IME-USP

Diagnóstico de influência bayesiano em modelos de regressão da família t-assimétrica

Diego Wesllen da Silva

sob orientação da **Profa. Dra. Márcia D'Elia Branco**

Maio/2017

Sumário

Introdução

Diagnóstico Bayesiano em Modelos de Regressão Simétricos

- Medidas de Diagnóstico
- Modelo de Regressão Normal
- Modelo de Regressão t-Student
- Aplicação

Diagnóstico Bayesiano em Modelos de Regressão Assimétricos

- Modelo de Regressão Normal Assimétrico
- Modelo de Regressão t-Student Assimétrico
- Aplicação

Conclusões

Sumário

Introdução

Diagnóstico Bayesiano em Modelos de Regressão Simétricos

- Medidas de Diagnóstico
- Modelo de Regressão Normal
- Modelo de Regressão t-Student
- Aplicação

Diagnóstico Bayesiano em Modelos de Regressão Assimétricos

- Modelo de Regressão Normal Assimétrico
- Modelo de Regressão t-Student Assimétrico
- Aplicação

Conclusões

Distribuição t-assimétrica

A família de distribuições t-assimétrica nos permite propor modelos alternativos ao modelo normal.

Denotada por $ST(\mu, \sigma^2, \nu, \lambda)$, temos os seguintes casos particulares:

- $N(\mu, \sigma^2)$, se $\nu \to \infty$ e $\lambda = 0$;
- $t(\mu, \sigma^2, \nu)$, se $\lambda = 0$;
- $SN(\mu, \sigma^2, \lambda)$, se $\nu \to \infty$.

Dessa forma, conseguimos propor, por exemplo, modelos mais robustos a observações discrepantes.

Contudo, é importante, sob estes modelos, identificar observações discrepantes e aferir a influência destas nas estimativas.

Como medida para identificar observações discrepantes utilizaremos a *conditional predictive ordinate* (CPO).

Para medir a influência das observações nas estimativas utilizaremos a divergência de Kullback-Leibler e a Norma L_1.

A influência será medida tanto de forma global, em todo vetor de parâmetros do modelo, quanto de forma marginal, em apenas parte deste vetor.

Neste trabalho nós tanto deduzimos o cálculo destas medidas quanto fornecemos os programas para obtenção das mesmas.

Sumário

Introdução

Diagnóstico Bayesiano em Modelos de Regressão Simétricos

- Medidas de Diagnóstico
- Modelo de Regressão Normal
- Modelo de Regressão t-Student
- Aplicação

Diagnóstico Bayesiano em Modelos de Regressão Assimétricos

- Modelo de Regressão Normal Assimétrico
- Modelo de Regressão t-Student Assimétrico
- Aplicação

Conclusões

Sumário

Introdução

Diagnóstico Bayesiano em Modelos de Regressão Simétricos

Medidas de Diagnóstico

Modelo de Regressão Normal

Modelo de Regressão t-Student

Aplicação

Diagnóstico Bayesiano em Modelos de Regressão Assimétricos

Modelo de Regressão Normal Assimétrico

Modelo de Regressão t-Student Assimétrico

Aplicação

Conclusões

Conditional Predictive Ordinate (CPO)

Validação cruzada

Considere

- $\boldsymbol{y} = (y_1, \ldots, y_n)$ uma amostra observada de um modelo $Y|\theta$;
- $I_1, \ldots, I_K$ uma partição do conjunto de índices $\{1, \ldots, n\}$.

Para cada $k = 1, \ldots, K$,

- obtemos uma estimativa de θ em $\boldsymbol{y}_{(I_k)} = (y_j)_{j \notin I_k}$;
- validamos esta estimativa em $\boldsymbol{y}_{I_k} = (y_j)_{j \in I_k}$.

Quando $K = n$, temos o método *leave-one-out* e $I_i = \{i\}$. Logo, para cada $i = 1, \ldots, n$,

- obtemos uma estimativa de θ em $\boldsymbol{y}_{(i)} = (y_j)_{j \neq i}$;
- validamos esta estimativa na observação y_i.

Conditional Predictive Ordinate (CPO)

A i-ésima ordenada preditiva condicional (CPO_i) é dada por

$$f(y_i|\boldsymbol{y}_{(i)}) = \int f(y_i|\theta)f(\theta|\boldsymbol{y}_{(i)})\,d\theta = E_{\theta|\boldsymbol{y}_{(i)}}[f(y_i|\theta)].$$

A caracterização alternativa baseada em $\theta|\boldsymbol{y}$

$$f(y_i|\boldsymbol{y}_{(i)}) = E_{\theta|\boldsymbol{y}}[f(y_i|\theta)^{-1}]^{-1}$$

pode ser aproximada, por meio da integração de Monte Carlo, da seguinte forma

$$CPO_i \approx \left[\frac{1}{L}\sum_{l=1}^{L}\frac{1}{f(y_i|\theta^{(l)})}\right]^{-1}.$$

onde $\theta^{(1)},\ldots,\theta^{(L)}$ é uma amostra de tamanho L da distribuição *a posteriori* de $\theta|\boldsymbol{y}$.

Conditional Predictive Ordinate (CPO)

Um baixo valor de CPO_i indica que a observação é discrepante sob o modelo proposto.

Para facilitar a inspeção gráfica, analisaremos $-\log(CPO_i)$.

Uma medida resumo para os CPO_i é a *Log pseudo marginal likelihood* (LMPL) *(Ibrahim, 2001)*

$$LPML = \sum_{i=1}^{n} \log(CPO_i).$$

que será considerada uma medida de ajuste, isto é, o modelo com o maior valor de $LPML$ será considerado o mais adequado aos dados.

Influência Global

A família de divergências apresentada em *Weiss (1996)* é dada por

$$D_\theta(g, i) = \int g\left(\frac{f(\theta|\boldsymbol{y}_{(i)})}{f(\theta|\boldsymbol{y})}\right) f(\theta|\boldsymbol{y})\, d\theta,$$

onde g é uma função convexa com $g(1) = 0$.

Propriedades

- $D_\theta(g, i) \geqslant 0$
- Se $f(\theta|\boldsymbol{y}) = f(\theta|\boldsymbol{y}_{(i)})$ então $D_\theta(g, i) = 0$
- Se g for estritamente convexa em 1, $D_\theta(g, i) = 0$ se, e somente se, $f(\theta|\boldsymbol{y}) = f(\theta|\boldsymbol{y}_{(i)})$ quase certamente.

A divergência $D_\theta(g, i)$ nos dará uma medida de similaridade entre as distribuições $f(\theta|\boldsymbol{y})$ e $f(\theta|\boldsymbol{y}_{(i)})$.

A CPO_i é relacionada com a divergência da seguinte forma

$$D_\theta(g, i) = \int g\left(\frac{f(\theta|\boldsymbol{y}_{(i)})}{f(\theta|\boldsymbol{y})}\right) f(\theta|\boldsymbol{y})\, d\theta = E_{\theta|\boldsymbol{y}}\left[g\left(\frac{CPO_i}{f(y_i|\theta)}\right)\right].$$

Logo,

$$D_\theta(g, i) \approx \frac{1}{L}\sum_{l=1}^{L} g\left(\frac{CPO_i}{f(y_i|\theta^{(l)})}\right),$$

com $\theta^{(1)}, \ldots, \theta^{(L)}$ uma amostra simulada de tamanho L da distribuição *a posteriori* de $\theta|\boldsymbol{y}$.

Este cálculo não depende de formas explícitas para $f(\theta|\boldsymbol{y})$ e $f(\theta|\boldsymbol{y}_{(i)})$.

Norma L_1

A norma L_1 é um caso particular da divergência com $g(a) = \frac{1}{2}|a-1|$

$$L_{1i,\theta} = \frac{1}{2} \int |f(\theta|\boldsymbol{y}_{(i)}) - f(\theta|\boldsymbol{y})| d\theta.$$

Propriedade adicional: $L_{1i,\theta} \leqslant 1$.

Caracterização alternativa: distância variacional total,

$$L_{1i,\theta} = \sup_B \int_B f(\theta|\boldsymbol{y}_{(i)}) - f(\theta|\boldsymbol{y})\, d\theta = \sup_B \int_B f(\theta|\boldsymbol{y}) - f(\theta|\boldsymbol{y}_{(i)})\, d\theta.$$

Kullback-Leibler

A divergência de Kullback-Leibler é um caso particular da divergência com $g(a) = a \log(a)$

$$K_{i,\theta} = E_{\theta|\boldsymbol{y}_{(i)}}[\log(f(\theta|\boldsymbol{y}_{(i)}))] - E_{\theta|\boldsymbol{y}_{(i)}}[\log(f(\theta|\boldsymbol{y}))].$$

Contudo, não é possível encontrar de uma forma simples um limite superior e é de difícil interpretação.

Influência Marginal

Objetivo: medir a influência de uma observação em parte do vetor de parâmetros $\theta = (\theta_1, \theta_2)$.

