), considera-se $F_{h,y} = (1-h)F+h\Delta_y$ a função de distribuição acumulada contaminada devido a introdução de uma contaminação infinitesimal h no ponto y. A FI de um estimador de interesse $\widehat{\theta}_M$ em F é definida por

$$\begin{aligned} FI(y;\widehat{\theta}_M, F) &= \frac{\partial}{\partial h}\left[\widehat{\theta}_M(F_{h,y})\right]\Big|_{h=0} \\ &= \lim_{h\to 0}\frac{\widehat{\theta}_M(F_{h,y}) - \widehat{\theta}_M(F)}{h} \\ &= \lim_{h\to 0}\frac{\widehat{\theta}_M\left((1-h)F + h\Delta_y\right) - \widehat{\theta}_M(F)}{h}\ , \end{aligned} \tag{2.3.19}$$

em que $\widehat{\theta}_M(F)$ denota o estimador avaliado sob a função de distribuição F e Δ_y é a medida de probabilidade que coloca toda a massa em y.

Da expressão (2.3.19), nota-se que a FI descreve o efeito ou o resultado causado no estimador de interesse $\widehat{\theta}_M$ em decorrência de uma contaminação infinitesimal h no ponto y, a qual é padronizada pela massa de contaminação h. Logo, a FI demonstra o comportamento infinitesimal do valor assintótico do estimador $\widehat{\theta}_M$, expressando o viés assintótico causado pela contaminação no ponto y (HAMPEL et al., 2011).

Dessa maneira, diz-se que um estimador $\widehat{\theta}_M$ é qualitativamente robusto quando possuir FI limitada para todo y no suporte da distribuição postulada para os dados. Como a contaminação é muito pequena, tendendo a zero, se o estimador $\widehat{\theta}_M$ apresentar uma alta variação após esta contaminação, isto será um indicativo de que $\widehat{\theta}_M$ é muito sensível a pequenos incrementos h no ponto y.

Segundo Hampel et al. (2011), para o MLE, $\widehat{\theta}_{ML}$, a FI é dada por

$$FI(y;\widehat{\theta}_{ML}, F) = -K(\theta, F)^{-1}U(y,\theta)$$

em que $-K(\theta, F)$ é a informação de Fisher e $U(y,\theta)$ é a função escore. Como apenas a função do escore depende de y, que está sofrendo a contaminação, se a função escore não for limitada, a FI do estimador de máxima verossimilhança não será limitada também e, portanto, $\widehat{\theta}_{ML}$ não será robusto. Logo, para verificar se um estimador de máxima verossimilhança sob um módulo F é robusto, bastaria ver se sua função escore é limitada.

Segundo Maronna et al. (2019), a FI de um M-estimador $\widehat{\theta}_M$ como definido em (2.3.17) é dada pela seguinte expressão

$$FI(y;\widehat{\theta}_M, F) = -\frac{\Psi(y,\theta)}{B(\theta,\Psi,F)}\ , \tag{2.3.20}$$

em que

$$B(\theta,\Psi,F) = E_F\left(\Psi'(y,\theta)\right).$$

Desde que $B(\theta, \Psi, F)$ esteja bem definida, a função de influência dada em (2.3.20) é limitada se a função $\Psi(y, \theta)$ for limitada, o que implica que um M-estimador $\widehat{\theta}_M$ é robusto ou não sensível a pequenas contaminações h no ponto y. Nesse contexto, note que os estimadores do tipo Huber e *bisquare* de Tukey são robustos pois nota-se pelas Figuras 2 e 3 que as funções $\psi(y)$ são limitadas. Contudo, a função $\psi(y) \propto y$ para o MLE sob a distribuição normal padrão não é limitada, logo, a função de influência do MLE sob normalidade não é limitada e, portanto, o MLE sob normalidade não é robusto. Em contrapartida, para a distribuição da exponencial dupla, tem-se que $\psi(y) \propto sign(y)$ é limitada, logo, o MLE sob essa distribuição é robusto.

2.4 Regressão Linear Robusta

Nesta seção, juntaremos os conceitos vistos nas seções anteriores para desenvolver estimadores para os parâmetros da regressão linear que sejam robustos.

Nesse sentido, considerando o modelo de regressão linear (2.1.1), para $x_{i1}, \ldots, x_{ik}$ fixos, tem-se a função de densidade dos erros aleatórios ε_i igual a

$$\frac{1}{\sigma} f_0 \left(\frac{\varepsilon_i}{\sigma} \right),$$

com parâmetro de escala σ. Além disso, as n variáveis respostas y_i, $i = 1, \ldots, n$, são independentes, mas não identicamente distribuídas, uma vez que y_i tem função de densidade dada por

$$f(y_i; \beta, \sigma) = \frac{1}{\sigma} f_0 \left(\frac{y_i - x_i^\top \beta}{\sigma} \right),$$

em que $x_i^\top = (1, x_{i1}, \ldots, x_{ik})^\top$ é o vetor $(k+1)$ dimensional das variáveis explicativas para a i-ésima observação, e $\beta = (\beta_0, \beta_1, \ldots, \beta_k)$ é o vetor $(k+1)$ dimensional dos parâmetros ou coeficientes da regressão.

Sob o modelo de regressão (2.1.1), supondo inicialmente σ fixo, segue que a função de verossimilhança de β é

$$\mathrm{L}(\beta) = \frac{1}{\sigma^n} \prod_{i=1}^{n} f_0 \left(\frac{y_i - x_i^\top \beta}{\sigma} \right).$$

Sob essa perspectiva, para se encontrar o MLE, denotado por $\widehat{\beta}_{ML}$, busca-se primeiro o logaritmo da função de verossimilhança de β dado por

$$\ell(\beta) = \log \left(\frac{1}{\sigma^n} \right) + \log \left(\prod_{i=1}^{n} f_0 \left(\frac{y_i - x_i^\top \beta}{\sigma} \right) \right) = -n \log(\sigma) + \sum_{i=1}^{n} \log \left(f_0 \left(\frac{y_i - x_i^\top \beta}{\sigma} \right) \right),$$

em que $\sigma > 0$ é um valor fixo. Dessa forma, para se encontrar o MLE de β, deve-se maximizar o logaritmo da função de verossimilhança $\ell(\beta)$. Para tal, pode-se reescrever

$\ell(\beta)$ em termos do erro $\varepsilon_i = y_i - x_i^\top \beta$. Então, queremos obter o valor de β que maximiza

$$\ell(\beta) = \sum_{i=1}^{n} \log\left(f_0\left(\frac{\varepsilon_i}{\sigma}\right)\right) - n\log(\sigma).$$

Adotando-se $\rho(y) = -\log(f_0(y))$, como feito na Subseção 2.3.1, tem-se que

$$-\ell(\beta) = \sum_{i=1}^{n} \rho\left(\frac{\varepsilon_i}{\sigma}\right) + n\log(\sigma).$$

