COMUNICAÇÃO

RESÍDUOS GENERALIZADOS DE COX-SNELL NA AVALIAÇÃO DO AJUSTE DE MODELOS

Cox-Snell generalized residuals in the evaluation of model fitting

Ana Lúcia Souza da Silva[1], Mário Javier Ferrua Vivanco[2], Fortunato Silva de Menezes[2]

RESUMO

Vários tipos de resíduos têm sido propostos para modelos de sobrevivência, sendo os mais adequados resultado dos resíduos generalizados de Cox e Snell (1968). O objetivo com este trabalho é avaliar a adequacidade de modelos por meio de gráficos de diagnósticos gerados a partir dos resíduos generalizados de Cox-Snell. Para ilustrar a teoria, foram feitas três aplicações. A primeira aplicação visou a ilustrar a lógica existente entre a plotagem dos resíduos ordenados de três distribuições, normal (0,1), logística (0,1) e valor extremo (0,1) *versus* as estatísticas de ordem esperadas desses resíduos de acordo com as distribuições assumidas. Para a segunda aplicação, foram utilizados dados de tempo de vida de isolantes, obtidos em Nelson (1990). A partir da verificação por meio dos gráficos de diagnósticos utilizando-se os resíduos generalizados de Cox-Snell, encontrou-se que o modelo apropriado para o tempo de vida dos isolantes era o log-normal. Para a terceira aplicação, foram analisados dados censurados referentes ao tempo de vida de pacientes, obtidos em Collett (1994). Avaliou-se a adequacidade de vários modelos por meio dos resíduos de Cox-Snell adaptados para dados de sobrevivência. Pelos resultados constatou-se que o modelo Weibull foi o mais adequado.

Termos para Indexação: Resíduos generalizados de cox-snell, adequacidade de modelos, gráficos de diagnósticos.

ABSTRACT

Several kinds of residuals have been proposed for survival models, the most suitable for this purpose are Cox and Snell (1968) generalized residuals. The objective of this work was to evaluate the adequacy of models by graphical diagnostics using Cox-Snell generalized residuals. To illustrate the theory three applications were considered. The first application sought to illustrate the heuristics by plotting ordered residuals from three distributions: normal (0,1), logistics (0,1) and extreme value (0,1), *versus* the expected order statistics of these residuals in consonance with the assumed distributions. The second application consisted of lifetime data of electric insulating, obtained by Nelson (1990). Starting from graphical diagnostics using Cox-Snell generalized residuals, it was found that the model appropriate for lifetime of electric insulating was log-normal. The third application referred to censored data of lifetime of patients, obtained by Collett (1994). The adequacy of several models was evaluated by Cox-Snell generalized residuals adapted for survival data. The results show that Weibull model was the most appropriate.

Index Terms: Cox-snell generalized residuals, adequacy of models, graphical diagnostics.

(Recebido para publicação em 28 de abril de 2003 e aprovado em 18 de agosto de 2004)

De maneira geral, a avaliação da adequacidade de ajuste de modelos a um conjunto de observações tem grande importância para validar a escolha do modelo a ser utilizado. Muitos procedimentos para essa checagem são baseados em quantidades conhecidas como resíduos.

Na análise de regressão, por exemplo, a partir da análise dos resíduos, pode-se detectar falhas no modelo, como as pressuposições distribucionais admitidas, a omissão incorreta do intercepto ou a necessidade de termos de ordem superior para a variável preditora.

Especificamente, em Análise de Sobrevivência, existem duas classes de modelos propostas na literatura como os modelos paramétricos e os modelos semiparamétricos, que também podem ser chamados de modelos de regressão de Cox. Dentro da classe dos modelos paramétricos, encontra-se a classe dos modelos de Tempo de Falha Acelerados (AFT), dados pela expressão:

$$LogT_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \sigma\varepsilon_i$$

para i = 1,2,...,n, em que T_i é o i-ésimo tempo até a ocorrência do evento correspondente à variável dependente, $\beta_0, \beta_1, \ldots, \beta_k$ e σ são os parâmetros desconhecidos, X_{ik} é a i-ésima observação correspondente à k-ésima variável independente e ε_i é o erro aleatório.

