Um estudo sobre aspectos probabilísticos e inferenciais frequentistas da distribuição t de Student generalizada de 3 parâmetros

Bolsista: Fernanda Ribeiro (nanda.ribeiro1990@gmail.com) **Orientador:** Prof. Dr. Caio Lucidius Naberezny Azevedo

Instituto de Matemática, Estatística e Computação Científica - IMECC

Projeto Financiado pela FAPESP (**Processo:** 2011/16158-7)

Palavras-Chave: Distribuição t de Student Generalizada - Aspectos Probabilísticos - Inferência Frequentista

Introdução

A distribuição t de Student generalizada (DTSG), como o nome sugere, é uma generalização da distribuição t de Student (DTS) na medida que considera um parâmetro de escala que oferece propriedades adicionais. Além de acomodar valores discrepantes, a velocidade de decaimento para as caudas, em relação a densidade, é controlada de modo mais flexível por esse segundo parâmetro.

No presente projeto consideramos a DTSG de 3 parâmetros (DTSG3P), que corresponde à DTSG adicionando-se um parâmetro de localização, correspondente à média. Apresentamos os principais aspectos probabilísticos, analítica e graficamente, e inferenciais, frequentista e Bayesiano, desta distribuição.

Metodologia

Seja X v.a. DTSG3P, $X \sim tg(\mu, 1, \nu_1, \nu_2)$ em que denotamos o parâmetro média por μ, o parâmetro escala por ν_2 e o grau de liberdade por ν_1.

A função densidade de probabilidade da v.a. X é dada por:

$$f_X(x) = \frac{1}{\sqrt{\pi}} \frac{\Gamma\left(\frac{(\nu_1+1)}{2}\right)}{\sqrt{\nu_2}\Gamma\left(\frac{\nu_1}{2}\right)} \left(1 + \frac{(x-\mu)^2}{\nu_2}\right)^{-\frac{(\nu_1+1)}{2}}$$

que tem a seguinte representação hierárquica:

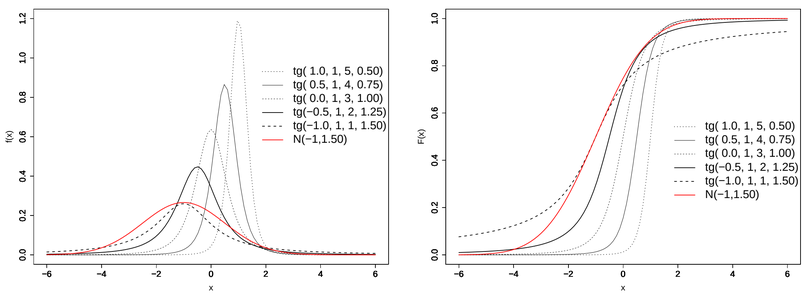
$$X_i|U_i = u_i \overset{indep.}{\sim} \mathrm{N}\left(\mu, \frac{1}{u_i}\right) \quad \text{e} \quad U_i \overset{indep.}{\sim} \mathrm{Gama}\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right), i = 1, \ldots, n$$

e

$$\mathbb{E}(X) = \mu \quad \text{e} \quad Var(X) = \frac{\nu_2}{\nu_1 - 2}, \quad \nu_1 > 2$$

Aspectos probabilísticos

A Figura 1 apresenta o comportamento da função densidade de probabilidade (fdp) e distribuição acumulada (fda). O controle da forma da densidade para a v.a. X, assim como a acomodação de valores extremos e caudas mais pesadas, diferem da distribuição Normal. Dessa maneira, com as caudas mais densas, valores extremos são mais prováveis de ocorrer com DTSG3P.



(a) Densidade (b) Acumulada

Figura 1: Gráfico para $X \sim tg(\mu, 1, \nu_1, \nu_2)$, variando todos os parâmetros

Temos que o parâmetro ν_1 é responsável pela forma da distribuição, em que para valores menores a densidade assume maior massa de probabilidade nos extremos e menor no centro, o contrário ocorre para valores maiores, em que a concentração de massa de probabilidade é maior no centro. O parâmetro ν_2 controla a velocidade de decaimento para as caudas de modo mais flexível, em que quanto menor o valor, maior a velocidade.

