Um aplicativo em *Shiny* para ajuste do modelo de regressão quantílica Chen

Alisson Rosa Pereira[1] **& Laís Helen Loose**[2]

[1]Graduando em Estatística, UFSM - alirpereira887@gmail.com

[2]Departamento de Estatística, UFSM - lais.loose@ufsm.br

24º Simpósio Nacional de Probabilidade e Estatística

Introdução

Na última década, novas distribuições foram propostas com o objetivo principal de fornecer modelos mais flexíveis que possibilitem melhores soluções quando aplicados a dados reais. Uma distribuição bastante flexível e pouco explorada na literatura é a distribuição Chen, a qual foi introduzida por [1] no contexto de análise de sobrevivência a fim de possibilitar ajustes em dados com taxa de falha crescentes e em formato de banheira.

Os modelos de regressão mais conhecidos e geralmente utilizados em ajustes a dados reais são baseados no pressuposto de normalidade. No entanto, esta suposição nem sempre é satisfeita na prática. Como consequência, tem crescido o interesse no desenvolvimento e análise de modelos não gaussianos. Apesar de uma ampla gama de novos modelos estarem sendo desenvolvidos nos últimos anos, a falta de ferramentas computacionais para ajuste desses modelos ainda é um limitador para o amplo uso de recentes propostas em áreas aplicadas

Assim, o presente trabalho tem como objetivo apresentar o modelo de regressão quantílico Chen e fornecer uma ferramenta *web* para ajuste do modelo. Para o desenvolvimento do aplicativo *web* utilizamos o pacote Shiny, disponível na linguagem R [3].

O modelo de regressão quantílico Chen: formulação, inferência e adequação do ajuste

Se Y é uma variável aleatória com distribuição Chen, então sua função densidade de probabilidade é dada pela forma abaixo [1]

$$f(y;\lambda,\delta) = \delta\lambda y^{\lambda-1}\exp\left\{\delta\left[1-\exp(y^{\lambda})\right]+y^{\lambda}\right\}, \quad y>0, \tag{1}$$

em que $\delta, \lambda > 0$ são os parâmetros de forma.

Consideremos $\mu = Q(\tau;\lambda,\delta)$ e que δ possa ser escrito como $\delta = \dfrac{\log(1-\tau)}{1-\exp(\mu^{\lambda})}$. Reescrevendo a função densidade dada pela equação (1) em termos da expressão de δ, a densidade reparametrizada em termos do quantil (μ) fica dada por

$$f(y;\lambda,\mu,\tau) = \frac{\log(1-\tau)}{1-\exp(\mu^{\lambda})}\lambda y^{\lambda-1}\exp\left[\frac{\log(1-\tau)}{[1-\exp(\mu^{\lambda})]}\left[1-\exp(y^{\lambda})\right]+y^{\lambda}\right]. \tag{2}$$

Dessa maneira podemos formalizar o modelo de regressão quantílico supondo $Y_1,\ldots,Y_n$ variáveis aleatórias independentes, em que cada Y_t, $t=1,\ldots,n$, segue a densidade em (2) com quantil μ_t, λ um parâmetro desconhecido e $\tau \in (0,1)$ conhecido. Assumimos então que o quantil (μ_t) de y_t pode ser escrito como

$$g(\mu_t) = \boldsymbol{x}_t^{\top}\boldsymbol{\beta} = \eta_t, \tag{3}$$

em que $\boldsymbol{\beta} = (\beta_0,\ldots,\beta_k)^{\top}$ é um vetor de parâmetros de regressão desconhecidos ($\boldsymbol{\beta} \in \mathbb{R}^{k+1}$) e $\boldsymbol{x}_t^{\top} = (1, x_{t_1},\ldots,x_{t_k})$ são observações sobre k covariáveis ($k<n$), que são assumidas como fixas e conhecidas. η_t é chamado preditor linear, $g(\cdot)$ é uma função de ligação, monótona e duas vezes diferenciável, tal que $g:\mathbb{R}^{+}\to\mathbb{R}$.

O modelo de regressão quantílico Chen é definido pelas expressões (2) e (3) [6]. Devido a restrição de que $\mu_t > 0$, a função de ligação mais usual nesse contexto é a logarítmica, $g(\mu_t) = \log(\mu_t)$.

Seja $y_1,\ldots,y_n$ uma amostra do modelo de regressão quantílico Chen proposto e $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\top},\lambda)^{\top}$ o vetor de parâmetros. A função de log-verossimilhança baseada nessa amostra é dada por

$$\ell(\boldsymbol{\theta}) = \sum_{t=1}^{n}\log\{f(y_t;\lambda,\mu_t,\tau)\}. \tag{4}$$

Os estimadores de máxima verossimilhança do vetor paramétrico $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\top},\lambda)^{\top}$ são obtidos pela maximização da log-verossimilhança.

