AVALIAÇÃO E CORREÇÃO DE VIÉS NO MODELO DE REGRESSÃO DE COX

J. G. N. Oliveira, F. R. B. Cruz, E. A. Colosimo
Departamento de Estatística - ICEx – UFMG
31270-901 - Belo Horizonte – MG
E-mail: {jangiovani,fcruz,enricoc}@ufmg.br

Resumo

O modelo de riscos proporcionais, proposto por Cox, é muito utilizado em diversas áreas do conhecimento e desempenha um papel fundamental na análise de dados de sobrevivência, pela sua flexibilidade em explorar associações entre covariáveis e taxas de falha. Entretanto, o estimador de máxima verossimilhança parcial (EMVP), normalmente utilizado para inferências neste modelo, é viesado para pequenas amostras. A presença de censuras, muito comum em situações reais, agrava este fenômeno. Correções neste viés são, portanto, de importância prática muito grande. Apresentamos resultados computacionais da avaliação do desempenho da correção de viés de segunda ordem do EMVP. As simulações Monte Carlo indicam viés menor, sem aumento de variância.

Palavras-chave: Dados censurados; estimador de máxima verossimilhança parcial; modelo de riscos proporcionais.

Abstract

The Cox regression model, employed in several areas of the human knowledge, has a crucial role in survival data analysis for its flexibility in exploring the association of covariates with failure rates. However, the maximum partial likelihood estimator (MPLE), based on which the inferences are usually taken, is biased for small sample sizes. Censoring, usually present in many practical situations, worsens the estimates so that searching for correction methods would be a boost for the applications. We present computational results for second-order bias evaluation and correction of MPLE's. Monte Carlos simulation results indicate smaller biases without variance inflation.

Keywords: Censored data; maximum partial likelihood estimates; proportional hazards model.

1. Introdução

Provavelmente, um dos métodos estatísticos mais importantes para a análise de dados censurados é o modelo de riscos proporcionais (MRP), proposto por Cox (1972), pela sua flexibilidade em explorar a associação entre covariáveis e taxas de falhas. Nos últimos anos, o MRP tem sido aplicado a diferentes situações práticas, da área médica à análise de dados

econômicos de ciclo de emprego e desemprego. O modelo de regressão de Cox, na forma exponencial da função risco, tem a seguinte expressão:

$$\lambda(t) = \lambda_0(t)\exp(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}), \qquad (1)$$

em que $\lambda_0(t)$ é a função de base, uma função do tempo, desconhecida e não-negativa, $\boldsymbol{\beta}^{\mathrm{T}} = (\beta_1, \beta_2, \ldots, \beta_p)$ é um vetor $p \times 1$ de parâmetros desconhecidos (a ser estimado), e $\mathbf{x}=(x_1,x_2,\ldots,x_p)^{\mathrm{T}}$ é o vetor de covariáveis.

A estimação dos coeficientes $\boldsymbol{\beta}$ na Equação (1) é baseada na função de verossimilhança parcial (Cox, 1975) a qual é viciada, tipicamente da ordem n^{-1}, sendo n o tamanho da amostra. Devido à sua natureza semi-paramétrica, os estimadores de máxima verossimilhança parcial (EMVP) são muito sensíveis a amostras de tamanho pequeno a médio, especialmente com alta proporção de censura. Infelizmente, em muitas das aplicações práticas nas quais o MRP é útil, o tamanho da amostra é muito pequeno. Na fase II de estudos clínicos, por exemplo, 20% de censuras reduziriam um grupo de 20 pacientes a apenas 16. Assim, em amostras de tamanho pequeno a moderado, como a da situação descrita, o viés pode ser consideravelmente alto. Portanto, a avaliação e a correção deste viés será de grande utilidade para as aplicações práticas. De fato, a avaliação e correção de estimativas por máxima verossimilhança têm recebido recentemente considerável atenção por parte dos pesquisadores (Ferrari et al., 1996; Cordeiro & McCullagh, 1991; Botter & Cordeiro, 1998; Colosimo et al., 2000; Montenegro et al., 2004, Cruz et al., 2004).

