Universidade Estadual de Campinas - UNICAMP
Instituto de Matemática, Estatística e Ciências da Computação - IMECC

O MÉTODO BOOTSTRAP E APLICAÇÕES À REGRESSÃO MÚLTIPLA

Damião Nóbrega da Silva

Orientadora : Profa. Dra. Gabriela Stangenhaus

CAMPINAS/SP
1995

O MÉTODO BOOTSTRAP E APLICAÇÕES À REGRESSÃO MÚLTIPLA

Este exemplar corresponde a redação final da tese devidamente corrigida e e defendida pelo Sr. Damião Nóbrega da Silva e aprovada pela comissão julgadora.

Campinas, 20 de Março de 1995.

Profa. Dra. Gabriela Stangenhaus
Orientadora

Dissertação apresentada ao Instituto de Matemática, Estatística e Ciência da Computação, UNICAMP, como requisito parcial para obtenção do título de Mestre em Estatística.

RESUMO

Neste trabalho será apresentada uma descrição básica de tópicos importantes do método Bootstrap (Efron, 1979a) para serem aplicados em problemas de Inferência Estatística cujas soluções analíticas são complicadas ou desconhecidas. Este texto é dirigido não só para estudantes de pós-graduação em Estatística, mas também para alunos de Bacharelado em Estatística que tenham cursado disciplinas básicas de Probabilidade e Inferência.

É vista, inicialmente, a implementação do método para estimar variâncias, tendenciosidades e construção de intervalos de confiança. Em seguida, esta metodologia é implementada em problemas de análise de regressão múltipla utilizando critérios de estimação como : mínimos quadrados, norma L_1 e outros métodos robustos.

Finalmente, é apresentada uma aplicação com dados de um experimento químico em bases de Schiff, que é um problema de regressão linear com restrição nos parâmetros, para mostrar a versatilidade do método em problemas mais complicados.

AGRADECIMENTOS

À Profa. Dra. Gabriela Stangenhaus pela orientação firme, persistente, objetiva, exigente, precisa e também pelo seu carinho e compreensão.

Ao Prof. Dr. Paulo César Formiga Ramos pela orientação, incentivo e apoio para realizar um Mestrado em Estatística.

A toda minha família pela apoio, principalmente, nas horas mais difíceis;

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES, em especial, ao Programa Especial de Treinamento - PET, pelo apoio financeiro e condições para parte de minha formação acadêmica.

À Universidade Estadual de Campinas - UNICAMP, pelas condições oferecidas para permitir a realização deste trabalho.

Aos Professores e funcionários do Departamento de Estatística da UNICAMP.

Aos funcionários da Biblioteca do IMECC.

Aos professores e funcionários do Departamento de Estatística da Universidade Federal do Rio Grande do Norte - UFRN, pelo apoio nas horas em que foram necessárias.

Aos colegas de mestrado pelo convívio, intercâmbio e aprendizado.

A Deus, aos meus pais, aos meus irmãos e à Linda.

SUMÁRIO

SUMÁRIO

Capítulo 1

Introdução

Existem, em toda a inferência estatística, muitos problemas cujas soluções analíticas não podem ser facilmente determinadas sendo, em algumas situações, até mesmo desconhecida. Por exemplo, considere o caso de um modelo de regressão linear simples