1. Mestranda em Agronomia/Estatística e Experimentação Agropecuária – Universidade Federal de Lavras/UFLA – Caixa Postal 3037 – 37200-000 – Lavras, MG, a_lss@bol.com.br
2. Professores Adjunto do Departamento de Ciências Exatas da UFLA, ferrua@ufla.br, fmenezes@ufla.br

Existem diversas propostas para avaliar a adequacidade desses modelos na literatura, como as sugeridas por Allison (1995) que propõe o estudo da adequacidade a partir do modelo Gama Generalizado e a partir dos gráficos de diagnósticos, isto é, por meio da análise de resíduos.

Vários tipos de resíduos têm sido propostos para modelos de sobrevivência, como os generalizados de Cox-Snell, os resíduos Deviance, Martingale e Score. Collett (1994) sugere como uma alternativa mais apropriada o uso dos Resíduos Generalizados de Cox-Snell.

Realizou-se este trabalho com o objetivo de avaliar a adequacidade de modelos de Tempo de Falha Acelerados (AFT) por meio dos gráficos de diagnósticos gerados a partir dos resíduos generalizados de Cox-Snell.

O desenvolvimento deste trabalho foi feito a partir das três aplicações:

Aplicação 01:

A primeira aplicação tem como objetivo mostrar a lógica da plotagem dos resíduos ordenados originados de uma distribuição *versus* as estatísticas de ordem esperadas pela distribuição.

Para esta aplicação, foram simuladas 4 amostras de tamanho 50, e consideradas como resíduos pertencentes a 3 distribuições, sendo uma amostra da distribuição normal (0,1), uma amostra da distribuição logistíca (0,1) e 2 amostras da distribuição valor extremo (0,1). Esses resíduos foram obtidos via simulação Monte Carlo.

Após a obtenção dos resíduos, o próximo passo foi ordená-los e calcular as estatísticas de ordem esperadas das distribuições propostas acima. As estatísticas de ordem esperadas das distribuições logistíca e valor extremo foram obtidas por meio do programa matemático MAPLE©V e as estatísticas de ordem esperadas da distribuição normal, via simulação.

Feita a plotagem dos resíduos ordenados *versus* as estatísticas de ordem esperadas, foram obtidos os seguintes resultados:

Pode-se observar nas Figuras 1,2,3 e 4, que na plotagem dos conjuntos dos resíduos para as distribuições sugeridas *versus* seus respectivos valores esperados das estatísticas de ordem, o resultado obtido foi próximo a uma linha reta, o que quer dizer que a distribuição assumida para os erros aleatórios, em cada uma das quatro situações, é a correta.

Cabe ressaltar que nas figuras 2, 3 e 4, a maioria dos pontos encontram-se abaixo de $E[e_{(i)}] = 2$, e a linha reta ajusta-se muito bem àqueles pontos. Os poucos pontos acima de $E[e_{(i)}] = 2$ desviam-se da linha reta, e para diagnosticar o porquê desse desvio, recomenda-se, num caso real, fazer uma análise de pontos influentes.

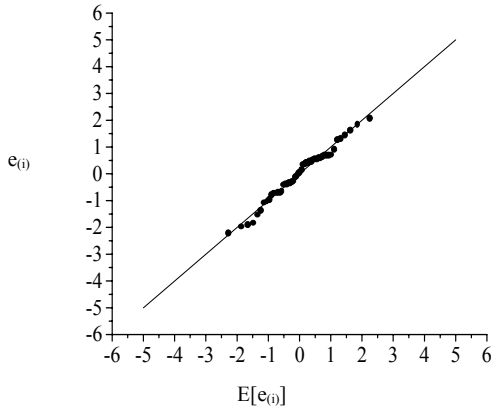

FIGURA 1 – Resíduos ordenados *versus* valores esperados das estatísticas de ordem, caso da distribuição normal (0,1). T tem distribuição Log-normal.