Uma medida para influência da i-ésima observação nas estimativas de θ_1 é

$$D_{\theta_1}(g, i) = \int g\left(\frac{f(\theta_1|\boldsymbol{y}_{(i)})}{f(\theta_1|\boldsymbol{y})}\right) f(\theta_1|\boldsymbol{y})\, d\theta_1$$

O Teorema 2, p. 742, de *Weiss (1996)* diz que

$$0 \leqslant D_{\theta_1}(g, i) \leqslant D_{\theta}(g, i)$$

Portanto, uma observação pode exercer influência de forma global, mesmo não tendo influência marginal.

A relação da divergência na versão marginal com o CPO é

$$D_{\theta_1}(g,i) = \int g\left(\frac{f(\theta_1|\boldsymbol{y}_{(i)})}{f(\theta_1|\boldsymbol{y})}\right) f(\theta_1|\boldsymbol{y})\, d\theta_1 = \int g\left(\frac{CPO_i}{f(y_i|\theta_1,\boldsymbol{y}_{(i)})}\right) f(\theta_1|\boldsymbol{y})\, d\theta_1$$

Nem sempre será fácil encontrar $f(y_i|\theta_1,\boldsymbol{y}_{(i)})$ de forma explícita.

Mas,

$$f(y_i|\theta_1,\boldsymbol{y}_{(i)}) \propto \int f(\boldsymbol{y}|\theta) f(\theta)\, d\theta_2$$

Logo, encontramos uma proporcionalidade para $f(y_i|\theta_1,\boldsymbol{y}_{(i)})$.

Sumário

Introdução

Diagnóstico Bayesiano em Modelos de Regressão Simétricos

Medidas de Diagnóstico

Modelo de Regressão Normal

Modelo de Regressão t-Student

Aplicação

Diagnóstico Bayesiano em Modelos de Regressão Assimétricos

Modelo de Regressão Normal Assimétrico

Modelo de Regressão t-Student Assimétrico

Aplicação

Conclusões

Estimação dos Parâmetros

O modelo de regressão linear normal é dado por

$$y_i = x_i^t \boldsymbol{\beta} + \epsilon_i,$$

onde, para cada $i = 1, \ldots, n$,

- y_i é a variável resposta da i-ésima observação
- $x_i^t = (1, x_{i1}, \ldots, x_{ip})$ é o vetor de variáveis explicativas
- $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^t$ é vetor de parâmetros
- $\epsilon_i \sim N(0, \sigma^2)$ independentes

Função de Verossimilhança

$$L(\boldsymbol{\beta}, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{j=1}^{n} (y_j - x_j^t \boldsymbol{\beta})^2 \right)$$

Distibuição *a priori*

$$f(\boldsymbol{\beta}, \sigma^2) \propto \frac{1}{\sigma^2}$$

Distribuição *a posteriori*

$$\begin{aligned} \boldsymbol{\beta} | \sigma^2, \boldsymbol{y} &\sim N_{p+1}(\hat{\boldsymbol{\beta}}, \sigma^2 (X^t X)^{-1}) \\ \sigma^2 | \boldsymbol{y} &\sim GI\left(\frac{n-(p+1)}{2}, \frac{(\boldsymbol{y} - X\hat{\boldsymbol{\beta}})^t (\boldsymbol{y} - X\hat{\boldsymbol{\beta}})}{2} \right) \end{aligned}$$

Algoritmo

Uma forma de obter uma amostra de tamanho L da distribuição *a posteriori* conjunta de $(\boldsymbol{\beta}, \sigma^2)|\boldsymbol{y}$ é utilizar o algoritmo a seguir:

Para cada $l = 1, \ldots, L$

1. Simular ${\sigma^2}^{(l)}$ da distribuição de $\sigma^2|\boldsymbol{y}$.
2. Simular $\boldsymbol{\beta}^{(l)}$ da distribuição de $\boldsymbol{\beta}|{\sigma^2}^{(l)}, \boldsymbol{y}$.

Assim, $(\boldsymbol{\beta}^{(1)}, {\sigma^2}^{(1)}), \ldots, (\boldsymbol{\beta}^{(L)}, {\sigma^2}^{(L)})$ é a amostra desejada

CPO

$$\widehat{CPO_i} = \left[\frac{1}{L} \sum_{l=1}^{L} \frac{1}{f(y_i|\boldsymbol{\beta}^{(l)}, \sigma^{2\,(l)})} \right]^{-1}$$

Influência Global

$$\widehat{D_{\boldsymbol{\beta},\sigma^2}(g, i)} = \frac{1}{L} \sum_{l=1}^{L} g \left(\frac{\widehat{CPO_i}}{f(y_i|\boldsymbol{\beta}^{(l)}, \sigma^{2\,(l)})} \right)$$

Influência Marginal

$$\widehat{D_{\boldsymbol{\beta}}(g, i)} = \frac{1}{L} \sum_{l=1}^{L} g \left(\frac{\widehat{CPO_i}}{f(y_i|\boldsymbol{\beta}^{(l)}, \boldsymbol{y}_{(i)})} \right)$$

Influência Marginal

$$f(y_i|\boldsymbol{\beta}, \boldsymbol{y}_{(i)}) \propto \int f(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) f(\boldsymbol{\beta}, \sigma^2)\, d\sigma^2$$

$$\propto \left[1 + \frac{1}{n-1}\left(\frac{y_i - x_i^t\boldsymbol{\beta}}{\sqrt{\frac{1}{n-1}\sum_{j\neq i}(y_j - x_j^t\boldsymbol{\beta})^2}}\right)^2\right]^{-n/2}$$

Logo,

$$y_i|\boldsymbol{\beta}, \boldsymbol{y}_{(i)} \sim t\left(x_i^t\boldsymbol{\beta}, \frac{1}{n-1}\sum_{j\neq i}(y_j - x_j^t\boldsymbol{\beta})^2, n-1\right)$$

Sumário

Introdução

Diagnóstico Bayesiano em Modelos de Regressão Simétricos

- Medidas de Diagnóstico
- Modelo de Regressão Normal
- Modelo de Regressão t-Student
- Aplicação

Diagnóstico Bayesiano em Modelos de Regressão Assimétricos

- Modelo de Regressão Normal Assimétrico
- Modelo de Regressão t-Student Assimétrico
- Aplicação

Conclusões

O modelo de regressão linear t-Student é dado por

$$y_i = x_i^t \boldsymbol{\beta} + \epsilon_i,$$

onde, para cada $i = 1, \ldots, n$,

- y_i é a variável resposta da i-ésima observação
- $x_i^t = (1, x_{i1}, \ldots, x_{ip})$ é o vetor de variáveis explicativas
- $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^t$ é vetor de parâmetros
- $\epsilon_i \sim t(0, \sigma^2, \nu)$ independentes e consideramos o parâmetro ν fixado.

Estimação dos Parâmetros

Representação hierárquica

Como $y_i|\boldsymbol{\beta}, \sigma^2 \sim t(x_i^t\boldsymbol{\beta}, \sigma^2, \nu)$, temos que

$$
\begin{aligned}
y_i|\boldsymbol{\beta}, \sigma^2, u_i &\sim N(x_i^t\boldsymbol{\beta}, \sigma^2/u_i) \\
u_i &\sim Gama(\nu/2, \nu/2)
\end{aligned}
$$

Função de Verossimilhança Aumentada

$$
L_A(\boldsymbol{\beta}, \sigma^2) \propto \frac{1}{\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} u_i(y_i - x_i^t\boldsymbol{\beta})^2 - \frac{\nu}{2}\sum_{i=1}^{n} u_i\right)\prod_{i=1}^{n} u_i^{\frac{\nu-1}{2}}
$$

Distibuição *a priori*

$$
f(\boldsymbol{\beta}, \sigma^2) \propto \frac{1}{\sigma^2}
$$

Distribuição *a posteriori*

$$
f(\boldsymbol{\beta}, \sigma^2, \boldsymbol{u}|\boldsymbol{y}) \propto L_A(\boldsymbol{\beta}, \sigma^2) f(\boldsymbol{\beta}, \sigma^2)
$$

Algoritmo de Gibbs

Distribuições condicionais completas

$$\beta|\sigma^2, \boldsymbol{u}, \boldsymbol{y} \sim N_{p+1}(\hat{\boldsymbol{\beta}}, \sigma^2(X^t UX)^{-1})$$
$$\sigma^2|\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{y} \sim GI\left(\frac{n}{2}, \frac{(Y - X\boldsymbol{\beta})^t U(Y - X\boldsymbol{\beta})}{2}\right)$$
$$u_i|\boldsymbol{\beta}, \sigma^2, \boldsymbol{y} \sim Gama\left(\frac{\nu+1}{2}, \frac{y_i - x_i^t\boldsymbol{\beta}}{2\sigma^2} + \frac{\nu}{2}\right) \text{ independentes}$$

com $\hat{\boldsymbol{\beta}} = (X^t UX)^{-1} X^t UY$ e $U = \mathrm{diag}(u_1, \ldots, u_n)$.

Proposição.