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, ..., n, \tag{1.1}$$

onde x_i, $i = 1, 2, ..., n$ são n valores pré-fixados de uma variável regressora, X; y_i, $i = 1, 2, ..., n$ são n observações de uma variável resposta, Y; ϵ_i, $i = 1, 2, ..., n$ são erros aleatórios independentes com mesma distribuição, esperança zero e variância constante e, β_0 e β_1 são parâmetros desconhecidos. Usualmente, deseja-se inferir sobre β_0 e β_1 a partir de procedimentos de estimação pontual, intervalos de confiança e testes de significância. Existem vários critérios para estimar pontualmente estes parâmetros como : mínimos quadrados, que minimiza a soma dos quadrados dos erros; o critério baseado na norma L_1, que minimiza a soma desvios absolutos dos erros; " least median of squares ", que minimiza a mediana dos quadrados dos erros etc... . O método dos mínimos quadrados é o mais utilizado, devido a sua facilidade matemática e computacional e as excelentes propriedades estatísticas em muitas situações. As estimativas fornecidas por este método têm expressão analítica fechada e são funções lineares dos dados y_i. Assim, supondo-se normalidade dos erros, y_i terá distribuição normal e, por conseguinte, os estimadores de mínimos quadrados também serão normalmente distribuídos. Desta forma, intervalos de confiança e testes de significância podem ser construídos facilmente pelos métodos tradicionais (método da quantidade pivotal, teste F etc.). Agora, para outros critérios de

estimação como o da norma L_1 ou o " least median of squares ", que são métodos que podem ser mais robustos que o de mínimos quadrados em um série de situações, mesmo que os erros sejam normalmente distribuídos, não se pode garantir a normalidade dos estimadores resultantes. Além disso, é difícil determinar a verdadeira distribuição desses estimadores, visto que eles são obtidos por métodos iterativos devido ao não conhecimento de expressões analíticas fechadas. Este fato dificulta ou torna impossível a construção de intervalos de confiança ou testes de significância exatos.

A solução clássica para contornar este problema é o uso de aproximações assintóticas. Sabe-se, que sob determinadas condições, os estimadores de mínimos quadrados, assim como os estimadores de norma L_1, têm distribuição assintótica normal, independente da distribuição dos erros, isto é, para um tamanho amostral suficientemente grande a distribuição destes estimadores estará próxima da distribuição normal. A prática de utilizar aproximações assintóticas é bastante comum em toda a Estatística. Seu grande problema é saber quão grande n deve ser para ter-se uma " boa " aproximação. Para um tamanho amostral que não seja suficientemente grande a validade das aproximações poderá ficar bastante comprometida.

Nas duas últimas décadas, vem-se desenvolvendo juntamente com o advento da computação de alta velocidade, com grandes capacidades de armazenamento de dados a custos mais acessíveis, uma nova metodologia para análise estatística conhecida como *computacionalmente intensiva.* Os métodos computacionalmente intensivos começaram a se tornar ferramentas atraentes e alternativas para os métodos estatísticos tradicionais, pois, visam substituir as complexidades analíticas ou uso de aproximações assintóticas de um determinado problema por computação maciça. Entre os métodos computacionalmente intensivos mais importantes, pode-se destacar o *Jackknife*, introduzido por Quenouille e Tukey por volta de 1950, e, o *Bootstrap* (Efron, 1979a).

O Jackknife é um método (não-paramétrico) que pode ser utilizado para avaliar

alguma medida de variabilidade, por exemplo, o erro padrão ou a tendenciosidade de uma determinada estimativa. Ele trabalha recalculando o valor da estatística de interesse em cada uma de *n* pseudo amostras de tamanho *n-1*, formadas a partir da amostra original, eliminado-se em cada, uma observação, da seguinte forma : a primeira pseudo amostra é formada com todas as observações da amostra original exceto a primeira, a segunda pseudo amostra é formada com todas as observações da amostra original exceto a segunda, e assim por diante, até a n-ésima pseudo amostra que é formada por todas as observações da amostra original execeto a n-ésima. Desta forma, pode-se ver o Jackknife como um método que trabalha com todas as *n* amostras extraídas *sem reposição*, de tamanho *n-1*, da amostra original. A variabilidade dos *n* valores da estatística de interesse é utilizada para estimar a quantidade desejada. Por exemplo, se é de interesse avaliar o erro padrão da estimativa obtida, então, toma-se como sua estimativa jackknife o desvio padrão das *n* estimativas calculadas em cada uma das *n* pseudos amostras. Neste trabalho não será discutido o uso Jackknife, pois o interesse principal é o método Bootstrap. Uma excelente revisão da metodologia Jackknife pode ser vista em Miller (1974).