Este artigo utiliza-se das expressões para a correção de vício de segunda ordem para o EMVP de $\boldsymbol{\beta}$, recentemente desenvolvidas por Montenegro et al. (2004). Seu principal objetivo é estender os experimentos computacionais apresentados por Montenegro et al. (2004), como evidência do bom desempenho da correção proposta, via realização de um estudo de simulação Monte Carlo adicional, para diferentes distribuições e um número de covariáveis maior. Na Seção 2 apresentaremos as expressões em forma matricial, para cálculo do vício de ordem n^{-1} do EMVP no modelo (1). São apresentadas, na Seção 3, as simulações Monte Carlo realizadas para comparação entre o EMVP com a versão corrigida. Conclusões finais encerram o artigo, com a confirmação de que a versão corrigida pode produzir inferências muito mais confiáveis que aquelas fornecidas pelo EMVP.

2. Correção de viés

Seja $l=l(\boldsymbol{\beta})$ a função de máxima verossimilhança parcial, para uma amostra de n indivíduos, na qual $k \leq n$ falhas ocorrem nos tempos $t_1 \leq t_2 \leq \cdots \leq t_k$. Na ausência de empates, esta função para o modelo (1) pode ser escrita como:

$$l = \sum_{i=1}^{n} \delta_i \left[\boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_i - \log\left(\sum_{j \in R_{(t_i)}} \exp\left(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_j\right) \right) \right], \tag{2}$$

em que $R_{(t_i)} = \{k : t_k \leq t_i\}$ é a função de risco no tempo t_i, δ_i é um indicador de falha (δ_i=1, para falhas, e δ_i=0, para observações censuradas) e $\mathbf{x}_i=(x_{i1},x_{i2},...,x_{ip})^{\mathrm{T}}$ corresponde ao vetor linha de covariáveis para o i-ésimo indivíduo. O EMVP para $\boldsymbol{\beta}$ (viesado) é obtido pela maximização da Eq. (2). O interesse é corrigir esta estimativa viesada.

2.1.Correção analítica

Para se obter o viés de ordem n^{-1} para $\hat{\boldsymbol{\beta}}$, $B\left(\hat{\boldsymbol{\beta}}\right)$, a expressão de Cox e Snell (1968) pode ser utilizada na sua versão simplificada, observando que $\kappa_{rs}^{(t)} = \kappa_{rst}$, uma vez que os cumulantes esperados e observados são idênticos, condicionados à história de falhas e censuras (Cox & Oakes, 1984):

$$B\left(\hat{\boldsymbol{\beta}}\right) = \tfrac{1}{2} \sum{}' \, \kappa^{ar} \kappa^{st} \kappa_{rst} \tag{3}$$

em que $\kappa_{rs} = E\left(\partial^2 l / \partial\beta_r \partial\beta_s\right)$, $\kappa_{rst} = E\left(\partial^3 l / \partial\beta_r \partial\beta_s \partial\beta_t\right)$, $\kappa_{rs,t} = E\left(\left(\partial^2 l / \partial\beta_r \partial\beta_s\right)\left(\partial l / \partial\beta_t\right)\right)$, $\kappa_{rs}^{(t)} = \partial^2 \kappa_{rs} / \partial\beta_t$, κ^{rs} é a inversa da matriz κ_{rs} e Σ' denota um somatório para todas as combinações de parâmetros $\beta_1,\ldots\beta_p$.