Por conseguinte, o MLE para β é calculado de $\widehat{\beta}_{ML} = \underset{\beta \in \mathbb{R}}{\operatorname{argmin}}[-\ell(\beta)]$. Com isso, para minimizar $-\ell(\beta)$, deve-se derivar o negativo dessa função com relação a β e igualar a zero, portanto, temos

$$\begin{aligned}
\frac{\partial}{\partial\beta}\left[-\ell(\beta)\right]\Big|_{\beta=\widehat{\beta}_{ML}} &= 0 \\
\sum_{i=1}^{n} \frac{1}{f_0\left(\frac{y_i - x_i^\top \widehat{\beta}_{ML}}{\sigma}\right)} f_0'\left(\frac{y_i - x_i^\top \widehat{\beta}_{ML}}{\sigma}\right) \cdot \left(\frac{-x_i}{\sigma}\right) &= 0 \\
\sum_{i=1}^{n} \frac{f_0'\left(\frac{\widehat{\varepsilon}_i}{\sigma}\right)}{f_0\left(\frac{\widehat{\varepsilon}_i}{\sigma}\right)} \cdot \left(\frac{-x_i}{\sigma}\right) &= 0 \\
\left(\frac{1}{\sigma}\right) \sum_{i=1}^{n} \frac{f_0'\left(\frac{\widehat{\varepsilon}_i}{\sigma}\right)}{f_0\left(\frac{\widehat{\varepsilon}_i}{\sigma}\right)} \cdot (-x_i) &= 0 \\
\sum_{i=1}^{n} -\frac{f_0'\left(\frac{\widehat{\varepsilon}_i}{\sigma}\right)}{f_0\left(\frac{\widehat{\varepsilon}_i}{\sigma}\right)} \cdot x_i &= 0 \cdot (\sigma) \\
\sum_{i=1}^{n} \psi\left(\frac{\widehat{\varepsilon}_i}{\sigma}\right) \cdot x_i &= 0, \qquad (2.4.1)
\end{aligned}$$

em que $\psi(y) = \rho'(y)$.

Supondo uma função de densidade $f_0(y)$ normal padrão, como definido pela equação (2.3.3), com os cálculos análogos ao da Subseção 2.3.1, pode-se provar que $\rho\left(\frac{\varepsilon_i}{\sigma}\right) \propto \left(\frac{\varepsilon_i}{\sigma}\right)^2 \propto (\varepsilon_i)^2$, para $\sigma > 0$ uma constante fixa. Daí, o MLE $\widehat{\beta}_{ML}$ de β será

$$\widehat{\beta}_{ML} = \underset{\beta \in \mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{n} \varepsilon_i^2,$$

o qual coincide com o LSE, isto é, $\widehat{\beta}_{ML} = \widehat{\beta}$ definido em (2.2.1).

Para o caso da função de densidade $f_0(y)$ como a exponencial dupla, definida na equação (2.3.4), analogamente ao calculado na Subseção 2.3.1, tem-se que $\rho\left(\frac{\varepsilon_i}{\sigma}\right) \propto \left|\frac{\varepsilon_i}{\sigma}\right| \propto |\varepsilon_i|$, com $\sigma > 0$ uma constante fixa. Portanto, o MLE $\widehat{\beta}_{ML}$ de β sob uma função de densidade exponencial dupla será

$$\widehat{\beta}_{ML} = \underset{\beta \in \mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{n} |\varepsilon_i|.$$

Neste caso, o estimador $\widehat{\beta}_{ML}$ é obtido minimizando a soma dos erros absolutos da regressão. Diferente do LSE, este estimador não possui forma fechada e deve ser obtido utilizando algum método de otimização não linear.

Na prática, σ é um valor desconhecido e, portanto, deve-se estimá-lo. Supondo que $\widehat{\sigma}_M$ é obtido anteriormente do procedimento inferencial para β, segundo Maronna et al. (2019), podemos definir os M-estimadores de regressão $\widehat{\beta}_M$ através de

$$\widehat{\beta}_M = \underset{\beta \in \mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{n} \rho\left(\frac{\varepsilon_i}{\widehat{\sigma}_M}\right). \tag{2.4.2}$$

em que $\widehat{\sigma}_M$ é um M-estimador de escala previamente calculado. Analogamente ao que foi feito para se chegar a expressão (2.4.1), também podemos obter os M-estimadores de regressão $\widehat{\beta}_M$ a partir do sistema de equações

$$\sum_{i=1}^{n} \psi\left(\frac{\widehat{\varepsilon}_i}{\widehat{\sigma}_M}\right) \cdot x_i \ = 0, \tag{2.4.3}$$

em que $\psi(y) = \rho'(y)$. Vale salientar que a função $\rho(y)$ é uma função que obedece as condições C_1 a C_4, descritas na Subseção 2.3.1, e a função $\psi(y)$ também é equivalente a definida nessa mesma subseção.

Para algumas situações, o estimador $\widehat{\sigma}_M$ é calculado previamente através de algum M-estimador robusto de escala. Contudo, $\widehat{\sigma}_M$ também pode ser calculado simultaneamente com $\widehat{\beta}_M$, seguindo a metodologia já discutida na Subseção 2.3.3. Considerando este cenário, deve-se resolver de forma simultânea o sistema de equações de estimação definido em (2.4.3) e o sistema de equações de estimação

$$\left(\frac{1}{n}\right) \sum_{i=1}^{n} \rho_{scale}\left(\frac{y_i - {x_i}^{\top}\widehat{\beta}_M}{\widehat{\sigma}_M}\right) \ = \delta, \tag{2.4.4}$$

em que $0 < \delta \leq 1$.

Neste trabalho, consideraremos os M-estimadores de regressão para o vetor de coeficientes β e para o parâmetro σ^2 obtidos de forma simultânea considerando as funções do tipo Huber e bisquare de Tukey para a função $\rho(y)$ em (2.4.2), e a função $\rho_{scale}(y) = \mathrm{I}(|y| > c)$ em (2.4.4). Estes M-estimadores não possuem forma fechada e devem ser obtidos utilizando algum método de otimização não linear.

Para realizar testes de hipóteses sobre os coeficientes da regressão linear baseando-se nas M-estimativas, podemos desenvolver um teste de hipóteses semelhante ao discutido na Seção 2.2. Sob algumas condições, vimos que os M-estimadores possuem uma distribuição aproximadamente normal (Subseção 2.3.4). Sendo assim, podemos considerar

que

$$\widehat{\beta}_{Mj} \overset{n\to\infty}{\sim} N\left(\beta_j, \sigma^2_{Mj}\right),$$

em que $\widehat{\beta}_{Mj}$ é o M-estimador de β_j e σ^2_{Mj} é a variância assintótica do M-estimador $\widehat{\beta}_{Mj}$ com $j = 0, 1, \ldots, k$ que pode ser obtida de forma análoga a expressão (2.3.18). Para mais detalhes sobre o cálculo da variância assintótica de M-estimadores sob o contexto de modelos de regressão lineares veja Maronna et al. (2019).

Considerando as hipóteses $H_0 : \beta_j = b_j$ contra $H_1 : \beta_j \neq b_j$, podemos definir a estatística

$$t_0 = \frac{\widehat{\beta}_{Mj} - b_j}{\sqrt{\widehat{\sigma}^2_{Mj}}} \underset{n\to\infty}{\overset{H_0}{\sim}} N\left(0, 1\right), \tag{2.4.5}$$

em que $\widehat{\sigma}^2_{Mj}$ é a variância assintótica estimada do M-estimador $\widehat{\beta}_{Mj}$. Assim, rejeita-se H_0 ao nível de significância de $\alpha \cdot 100\%$, com $0 < \alpha < 1$, se $t_0 < -z_c$ ou $t_0 > z_c$, com z_c obtido tal que $P(Z > z_c) = \frac{\alpha}{2}$ com $Z \sim N(0, 1)$. Ainda, em termos do p-valor, rejeita-se H_0 se p-valor $= 2P(Z > |t_0|) \leq \alpha$. Note este p-valor é aproximado pois é obtido baseado na distribuição aproximada do M-estimador $\widehat{\beta}_{Mj}$. Sendo assim, torna-se importante certo cuidado na utilização dos mesmos, uma vez que, dado certo valor do tamanho amostral n, podemos não ter uma boa aproximação da distribuição verdadeira da estatística de teste t_0 sob H_0.