O Bootstrap é um método mais versátil que o Jackknife, que pode ser implementado facilmente, tanto de forma não-paramétrica quanto paramétrica dependendo do conhecimento do problema, para uma grande variedade de outras situações com estruturas de dados para modelos lineares e não-lineares, problemas de análise discriminante, dados censurados etc. Em síntese, no caso não paramétrico, o método Bootstrap trabalha, ao contrário do método Jackknife, retirando-se uma amostra *com reposição*, de tamanho *n*, da amostra original. Esta amostra é denominada amostra Bootstrap. No caso paramétrico, quando se tem infomação suficiente sobre a forma da distribuição dos dados, a amostra Bootstrap é formada realizando-se a amostragem diretamente nesta distribuição com os parâmetros desconhecidos substituídos por estimativas paramétricas. A distribuição, condicional aos dados observados, da estatística de interesse aplicada aos valores da amostra Bootstrap em lugar dos valores da amostra original é definida como a distribuição Bootstrap desta estatística. Esta distribuição fornece uma aproximação para a verdadeira

distribuição da estatística de interesse. Logo, características desta distribuição são estimadas pelas respectivas características na distribuição Bootstrap, isto é, a variância de uma determinada estimativa é estimada pela variância da distribuição Bootstrap da estatística de interesse. Esta idéia é bastante antiga, porém, só recentemente foi implementada pois, em geral, não é fácil obter a distribuição Bootstrap de uma determinada estatística. Mas, com a ajuda de processos de simulação de Monte Carlo pode-se sempre obter uma aproximação para a distribuição Bootstrap retirando-se, em vez de uma, um grande número de amostras Bootstrap. A distribuição empírica dos valores da estatística de interesse calculada em cada uma das amostras Bootstrap é utilizada como uma aproximação para a verdadeira distribuição Bootstrap. Devido a este procedimento, que é utilizado na maioria das vezes, é que se considera o método Bootstrap um método que necessita mais intensamente da potência computacional que o método Jackknife.

O método Bootstrap ainda é um método pouco difundido. A finalidade deste trabalho consiste em realizar uma descrição básica da metodologia Bootstrap para estimação de variância, erro padrão, tendenciosidade, construção de intervalos de confiança etc., aplicando-a, mais especificamente, em problemas de análise de regressão linear múltipla, sem a suposição de normalidade da distribuição dos erros, com critérios de estimação de mínimos quadrados e outros robustos e, também, como um método alternativo para os procedimentos baseados no uso de aproximações assintóticas. Pretende-se que este texto possa servir como notas de aula para o ensino do método Bootstrap não só em cursos de estatística em nível de mestrado tais como inferência, análise de regressão, amostragem, estatística computacional, métodos não-paramétricos etc., mas também, que alunos de bacharelado que já tenham cursado disciplinas básicas de probabilidade e inferência possam acompanhar boa parte do material.

No capítulo 2 serão descritas e discutidas as etapas do processo Bootstrap para estruturas de dados independentes, conduzido de formas paramétrica, não-paramétrica e

semi-paramétrica. Será enfocado a estimação Bootstrap da variância e da tendenciosidade de um estimador, com cálculos exatos e aproximados por simulação de Monte Carlo.

No capítulo 3, será discutida a metodologia Bootstrap para construção de intervalos de confiança para um determinado parâmetro, assim como comentários sobre algumas de suas propriedades. Este capítulo será de grande importância para os capítulos 4 e 5, visto que a construção de intervalos de confiança ou de testes de hipóteses via inversão de um intervalo são objetivos fundamentais em análise de regressão. Porém, como se sabe, nem sempre é possível realizar estes procedimentos pelos métodos tradicionais quando não se supõe normalidade dos erros aleatórios para o caso do estimador de mínimos quadrados do vetor de parâmetros ou quando se utiliza um critério de estimação alternativo ao de mínimos quadrados que, em geral, só é possível obter estimativas dos parâmetros e expressões para intervalos de confiança e testes de significância não são conhecidas.