Após alguma álgebra, pode ser mostrado que $B\left(\hat{\boldsymbol{\beta}}\right)$, o vetor $p \times 1$ de vieses, se reduz a:

$$B\left(\hat{\boldsymbol{\beta}}\right) = \left(X^{\mathrm{T}} W X\right)^{-1} X^{\mathrm{T}} W \xi, \tag{4}$$

em que $\left(X^{\mathrm{T}} W X\right)$ é a matriz de informação de Fisher para $\boldsymbol{\beta}$ e X é uma matriz $p \times 1$ de regressores fixos com rank completo. Adicionalmente, $W = \Delta - \Delta^{(2)}$, $\Delta = \sum_{i=1}^{n} \Delta_i$, $\Delta_i = \mathrm{diag}\left\{\delta_i \gamma_{ji} \exp\left(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_j\right) / s_i\right\}$, com $\gamma_{ji} = 1$, se $t_j \geq t_i$, e $\gamma_{ji} = 0$, se $t_j \leq t_i$, $s_i = \sum_{j=1}^{n} \gamma_{ji} \exp\left(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_j\right)$,

$\Delta^{(2)} = \sum_{i=1}^{n} \Delta_i \mathbf{1}\mathbf{1}^{\mathrm{T}} \Delta_i$, $\mathbf{1}$ é um vetor $n\times1$ de uns e ξ é um vetor também $n\times1$, definido como $\xi = \frac{1}{2} W^{-1}\left(\overline{\Delta} + 2M - \Delta Z_d - 2\dot{\Delta} \right)\mathbf{1}$. Aqui, $\overline{\Delta} = \sum_{i=1}^{n} t_i \Delta_i$, $t_i = \mathbf{1}^{\mathrm{T}} Z_d \Delta_i \mathbf{1}$, $Z_d = \mathrm{diag}\{Z\}$, $Z = X\left(X^{\mathrm{T}} X\right)^{-1} X^{\mathrm{T}}$, $M = \sum_{i=1}^{n} \Delta_i Z \Delta_i$, $\dot{\Delta} = \sum_{i=1}^{n} v_i \Delta_i$, e $v_i = \mathbf{1}^{\mathrm{T}} \Delta_i Z \Delta_i \mathbf{1}$. O leitor interessado pode encontrar todos os detalhes no recente artigo de Montenegro et al. (2004).

Podemos inserir no lado direito da Equação (4), que é de ordem n^{-1}, uma estimativa do parâmetro $\boldsymbol{\beta}$, para definir o EMVP corrigido:

$$\tilde{\boldsymbol{\beta}}_{\mathrm{C}} = \hat{\boldsymbol{\beta}} - B\left(\hat{\boldsymbol{\beta}}\right). \tag{5}$$

Em seguida, verificaremos, via simulações Monte Carlo, o desempenho da estimativa corrigida $\tilde{\boldsymbol{\beta}}_{\mathrm{C}}$, cujas propriedades amostrais se esperam ser melhores que a da versão não corrigida $\hat{\boldsymbol{\beta}}$.

3. Simulações Monte Carlo

Os estimadores foram implementados em Fortran e estão disponíveis a pedido diretamente dos autores. Simulações Monte Carlo foram realizadas para se comparar o EMVP usual com a sua versão corrigida, dada pela Equação (5). As simulações foram realizadas em um PC, 1.2 GHz, 256 MB RAM, sistema operacional *Windows*™ NT 4.0, e o compilador *Compaq Visual Fortran Professional Edition* 6.5.0. O estudo de simulação foi baseado no modelo de regressão de Weibull com três variáveis explicativas. Para cada experimento, estimativas pela máxima verossimilhança parcial $\hat{\boldsymbol{\beta}}$ e pela correção analítica $\tilde{\boldsymbol{\beta}}_{\mathrm{C}}$ foram calculadas.

Para cada uma das replicações, dois conjuntos de variáveis aleatórias independentes $T^{\mathrm{T}}=(T_1,T_2,\ldots T_n)$ e $U^{\mathrm{T}}=(U_1,U_2,\ldots U_n)$, foram geradas e o tempo de vida $\min(T_i,U_i)$ e δ_i, calculados. T_i é a realização de uma distribuição $\mathrm{Weibull}\left[\rho, \exp\left(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}_j\right)\right]$ com três parâmetros, sendo que $U_i \sim U(0,\theta)$ corresponde ao mecanismo aleatório de censura. O conjunto de valores das covariáveis $\mathbf{x}_i$ foi mantido constante durante todas as replicações. As covariáveis utilizadas foram geradas de distribuições normais com média zero e variâncias 1, 4 e 9, isto é, $X_1 \sim N(0,1)$, $X_2 \sim N(0,4)$, $X_3 \sim N(0,9)$.