O vetor escore $\boldsymbol{U}(\boldsymbol{\theta}) = (\boldsymbol{U}_{\boldsymbol{\beta}}(\boldsymbol{\theta})^{\top}, U_{\lambda}(\boldsymbol{\theta}))^{\top}$ é obtido pela diferenciação da função de log verossimilhança em relação a cada elemento de $\boldsymbol{\theta}$. Os vetores escore relativos a cada um dos parâmetros $\beta_i, i=1,\ldots,k+1$, do vetor da estrutura de regressão $\boldsymbol{\beta}$ e λ, são dados, respectivamente, por

$$U_{\beta_i}(\boldsymbol{\theta}) = \frac{\partial\ell(\boldsymbol{\theta})}{\partial\beta_i} = \sum_{t=1}^{n}\frac{\partial\ell_t(\mu_t,\lambda)}{\partial\mu_t}\frac{d\mu_t}{d\eta_t}\frac{\partial\eta_t}{\partial\beta_i} \quad \text{e} \quad U_{\lambda}(\boldsymbol{\theta}) = \frac{\partial\ell(\boldsymbol{\theta})}{\partial\lambda} = \sum_{t=1}^{n}\frac{\partial\ell_t(\mu_t,\lambda)}{\partial\lambda}.$$

Os estimadores de máxima verossimilhança $\widehat{\boldsymbol{\beta}}$ e $\widehat{\lambda}$ dos parâmetros $\boldsymbol{\beta}$ e λ são obtidos resolvendo o seguinte sistema de equações não lineares:

$$\begin{cases}\boldsymbol{U}_{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \boldsymbol{0},\\ U_{\lambda}(\boldsymbol{\theta}) = 0.\end{cases} \tag{5}$$

Este sistema não possui solução analítica em forma fechada, sendo necessário o uso de algoritmos de otimização não lineares. No presente trabalho utilizamos o método de Nelder-Mead.

A matriz de informação esperada é dada por

$$\boldsymbol{K}(\boldsymbol{\theta}) = E\left[-\frac{\partial^2\ell(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\top}}\right] = E[\boldsymbol{J}(\boldsymbol{\theta})].$$

A distribuição Chen não tem expressões analíticas em forma fechada para os momentos, logo não é possível obter expressões analíticas para a matriz de informação de Fisher. A matriz de informação observada $\boldsymbol{J}(\widehat{\boldsymbol{\theta}})$ é um estimador consistente para $\boldsymbol{K}(\boldsymbol{\theta})$. Sendo assim, no presente trabalho utilizaremos a matriz de informação observada.

Sob condições usuais de regularidade, os estimadores de máxima verossimilhança $\widehat{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}$ são consistentes, com distribuição aproximada normal $(k+2)$-variada, com vetor de médias $\boldsymbol{\theta}$ e matriz de variância e covariâncias $\boldsymbol{K}(\boldsymbol{\theta})^{-1}$ em grandes amostras, ou seja,

$$\begin{pmatrix}\widehat{\boldsymbol{\beta}}\\ \widehat{\lambda}\end{pmatrix} \sim \mathcal{N}_{k+2}\left(\begin{pmatrix}\boldsymbol{\beta}\\ \lambda\end{pmatrix}, \boldsymbol{K}(\boldsymbol{\theta})^{-1}\right),$$

em que $\widehat{\boldsymbol{\beta}}$ e $\widehat{\lambda}$ são os estimadores de $\boldsymbol{\beta}$ e λ, respectivamente.

Para a avaliação do modelo utilizamos o coeficiente de determinação generalizado, definido por:

$$R^2 = 1-\exp\left[-\frac{2}{n}\left[\ell(\widehat{\boldsymbol{\theta}})-\ell(\widehat{\boldsymbol{\theta}}_0)\right]\right], \tag{6}$$

em que $\ell(\widehat{\boldsymbol{\theta}}_0)$ é a log verossimilhança maximizada do modelo sem regressores (nulo), $\ell(\widehat{\boldsymbol{\theta}})$ é a log verossimilhança maximizada do modelo ajustado e n é o tamanho da amostra.

Ainda, para a avaliação do ajuste do modelo utilizamos o resíduo quantílico que é definido em [4] por:

$$r_t = \Phi^{-1}\left\{\int_0^{y_t} f(u_t;\widehat{\lambda},\widehat{\mu}_t,\tau)du_t\right\},$$

em que $\Phi(\cdot)$ é a função de distribuição acumulada da Normal padrão.

Dada a estrutura aqui citada elaboramos um aplicativo em Shiny que possibilita o ajuste e verificação dos pressupostos básicos do modelo.