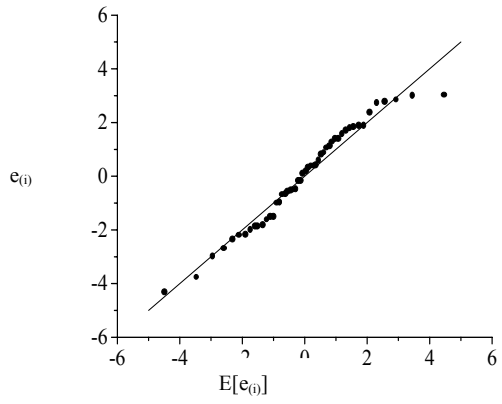

FIGURA 2 – Resíduos ordenados *versus* valores esperados das estatísticas de ordem, caso da distribuição logística (0,1). T tem distribuição Log-logística.

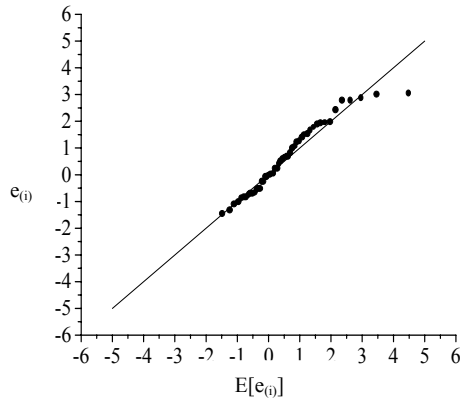

FIGURA 3 – Resíduos ordenados *versus* valores esperados das estatísticas de ordem, caso da distribuição valor extremo (0,1). T tem distribuição Weibull.

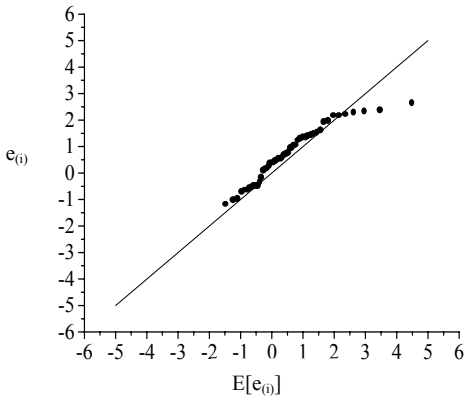

FIGURA 4 – Resíduos ordenados *versus* valores esperados das estatísticas de ordem, caso da distribuição valor extremo (0,1) de um parâmetro.T tem distribuição Exponencial.

Aplicação 02:

Com a segunda aplicação teve-se por finalidade ilustrar, com dados reais, a teoria dos resíduos generalizados de Cox-Snell, a partir de dados de tempo de vida de isolantes de sistemas de uma nova classe H, analisados por Nelson (1990), em que se aplicou um teste acelerado de vida nesses isolantes. Tais dados referem-se ao tempo, em horas, que os isolantes levaram para se tornarem defeituosos. Os testes foram realizados em pequenos motores, a temperaturas elevadas. Dez motores foram colocados para trabalhar a temperaturas de 190, 220, 240 e 260^0C e inspecionados periodicamente para detectar a ocorrência de falha. Segundo Nelson (1990), para o ajuste dos tempos de vida, foi considerado o modelo de regressão log-normal, como é dado a seguir:

$$LogT_{ij} = \beta_0 + \beta_1 x_i + \varepsilon_{ij}$$

para i = 1,2,...4 e j = 1,2,...10, em que $\log T_{ij}$ é o log do tempo de vida dos isolantes, x_i é a temperatura em graus centígrados, β_0 e β_1 são os parâmetros do modelo e ε_{ij} são erros aleatórios, que, segundo Nelson (1990), têm distribuição normal.