Os valores "verdadeiros" para o vetor de parâmetros $\boldsymbol{\beta}^{\mathrm{T}} = (\beta_1, \beta_2, \beta_3)$ foram fixados em (1,0; 1,0; 1,0). As simulações foram realizadas para diversas combinações de ρ (0,5; 0,5; 1,0;

2,0), proporções de censura F (0%, 20%, 40%) e tamanhos de amostras n (10, 20, 30). A proporção de censura nominal, $P(U_i<T_i)$, foi obtida por meio de ajuste adequado do parâmetro θ da distribuição uniforme. Para cada combinação $\rho \times F \times n$, foram realizadas 10.000 replicações, conforme apresentado na Tabela 1.

Tabela 1:Resultado das simulações Monte Carlo.

ρ	F	θ	n	$\hat{\boldsymbol{\beta}}$	REQM	$\tilde{\boldsymbol{\beta}}_C$	REQM
0,2	0,1	1000000	10	1,207	3,355	0,678	2,531
	0,5	1000000	20	1,105	1,529	1,002	1,441
	0,8	1000000	30	1,080	1,250	1,046	1,216
	19,8	50	10	1,309	3,877	0,652	2,875
	20,8	172	20	1,110	1,622	0,990	1,509
	20,6	180	30	1,059	1,344	1,016	1,299
	40,2	1,6	10	1,900	6,669	0,814	6,120
	40,0	0,25	20	1,145	1,998	0,975	1,802
	40,0	1,5	30	1,060	1,585	0,996	1,510
0,5	0,0	1000000	10	1,418	1,712	0,848	1,435
	0,0	1000000	20	1,143	0,719	1,033	0,645
	0,0	1000000	30	1,084	0,552	1,038	0,524
	20,2	25	10	1,461	1,995	0,849	1,644
	20,3	120	20	1,164	0,791	1,035	0,695
	20,7	129	30	1,086	0,604	1,028	0,567
	39,7	2,5	10	1,673	3,002	0,878	2,714
	40,3	0,35	20	1,213	1,019	1,034	0,873
	39,9	4,0	30	1,101	0,725	1,019	0,666
1,0	0,0	1000000	10	1,504	1,246	0,586	2,624
	0,0	1000000	20	1,183	0,549	1,052	0,483
	0,0	1000000	30	1,097	0,367	1,034	0,330
	19,8	30	10	1,464	1,226	0,599	2,433
	20,5	122	20	1,225	0,634	1,070	0,562
	20,0	170	30	1,112	0,410	1,036	0,362
	40,0	3,0	10	1,461	1,256	0,669	1,266
	39,8	0,5	20	1,319	0,861	1,117	0,788
	40,4	6,0	30	1,156	0,530	1,053	0,479
2,0	0,0	1000000	10	1,325	0,668	0,261	1,350
	0,0	1000000	20	1,233	0,530	1,182	0,543
	0,0	1000000	30	1,137	0,353	1,048	0,322
	19,9	33	10	1,219	0,590	0,249	1,366
	20,5	138	20	1,281	0,606	1,224	0,620
	20,6	190	30	1,167	0,408	1,080	0,390
	39,7	3,5	10	1,010	0,340	0,281	0,872
	40,1	0,55	20	1,300	0,646	1,227	0,669
	40,0	9,0	30	1,233	0,522	1,190	0,531

A Figura 1 apresenta as médias das estimativas em função dos parâmetros ρ, F e n. Podemos notar que o viés do EMVP padrão pode ser bem significativo para amostras pequenas e alta proporção de censura. Há uma redução substancial do viés para a versão corrigida do estimador, quando comparado com o EMVP. Outras conclusões que podem ser tiradas da Figura 1 são que (i) o EMVP padrão sempre superestima o valor verdadeiro do parâmetro β, (ii) conforme esperado, o viés aumenta com o aumento da proporção de censura F e com o decréscimo do tamanho da amostra n, (iii) o viés é realmente alto para n=10, (iv) em geral, o estimador corrigido tem um desempenho em zig-zag (sub-estimativas seguidas por super-estimativas) que tende para o valor verdadeiro com o aumento de n.

