3
A técnica de computação intensiva *Bootstrap*

O termo *Bootstrap* tem origem na expressão de língua inglesa "*lift oneself by pulling his/her bootstrap*", ou seja, alguém levantar-se puxando seu próprio cadarço de bota. O que está por trás desta expressão é o fato de que, pelo método, é possível obter-se propriedades de grandes amostras com base em um número reduzido de observações.

O *Bootstrap*, introduzido por EFRON (1993), é uma técnica estatística não paramétrica computacionalmente intensiva que permite a avaliação da variabilidade de estimadores com base nos dados de uma única amostra existente.

A técnica é indicada para problemas nos quais os procedimentos estatísticos convencionais sejam de difícil aplicação. Em geral, apresenta vantagens se usado em situações de amostras pequenas ou grandes, desde que forneça resultados próximos aos obtidos por meios assintóticos usuais em grandes amostras ou superior a amostras reduzidas.

Operacionalmente, a técnica consiste de um sorteio com reposição dos elementos de uma amostra aleatória, gerando uma "amostra *Bootstrap*", de tamanho igual à original. Extrai-se um número suficiente de amostras a fim de se obter a "distribuição *Bootstrap*" de qualquer estatística de interesse do pesquisador. Desta forma, o conjunto de observações *Bootstrap* corresponde a uma estimativa da verdadeira distribuição amostral da estatística em questão. Como mostrado em EFRON (1993), à medida que o tamanho da amostra tende ao infinito, a distribuição *Bootstrap* converge para a distribuição verdadeira da estatística.

Considere $X = (x_1, x_2, ..., x_{n-1}, x_n)$ a amostra original finita de tamanho n obtida de um modelo probabilístico desconhecido descrito por sua função de distribuição acumulada F e $\theta = S(X)$ a estatística de interesse. Denota-se X_i^*, i = 1, 2,..., B como sendo a i-ésima amostra *Bootstrap* de tamanho n obtida da amostra original X por meio de sorteios com reposição. Para cada amostra tem-se a correspondente estimativa *Bootstrap* da estatística de interesse, isto é, $\theta_i^* = S(X_i^*)$.

Define-se a média, variância e erro padrão do estimador *Bootstrap* de θ por, respectivamente:

$$\theta^* = \frac{\sum_{i=1}^{B} \theta_i^*}{B} \tag{3.1}$$

$$Var(\theta_i^*) = \frac{\sum_{i=1}^{B} (\theta_i^* - \theta^*)^2}{B-1} \tag{3.2}$$

$$SE_{boot} = \sqrt{Var(\theta_i^*)} \tag{3.3}$$

Sem maiores detalhes matemáticos, é possível provar, ver EFRON (1993), que:

$$Var(SE_{boot}) \cong \frac{C_1}{n^2} + \frac{C_2}{nB} \tag{3.4}$$

C_1 e C_2 são constantes que dependem da distribuição populacional F, mas não dependem de n e B. Portanto, a incerteza associada ao estimador *Bootstrap* dependerá, em última análise, do tamanho da amostra original n, isto é, mesmo que se gere uma infinidade de amostras *Bootstrap*, a incerteza do estimador não vai a zero.

3.1 ***Bootstrap* em séries temporais**

Segundo NETO (1996), a utilização da técnica de *Bootstrap* em séries temporais surgiu em 1984, quando a metodologia foi aplicada num contexto econométrico para previsão de demanda de energia no mercado americano. Mais tarde outros trabalhos surgiram e estudou-se a estimação do erro-padrão das estimativas dos parâmetros dos modelos de previsão. Atualmente o *Bootstrap* vem sendo aplicado em uma série de situações, com os mais variados objetivos. A técnica tem se mostrado muito robusta e os resultados encontrados bastante satisfatórios.

No contexto de séries temporais, há basicamente duas maneiras para aplicação: *Bootstrap* nos resíduos e o chamado método "*moving blocks*", SOUZA (2004).

No primeiro caso, garantindo a hipótese de independência dos resíduos, estes são usados para gerar novas séries "Bootstrapadas". Inicialmente ajusta-se um modelo para a série em questão e, a partir dos resíduos obtidos, são geradas *B* amostras *Bootstrap*, sorteando *B* vezes com reposição dos resíduos. Por fim, geram-se *B* séries *Bootstrap* utilizando os parâmetros do modelo original e os resíduos *Bootstrap*. Portanto, têm-se *B* séries temporais *Bootstrap*.

No caso do método *"moving blocks"*, proposto por EFRON (1993), nenhum modelo é ajustado a priori. A partir dos dados, são construídos blocos de tamanho *M* da série original. Sorteiam-se *k* blocos amostrando com reposição, agregando-os para formar a amostra *Bootstrap*. Repete-se este passo *B* vezes, gerando *B* novas séries *Bootstrap*. Neste método, a escolha do tamanho *M* dos blocos é um problema não totalmente resolvido e fonte de diversas críticas na literatura. O procedimento requer, evidentemente, estacionariedade (de segunda ordem) da série original.