Calculando os resíduos modificados de acordo com a expressão a seguir:

$$R_i' = (1 + K_i)R_i + l_i$$

em que K_i, l_i e c_{ii}^+ são constantes definidas em Cox e Snell (1968) e R_i é o i-ésimo resíduo generalizado, obtêm-se os seguintes resultados mostrados na Tabela 1.

Cabe ressaltar que para o modelo proposto por Nelson (1990), os valores de l_i, que serviram para gerar os R_i', foram todos iguais a zero.

Após a obtenção dos resíduos generalizados de Cox-Snell, foram obtidas as estatísticas de ordem esperadas de uma normal, via simulação, e de uma exponencial, dada pela expressão $E(e_{(i)}) = \sum_{l=1}^{n} (n - l + 1)^{-1}$, segundo Lawless (1982).

Feita a plotagem dos resíduos modificados *versus* valores esperados das estatísticas de ordem das distribuições normal e exponencial, obtiveram-se os resultados representados nas Figuras 5 e 6.

Na Figura 6, verifica-se que os resíduos modificados não apresentaram a mesma distribuição dos erros aleatórios, pois esses tendem a afastar-se da reta, e, portanto, não seguem a distribuição exponencial.

É possível observar pela Figura 5 que os resíduos modificados têm distribuição equivalente à distribuição dos erros do modelo proposto para os tempos de vida dos isolantes. Dessa forma, pode-se confirmar que o modelo proposto por Nelson (1990), que assumia uma distribuição normal para os erros, está correto.

Sendo assim, verifica-se que os resíduos modificados de Cox-Snell fornecem resultados equivalentes aos resíduos de um modelo de regressão clássico, em que é assumida uma distribuição normal para os erros. Nesse caso, o uso dos resíduos modificados de Cox-Snell é equivalente ao uso dos resíduos clássicos num modelo de regressão.

TABELA 1 – Resíduos modificados obtidos para o ensaio do tempo de vida dos isolantes.

Obs.	c_{ii}^{+}	k_i	R_i	R'_i	Obs.	c_{ii}^{+}	k_i	R_i	R'_i
1	0.0041	-0.0267	-0.0647	-0.0630	21	0.0049	-0.0319	-0.3599	-0.3484
2	0.0041	-0.0267	-0.0647	-0.0630	22	0.0049	-0.0319	-0.3599	-0.3484
3	0.0041	-0.0267	-0.0647	-0.0630	23	0.0049	-0.0319	-0.1018	-0.0985
4	0.0041	-0.0267	0.0913	0.0888	24	0.0049	-0.0319	-0.0707	-0.0684
5	0.0041	-0.0267	0.1729	0.1683	25	0.0049	-0.0319	-0.0405	-0.0393
6	0.0041	-0.0267	0.1729	0.1683	26	0.0049	-0.0319	-0.0113	-0.0109
7	0.0041	-0.0267	0.1729	0.1683	27	0.0049	-0.0319	-0.0113	-0.0109
8	0.0041	-0.0267	0.1729	0.1683	28	0.0049	-0.0319	0.0171	0.0166
9	0.0041	-0.0267	0.3098	0.3015	29	0.0049	-0.0319	0.0448	0.0433
10	0.0041	-0.0267	0.3098	0.3015	30	0.0049	-0.0319	0.1482	0.1435
11	0.0045	-0.0295	-0.5622	-0.5456	31	0.0052	-0.0338	-0.4233	-0.4090
12	0.0045	-0.0295	-0.2394	-0.2323	32	0.0052	-0.0338	-0.2082	-0.2012
13	0.0045	-0.0295	-0.2394	-0.2323	33	0.0052	-0.0338	-0.2082	-0.2012
14	0.0045	-0.0295	-0.2394	-0.2323	34	0.0052	-0.0338	-0.2082	-0.2012
15	0.0045	-0.0295	-0.2394	-0.2323	35	0.0052	-0.0338	-0.0046	-0.0045
16	0.0045	-0.0295	-0.2394	-0.2323	36	0.0052	-0.0338	0.2079	0.2009
17	0.0045	-0.0295	0.0042	0.0041	37	0.0052	-0.0338	0.3651	0.3528
18	0.0045	-0.0295	0.0042	0.0041	38	0.0052	-0.0338	0.4687	0.4528
19	0.0045	-0.0295	0.0042	0.0041	39	0.0052	-0.0338	0.5625	0.5435
20	0.0045	-0.0295	0.0042	0.0041	40	0.0052	-0.0338	0.7272	0.7027