De forma análoga, um intervalo para o coeficiente β_j com coeficiente de confiança $(1 - \alpha) \cdot 100\%$ aproximado pode ser obtido de

$$\widehat{\beta}_{Mj} \pm z_c \cdot \sqrt{\widehat{\sigma}^2_{Mj}}.$$

Novamente, vale ressaltar que este intervalo de confiança para β_j é aproximado e deve ser utilizado com cuidado.

3 Aplicações

Para ilustrar o comportamento de alguns dos estimadores discutidos no Capítulo 2, e a respectiva utilização destes em análise de dados reais, neste capítulo, serão aplicadas técnicas computacionais em `R`, por meio do ambiente de desenvolvimento integrado (IDE) RStudio, criado pela empresa Posit. O IDE que é aplicado para a linguagem de programação `R` está disponível em https://posit.co/products/open-source/rstudio/. Serão considerados dois conjuntos de dados reais.

Para realizar o ajuste de modelos de regressão linear sob os métodos de estimação discutidos foram utilizados principalmente, os seguintes pacotes e funções do `R`:

- Pacote `stats` - função `lm()`: ajuste dos dados sob modelo de regressão linear via o método dos mínimos quadrados;
- Pacote `MASS` - função `rlm()`: ajuste dos dados sob modelo de regressão linear via os métodos de estimação robustos.

3.1 Aplicação 1 - Experimento com ratos

Nessa primeira aplicação, utilizaremos um conjunto de dados disponível no pacote do `R` chamado `RobStatTM`. Estes dados também são discutidos e apresentados por Maronna et al. (2019, p.87). Esta aplicação foi brevemente discutida no Capítulo 1 deste trabalho.

Os dados são referentes a um experimento sobre a velocidade de aprendizagem com que 16 ratos atravessavam uma caixa, no qual foram registrados os tempos em diversas tentativas. Se o rato levasse mais de 5 segundos para realizar a travessia, este recebia um choque elétrico antes da próxima tentativa. As variáveis em estudo são o número de choques recebidos por cada rato e o tempo médio de travessia das tentativas. Assim, para entender como o número de choques levados pelo rato afeta o tempo médio de travessia, consideraremos a regressão linear simples definida por

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \ , \ i \in \{1, 2, \ldots, 16\}, \tag{3.1.1}$$

em que a variável resposta y_i é o tempo médio de travessia, em segundos, de todas as tentativas que o i-ésimo rato fez para atravessar a caixa, e a variável explicativa x_i é o número de choques no total levado pelo i-ésimo rato. Supõe-se ainda que $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$, o que implica que $y_i \overset{ind}{\sim} N(\mu_i, \sigma^2)$.

Inicialmente, foi feita uma análise unidimensional dos possíveis *outliers* que estariam presentes na variável resposta y por meio do método de Tukey (TUKEY et al.,

1977), no qual considera-se que

$$\begin{cases} \text{O.I.} < Q_1 - K \cdot (Q_3 - Q_1) \cdot F; \\ \text{O.S.} > Q_3 + K \cdot (Q_3 - Q_1) \cdot F; \end{cases} \tag{3.1.2}$$

em que Q_1 e Q_3 são o primeiro e o terceiro quartis amostrais, respectivamente, K é o fator de Tukey, comumente fixado em $K = 1,5$, F é um valor que pode ser fixado ou depender de outros fatores como, por exemplo, a assimetria dos dados, e O.I. e O.S. são os valores de y que são identificados como *outliers* inferiores e superiores, respectivamente. Para esta aplicação, o método de Tukey et al. (1977) foi realizado tomando $F = 1$ e $K = 1,5$. Identificamos três possíveis *outliers* em y, sendo estes as observações associadas aos ratos 1, 2 e 4. Note que estas são as observações previamente identificadas na Figura 1.

O modelo de regressão (3.1.1) foi ajustado, via linguagem estatística `R`, sob quatro cenários:

- via mínimos quadrados considerando os dados completos (LS);
- via mínimos quadrados para os dados incompletos excluindo os *outliers* identificados via o método de Tukey e via o gráfico de dispersão na Figura 1 (LS-);
- via o M-estimador com a função $\psi(y)$ do tipo Huber considerando os dados completos (Huber);
- via o M-estimador com a função $\psi(y)$ do tipo *bisquare* de Tukey considerando os dados completos (Bisquare).

As estimativas associadas ao modelo de regressão linear simples ajustado sob cada método de estimação estão apresentadas na Tabela 1.

Tabela 1: Estimativas obtidas pelos métodos de estimação considerados sob o modelo de regressão linear simples definido em (3.1.1).

Métodos de estimação	Estimativas dos parâmetros	Erro padrão das estimativas	Estatísticas do teste	p-valores do teste
LS	$\widehat{\beta}_0 = 10,4846$	$1,0776$	$9,7300$	$< 0,001$
	$\widehat{\beta}_1 = -0,6129$	$0,1224$	$-5,0070$	$0,0002$
LS-	$\widehat{\beta}_0 = 7,2152$	$0,7636$	$9,4490$	$< 0,001$
	$\widehat{\beta}_1 = -0,3198$	$0,0785$	$-4,0740$	$0,0018$
Huber	$\widehat{\beta}_{M0} = 9,8174$	$0,8777$	$11,1860$	$< 0,001$
	$\widehat{\beta}_{M1} = -0,5719$	$0,0997$	$-5,7366$	$< 0,001$
Bisquare	$\widehat{\beta}_{M0} = 7,9164$	$0,3202$	$24,7219$	$< 0,001$
	$\widehat{\beta}_{M1} = -0,4137$	$0,0364$	$-11,3738$	$< 0,001$

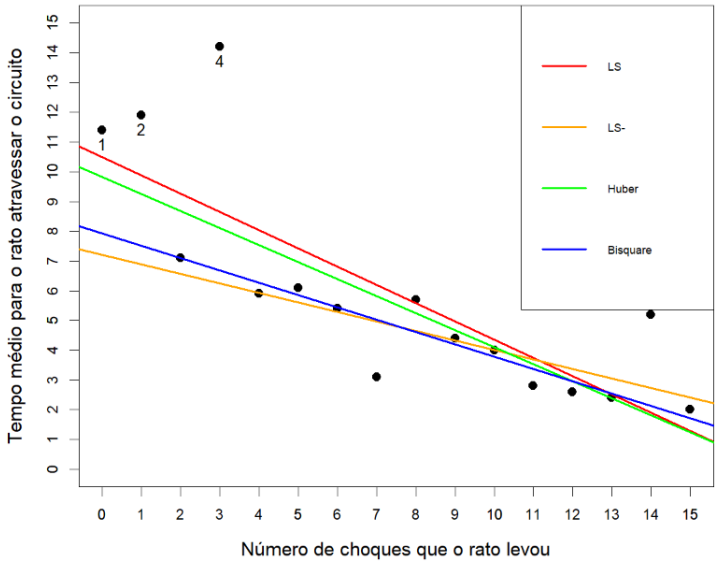
A partir das informações apresentadas na Tabela 1, pode-se observar que a estimativa $\widehat{\beta}_1$ via o LSM para os dados completos é $91,65\%$ maior que essa mesma estimativa via o LSM aplicado aos dados sem *outliers*. Apesar do p-valor associado ao teste H_0: $\beta_1 = 0$ sob o LSM sem *outliers* ter ficado maior do que o obtido via o LSM com dados completos, fixando um nível de significância de 10%, em ambos os casos β_1 se manteve estatisticamente diferente de zero. Com relação ao intercepto β_0, observa-se que este se manteve estatisticamente significante para os dois ajustes sob o LSM, com um aumento, em sua estimativa, de $45,31\%$ quanto a estimativa desse parâmetro obtida via método robusto sem os *outliers*. Portanto, notamos que a reta de regressão ajustada via o LSM é muito afetada na presença das observações referentes aos Ratos 1, 2 e 4.