Por sua vez, a Figura 2 apresenta as raízes dos erros quadráticos médios em função de ρ, F e n. Como esperado, os erros diminuem com o crescimento do parâmetro ρ e do tamanho da amostra n e com o decrescimento da proporção de censura F. Entretanto, o mais relevante a ser observado nestes gráficos é que não há aumento da variância no estimador corrigido. Na maioria dos casos, a correção analítica produz os mais baixos erros quadráticos médios.

4. Conclusões e observações finais

O principal objetivo deste artigo foi aprofundar o estudo da técnica analítica desenvolvida recentemente por Montenegro et al. (2004) para correção de viés de segunda ordem em estimadores de máxima verossimilhança parcial (EMVP) e apresentar os resultados de um estudo de simulações Monte Carlo para comparar o seu desempenho com os EMVP's usuais, em situações não anteriormente estudadas. Os resultados confirmam o bom desempenho da correção analítica, também para as situações aqui consideradas. Na maioria dos casos, o estimador corrigido produziu inferências com menores erros e sem aumento na variância.

Por evitar, a um custo de processamento computacional bem maior, as sofisticações e dificuldades dos métodos analíticos, as técnicas de reamostragem (Efron & Tibshirani, 1993) têm se mostrado como uma opção competitiva na resolução de problemas de correção de viés de estimadores. Com a popularização de computadores, com capacidades crescentes de processamento, possíveis direções para trabalhos futuros incluem o desenvolvimento de métodos de correção baseados em reamostragem e sua comparação com a técnica analítica aqui descrita.