No capítulo 4, os conceitos apresentados nos capítulos 2 e 3, serão aplicados diretamente ao problema de análise de regressão linear múltipla, com estimação de mínimos quadrados ordinários. Também será mostrado como utilizar o método para realizar testes de significância.

No capítulo 5, será discutida a aplicação dos métodos Bootstrap para regressão robusta, conforme realizada no capítulo 4, porém, com ênfase na estimação pelo critério baseado na norma L_1. De forma similar ao capítulo 4, a construção de intervalos de confiança e testes de significância será descrita. Adicionalmente, será aplicada a metodologia para estimação de tendenciosidade, descrita no capítulo 2.

No capítulo 6, será apresentada uma aplicação do método Bootstrap para estimar uma elipsóide de concentração em um problema de estimação de um ponto de interseção de um feixe de retas, mostrando a versatilidade do método em problemas mais complicados.

A bibliografia utilizada, descrita no final deste trabalho, inclui algumas referências adicionais sobre o método Bootstrap e temas relacionados, para leitores interessados no assunto.

Capítulo 2

Introdução ao Método Bootstrap

2.1 Introdução

O método Bootstrap, introduzido por Efron (1979a), é um conjunto de técnicas de reamostragem[1] para se obter informações de características da distribuição de alguma variável aleatória, que não podem ser facilmente avaliadas por métodos analíticos tradicionais ou cuja aproximação existente tenha suposições questionáveis em alguma situação. Por exemplo, considere o coeficiente de correlação amostral de Pearson, que é dado por:

$$\hat{\rho} = t(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2 \sum_{i=1}^{n} (Y_i - \bar{Y})^2}}, \qquad (2.1.1)$$

onde, $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ e $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ são tais que:

---

[1] O termo reamostragem empregado no contexto Bootstrap não se refere à amostragem adicional de observações da população em que a amostra original foi retirada. Os tipos de reamostragem Bootstrap serão discutidos na seção 2.2.

$$(X_1, Y_1)', \ (X_2, Y_2)', \ldots, \ (X_n, Y_n)' \underset{iid}{\sim} F \ , \tag{2.1.2}$$

e F é uma função de distribuição acumulada bivariada. Para estimar seu erro padrão, que é dado por:

$$\sigma(\hat{\rho}) = [Var_F(\hat{\rho})]^{1/2} \ , \tag{2.1.3}$$

constata-se que este não é um cálculo simples de efetuar, devido a expressão complicada de $\hat{\rho}$ em (2.1.1). Usualmente, uma solução é obtida sob a suposição de que F é normal bivariada, isto é,

$$\begin{pmatrix} X_i \\ Y_i \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix} \right), \quad i = 1, 2, \ldots, n. \tag{2.1.4}$$

Neste caso, uma boa aproximação para $\sigma(\hat{\rho})$ (Johnson e Kotz, 1970), é dada por:

$$\sigma(\hat{\rho}) \approx \sigma_N(\rho) = \frac{1 - \rho^2}{\sqrt{n-3}} \ , \tag{2.1.5}$$

de forma que uma estimativa para (2.1.3), é dada por:

$$\hat{\sigma}_N(\hat{\rho}) = \frac{1 - \hat{\rho}^2}{\sqrt{n-3}} \ . \tag{2.1.6}$$

Mas, quando não for possível supor a normalidade dos dados, como se pode estimar $\sigma(\hat{\rho})$, ou outras características da distribuição de $\hat{\rho}$?