Comparando as estimativas obtidas via o M-estimador *bisquare* com as estimativas obtidas via o LSM sem os *outliers*, obtemos que as estimativas de β_0 e β_1 via *bisquare* são 9,7% e 29,4% maiores, respectivamente. Assim, via o M-estimador *bisquare*, obtivemos estimativas mais próximas daquelas obtidas via o LSM sem as observações discrepantes. Vale ressaltar que as estimativas obtidas via o método robusto não necessariamente devem coincidir com as estimativas obtidas pelo método de estimação clássico sem os *outliers*. Isto de deve ao fato de que, mesmo sendo observações discrepantes, o procedimento inferencial robusto atribui algum peso (que pode ser zero ou não) a estas observações. Comparando as estimativas obtidas via o M-estimador de Huber com as estimativas obtidas via o LSM sem os *outliers*, obtemos que as estimativas de β_0 e β_1 via Huber são 36,1% e 78,8% maiores, respectivamente. Assim, nota-se que as estimativas obtidas pelo método via função Huber não foram próximas àquelas obtidas via LSM sem *outliers*, e portanto, este método não se mostrou robusto para o cenário estudado.

Ainda, nota-se que os coeficientes da regressão foram estatisticamente significantes sob todos os modelos de regressão ajustados para um nível de significância de 10%.

Para ilustrar o desempenho dos estimadores considerados, apresentamos a Figura 4 que contém o gráfico de dispersão para os dados com as diferentes retas de regressão ajustadas de acordo com cada método de estimação. Analisando-a, nota-se que as retas obtidas via os métodos de estimação LS e LS- se encontram nos extremos superior e inferior, respectivamente, devido a influência dos *outliers* 1, 2 e 4, os quais alavancam a reta de regressão linear simples para cima quando incluídos no modelo. Ademais, fica claro como o método de estimação com a função $\psi(y)$ do tipo *bisquare* robustifica melhor o processo de estimação para o modelo de regressão linear simples, que implicitamente atribui pesos mais baixos para esses *outliers* do que com a função $\psi(y)$ do tipo Huber, ajustada mais próxima da reta obtida via o LSM com os dados completos. Ainda, pode-se concluir que, conforme os ratos levavam mais choques, fazendo mais tentativas, menor ficava o seu tempo médio para atravessar a caixa. A partir das M-estimativas do método *bisquare*, a cada choque levado, o tempo para travessia diminui, em média, 0,41 segundos.

Figura 4: Gráfico de dispersão do tempo médio para travessia versus número de choques levados pelo rato juntamente com retas de regressão ajustadas via cada método de estimação.

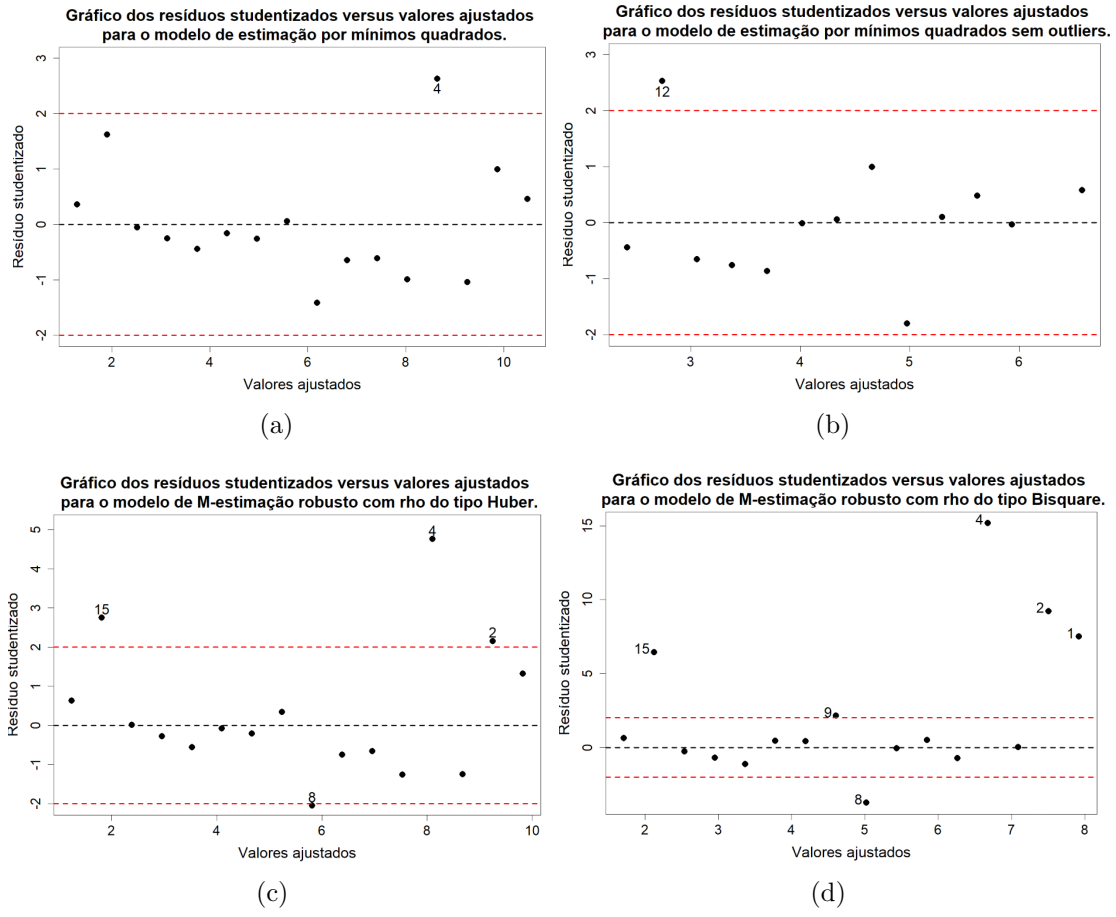


Após o ajuste de um modelo de regressão linear se faz necessário uma análise de resíduos com o objetivo de verificar se as suposições apresentadas na Subseção 2.1.3 são válidas. Sendo assim, na Figura 5 apresentamos os gráficos dos resíduos studentizados internos versus os valores ajustados sob cada método de estimação. Aqui, vale pontuar que sob as estimativas robustas, em tese, podemos calcular os resíduos studentizados internos t_i e os resíduos externos t_i^* discutidos na Seção 2.2. Entretanto, o cálculo de forma rápida do resíduo t_i^* e sua respectiva distribuição de probabilidades são válidos apenas sob as estimativas obtidas via o LSM. Portanto, para realizar as comparações entre os resíduos obtidos via todos os métodos de estimação utilizamos o resíduo studentizado interno t_i. Ainda, lembre-se que esperamos que os resíduos t_i possuam média zero e variância constante igual a 1. A partir destes gráficos, nós observamos fortes indícios de que a suposição de homoscedasticidade não é adequada. Entretanto, sob o ajuste LS com dados completos e o ajuste via o M-estimador de Huber vemos que os resíduos não parecem estar distribuídos de forma parecida em torno da média zero. Este último comportamento fica evidenciado ao observar Figura 5(a) e Figura 5(c). Vale ressaltar que estas conclusões devem ser tomadas com cautela pois estamos considerando um tamanho amostral de apenas $n = 16$ observações que dificulta a análise gráfica.