Uma estimativa de Monte Carlo da divergência para a influência marginal em β no modelo t-Student é dada por

$$\widehat{D_\beta(g,i)} = \frac{1}{M}\sum_{m=1}^{M} g\left(\frac{\widehat{CPO_i}}{\hat{f}(y_i|\boldsymbol{\beta}^{(m)}, \boldsymbol{y}_{(i)})}\right)$$

onde $\widehat{CPO_i}$ *é a estimativa para o* CPO_i *e*

$$\hat{f}(y_i|\boldsymbol{\beta}, \boldsymbol{y}_{(i)}) = \frac{\Gamma\left(\frac{n}{2}\right)\frac{1}{L}\sum_{l=1}^{L}\left[\sum_{j=1}^{n} u_j^{(l)}(y_j - x_j^t\boldsymbol{\beta})^2\right]^{-\frac{n}{2}}}{\Gamma\left(\frac{n-1}{2}\right)\sqrt{\pi}\frac{1}{L}\sum_{l=1}^{L}\frac{1}{\sqrt{u_i^{(l)}}}\left[\sum_{j\neq i} u_j^{(l)}(y_j - x_j^t\boldsymbol{\beta})^2\right]^{-\frac{n-1}{2}}}$$

onde, para cada $j = 1, \ldots, n$ *e* $l = 1, \ldots, L$, $u_j^{(l)} \sim Gama(\frac{\nu+1}{2}, \frac{\nu}{2})$.

Sumário

Introdução

Diagnóstico Bayesiano em Modelos de Regressão Simétricos

- Medidas de Diagnóstico
- Modelo de Regressão Normal
- Modelo de Regressão t-Student
- Aplicação

Diagnóstico Bayesiano em Modelos de Regressão Assimétricos

- Modelo de Regressão Normal Assimétrico
- Modelo de Regressão t-Student Assimétrico
- Aplicação

Conclusões

Dados

Presente em *Weiss e Cho (1998)*, os dados são parte de um estudo sobre ocorrência de doença cardíaca cianótica em 21 crianças, conduzido pela UCLA. Utilizamos as seguintes variáveis

- X: idade (meses) em que a criança disse sua primeira palavra
- Y: *Gesell adaptive score*

O *Gesell adaptive score* é uma medida de desenvolvimento cognitivo mensurada em uma idade mais avançada da criança.

Dados

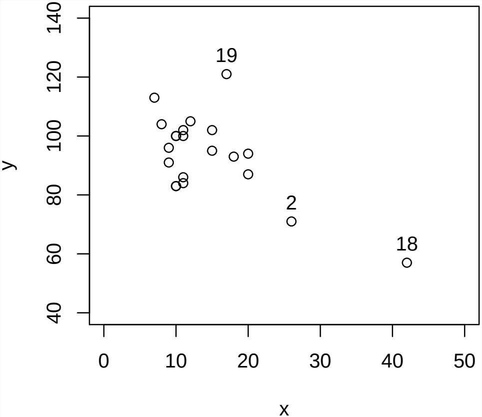


Figura 1: Diagrama de dispersão com pontos discrepantes destacados

Modelos propostos

	Modelo	
Parâmetros	Normal	t-Student
β_0	109.82	110.42
β_1	-1.12	-1.18
$Var(\varepsilon)$	131.79	143.04
$LPML$	-82.99	-82.05

Tabela 1: Estimativas dos parâmetros dos modelos

CPO

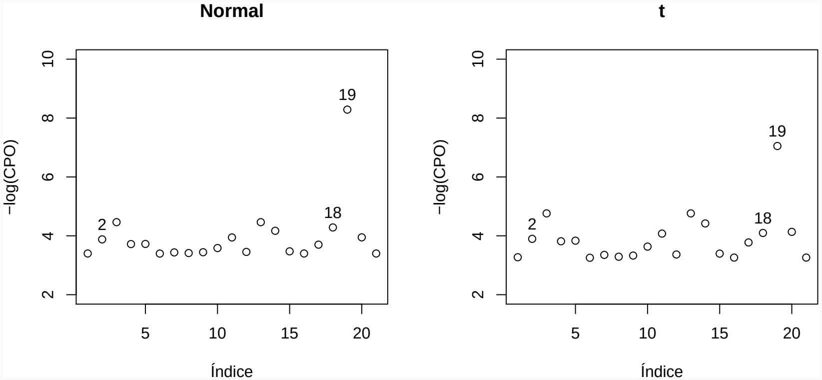


Figura 2: $-\log(CPO)$ para os modelos normal e t-Student

Diagnóstico global - Norma L_1

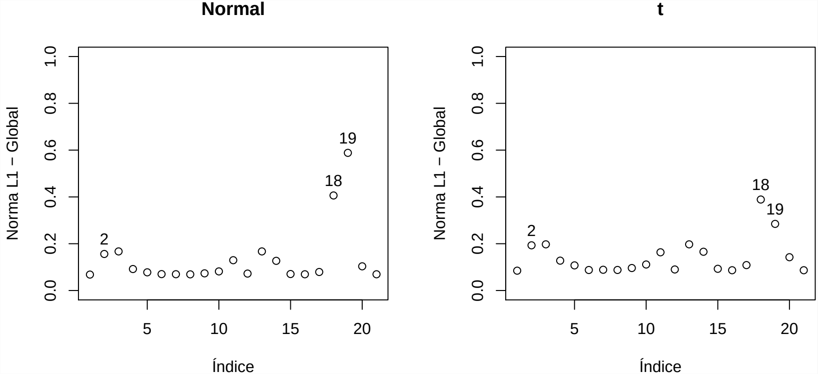


Figura 3: Influência global via norma L_1

Diagnóstico global - Kullback-Leibler

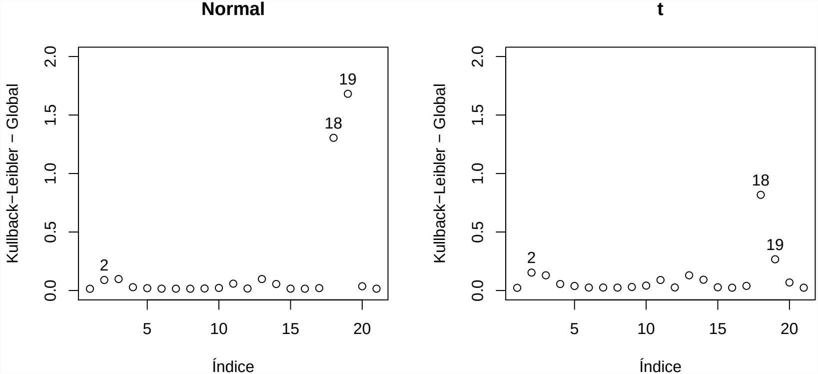


Figura 4: Influência global via Kullback-Leibler

Diagnóstico marginal - Norma L_1

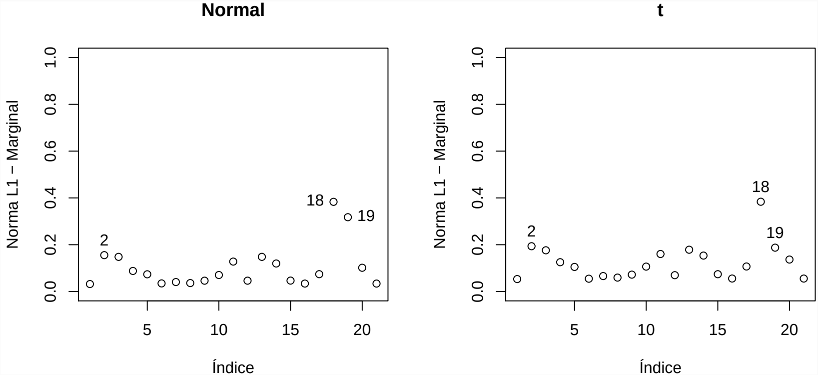


Figura 5: Influência marginal via norma L_1

Diagnóstico marginal - Kullback-Leibler

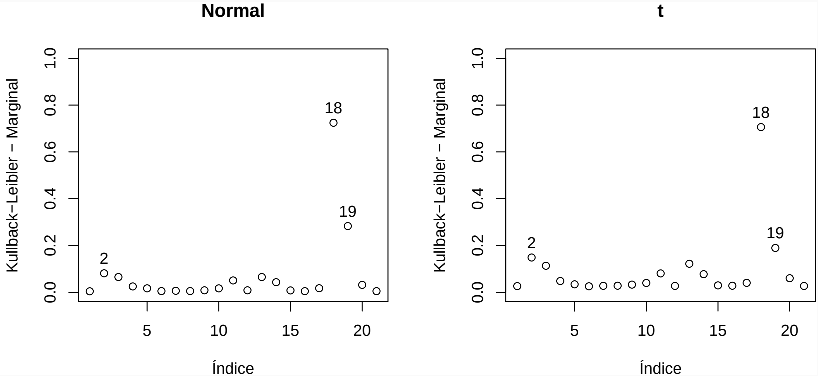


Figura 6: Influência marginal via Kullback-Leibler

Sumário

Introdução

Diagnóstico Bayesiano em Modelos de Regressão Simétricos

Medidas de Diagnóstico

Modelo de Regressão Normal

Modelo de Regressão t-Student

Aplicação

Diagnóstico Bayesiano em Modelos de Regressão Assimétricos

Modelo de Regressão Normal Assimétrico

Modelo de Regressão t-Student Assimétrico

Aplicação

Conclusões

Sumário

Introdução

Diagnóstico Bayesiano em Modelos de Regressão Simétricos

- Medidas de Diagnóstico
- Modelo de Regressão Normal
- Modelo de Regressão t-Student
- Aplicação

Diagnóstico Bayesiano em Modelos de Regressão Assimétricos

- Modelo de Regressão Normal Assimétrico
- Modelo de Regressão t-Student Assimétrico
- Aplicação

Conclusões

Estimação dos Parâmetros

O modelo de regressão linear normal assimétrico é dado por

$$y_i = x_i^t \boldsymbol{\beta} + \epsilon_i,$$

onde, para cada $i = 1, \ldots, n$,

- y_i é a variável resposta da i-ésima observação
- $x_i^t = (1, x_{i1}, \ldots, x_{ip})$ é o vetor de variáveis explicativas
- $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^t$ é vetor de parâmetros
- $\epsilon_i \sim SN(0, \sigma^2, \lambda)$ independentes

Estimação dos Parâmetros

Representação hierárquica

Dada por *Bayes (2005)*,

$$y_i|v_i \sim N(x_i^t\boldsymbol{\beta} + \sigma\delta v_i, \sigma^2(1-\delta^2))$$
$$v_i \sim HN(0,1)$$

onde $\delta = \frac{\lambda}{\sqrt{1+\lambda^2}}$.