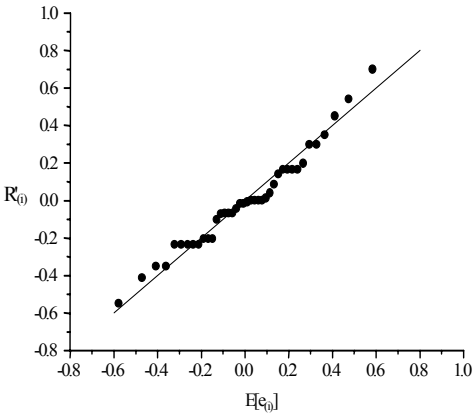

FIGURA 5 – Resíduos modificados *versus* valores esperados das estatísticas de ordem deuma distribuição normal.

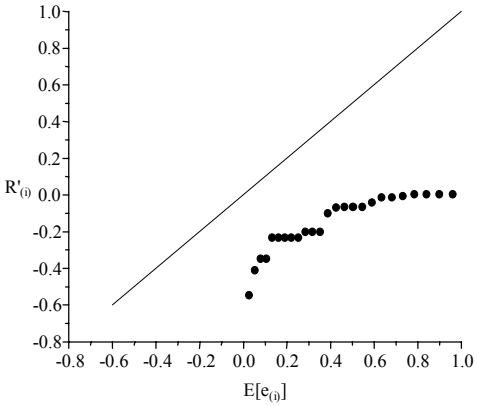

FIGURA 6 – Resíduos modificados *versus* valores esperados das estatísticas de ordem de uma distribuição exponencial.

Aplicação 03:

Para a terceira aplicação, foram utilizados dados com presença de censura, em que, para a obtenção dos resíduos de Cox-Snell, foi utilizada uma sub-rotina apresentada em SAS® por Allison (1994).

Os dados tratam do estudo de tempo de vida de pacientes com mieloma múltiplo, doença caracterizada pelo acúmulo de células plasmáticas anormais na medula, realizado pelo Centro Médico da Universidade da Virgínia do Oeste, USA, cuja finalidade era examinar a associação entre os valores de certas covariáveis e o tempo de sobrevivência dos pacientes. No estudo, a variável-resposta foi o tempo, em meses, a partir do diagnóstico antes da morte do paciente em conseqüência da doença. Os dados foram obtidos por Krall et al. (1975), em que foram registrados 48 pacientes com idade entre 50 e 80 anos. Durante o tempo de diagnóstico, os valores de um número de covariáveis foram registrados para cada paciente, incluindo a idade dos pacientes em anos, o sexo, sendo 1 para homem e 2 para mulher, o nível de uréia nitrogenada no sangue, cálcio e hemoglobina, a porcentagem de células plasmáticas na medula e um indicador da variável BC que denota a presença ou não de proteína Bence-Jones, dada por 0 se estiver ausente na urina e 1, caso contrário. Os pacientes que não morreram até o tempo em que o estudo foi completado tiveram o tempo de sobrevivência censurado (censura tipo I à direita) indicado por 0 e os pacientes que morreram a partir da miolema múltiplo receberam 1.

O modelo ajustado é um AFT (tempo de falha acelerado) dado por:

$$Log T_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \\ + \beta_6 X_{i6} + \beta_7 X_{i7} + \sigma\varepsilon_i$$

em que T_i é o i-ésimo tempo de sobrevivência para o i-ésimo paciente; $\beta_0, \beta_1, ..., \beta_7$ e σ são os parâmetros do modelo; $X_{i1}, X_{i2}, ..., X_{i7}$ correspondem às covariáveis do modelo e ε_i é o erro aleatório que, segundo Collett (1994), tem distribuição valor extremo.