Em particular, observamos que alguns resíduos assumem valores discrepantes sob o ajuste via os M-estimadores de Huber e *bisquare*. Isto se deve ao fato de que o método de estimação robusta objetiva ajustar bem a maior parte dos dados e aquelas observações que são identificadas pelo procedimento como atípicas recebem peso baixo e, portanto, não ficarão bem ajustadas. A partir da Figura 5(d), percebe-se que o M-estimador *bisquare* identifica como observações influentes àquelas associadas aos Ratos 1, 2, 4, 8 e 15. Assim, este método de estimação além de dar pouca contribuição para os *outliers* identificados anteriormente, também atribui menor peso para outras duas observações (8

e 15). Portanto, observamos que os gráficos dos resíduos studentizados sob estimadores robustos além de ajudar na avaliação das suposições, estes nos ajudam na identificação de observações que receberam peso menor no processo de estimação devido sua influência desproporcional.

Figura 5: Gráficos dos resíduos studentizados internos versus os valores ajustados da regressão sob cada método de estimação.



3.2 Aplicação 2 - Custo de aeronaves monomotoras

Para esta segunda aplicação, usaremos um conjunto de dados disponibilizado no pacote do `R` chamado `robustbase` e apresentados por Rousseeuw e Leroy (2005, p.154).

Os dados são referentes a 23 aeronaves monomotoras construídas ao longo dos anos de 1947 a 1979, dados estes que foram obtidos da fonte Escritório de Pesquisa Naval dos Estados Unidos. Neste caso, a variável resposta y é o custo da aeronave (em unidades de US$ 100.000) e as variáveis explicativas são a proporção de aspecto da aeronave (x_1), a relação de sustentação/arrasto (x_2), o peso do avião em libras (x_3), e o impulso máximo que ele alcança (x_4). Consideramos a modelagem desses dados por meio de uma regressão linear múltipla, ou seja,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i \ , \ i \in \{1, 2, \ldots, 23\}, \tag{3.2.1}$$

em que a variável resposta y_i é o custo da i-ésima aeronave $x_{i1},\ x_{i2},\ x_{i3}$ e x_{i4} são os valores das variáveis explicativas para a i-ésima aeronave. Ainda, asssumimos que $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$, o que implica que $y_i \overset{ind}{\sim} N(\mu_i, \sigma^2)$.

De forma análoga ao que foi feito na Aplicação 1, realizamos uma análise unidimensional para identificação de possíveis *outliers* com relação a y por meio do método de Tukey et al. (1977). Identificou-se previamente que a Aeronave 22 é tida como *outlier*. Adicionalmente, observando os gráficos de dispersão de y versus x_1 e x_2 apresentados na Figuras 6 e 7, respectivamente, nota-se um comportamento atípico da Aeronave 16. Então, consideraremos as observações relacionadas as Aeronaves 16 e 22 como dois possíveis *outliers* em y. Conforme feito na aplicação anterior, serão realizados o ajuste do modelo de regressão linear mútiplo (3.2.1) considerando os quatro cenários explicitados na Seção 3.1: LS, LS-, Huber e Bisquare.

Aplicando-se a modelagem de regressão linear múltipla aos dados, via linguagem estatística `R`, obtivemos as estimativas dos parâmetros do modelo (3.2.1) ajustado sob cada método de estimação, conforme apresentado na Tabela 2.

Tabela 2: Estimativas obtidas pelos métodos de estimação considerados sob o modelo de regressão linear múltiplo definido em (3.2.1).

Métodos de estimação	Estimativas dos parâmetros	Erro padrão das estimativas	Estatísticas do teste	p-valores do teste
LS	$\widehat{\beta}_0 = -3,7914$	$10,1157$	$-0,3750$	$0,7122$
	$\widehat{\beta}_1 = -3,8529$	$1,7630$	$-2,1850$	$0,0423$
	$\widehat{\beta}_2 = \ \ 2,4883$	$1,1868$	$2,0970$	$0,0504$
	$\widehat{\beta}_3 = \ \ 0,0035$	$0,0005$	$7,3050$	$< 0,001$
	$\widehat{\beta}_4 = -0,0020$	$0,0005$	$-3,9180$	$0,0010$
LS-	$\widehat{\beta}_0 = \ \ 9,5007$	$5,5775$	$1,703$	$0,1078$
	$\widehat{\beta}_1 = -3,0488$	$0,9191$	$-3,317$	$0,0044$
	$\widehat{\beta}_2 = \ \ 1,2100$	$0,6492$	$1,864$	$0,0808$
	$\widehat{\beta}_3 = \ \ 0,0014$	$0,0004$	$3,519$	$0,0029$
	$\widehat{\beta}_4 = -0,0006$	$0,0003$	$-1,691$	$0,1102$
Huber	$\widehat{\beta}_{M0} = -1,2850$	$8,6035$	$-0,1494$	$0,8813$
	$\widehat{\beta}_{M1} = -3,4214$	$1,4994$	$-2,2818$	$0,0225$
	$\widehat{\beta}_{M2} = \ \ 2,2160$	$1,0093$	$2,1955$	$0,0281$
	$\widehat{\beta}_{M3} = \ \ 0,0029$	$0,0004$	$7,2207$	$< 0,001$
	$\widehat{\beta}_{M4} = -0,0016$	$0,0004$	$-3,6940$	$< 0,001$
Bisquare	$\widehat{\beta}_{M0} = \ \ 8,7164$	$6,5184$	$1,3372$	$0,1812$
	$\widehat{\beta}_{M1} = -3,1804$	$1,1361$	$-2,7995$	$0,0051$
	$\widehat{\beta}_{M2} = \ \ 1,3792$	$0,7647$	$1,8035$	$0,0713$
	$\widehat{\beta}_{M3} = \ \ 0,0016$	$0,0003$	$5,1288$	$< 0,001$
	$\widehat{\beta}_{M4} = -0,0007$	$0,0003$	$-2,2088$	$0,0272$

Analisando os resultados apresentados na Tabela 2, nota-se que, para um nível de significância de 10%, alguns coeficientes da regressão foram estatisticamente iguais a zero, pois possuem p-valores que não rejeitam H_0: $\beta_j = 0$. Por exemplo, observe os p-valores associados ao coeficiente de β_0 obtidos a partir dos quatro métodos de estimação. Todos são maiores do que 0,10. Assim, a comparação entre estimativas de β_0 de acordo com os métodos de estimação considerados não é relevante, tendo em vista que podemos afirmar que β_0 é estatisticamente igual a zero sob os quatro cenários.

Comparando a estimativa do coeficiente β_2 obtida via o LSM com todas as observações com relação a estimativa obtida via LSM sem os *outliers*, pode-se notar que a estimativa $\widehat{\beta}_2$ obtida do primeiro é 105,64% maior que a obtida do segundo. Ademais, a estimativa de β_2 calculada via o M-estimador de Huber é 83,14% maior que para o método de estimação LSM sem os *outliers*, enquanto que a estimativa do β_2 via o M-estimador *bisquare* é apenas 13,98% maior do que esta última. Portanto, via o M-estimador *bisquare*, obteve-se uma estimativa de β_2 mais próxima da obtida via o LSM sem as observações discrepantes no modelo, o que sugere que este método de estimação robusta atribuiu um peso menor para esses *outliers* durante o processo inferencial dos parâmetros.