Reparametrização

Com $\eta = \sigma\delta$ e $\tau = \sigma\sqrt{1-\delta^2}$, teremos

$$y_i|v_i \sim N(x_i^t\beta + \eta v_i, \tau^2)$$
$$v_i \sim HN(0,1)$$

Função de Verossimilhança Aumentada

$$L_A(\boldsymbol{\beta}, \tau, \eta) = \frac{1}{\tau^n} \exp\left(-\frac{1}{2\tau^2}\sum_{i=1}^{n}(y_i - x_i^t\boldsymbol{\beta} - \eta v_i)^2\right)$$
$$\exp\left(-\frac{1}{2}\sum_{i=1}^{n} v_i^2\right)\prod_{i=1}^{n}\mathbb{I}_{[0,\infty[}(v_i)$$

Distibuição *a priori*

$$f(\boldsymbol{\beta}, \sigma, \lambda) \propto \frac{1}{\sigma} f(\lambda)$$

com $\lambda \sim t(0, \sigma_t^2, k)$.

- $k = \frac{1}{2}$ e $\sigma_t^2 = \frac{\pi^2}{4}$ $\Rightarrow$ aproximação da *priori* de Jeffreys
- $k = 2$ e $\sigma_t^2 = \frac{1}{2} \Rightarrow \delta \sim U(-1, 1)$

Distibuição *a priori*

Como $\lambda \sim t(0, \sigma_t^2, k)$, temos que

$$\lambda|\omega \sim N\left(0, \frac{\sigma_t^2}{\omega}\right)$$
$$\omega \sim Gama\left(\frac{k}{2}, \frac{k}{2}\right).$$

Consequentemente,

$$f(\boldsymbol{\beta}, \sigma, \lambda, \omega) \propto \frac{1}{\sigma} \exp\left(-\frac{\lambda^2 \omega}{2\sigma_t^2}\right) \omega^{\frac{k+1}{2}-1} \exp\left(-\frac{k}{2}\omega\right).$$

e sob a reparametrização,

$$f(\boldsymbol{\beta}, \tau, \eta, \omega) \propto \frac{1}{\tau^2} \exp\left(-\frac{\eta^2 \omega}{2\tau^2 \sigma_t^2}\right) \omega^{\frac{k+1}{2}-1} \exp\left(-\frac{k}{2}\omega\right).$$

Estimação dos Parâmetros

Distribuição *a posteriori*

$$f(\boldsymbol{\beta}, \sigma^2, \boldsymbol{u}|\boldsymbol{y}) \propto L_A(\boldsymbol{\beta}, \tau, \eta) f(\boldsymbol{\beta}, \tau, \eta, \omega)$$

Distribuições condicionais completas

$$\boldsymbol{\beta}|\tau, \eta, \omega, \boldsymbol{v}, \boldsymbol{y} \sim N_{p+1}\left((X^t X)^{-1} X^t (Y - \eta V), \tau^2 (X^t X)^{-1}\right)$$

$$\frac{1}{\tau^2}|\boldsymbol{\beta}, \eta, \omega, \boldsymbol{v}, \boldsymbol{y} \sim Gama\left(\frac{n+1}{2}, \frac{1}{2}\left((Y - \eta V - X\boldsymbol{\beta})^t (Y - \eta V - X\boldsymbol{\beta}) + \frac{\eta^2 \omega}{\sigma_t^2}\right)\right)$$

$$\eta|\boldsymbol{\beta}, \tau, \omega, \boldsymbol{v}, \boldsymbol{y} \sim N\left(\frac{(Y - X\boldsymbol{\beta})^t V}{V^t V + \frac{\omega}{\sigma_t^2}}, \frac{\tau^2}{V^t V + \frac{\omega}{\sigma_t^2}}\right)$$

$$\omega|\boldsymbol{\beta}, \tau, \eta, \boldsymbol{v}, \boldsymbol{y} \sim Gama\left(\frac{k+1}{2}, \frac{1}{2}\left(k + \frac{\eta^2}{\tau^2 \sigma_t^2}\right)\right)$$

$$v_i|\boldsymbol{\beta}, \tau, \eta, \omega, \boldsymbol{v}, \boldsymbol{y} \sim N\left(\frac{\eta(y_i - x_i^t \boldsymbol{\beta})}{\eta^2 + \tau^2}, \frac{\tau^2}{\eta^2 + \tau^2}\right) \mathbb{I}_{[0,\infty[}(v_i), \;\; i = 1, \ldots, n$$

Proposição.

Uma estimativa de Monte Carlo da divergência para a influência marginal no modelo normal assimétrico é dada por

$$\widehat{D_\beta(g,i)} = \frac{1}{M}\sum_{m=1}^{M} g\left(\frac{\widehat{CPO_i}}{\hat{f}(y_i|\boldsymbol{\beta}^{(m)},\boldsymbol{y}_{(i)})}\right)$$

onde $\widehat{CPO_i}$ *é uma estimativa para o* CPO_i *e*

$$\hat{f}(y_i|\boldsymbol{\beta},\boldsymbol{y}_{(i)}) = \frac{\Gamma(\frac{n-1}{2})\frac{1}{L}\sum_{l=1}^{L}\left[(\Omega_{(l)}+V_{(l)})A_{(l)}^{n-1}\right]^{-\frac{1}{2}}}{\Gamma(\frac{n-2}{2})\sqrt{\pi}\frac{1}{L}\sum_{l=1}^{L}\left[\left(\Omega_{(l)}+\sum_{k\neq i}{v_k^{(l)}}^2\right)K_{(l)}^{n-2}\right]^{-\frac{1}{2}}}$$

com

Proposição.

$$
\begin{aligned}
z_i &= y_i - x_i^t\boldsymbol{\beta}\\
A_{(l)} &= \sum_{j=1}^{n} z_j^2 - \frac{1}{\Omega_{(l)} + V_{(l)}}\left(\sum_{k=1}^{n} v_k^{(l)} z_k\right)^2\\
K_{(l)} &= \sum_{j\neq i} z_j^2 - \frac{1}{\Omega_{(l)} + \sum_{k\neq i} {v_k^{(l)}}^2}\left(\sum_{j\neq i} v_j^{(l)} z_j\right)^2\\
V_{(l)} &= \sum_{j=1}^{n} {v_j^{(l)}}^2\\
\Omega_{(l)} &= \frac{\omega_{(l)}}{\sigma_t^2}
\end{aligned}
$$

onde, para cada $j = 1, \ldots, n$ e $l = 1, \ldots, L$, $v_j^{(l)} \sim HN(0,1)$ e $\omega^{(l)} \sim Gama(\frac{k+1}{2}, \frac{k}{2})$.

Sumário

Introdução

Diagnóstico Bayesiano em Modelos de Regressão Simétricos

Medidas de Diagnóstico

Modelo de Regressão Normal

Modelo de Regressão t-Student

Aplicação

Diagnóstico Bayesiano em Modelos de Regressão Assimétricos

Modelo de Regressão Normal Assimétrico

Modelo de Regressão t-Student Assimétrico

Aplicação

Conclusões

Estimação dos Parâmetros

O modelo de regressão linear t-assimétrico é dado por

$$y_i = x_i^t \boldsymbol{\beta} + \epsilon_i,$$

onde, para cada $i = 1, \ldots, n$,

- y_i é a variável resposta da i-ésima observação
- $x_i^t = (1, x_{i1}, \ldots, x_{ip})$ é o vetor de variáveis explicativas
- $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^t$ é vetor de parâmetros
- $\epsilon_i \sim ST(0, \sigma^2, \nu, \lambda)$ independentes, ν fixado

Representação hierárquica

Dada em *Godoi (2007)*

$$
\begin{aligned}
y_i|u_i, v_i &\sim N\left(x_i^t\beta + \sigma\delta\frac{v_i}{\sqrt{u_i}}, \frac{\sigma^2(1-\delta^2)}{u_i}\right) \\
u_i &\sim Gama\left(\frac{\nu}{2}, \frac{\nu}{2}\right) \\
v_i &\sim HN(0,1)
\end{aligned}
$$

onde $\delta = \frac{\lambda}{\sqrt{1+\lambda^2}}$.