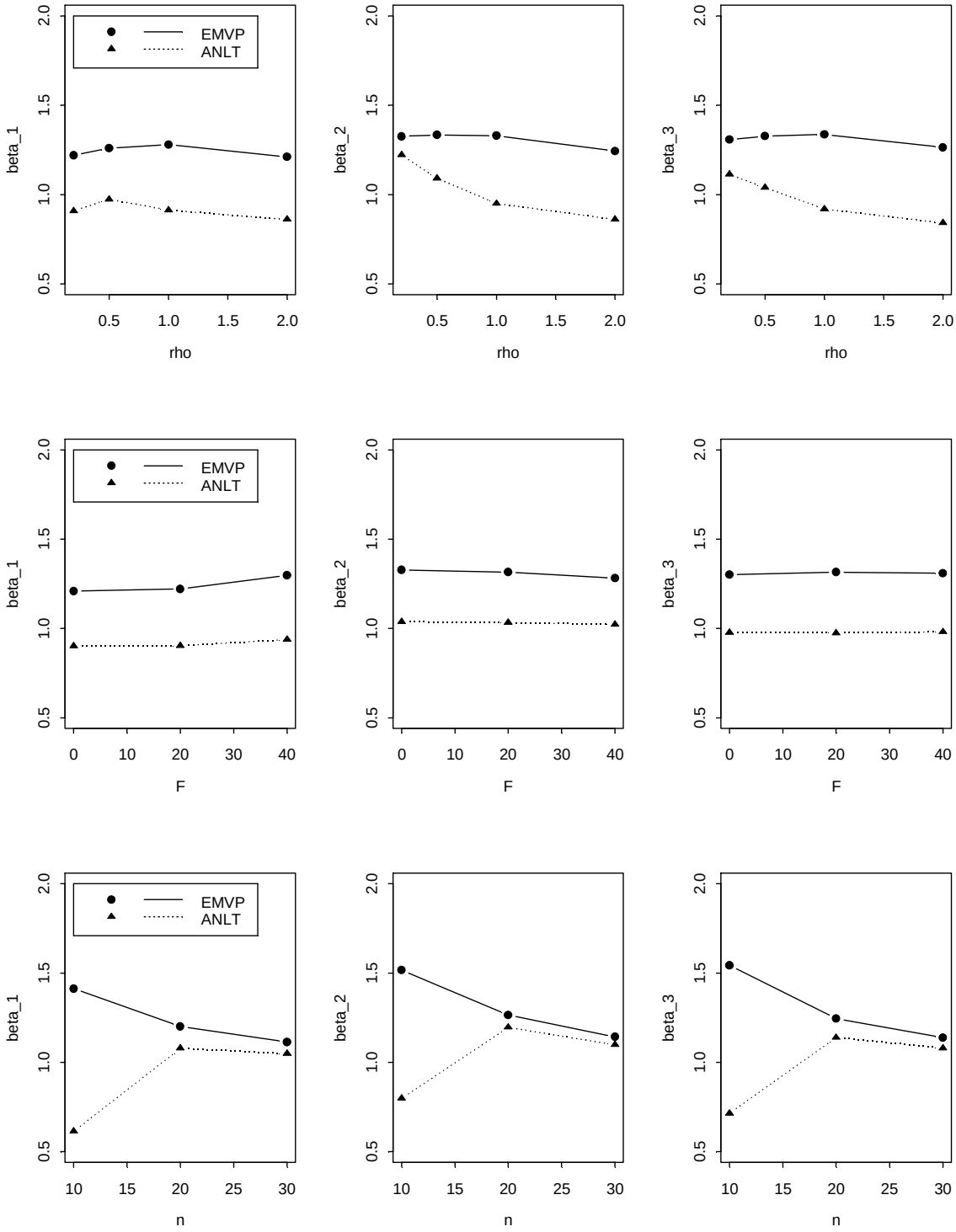

Figura 1: Estimativas médias.