Observando as estimativas obtidas para o coeficiente β_3, que é estatisticamente diferente de zero para um nível de significância de 10% sob a regressão ajustada via todos os diferentes métodos de estimação, notamos que $\widehat{\beta}_3$ via LSM com todas as observações é 150% maior do que a estimativa calculada via esse método sem os *outliers*. Ainda, a estimativa de β_3 via o M-estimador de Huber é 107,14% maior que a obtida via LSM sem os *outliers*. Em contrapartida, a M-estimativa via *bisquare* de β_3 é apenas 14,28% maior do que este último método de estimação clássico sem as observações discrepantes.

Portanto, nota-se que, no geral, as estimativas dos parâmetros do modelo de regressão obtidas via o método de estimação robusto *bisquare* são mais próximas àquelas calculadas via LSM sem os *outliers*, ou seja, observa-se que o método de M-estimação via função *bisquare* robustificou o procedimento de estimação. Esta mesma conclusão não pode ser feita sobre o método de M-estimação via função de Huber, uma vez que este último apresentou desempenho próximo ao LSM com os dados completos.

Para ilustrar graficamente as diferenças entre os ajustes dos modelos de regressão via os diferentes métodos de estimação, nas Figuras 6 e 7 construímos os gráficos de dispersão da variável resposta y versus as covariáveis x_1 e x_3 juntamente com as retas de regressão ajustadas. Vale ressaltar que para construir as retas de regressão ajustadas para cada método de estimação, as covariáveis não consideradas nos eixo das abcissas do gráficos de dispersão foram fixadas em seus valores médios amostrais.

Figura 6: Custo da aeronave versus a proporção de seu aspecto (x_1), com as retas ajustadas por cada método de estimação.

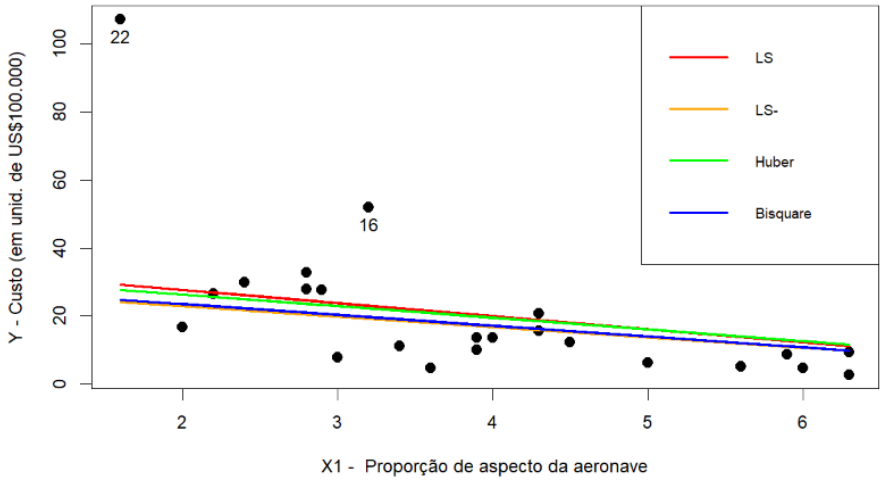
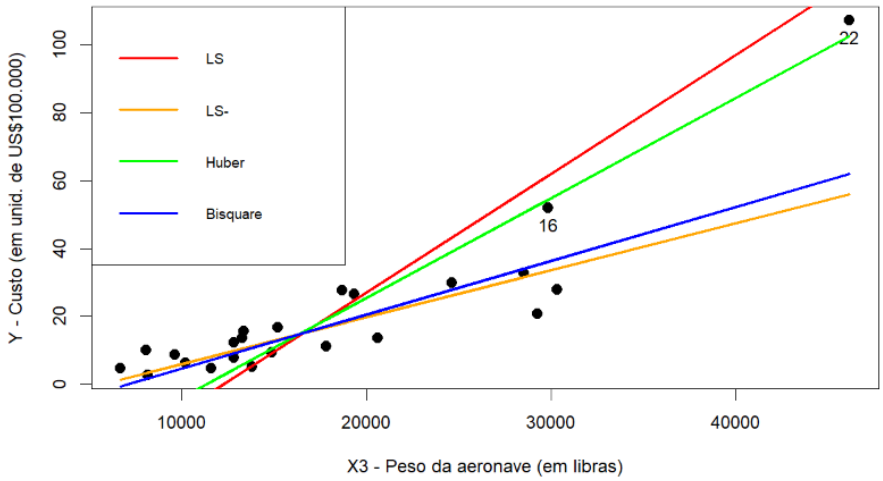


Figura 7: Custo da aeronave versus seu peso (x_3), com as retas ajustadas por cada método de estimação.