Reparametrização

Com $\eta = \sigma\delta$, $\tau = \sigma\sqrt{1-\delta^2}$ e $t_i = \frac{v_i}{\sqrt{u_i}}$, teremos

$$
\begin{aligned}
y_i|u_i, t_i &\sim N\left(x_i^t\beta + \eta t_i, \frac{\tau^2}{u_i}\right) \\
t_i|u_i &\sim HN(0, 1/u_i) \\
u_i &\sim Gama\left(\frac{\nu}{2}, \frac{\nu}{2}\right)
\end{aligned}
$$

Função de Verossimilhança Aumentada

$$L_A(\boldsymbol{\beta}, \tau, \eta) \propto \frac{1}{\tau^n} \exp\left(-\frac{1}{2\tau^2} \sum_{i=1}^{n} u_i (y_i - x_i^t \beta - \eta t_i)^2 \right)$$
$$\prod_{i=1}^{n} \left[u_i^{\frac{\nu-1}{2}} e^{-\frac{\nu}{2} u_i} e^{-\frac{u_i t_i^2}{2}} \mathbb{I}_{[0,\infty[}(u_i) \mathbb{I}_{[0,\infty[}(t_i) \right]$$

Distibuição *a priori*

$$f(\boldsymbol{\beta}, \tau, \eta, \omega) \propto \frac{1}{\tau^2} \exp\left(-\frac{\eta^2 \omega}{2\tau^2 \sigma_t^2} \right) \omega^{\frac{k+1}{2}-1} \exp\left(-\frac{k}{2} \omega \right)$$

Distribuição *a posteriori*

$$f(\boldsymbol{\beta}, \tau, \eta, \omega, \boldsymbol{u}, \boldsymbol{t} | \boldsymbol{y}) \propto L_A(\boldsymbol{\beta}, \tau, \eta) f(\boldsymbol{\beta}, \tau, \eta, \omega)$$

Distribuições condicionais completas

$$\beta|\tau, \eta, \omega, \boldsymbol{u}, \boldsymbol{t}, \boldsymbol{y} \sim N_{p+1}\left((X^t UX)^{-1} X^t UZ, \tau^2 (X^t UX)^{-1}\right)$$

$$\frac{1}{\tau^2}|\mu, \eta, \omega, \boldsymbol{u}, \boldsymbol{t}, \boldsymbol{y} \sim Gama\left(\frac{n+1}{2}, \frac{1}{2}\left((Z - X\boldsymbol{\beta})^t U(Z - X\boldsymbol{\beta}) + \frac{\eta^2\omega}{\sigma_t^2}\right)\right)$$

$$\eta|\mu, \tau, \omega, \boldsymbol{u}, \boldsymbol{t}, \boldsymbol{y} \sim N\left(\frac{(Y - X\boldsymbol{\beta})^t UT}{T^t UT + \frac{\omega}{\sigma_t^2}}, \frac{\tau^2}{T^t UT + \frac{\omega}{\sigma_t^2}}\right)$$

$$\omega|\mu, \tau, \eta, \boldsymbol{u}, \boldsymbol{t}, \boldsymbol{y} \sim Gama\left(\frac{k+1}{2}, \frac{1}{2}\left(k + \frac{\lambda^2}{\tau^2\sigma_t^2}\right)\right)$$

$$u_i|\mu, \tau, \eta, \omega, \boldsymbol{t}, \boldsymbol{y} \sim Gama\left(\frac{\nu+1}{2}, \frac{1}{2}\left(t_i^2 + \nu + \frac{1}{\tau^2}(y_i - x_i^t\beta - \eta t_i)^2\right)\right)$$

$$t_i|\mu, \tau, \eta, \omega, \boldsymbol{u}, \boldsymbol{y} \sim N\left(\frac{\eta(y_i - x_i^t\boldsymbol{\beta})}{\eta^2 + \tau^2}, \frac{\tau^2}{u_i(\eta^2 + \tau^2)}\right)\mathbb{I}_{[0,+\infty[}(t_i)$$

Proposição.

Uma estimativa de Monte Carlo da divergência para a influência marginal no modelo t-assimétrico é dada por

$$\widehat{D_\beta(g,i)} = \frac{1}{M}\sum_{m=1}^{M} g\left(\frac{\widehat{CPO_i}}{\hat{f}(y_i|\boldsymbol{\beta}^{(m)}, \boldsymbol{y}_{(i)})}\right)$$

onde $\widehat{CPO_i}$ *é a estimativa para o* CPO_i

$$\hat{f}(y_i|\boldsymbol{\beta}, \boldsymbol{y}_{(i)}) = \frac{\Gamma(\frac{n-1}{2})\frac{1}{L}\sum_{l=1}^{L}\left[(\Omega_{(l)} + V_{(l)})A_{(l)}^{n-1}\right]^{-\frac{1}{2}}}{\Gamma(\frac{n-2}{2})\sqrt{\pi}\frac{1}{L}\sum_{l=1}^{L}\left[\left(\Omega_{(l)} + \sum_{k\neq i} {v_k^{(l)}}^2\right) K_{(l)}^{n-2}\right]^{-\frac{1}{2}}}$$

com

Proposição.

$$z_j^{(l)} = \sqrt{u_j^{(l)}}(y_j - x_j^t\boldsymbol{\beta}) \quad , \qquad v_j^{(l)} = \sqrt{u_j^{(l)}}\,t_j^{(l)}$$

$$A_{(l)} = \sum_{j=1}^{n} {z_j^{(l)}}^2 - \frac{1}{\Omega_{(l)} + V_{(l)}} \left(\sum_{k=1}^{n} v_k^{(l)} {z_k^{(l)}}^2 \right)^2$$

$$K_{(l)} = \sum_{j \neq i} {z_j^{(l)}}^2 - \frac{1}{\Omega_{(l)} + \sum_{k \neq i} {v_k^{(l)}}^2} \left(\sum_{j \neq i} v_j^{(l)} {z_j^{(l)}}^2 \right)^2$$

$$V_{(l)} = \sum_{j=1}^{n} {v_j^{(l)}}^2 \quad , \qquad \Omega_{(l)} = \frac{\omega_{(l)}}{\sigma_t^2}$$

sendo para cada $l = 1, \ldots, L$ *e* $i = 1, \ldots, n$,

$$u_i^{(l)} \sim Gama\left(\frac{\nu+1}{2}, \frac{\nu}{2}\right),\ t_i^{(l)} \sim HN(0, 1/u_i),\ \omega^{(l)} \sim Gama\left(\frac{k+1}{2}, \frac{k}{2}\right)$$

Sumário

Introdução

Diagnóstico Bayesiano em Modelos de Regressão Simétricos

- Medidas de Diagnóstico
- Modelo de Regressão Normal
- Modelo de Regressão t-Student
- Aplicação

Diagnóstico Bayesiano em Modelos de Regressão Assimétricos

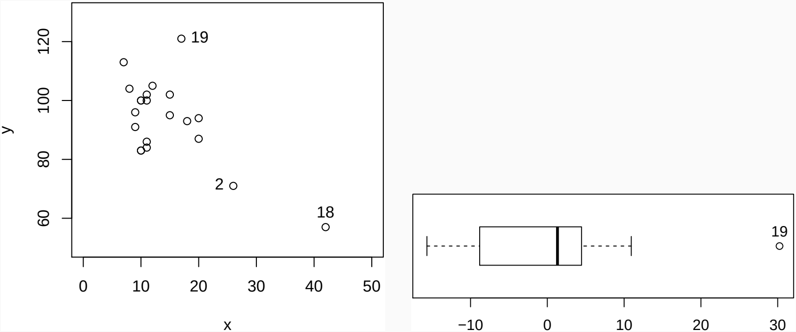
- Modelo de Regressão Normal Assimétrico
- Modelo de Regressão t-Student Assimétrico
- Aplicação

Conclusões

Dados originais

Normal	t-Student	Normal assimétrica	t-assimétrica
-82.75	-81.75	-82.5	-84.82

Tabela 2: LPML dos modelos nos dados originais



(a) (b)

Figura 7: Diagrama de dispersão e boxplot dos resíduos do modelo de regressão normal nos dados originais