Os resultados são apresentados nas Figuras 7, 8 e 9.

O tempo de vida dos pacientes com mieloma múltiplo demonstra melhor ajuste pelo modelo prababilístico Weibull (Figura 8), uma vez que os erros são distribuídos segundo a distribuição valor extremo, como proposto por Collet (1994).

Nas três aplicações apresentadas, observa-se que, de acordo com a teoria dos resíduos generalizados de Cox-Snell, é possível determinar uma expressão para os resíduos, de forma que eles tivessem uma distribuição muito próxima da distribuição dos erros aleatórios. Dessa forma, a verificação da adequacidade, a partir dos gráficos de diagnósticos, pode ser feita com a plotagem dos resíduos generalizados modificados de Cox-Snell ordenados *versus* as estatísticas de ordem esperadas da distribuição assumida pelos erros no início da avaliação. Sendo assim, se o resultado da plotagem for uma linha reta, isso significa que a distribuição proposta para os erros aleatórios está correta. Caso o resultado não seja uma linha reta, deve-se assumir uma outra distribuição para esses erros aleatórios.

Modificando os resíduos generalizados de Cox-Snell, pode-se obter os resíduos de Cox-Snell, conforme Allison (1995), para modelos de sobrevivência, em que qualquer que seja a distribuição dos erros, se ela for a distribuição correta, então a quantidade $e_i = -\log \hat{S}(t_i | x_i)$ tem distribuição exponencial com λ=1.

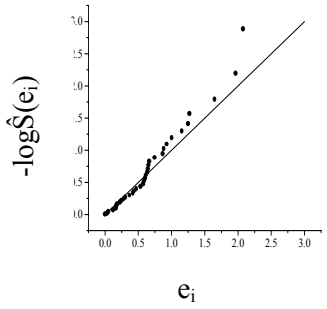

FIGURA 7 – Ajuste Log-normal.

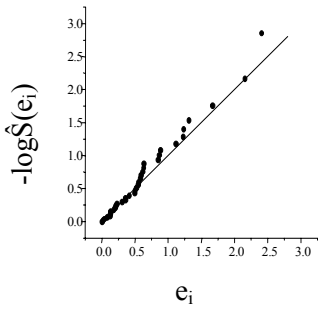

FIGURA 8 – Ajuste Weibull.

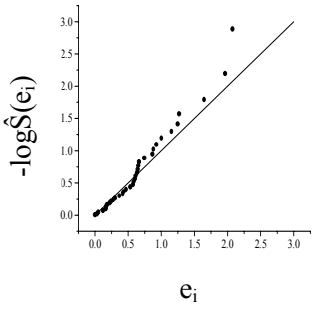

FIGURA 9 – Ajuste Exponencial.

REFERÊNCIAS BIBLIOGRÁFICAS

ALLISON, P. D. **Survival analysis using the SAS system**: a pratical guide. Cary: SAS Institute, 1995. 292 p.

COLLETT, D. **Modelling survival data medical research**. London: Chapman and Hall, 1994. 347 p.

COX, D. R.; SNELL, E. J. A general definition of residuals. **Journal of the Royal Statistical Society B**, London, v. 30, n. 2, p. 248-254, Mar. 1968.

KRALL, J. M.; UTHOFF, V. A.; HARLEY, J. B. A step-up procedure for selecting variables associated with survival. **Biometrics**, Washington, v. 31, n. 282, p. 49-57, June 1975.

LAWLESS, J. F. **Statistical models and methods for lifetime data**. New York: J. Wiley, 1982. 580 p.

NELSON, W. **Accelerated testing, statistical models, test plans and data analyses.** New York: J. Wiley, 1990. 621 p.