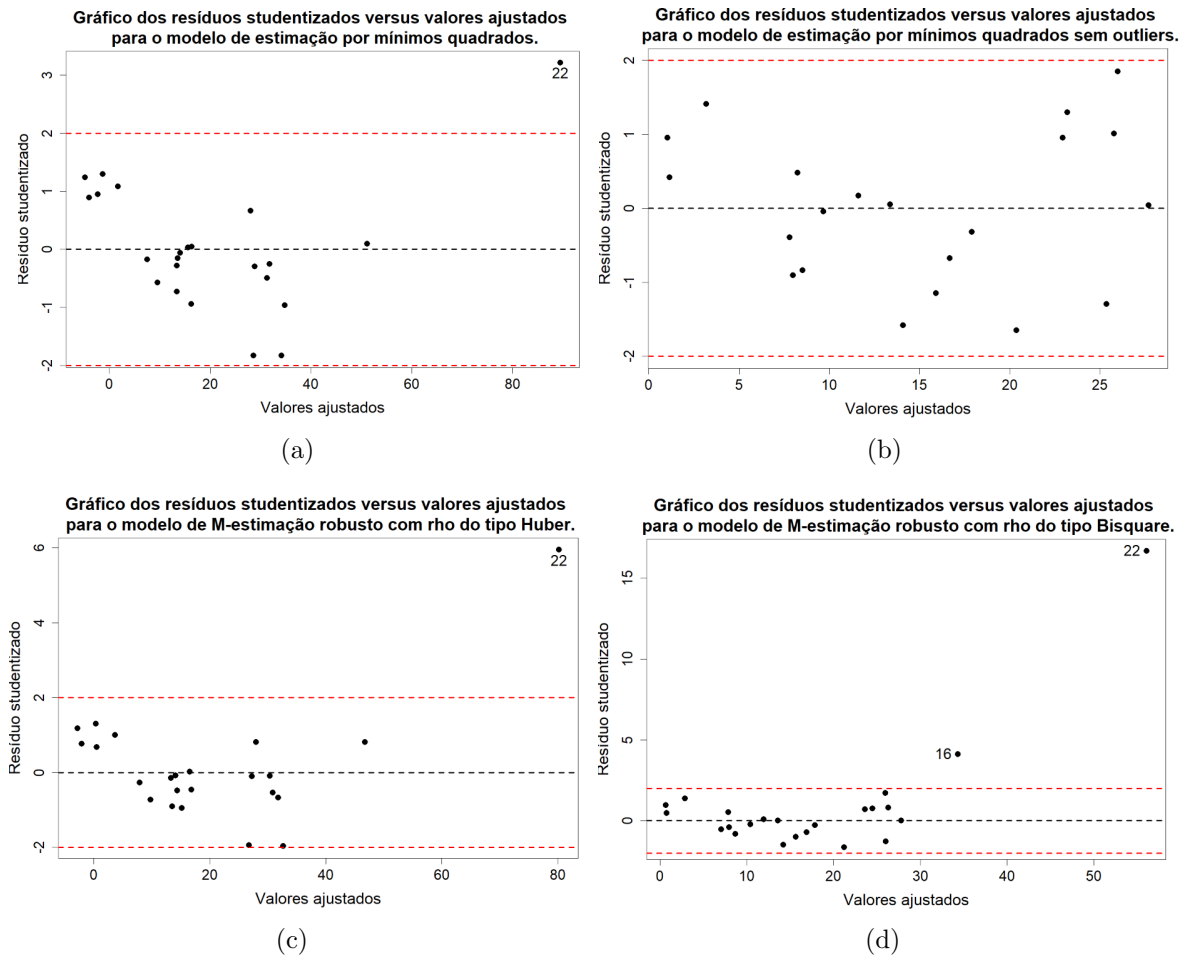


A partir das Figuras 6 e 7, observamos comportamentos distintos para as retas de regressão ajustadas a partir de cada método de estimação. Notamos o mesmo padrão discutido na Aplicação 1, isto é, as retas obtidas via os métodos de estimação LS e LS- nos extremos superior e inferior, respectivamente, devido a influência dos *outliers* 16 e 22, os quais alavancam a reta de regressão ajustada pelo LSM. Além disso, fica evidente como o método de M-estimação com a função *bisquare* produz um melhor ajuste a maioria dos dados quando comparado ao M-estimador com a função do tipo Huber. Esta última possui uma inclinação muito próximo a reta LS com dados completos, o que indica maior influência das observações discrepantes sobre a respectiva reta de regressão ajustada. Ainda, vemos que conforme a proporção de aspecto da aeronave (x_1) aumenta, o seu custo médio diminui, enquanto para valores maiores de peso da aeronave (x_3), esse custo médio é maior.

Por fim, avaliaremos o comportamento dos resíduos studentizados sob os modelos de regressão ajustados. Para tanto, inicialmente construimos os gráficos dos resíduos

studentizados internos para cada método de estimação que estão apresentados na Figura 8. A partir destes gráficos observamos que o ajuste via LSM sem os *outliers* e o ajuste via o M-estimador *bisquare* produzem um melhor ajuste comparado aos ajustes obtidos via LSM com dados completos e M-estimador de Huber. Ainda, observamos que o método de estimação com a função *bisquare* atribui pesos menores para as observações atípicas 16 e 22, desde que estas observações possuem resíduos discrepantes.

Figura 8: Gráficos dos resíduos studentizados internos versus os valores ajustados da regressão sob cada método de estimação.



Para finalizar esta análise, na Figura 9 estão apresentados os gráficos de probabilidade normal dos resíduos studentizados com envelope simulado sob cada regressão ajustada. Este tipo de gráfico é muito útil para avaliar a suposição de normalidade do erro aleatório. Esta suposição é importante para que possamos utilizar seguramente os p-valores associados aos testes de significância dos coeficientes. Se os resíduos studentizados se encontrarem dentro do envelope simulado sem padrão sistemático, teremos indícios de que a suposição de normalidade está adequada. É importante pontuar que esta avaliação não é a mesma quando o gráfico é construído sob estimativas obtidas a partir de uma inferência robusta. O objetivo da estimação robusta é modelar bem a maior parte dos dados, exceto observações discrepantes que possuem influência desproporcional no

procedimento inferencial. Assim, sob a inferência robusta, esperamos que a maior parte dos resíduos se encontrem dentro do envelope simulado sem padrão sistemático, porém não é necessário que todas as observações estejam dentro destes limites. Na verdade, as observações atípicas devem ser destacadas neste gráfico pois devem possuir resíduo discrepante.

A partir do gráfico sob a estimação via o LSM com dados completos e M-estimação de Huber, não temos fortes indícios contra a suposição de normalidade dos erros. A observação 22 é destacada em ambos os gráficos. Nota-se que a maioria dos resíduos obtidos via o método de estimação com a função *bisquare* se encontram dentro do envelope simulado, enquanto que os resíduos das observações 16 e 22 estão claramente destacadas. Para observar isso mais detalhadamente, contruimos a versão ampliada deste gráfico que está apresentada na Figura 10. Portanto, o ajuste da regressão linear sobre o M-estimador *bisquare* produziu resíduos studentizados que indicam adequação da suposição de normalidade para a maior parte dos dados, exceto as observações 16 e 22 que receberam peso baixo no procedimento de estimação.

Figura 9: Gráficos de probabilidade normal dos resíduos studentizados com envelope simulado sob cada método de estimação.