Dados originais

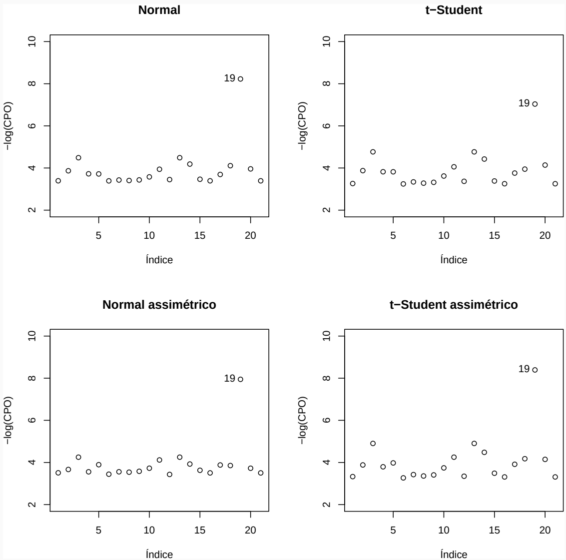


Figura 8: $-\log(CPO)$ dos modelos nos dados originais

Dados originais

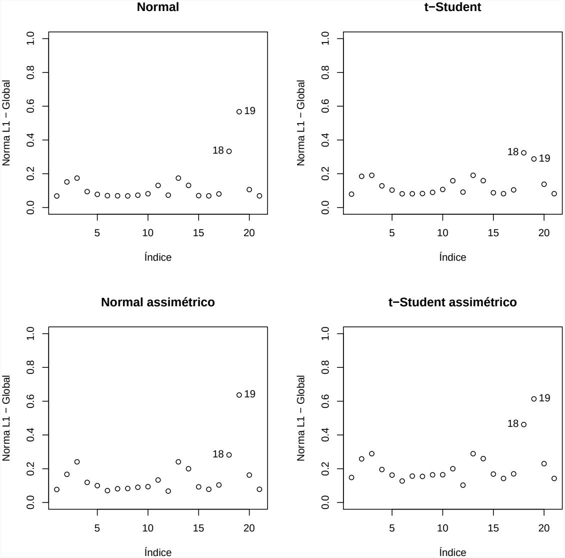


Figura 9: Influência global via norma L_1 dos modelos nos dados originais

Dados originais

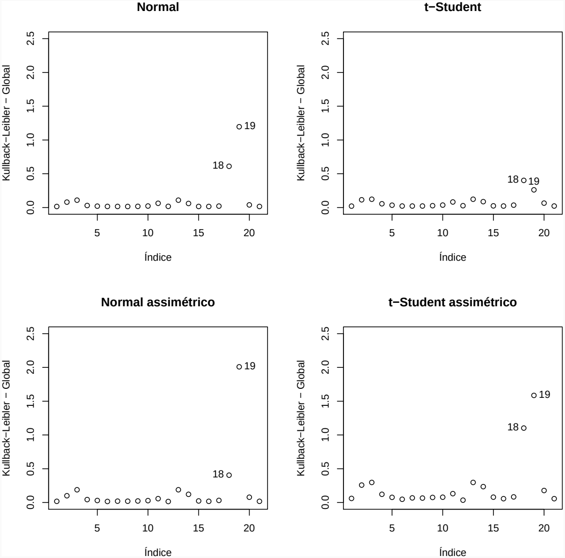


Figura 10: Influência global via divergência de Kullback-Leibler dos modelos nos dados originais

Dados originais

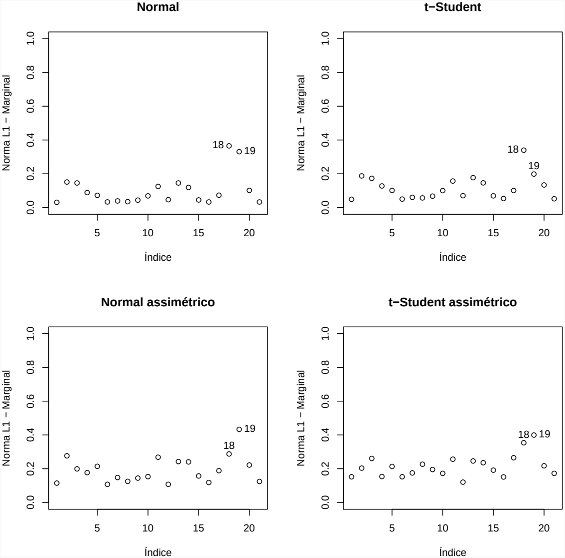


Figura 11: Influência marginal via norma L_1 dos modelos nos dados originais

Dados originais

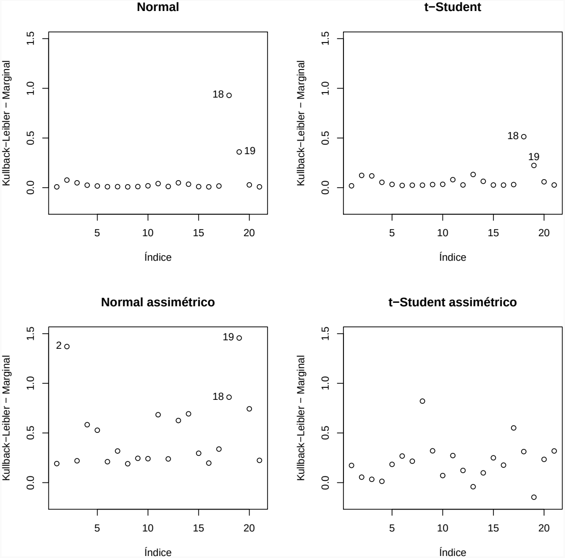
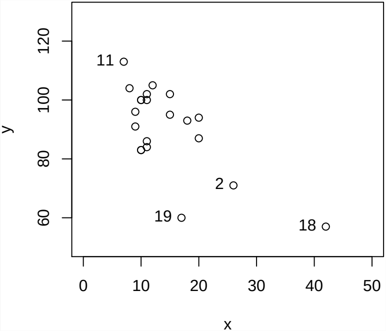


Figura 12: Influência marginal via divergência de Kullback-Leibler dos modelos nos dados originais

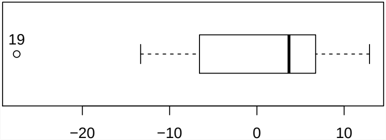
Contaminação da observação 19

Normal	t-Student	Normal assimétrica	t-assimétrica
-81.6	-81.38	-78.73	-79.7

Tabela 3: LPML dos modelos contaminando a observação 19



(a)



(b)

Figura 13: Diagrama de dispersão e boxplot dos resíduos do modelo de regressão normal contaminando a observação 19

Contaminação da observação 19

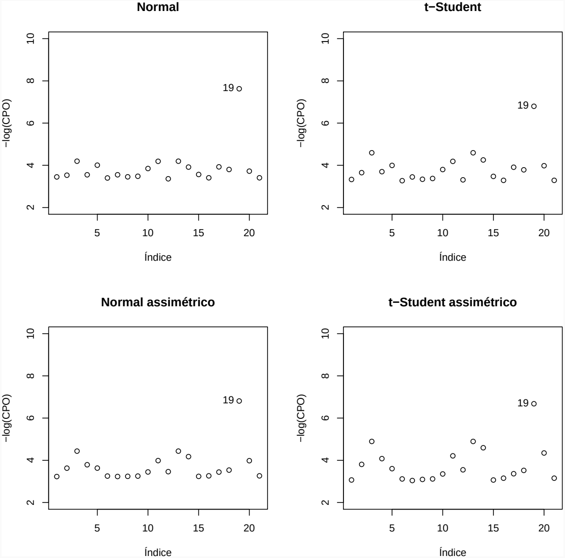


Figura 14: $-\log(CPO)$ dos modelos contaminando a observação 19

Contaminação da observação 19

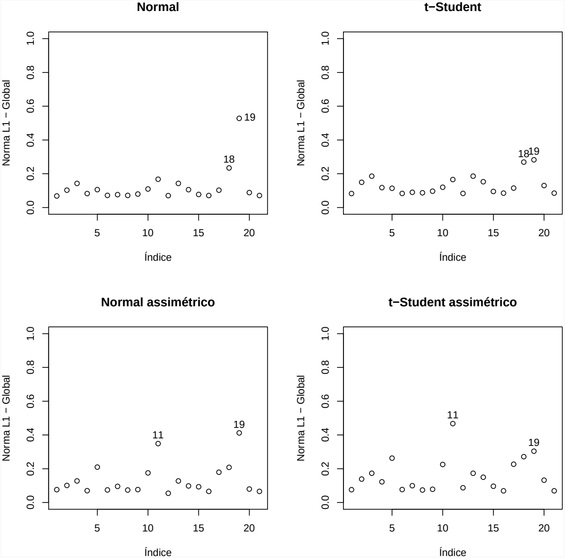


Figura 15: Influência global via norma L_1 dos modelos contaminando a observação 19

Contaminação da observação 19

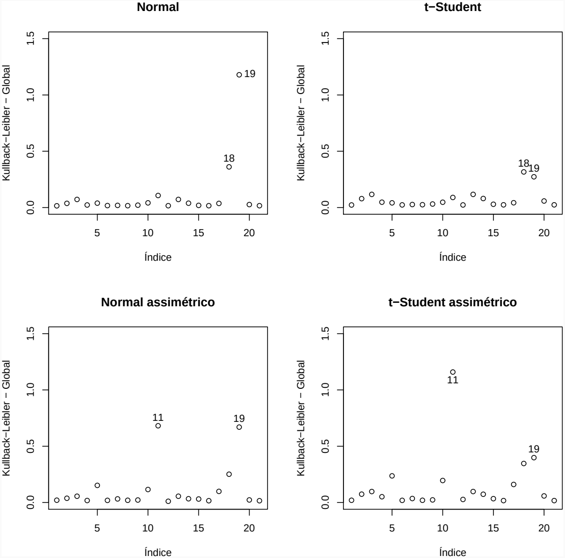


Figura 16: Influência global via divergência de Kullback-Leibler dos modelos contaminando a observação 19