Aplicativo Shiny - Estrutura

Para o ajuste do modelo em Shiny, são necessários alguns *inputs* que devem ser fornecidos/escolhidos pelo usuário, sendo eles:

- **Dados**: O usuário pode fazer *upload* de um arquivo em formato `csv`, ou então usar dados simulados.
- **Escolha da variável resposta (Y)**: O usuário pode selecionar a variável resposta dentre as variáveis disponíveis nos dados inicialmente fornecidos.
- **Seleção de covariáveis ($\boldsymbol{X}$):** As variáveis explicativas podem ser selecionadas usando a sintaxe de fórmulas disponível no R, ou seja, se deseja, por exemplo, usar as covariáveis X_2 e X_3 então usa-se `X2 + X3`.
- **Escolha da função de ligação**: Pode-se escolher entre a função logarítmica (log) e raiz quadrada (sqrt).
- **Escolha do quantil**: A escolha do quantil pode ser feita escolhendo qualquer valor no intervalo (0,1) espaçado por 0.01.

Após a inserção de todos os *inputs* necessários deve-se clicar em fit para que o modelo seja ajustado.

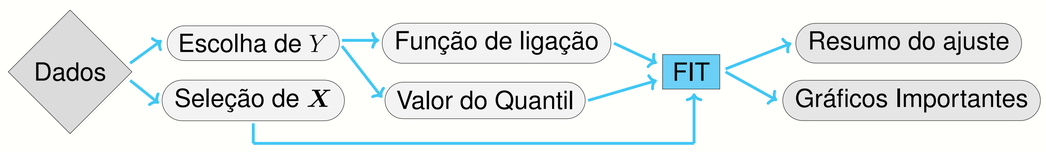

Figura 1: **Fluxograma da Estrutura**

Resumo do Ajuste

Após o modelo ser ajustado será fornecido um resumo do ajuste semelhante ao exemplo a seguir.

	estimate	std_error	z_value	p_value
lambda	0.969	0.0162	59.9251	<0.001
(Intercept)	-0.1632	0.0254	6.4333	<0.001
V2	2.5776	0.0179	144.189	<0.001
V3	-2.0559	0.0156	131.487	<0.001
V4	3.106	0.0218	142.7914	<0.001

Tabela 1: **Resumo do ajuste**

A Tabela contém um resumo do ajuste incluindo as estimativas dos parâmetros, erro padrão e p-valores associados ao testes de hipóteses. Tem-se também algumas métricas de seleção de modelos, como AIC, BIC e RMSE, que são fornecidas no aplicativo.

São fornecidas outras quantidades de interesse, sendo elas:

- **Number of iteractions**: Indica a quantidade de iterações que foram necessárias para a convergência do método de otimização.
- **Residuals**: Medidas descritivas básicas dos resíduos.
- **R-squared**: Informa o valor do coeficiente de determinação generalizado, definido em (6)

Uma tabela indicando medidas descritivas básicas sobre o banco de dados também é fornecida.

Gráficos Importantes

Os gráficos fornecidos para avaliar a qualidade do ajuste do modelo são:

- **Resíduos vs Valores Preditos:** Útil para detecção de *outliers*, em modelos bem ajustados os resíduos devem estar em torno de zero.
- **Densidade do Resíduo:** Espera-se que a densidade dos resíduos se comporte semelhante a uma normal padrão.
- **Resíduos vs Índices:** Se o modelo estiver bem ajustado os resíduos quantílicos possuem distribuição normal padrão, assim espera-se poucos valores fora do intervalo de -3 e 3.
- **Envelope simulado:** Se a distribuição assumida é adequada, espera-se que os pontos se encontrem dentro das bandas de confiança.

Considerações finais

- Atualmente está sendo desenvolvido um pacote em linguagem R para o uso do modelo de regressão quantílico Chen, onde pode-se ajustar o modelo e também avaliar a qualidade do ajuste. Porém evidentemente tal abordagem necessita que o usuário tenha conhecimentos básicos da linguagem R.
- O aplicativo em Shiny é uma boa alternativa para aqueles não tem ou não desejam ter conhecimento na linguagem R, mas desejam utilizar o modelo.

Acesse o Shiny aqui

Referências

[1] Zhenmin Chen. A new two-parameter lifetime distribution with bathtub shape or increasing failure rate function. *Statistics & Probability Letters*, 49(2):155–161, 2000.

[2] J. G. Chen M.H. & Ibrahim. Maximum likelihood methods for cure rate models with missing covariates. *Biometrics*, 57(1):43–52, 2001.

[3] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021.

[4] Dunn P. K. Smyth GK. Randomized quantile residuals. *Journal of Computational and Graphical Statistics. 1996.*

[5] Hadley Wickham. *Mastering shiny*. "O'Reilly Media, Inc.", 2021.

[6] Souza, G. de. Modelo de regressão quantílico Chen. Trabalho de Conclusão de Curso do Bacharelado em Estatística - UFSM, 2021.

Agradecimentos

Os autores agradecem ao Fundo de Incentivo à Pesquisa da UFSM, ao CNPq e a Sigma Jr Consultoria Estatística pelo auxílio financeiro recebido.