Será apresentado, neste capítulo, o método Bootstrap como uma metodologia geral para responder questões, como a anterior, sobre o coeficiente de correlação. A idéia básica deste método consiste em estimar características desejadas reproduzindo-se o mecanismo probabilístico gerador dos dados originais, com a distribuição de probabilidade desconhecida destes dados sendo substituída por uma outra conhecida que possa aproximá-la. Características que não poderiam ser avaliadas na estrutura original do problema são estimadas pelas respectivas características calculadas em uma pseudo-estrutura, criada pelo processo de reprodução, sob a distribuição estimada escolhida para aproximar a original. O termo " Bootstrap " foi derivado da frase " *to pull oneself up by one's bootstrap* " , que é inspirada em uma situação de uma pessoa que está afundando em um lago e achando que tudo está perdido, pensa que conseguirá emergir puxando pelos cadarços dos sapatos. De uma certa forma, o sentido desta frase é : em *situações de dificuldade* tentar realizar o *impossível.* Na Estatística, as " situações de dificuldade " podem ser vistas como os problemas de solucões analíticas complexas e, o " impossível ", é a utilização de uma metodologia que pode precisar de uma quantidade muito grande de cálculos, mesmo para analisar um pequeno conjunto de dados, mas que pode fornecer uma solução nestes casos.

De uma forma não muito precisa, porém informativa, pode-se dizer que o método Bootstrap pode ser usado para calcular estimativas de medidas de dispersão, tendenciosidade, limites de confiança etc. Testes de significância podem também ser realizados, aproximando-se a distribuição de uma estatística de teste ou invertendo-se um intervalo de confiança Bootstrap, pode-se formular um teste de hipótese. Um escopo de muitas aplicações do Bootstrap pode ser encontrado, principalmente, em Efron e Tibshirani (1986) e Hinkley (1988).

Também, neste capítulo, será visto o Bootstrap como um método computacionalmente intensivo, pois, devido à generalidade dos problemas em que ele pode atuar, seus cálculos podem necessitar de aproximação numérica, via simulação de Monte

Carlo, onde o poder analítico de análises teóricas é substituído pela potência computacional de computadores que hoje em dia se tornam cada vez mais velozes a custos mais reduzidos. Como será visto nas seções 2.2 e 2.3, outra situação em que o método Bootstrap se torna computacionalmente intensivo é quando suas estimativas podem ser calculadas de forma exata (e não com a aproximação de Monte Carlo), a partir da extração de todas as amostras simuladas da distribuição escolhida para aproximar a original.

Devido aos objetivos deste trabalho, não se pretende, neste capítulo, discutir todos os métodos e aplicações do Bootstrap, mais, sim, fornecer uma quantidade razoável de informações sobre os princípios fundamenteis do processo Bootstrap. Em virtude disto, não serão tratadas aqui estruturas estocásticas envolvendo *dados dependentes*, que é, sem dúvida, uma importante área na tecnologia Bootstrap que merece ser pesquisada mais intensamente. O leitor interessado neste tópico pode consultar o artigo de Léger, Politis e Romano (1992), que fazem uma discussão do problema da dependência, incluindo esquemas de reamostragem em certas classes de modelos de séries temporais. Referências adicionais são fornecidas neste trabalho. Efron (1993, cap. 26) também fornece uma discussão do problema de dependência.

Na seção 2.2, serão apresentadas idéias básicas do método Bootstrap no problema da estimação da variância de uma determinada estatística, com cálculos Bootstrap não-paramétricos, paramétricos e suavizados, e também, com a aproximação de Monte Carlo. Na seção 2.3, serão aplicadas estas mesmas idéias, para a estimação da tendenciosidade de uma dada estatística. Em seguida, os conceitos envolvidos no processo Bootstrap serão estendidos na seção 2.4 para estruturas de dados mais gerais.