3.2
Bootstrap **na identificação da ordem dos modelos**

Especificamente no emprego da metodologia Box & Jenkins, outra utilização da técnica de *Bootstrap* é na fase de identificação de estruturas no contexto destes modelos, isto é, das ordens empregadas. Classicamente, conforme já comentado, essas ordens são identificadas (principalmente, apesar de haver outras maneiras menos usuais, ver OBEYSEKERA (1982) para procedimentos alternativos) por meio da significância das funções de autocorrelação e autocorrelação parcial, com base nos intervalos de confiança assintóticos definidos no item 2.2.2. A proposta deste trabalho é a construção da distribuição de probabilidades para cada um dos *lags* de cada período para os coeficientes das funções referenciadas anteriormente.

O procedimento é análogo ao mostrado no item 2.2.2, em que serão testadas as hipóteses $H_{01}: \rho_k = 0$ e $H_{02}: \varphi_{kk} = 0$. A ideia do *Bootstrap* é estimar os parâmetros de interesse e uma medida de precisão dos mesmos, utilizando-os na definição dos intervalos de confiança para testas as hipóteses definidas em substituição ao resultado assintótico mostrado anteriormente.

Para construção da distribuição *Bootstrap* de ρ_k e φ_{kk} é necessário um algoritmo que preserve a estrutura de autocorrelação da série. De posse da série histórica de dados, obtém-se ρ_k^{*l}, a estimativa de ρ_k na *l*-ésima replicação *Bootstrap* tomando-se, com reposição, $(n - k)$ pares da amostra original de pares e construindo-se com eles a amostra *Bootstrap*. A seguir, calcula-se a estimativa pelo método usual. Repetindo-se o processo B vezes, tem-se o estimador *Bootstrap* de ρ_k:

$$\rho_k^* = \sum_{l=1}^{\mathbf{B}} \rho_k^{*l} \Big/ B \tag{3.5}$$

O conjunto de pares de observações com defasagem k fornece as informações para a estimativa do parâmetro ρ_k. A estimativa desta autocorrelação, obtida deste conjunto, é um elemento da distribuição amostral do estimador clássico. Deste conjunto pode-se obter uma amostra *Bootstrap*, de índice l, e a respectiva estimativa do parâmetro ρ_k. Um conjunto muito grande de estimativas ρ_k^{*l} constitui uma aproximação da distribuição amostral de ρ_k. Pela Lei dos Grandes Números, EFRON (1993), tem-se que:

$$\rho_k^* = \left[\sum_{l=1}^{\mathbf{B}} \rho_k^{*l} \Big/ B \right] \to E[\rho_k] \tag{3.6}$$

Dispondo-se da distribuição *Bootstrap* do estimador ρ_k, é possível construir a distribuição *Bootstrap* de φ_{kk} em função da autocorrelação *Bootstrap* de defasagem k e anteriores, da maneira usual.

Neste momento, após execução do algoritmo (detalhado no Capítulo 4), são obtidas as distribuições *Bootstrap* de ρ_k e φ_{kk}. Uma medida de precisão destas estatísticas pode ser calculada com base nestes conjuntos. Os erros-padrão *Bootstrap* de ρ_k e φ_{kk} são, respectivamente:

$$s^*(\rho_k) = \sqrt{\frac{\left[\sum_{l=1}^{B}(\rho_k^{*l} - \rho_k^*)^2\right]}{B}} \quad (3.7)$$

$$s^*(\varphi_{kk}) = \sqrt{\frac{\left[\sum_{l=1}^{B}(\varphi_{kk}^{*l} - \varphi_{kk}^*)^2\right]}{B}} \quad (3.8)$$

Portanto, estendendo o resultado mostrado em 2.2.2, se $\rho_k = 0$ e $\varphi_{kk} = 0$ e N é suficientemente grande, tem-se $\rho_k \sim N[0,(s^*(\rho_k))^2]$ e $\varphi_{kk} \sim N[0,(s^*(\varphi_{kk}))^2]$.

Desta forma, intervalos de confiança *Bootstrap* podem ser construídos sem hipótese de normalidade, tanto para ρ_k quanto para φ_{kk} e é possível verificar se os parâmetros de interesse são estatisticamente iguais a zero. Para que as hipóteses nulas não sejam rejeitadas, o valor zero deve estar contido no intervalo de confiança ao nível (1- α). Neste caso, tem-se disponível, segundo NETO (1991):

$$I_{\rho_k} = \left[\rho_k \pm t_{gl,(1-\alpha)} s^*(\rho_k)\right] \quad (3.9)$$

$$I_{\varphi_{kk}} = \left[\varphi_{kk} \pm t_{gl,(1-\alpha)} s^*(\varphi_{kk})\right] \quad (3.10)$$