Aspectos Inferenciais

Em relação aos principais aspectos inferenciais, desenvolvemos os métodos de estimação pontual com o objetivo de estimar todos os parâmetros envolvidos, $\boldsymbol{\theta} = (\mu, \nu_1, \nu_2)$, tais como: método dos momentos (MM), máxima verossimilhança via algoritmo *Newton-Raphson* (NR), implementando o algoritmo (NR^I) e considerando a função `maxLik` disponível no *software* **R** (NR^R), Escore de *Fisher* (EF), EM e BFGS (função `optim`, no **R**).

Além disso utilizamos o método de Monte Carlo via cadeias de Markov (MCMC). Nesse caso, para a obtenção das estimativas dos parâmetros são utilizadas técnicas de simulação iterativa baseadas em cadeias de Markov utilizando o *software* **WinBUGS**. Apresentamos a seguir uma breve descrição dos métodos implementados:

- **Método dos momentos:** os estimadores por métodos dos momentos são encontrados igualando os k primeiros momentos amostrais com os correspondentes k momentos populacionais, e resolvendo simultaneamente o sistema de equações resultantes. Temos que os estimadores por métodos dos momentos para os parâmetros que envolvem a DTSG3P, são:

$$\hat{\mu} = \sum_{i=1}^{n} \frac{X_i}{n} = \bar{X}_n$$

$$\hat{\nu}_1 = \frac{-4\left(\sum_{i=1}^n \frac{X_i^4}{n} - 6\sum_{i=1}^n \frac{X_i^2}{n}\bar{X}_n^2 + 5\bar{X}_n^4\right) + 6\left(\sum_{i=1}^n \frac{X_i^2}{n} - \bar{X}_n^2\right)^2}{3\left(\sum_{i=1}^n \frac{X_i^2}{n} - \bar{X}_n^2\right)^2 - \left(\sum_{i=1}^n \frac{X_i^4}{n} - 6\sum_{i=1}^n \frac{X_i^2}{n}\bar{X}_n^2 + 5\bar{X}_n^4\right)}$$

e

$$\hat{\nu}_2 = \left(\frac{-4\left(\sum_{i=1}^n \frac{X_i^4}{n} - 6\sum_{i=1}^n \frac{X_i^2}{n}\bar{X}_n^2 + 5\bar{X}_n^4\right) + 6\left(\sum_{i=1}^n \frac{X_i^2}{n} - \bar{X}_n^2\right)^2}{3\left(\sum_{i=1}^n \frac{X_i^2}{n} - \bar{X}_n^2\right)^2 - \left(\sum_{i=1}^n \frac{X_i^4}{n} - 6\sum_{i=1}^n \frac{X_i^2}{n}\bar{X}_n^2 + 5\bar{X}_n^4\right)} - 2\right)\left(\sum_{i=1}^n \frac{X_i^2}{n} - \bar{X}_n^2\right)$$

- **Máxima Verossimilhança via algoritmo NR e EF:** devido a dificuldade em encontrar um estimador, já que em muitos casos não se encontra uma forma analítica "fechada" para o cálculo, surge a necessidade desses métodos de otimização não-linear, ou seja, métodos iterativos que maximizam numericamente a função de log-verossimilhança.

Algumas vantagens na utilização desses métodos são descritos em Jennrich e Sampson (1976), como por exemplo, os erros padrão para as estimativas dos parâmetros cair automaticamente em ambos os algoritmos, além da vantagem prática do algoritmo EF, por sua robustez para valores iniciais "pobres".

- **Máxima Verossimilhança via algoritmo EM:** trata-se de um algoritmo iterativo, que envolve em cada iteração duas etapas, chamadas de passo E (esperança) e passo M (maximização). Considerando $\boldsymbol{\theta}^{(k)} = \left(\mu^{(k)}, \nu_1^{(k)}, \nu_2^{(k)}\right)^T$ a estimativa de $\boldsymbol{\theta}$ para a k-ésima iteração, temos que a esperança com respeito a $\boldsymbol{u}$, condicionada a $\boldsymbol{x}$, da função log-verossimilhança completa, $\mathbb{E}\left[l_c(\boldsymbol{\theta}|\boldsymbol{y_c})\,|\boldsymbol{y}, \hat{\boldsymbol{\theta}}^{(\boldsymbol{k})}\right]$, é dada por:

$$Q\left(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(k)}\right) \propto -\frac{1}{2}\sum_{i=1}^n \hat{u}_i\left(x_i - \mu^{(k)}\right)^2 - n\ln\Gamma\left(\frac{\nu_1^{(k)}}{2}\right) + \frac{n\nu_1^{(k)}}{2}\ln\left(\frac{\nu_2^{(k)}}{2}\right) - \frac{\nu_2^{(k)}}{2}\sum_{i=1}^n \hat{u}_i + \frac{\nu_1^{(k)}}{2}\sum_{i=1}^n \widehat{\ln u_i}^{(k)}$$

em que obtemos o seguinte algoritmo EM:

- Passo E: Dado $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^{(k)}$, calcule $\hat{u}_i^{(k)}$ e $\widehat{\ln(u)}_i^{(k)}$, para $i = 1, 2, \ldots, n$.
- Passo M: Atualize $\hat{\boldsymbol{\theta}}^{(k+1)}$ maximizando $Q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}\right)$ em $\boldsymbol{\theta}$.

Os passos são repetidos até que haja convergência. Utiliza-se o seguinte critério de parada: $|l\left(\boldsymbol{\theta}^{(k+1)}\right) - l\left(\boldsymbol{\theta}^{(k)}\right)| < \epsilon$, em que $l(.)$ representa a log-verossimilhança e ϵ é um valor maior que zero.

- **Monte Carlo via cadeias de Markov:** conforme Ehlers (2007), os métodos MCMC são uma alternativa aos métodos não iterativos em problemas complexos. De modo resumido, consiste em obter uma amostra da distribuição a posteriori e calcular as estimativas amostrais de características desta distribuição. No **WinBUGS**, com base nas funções proposta o algoritmo de amostragem mais eficiente é selecionado para a simulação. Nesse caso, consideramos as seguintes distribuições prioris vagas dos componentes aleatórios,

$$\mu \sim N(0, 20), \quad \nu_1 \sim \mathrm{Gama}(0, 1; 0, 1) \quad \text{e} \quad \nu_2 \sim \mathrm{Gama}(0, 1; 0, 0001)$$

Resultados e Discussão

Com o objetivo de comparar os métodos propostos realizamos um estudo de simulação no qual replicou-se o processo de estimação 50 vezes. Foram gerados amostras de tamanho 100 com distribuição $tg(2; 1; 6; 0, 2)$. No caso do método MCMC foram realizadas 50000 iterações no qual foram descartadas as 20000 primeiras para o período de aquecimento e considerado salto igual a 30.

Na Figura 2 apresentamos os gráficos para as estimativas dos parâmetros em relação à cada método. A Tabela 1 apresenta as estatísticas relativas às estimativas do parâmetros, como por exemplo, o cálculo do viés, que avalia se os parâmetros são superestimados (viés positivo) ou subestimado (viés negativo) e o REQM, medida de precisão que aumenta com a diferença entre os valores observados e verdadeiros.