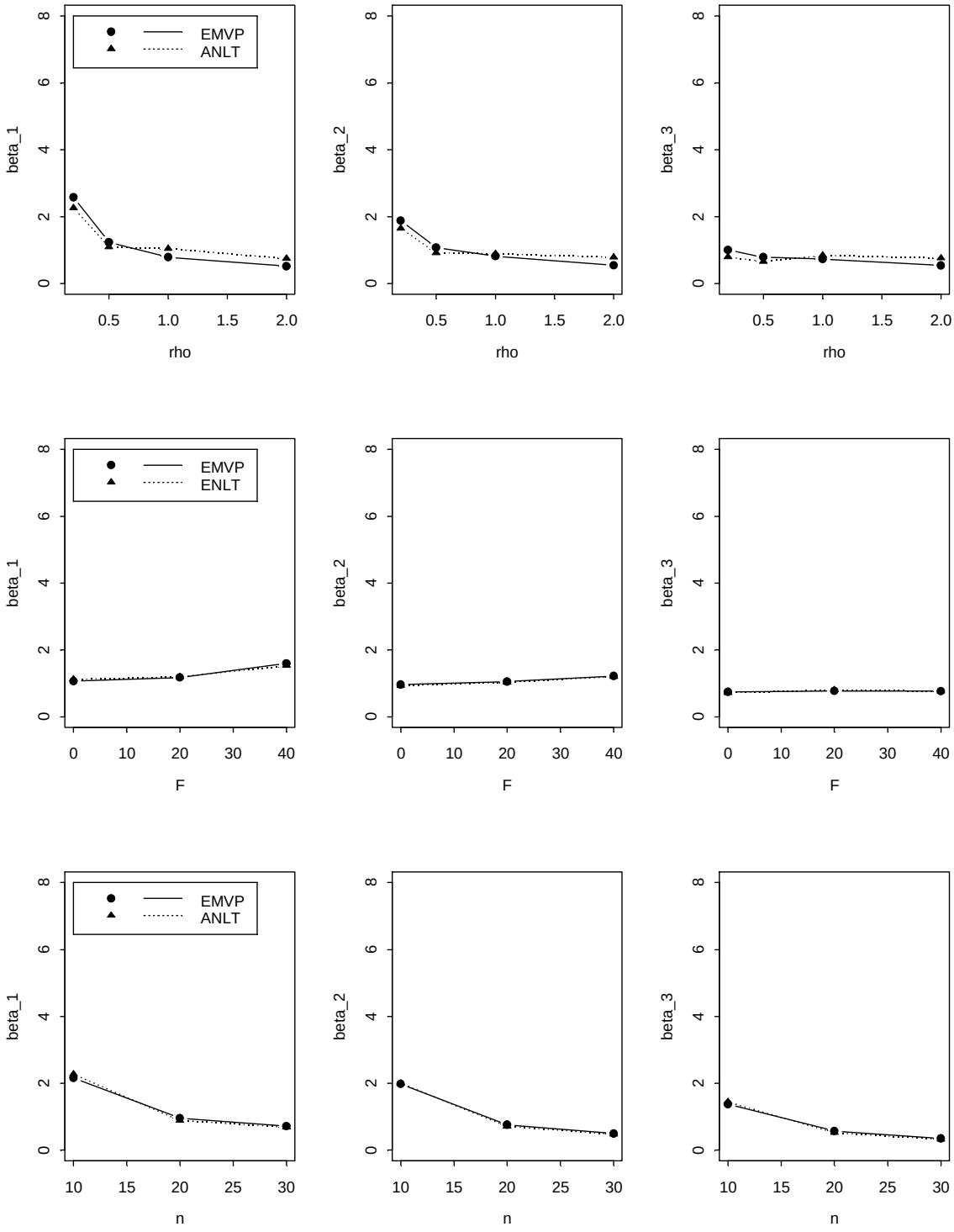

Figura 2: Raízes dos erros quadráticos médios.

Agradecimentos

A pesquisa de. Frederico R. B. Cruz é financiada pelo CNPq, processos 201046/1994-6, 301809/1996-8, 307702/2004-9, 472066/2004-8 e 472877/2006-2, pela Fundação de Amparo à Pesquisa do Estado de Minas Gerais, FAPEMIG, processos CEX-289/98 e CEX-855/98, e pela Pró-Reitoria de Pesquisa da UFMG, PRPq-UFMG, processo 4081-UFMG/RTR/FUNDO/PRPq/99.

Enrico A. Colosimo recebe apoio financeiro do CNPq, processo 300582/2003-0.

5. Referências bibliográficas

Botter, D. A. & Cordeiro, G. M. (1998) Improved estimators for generalized linear models with dispersion covariates, *Journal of Statistical Computation and Simulation*, 62, 91-104.

Colosimo, E. A., Silva, A. F. & Cruz, F. R. B. (2000). Bias evaluation in the proportional hazards model, *Journal of Statistical Computation and Simulation*, 65, 191-201.

Cordeiro, G. M. & McCullagh, P. (1991). Bias correction in generalized linear models, *Journal of the Royal Statistical Society B*, 53, 629-643.

Cox, D. R. & Oakes, D. (1984). *Analysis of Survival Data*, Chapman & Hall, London, UK.

Cox, D. R. & Snell, E. J. (1968) A general definition of residuals (with discussion), *Journal of the Royal Statistical Society B*, 30, 248-275.

Cox, D. R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society B*, 34, 187-220.

Cox, D. R. (1975). Partial likelihood, *Biometrika*, 62, 269-276.

Cruz, F. R. B.; Colosimo, E. A. & MacGregor Smith, J. (2004) Sample Size Corrections for the Maximum Partial Likelihood Estimator, *Communications in Statistics - Simulation & Computation*, 33(1), 35-47.

Efron, B. & Tibshirani, R. (1993) *An Introduction to the Bootstrap*, Chapman & Hall, London, UK.

Ferrari, S. L. P.; Botter, D. A.; Cordeiro, G. M. & Cribari-Neto, F. (1996) Second and third order bias reduction for one-parameter family models, *Statistics and Probability Letters*, 30, 339-345.

Montenegro, L. C. C.; Colosimo, E. A.; Cordeiro, G. M.; Cruz, F. R. B. (2004) Bias correction in the Cox regression model, *Journal of Statistical Computation and Simulation*, 74(5), 379-386.