Contaminação da observação 19

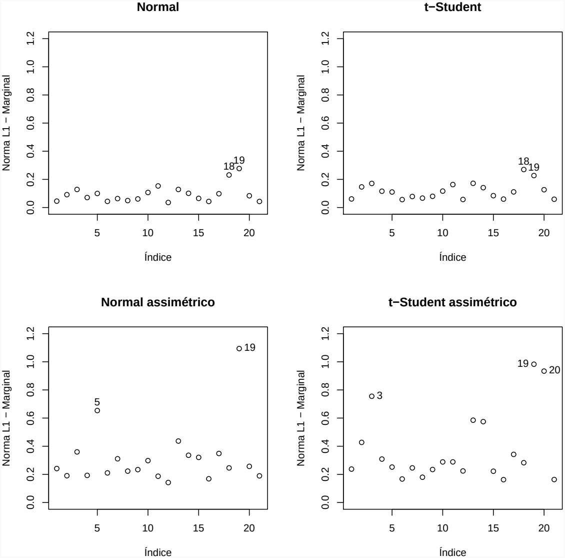


Figura 17: Influência marginal via norma L_1 dos modelos contaminando a observação 19

Contaminação da observação 19

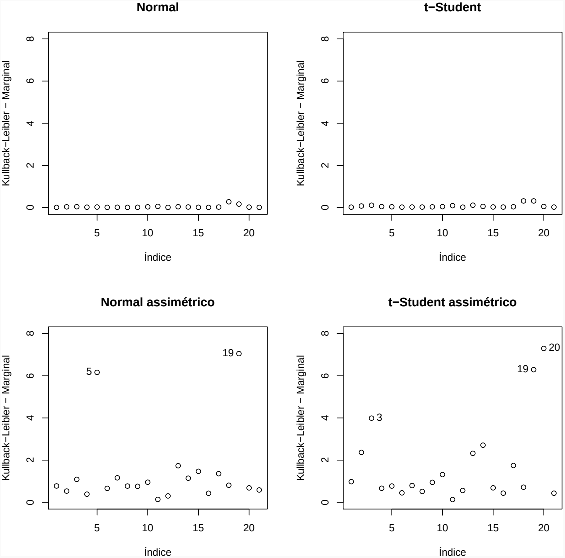
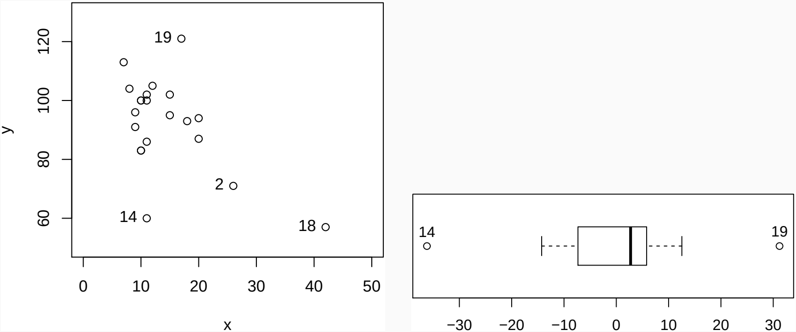


Figura 18: Influência marginal via divergência de Kullback-Leibler dos modelos contaminando a observação 19

Contaminação da observação 14

Normal	t-Student	Normal assimétrica	t-assimétrica
-87.49	-85.31	-87.8	-90.68

Tabela 4: LPML dos modelos contaminando a observação 14



(a) (b)

Figura 19: Diagrama de dispersão e boxplot dos resíduos do modelo de regressão normal contaminando a observação 14

Contaminação da observação 14

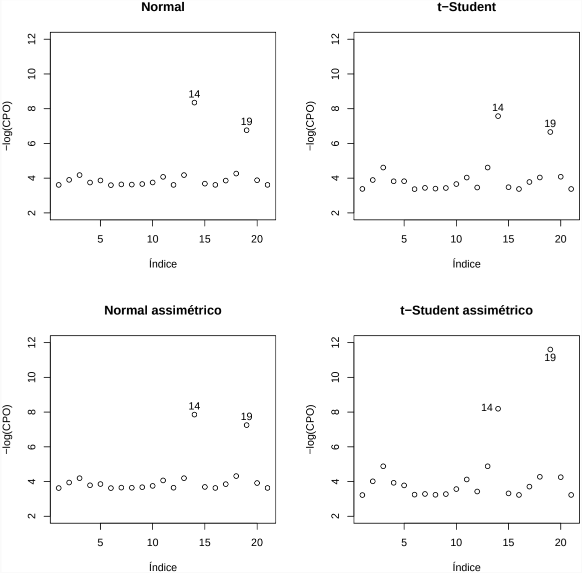


Figura 20: $-\log(CPO)$ dos modelos contaminando a observação 14

Contaminação da observação 14

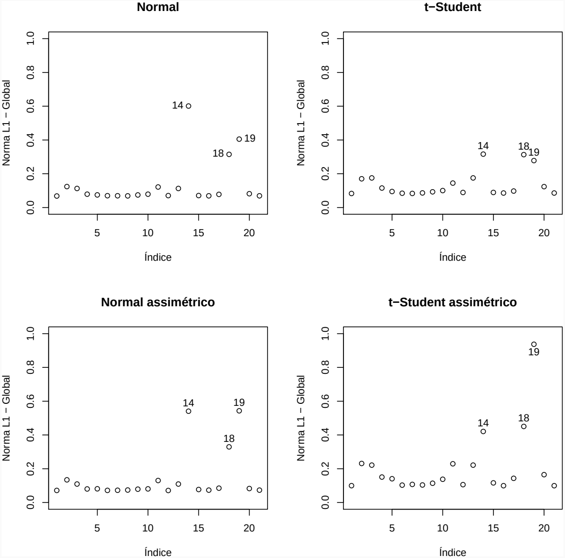


Figura 21: Influência global via norma L_1 dos modelos contaminando a observação 14

Contaminação da observação 14

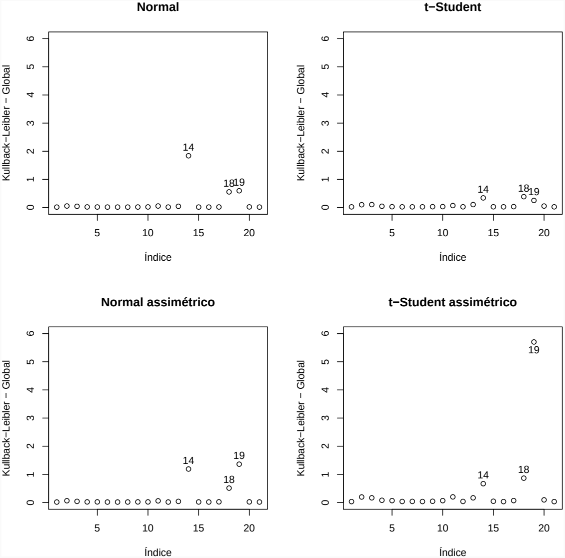


Figura 22: Influência global via divergência de Kullback-Leibler dos modelos contaminando a observação 14

Contaminação da observação 14

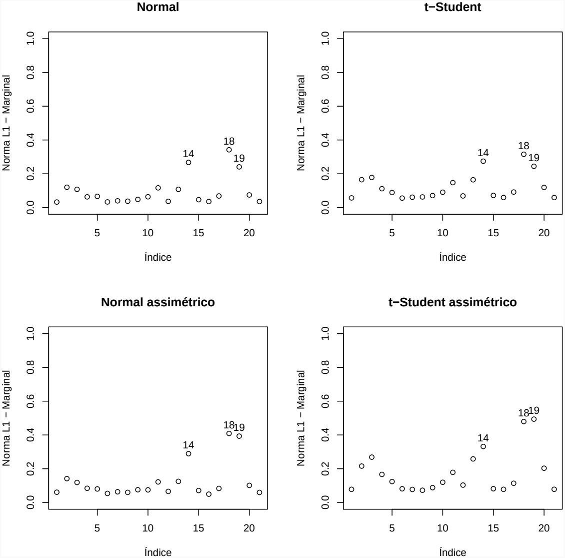


Figura 23: Influência marginal via norma L_1 dos modelos contaminando a observação 14

Contaminação da observação 14

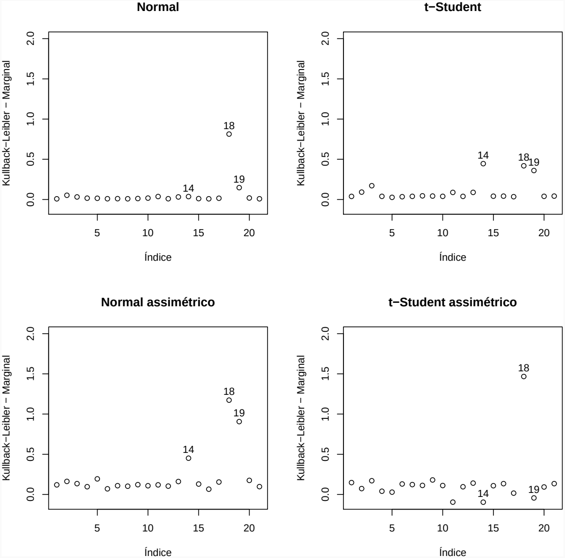
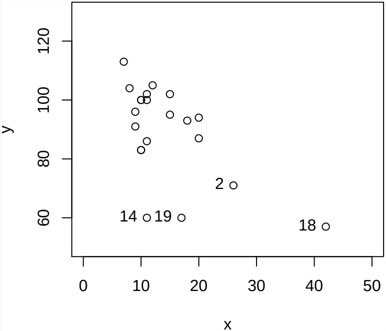