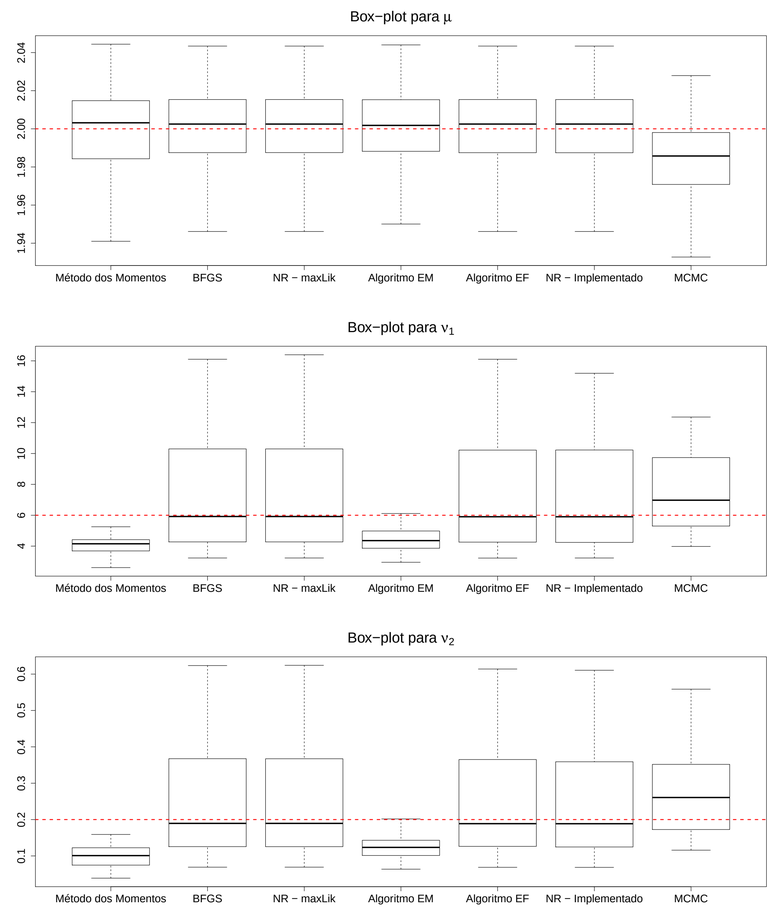


Figura 2: Box-plot para as estimativas dos parâmetros

Tabela 1: Estatísticas relativas às estimativas dos parâmetros

Parâmetros	Estatísticas	Métodos						
		MM	BFGS	NR^R	EM	EF	NR^I	MCMC
μ	Média	1,9991	1,9995	1,9995	1,9993	1,9995	1,9995	1,9822
	Máximo	2,0444	2,0434	2,0434	2,0440	2,0434	2,0434	2,0279
	Variância	0,0006	0,0005	0,0005	0,0005	0,0005	0,0005	0,0005
	Viés	-0,0009	-0,0005	-0,0005	-0,0007	-0,0005	-0,0005	-0,0178
	REQM	0,0240	0,0222	0,0222	0,0223	0,0222	0,0222	0,0290
	Erro relativo	0,0005	0,0003	0,0003	0,0003	0,0003	0,0003	0,0089
	% Outlier	0	0	0	2	0	0	2
ν_1	Média	4,1046	8,7637	10,8777	4,4613	20,5649	11,9353	7,5280
	Máximo	5,9984	47,4545	128,9937	6,1134	619,8693	204,0730	12,3588
	Variância	0,4494	69,1364	369,9684	0,5151	7550,2726	816,0675	6,0208
	Viés	-1,8954	2,7637	4,8777	-1,5387	14,5649	5,9353	1,5280
	REQM	2,0105	8,7621	19,8434	1,6979	88,1045	29,1770	2,8906
	Erro relativo	0,3159	0,4606	0,8130	0,2565	2,4275	0,9892	0,2547
	% Outlier	4	6	6	0	6	6	0
ν_2	Média	0,1014	0,3035	0,3799	0,1260	0,7295	0,4169	0,2740
	Máximo	0,2232	1,6550	4,5973	0,2246	22,3389	7,3044	0,5586
	Variância	0,0013	0,0976	0,4824	0,0013	9,8244	1,0574	0,0132
	Viés	-0,0986	0,1035	0,1799	-0,0740	0,5295	0,2169	0,0740
	REQM	0,1048	0,3291	0,7175	0,0821	3,1788	1,0509	0,1367
	Erro relativo	0,4928	0,5175	0,8996	0,3700	2,6475	1,0844	0,3699
	% Outlier	2	6	6	2	6	6	0

[0] Raiz do erro quadrático médio

Conclusões

Comparando os métodos na Figura 2, nota-se um melhor desempenho para os algoritmos BFGS, NR e EF. Além disso, observa-se que os métodos EM e MM tendem a subestimar os parâmetros ν_1 e ν_2, enquanto que os demais métodos superestimam.

Referências

FONSECA, T. C. O. da. *Análise Bayesiana de Referência para a classe de Distribuições Hiperbólicas Generalizadas*. Tese (Doutorado) – Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2004.

FONSECA, T. C. O. da; FERREIRA, M.A.; MIGON, H. S. Objective Bayesian analysis for the Student-t regression model. *Biometrika*, v. 95, p. 325-333, 2008.

JOHNSON, N. L.; KOTZ, S.; BALAKRISHNAN, N. *Continuos Univariate Distributions*. [S.l.]: John Wiley & Sons, 1995.