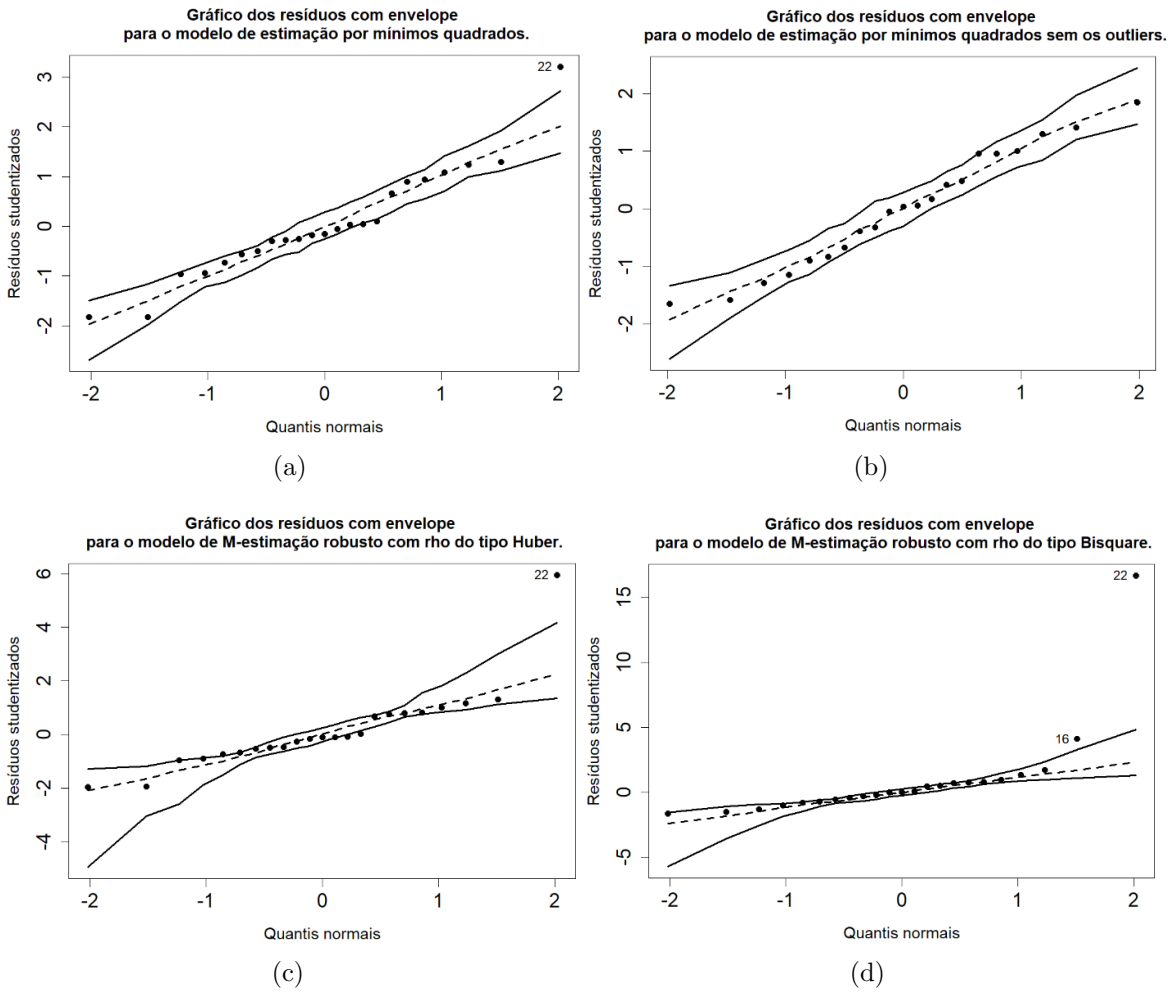
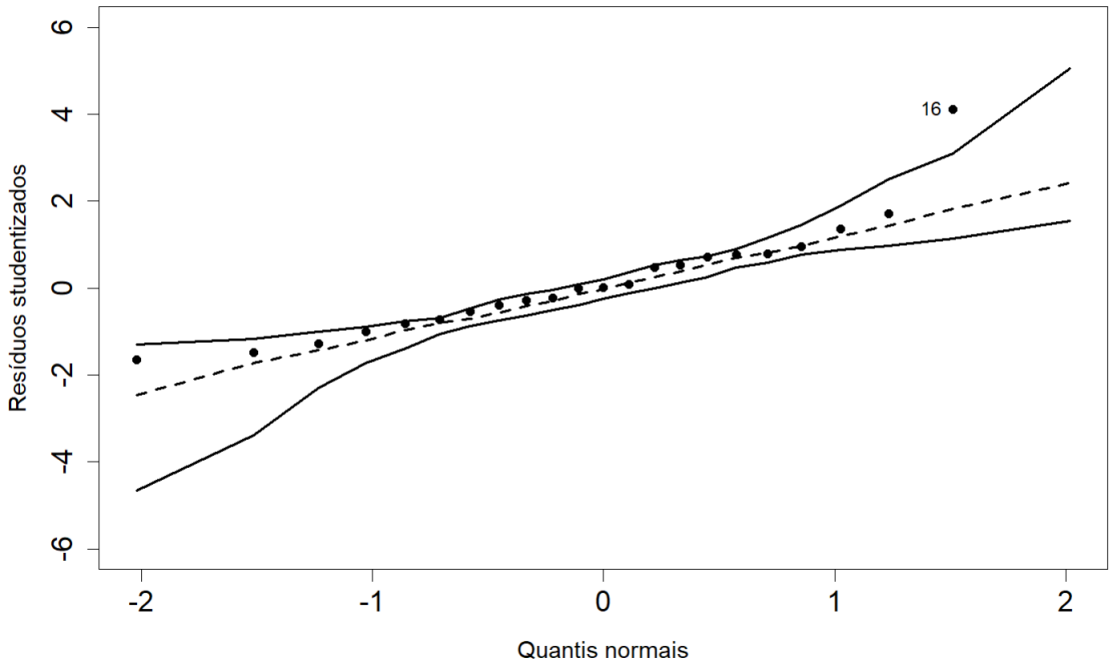


Figura 10: Versão ampliada do gráfico de probabilidade normal dos resíduos studentizados com envelope simulado para os resíduos sob método robusto do tipo *bisquare*.



4 Considerações Finais

O presente trabalho dedicou-se a fazer uma introdução à regressão linear robusta, voltada para o contexto de M-estimadores sob modelos de localização e escala, os quais possuem boas propriedades como Fisher-consistência e normalidade assintótica. Foram aplicados métodos estatísticos robustos na análise de dados sob modelos de regressão linear simples e múltiplo, com foco em situações em que a presença de *outliers* pode impactar significativamente os resultados obtidos pelo método clássico de estimação via mínimos quadrados.

A estrutura do trabalho foi dividida em capítulos que abordaram desde a fundamentação teórica com detalhes até a aplicação prática para comparação dos métodos de estimação estudados. No Capítulo 2 foi feita uma revisão sobre modelos de regressão lineares e os respectivos métodos de estimação que foram utilizados. Os métodos de estimação para os parâmetros da regressão linear foram discutidos e comparados a partir de suas diferenças de obtenção e propriedades. Em particular, a classe de M-estimadores, conhecida por produzir estimadores robustos, foi introduzida sob diferentes contextos.

Nas aplicações práticas, apresentadas no Capítulo 3, foram abordados dois conjuntos de dados diferentes para ilustrar o uso e as vantagens dos métodos de estimação robustos quando comparados ao LSM. A partir das aplicações foi possível aplicar os conceitos e métodos estudados no Capítulo 2, destacando a sensibilidade dos resultados obtidos a partir do LSM na presença de observações atípicas. Foi observado como os métodos de estimação robustos podem oferecer um ajuste melhor a maior parte dos dados, atribuindo peso baixo às observações discrepantes.

É de suma importância o estudo de métodos robustos para lidar com cenários em que a presença de *outliers* podem comprometer a validade das análises estatísticas tradicionais. Nesse sentido, reforça-se a necessidade contínua de avançar nas técnicas estatísticas para enfrentar os desafios presentes na análise de dados. Os métodos robustos apresentados neste trabalho são de grande relevância para pesquisadores, analistas e profissionais que buscam interpretações mais confiáveis e robustas a partir de seus dados, proporcionando interpretações mais consistentes diante de dados que fogem dos padrões usuais.

Em síntese, este trabalho proporcionou uma visão introdutória sobre a teoria e aplicação de métodos de estimação robustos sob modelos de regressão linear, contribuindo para a compreensão e utilização dessas técnicas em contextos diversos, ressaltando sua relevância em promover uma análise estatística mais sólida e confiável.

Referências

CHARNET, R. et al. *Análise de modelos de regressão linear com aplicações*. [S.l.: s.n.], 1999.

DRAPER, N. R.; SMITH, H. *Applied regression analysis*. [S.l.]: John Wiley & Sons, 1998. v. 326.

HAMPEL, F. R. et al. *Robust Statistics: The Approach Based on Influence Functions*. [S.l.]: New York: John Wiley & Sons, 2011. v. 196.

HUBER, P. *Robust Statistics*. [S.l.]: New York: John Wiley & Sons, 1981.

HUBER, P. J. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, JSTOR, v. 35, p. 73–101, 1964.

HUBER, P. J. Robust regression: asymptotics, conjectures and monte carlo. *The annals of statistics*, JSTOR, p. 799–821, 1973.

MARONNA, R. A. et al. *Robust statistics: theory and methods (with R)*. [S.l.]: John Wiley & Sons, 2019.

MICHAEL, H. et al. Applied linear statistical models. *McGraw-Hil*, 2004.

PAULA, G. A. *Modelos de regressão: com apoio computacional*. [S.l.]: IME-USP São Paulo, 2023.

ROUSSEEUW, P. J.; LEROY, A. M. *Robust regression and outlier detection*. [S.l.]: John wiley & sons, 2005.

ROUSSEEUW, P. J.; YOHAI, V. J. Robust regression by means of S-estimators. *Robust and Nonlinear Time Series Analysis*, Springer, v. 26, p. 256–272, 1984.

TUKEY, J. W. et al. *Exploratory data analysis*. [S.l.]: Reading, MA, 1977. v. 2.

WEISBERG, S. *Applied Linear Regression*. [S.l.]: John Wiley & Sons, 2005. v. 528.

YOHAI, V. J. High breakdown-point and high efficiency robust estimates for regression. *The Annals of statistics*, JSTOR, p. 642–656, 1987.