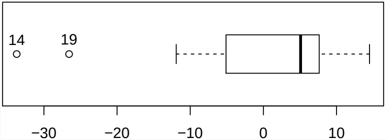
Figura 24: Influência marginal via divergência de Kullback-Leibler dos modelos contaminando a observação 14

Normal	t-Student	Normal assimétrica	t-assimétrica
-86.15	-84.95	-80.92	-81.58

Tabela 5: LPML dos modelos contaminando as observações 14 e 19



(a)



(b)

Figura 25: Diagrama de dispersão e boxplot dos resíduos do modelo de regressão normal contaminando as observações 14 e 19

Contaminação das observações 14 e 19

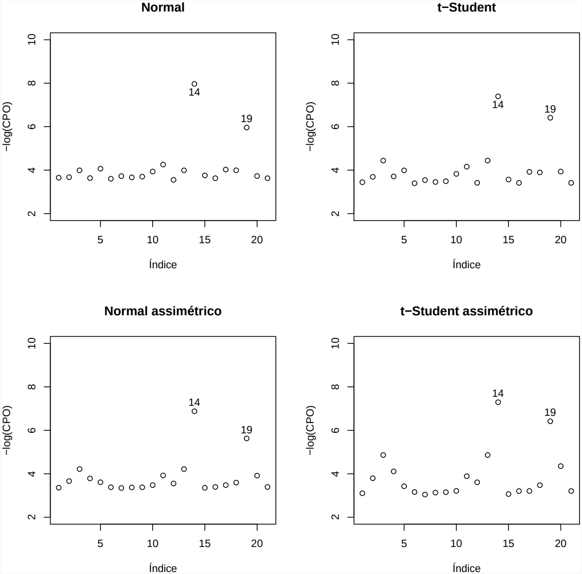


Figura 26: $-\log(CPO)$ dos modelos contaminando as observações 14 e 19

Contaminação da observações 14 e 19

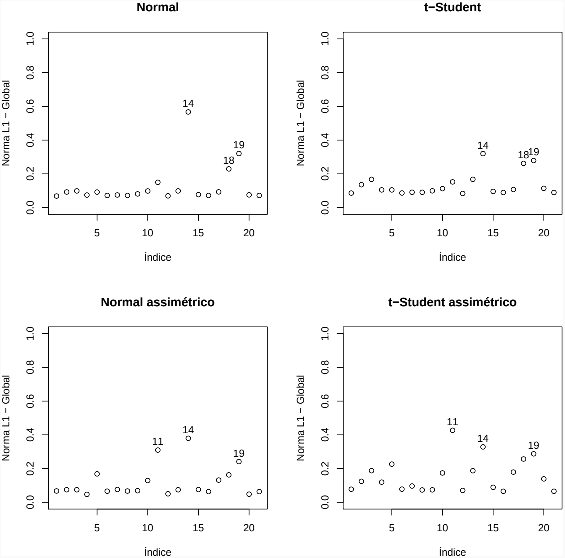


Figura 27: Influência global via norma L_1 dos modelos contaminando a observações 14 e 19

Contaminação da observações 14 e 19

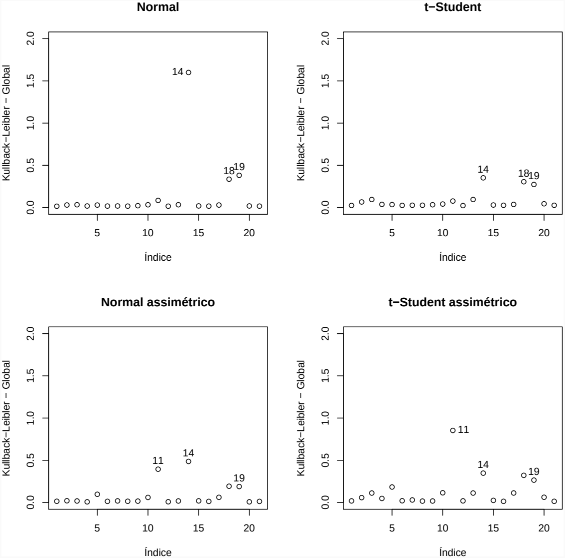


Figura 28: Influência global via divergência de Kullback-Leibler dos modelos contaminando a observações 14 e 19

Contaminação da observações 14 e 19

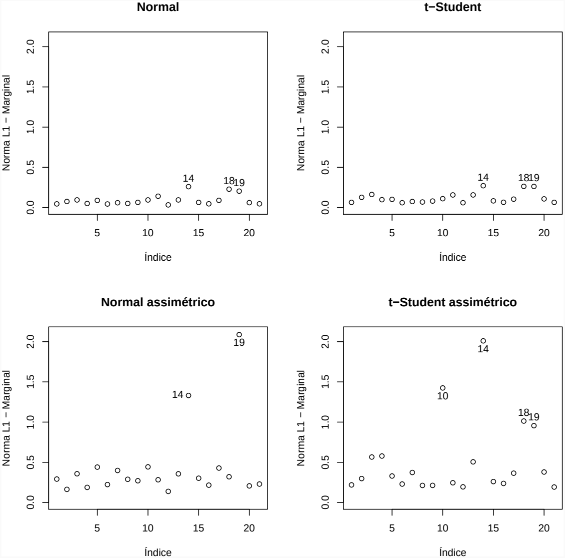


Figura 29: Influência marginal via norma L_1 dos modelos contaminando a observações 14 e 19

Contaminação da observações 14 e 19

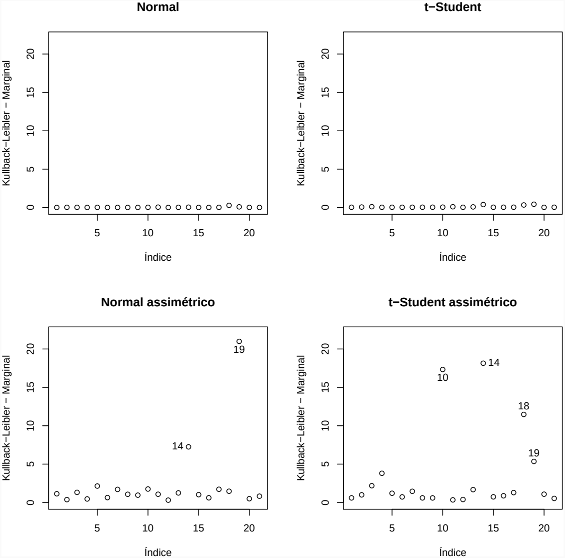


Figura 30: Influência marginal via divergência de Kullback-Leibler dos modelos contaminando a observações 14 e 19

Sumário

Introdução

Diagnóstico Bayesiano em Modelos de Regressão Simétricos

- Medidas de Diagnóstico
- Modelo de Regressão Normal
- Modelo de Regressão t-Student
- Aplicação

Diagnóstico Bayesiano em Modelos de Regressão Assimétricos

- Modelo de Regressão Normal Assimétrico
- Modelo de Regressão t-Student Assimétrico
- Aplicação

Conclusões

Comentários gerais

- Em geral, o modelo t-Student é uma alternativa robusta ao modelo normal.
- A melhor aplicação para o modelo t-Student é quando os resíduos possuem uma certa simetria.
- Nas aplicações estudadas, as estimativas dos coeficientes regressores foram pouco influenciadas.
- O modelo t-assimétrico não é, em geral, uma alternativa robusta ao modelo normal.
- O modelo t-assimétrico será mais robusto que o normal assimétrico caso os pontos discrepantes estejam na cauda de maior peso dos resíduos.
- As principais conclusões deste trabalho baseiam-se na análise de influência global pois o cálculo da influência marginal mostrou-se instável.

Contribuições

- Obtenção das medidas de diagnóstico global e marginal calculadas para os modelos t-Student, normal assimétrico e t-assimétrico.
- Dedução teórica destas medidas estendendo o trabalho de *Weiss e Cho (1998)*, que abordou o caso normal, e também apresentamos como obter estimativas destas medidas por meio da integração de Monte Carlo.
- Adição ao trabalho de *Godoi (2007)* exibindo as distribuições condicionais completas da distribuição t-assimétrica possibilitando o uso do algoritmo de Gibbs para obtenção de uma amostra da distribuição *a posteriori*, no caso de ν fixado.

Pesquisas futuras

- Considerar os graus de liberdade, ν, da distribuição desconhecidos e obter as condicionais completas explicitamente para este caso.
- Estudar a sensibilidade do cálculo das estimativas de influência de acordo com o tamanho da amostra de Monte Carlo considerada.
- Estudar outras divergências que não norma L_1 e divergência de Kullback-Leibler.
- Para os modelos normal e t-Student, obter estimativas de influência marginal para as componentes do vetor $\boldsymbol{\beta}$ e para σ^2.
- Para os modelos normal assimétrico e t-assimétrico, obter estimativas de influência marginal para as componentes do vetor $\boldsymbol{\beta}$, para σ^2 e para λ.