2.2 A estimativa Bootstrap da variância

Suponha que um conjunto de dados observados $\mathbf{x} = (x_1, x_2,..., x_n)$ consiste de n observações independentes de uma mesma população, com média μ e variância v^2. Seja F a função de distribuição acumulada (FDA) desta população, isto é,

$$F(x) = Pr \{ X_i \leq x \}. \tag{2.2.1}$$

A notação $\mathbf{x} = (x_1, x_2,..., x_n)$ indica que $x_1, x_2,..., x_n$ representam, respectivamente, realizações das n variáveis aleatórias $X_1, X_2,..., X_n$, onde

$$X_1, X_2,..., X_n \underset{iid}{\sim} F. \tag{2.2.2}$$

Sejam

$$\mu = E_F(X_i) \tag{2.2.3}$$

e

$$\mathbf{v}^2 = \mathbf{v}^2(F) = Var_F(X_i) = E_F(X_i^2) - (E_F(X_i))^2, \; i = 1,2,...,n \; . \tag{2.2.4}$$

A estimativa natural da média populacional μ, a média amostral $\bar{x} = (1/n)\sum_{i=1}^{n} x_i$, tem variância dada por :

$$\sigma^2(F) = Var_F(\bar{X}) = \frac{\sum_{i=1}^{n} Var_F(X_i)}{n^2} = \frac{n\mathbf{v}^2(F)}{n} = \frac{\mathbf{v}^2(F)}{n} \; , \tag{2.2.5}$$

que é estimada não tendenciosamente por :

$$\hat{\sigma}^2_{\bar{x}} = \frac{s^2}{n} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n(n-1)} \quad . \tag{2.2.6}$$

Observe que (2.2.6) representa uma estimativa da medida de variação das estimativas $\bar{x}$, entre todas as amostras de n observações *iid*, extraídas na mesma população cuja FDA é F. Isto quer dizer que, se houvesse possibilidade de se retirar todas estas amostras, poderíamos calcular o valor da expressão dada em (2.2.5) diretamente pela variância dos valores das médias amostrais de cada uma destas amostras.

Naturalmente, devido à expressão (2.2.5) e ao estimador dado por (2.2.6), para o caso da média $\bar{x}$, não é necessário aplicar nenhum outro princípio para estimar sua variância. A dificuldade aparecerá em situações onde o interesse consiste em uma estatística de forma mais complicada que $\bar{x}$, como a mediana, a trimédia, o coeficiente de correlação, etc., cujas expressões analíticas para suas variâncias não são fáceis de deduzir. De volta ao princípio anterior, para estimação da variância, é impossível extrair repetidas amostras da população descrita pela FDA desconhecida F, porém, é possível obter amostras de uma outra população, cuja FDA aproxima a verdadeira F. Esta é a idéia do Bootstrap: *trocar a FDA desconhecida F, que descreve uma população que não pode ser reamostrada, por um estimador $\hat{F}$ de F, que descreva uma população que pode ser reamostrada exaustivamente*, visando com isto tornar possível o estudo direto da variabilidade da característica de interesse.

Tabela 2.1. *Amostra observada* $x = (x_1, x_2, \ldots, x_{10})$

i	x_i
1	3,5921
2	3,1255
3	2,2240
4	-0,0204
5	-0,0210
6	5,4685
7	1,3322
8	3,1949
9	-1,4188
10	2,4099

Será visto, a partir de agora, como se efetuam os cálculos Bootstrap. A idéia inicial é usar toda informação dísponível para definir uma boa escolha para $\hat{F}$, isto é, um estimador que melhor aproxime a FDA verdadeira F. É importante observar que as escolhas que serão apresentadas no decorrer desta seção serão baseadas no grau de conhecimento que se tenha sobre F. Os dados da tabela 2.1 representam uma realização de uma amostra de n=10 observações de uma distribuição normal com média média μ = 2 e variância v^2 = 4. Estes dados, cuja média e variância amostral é $\bar{x}$ = 1,9886 e s^2 = 4,1974, respectivamente, serão usados para ilustração das etapas do processo Bootstrap.

Dado um mínimo de suposições sobre F, uma escolha para $\hat{F}$ é o *estimador de máxima verossimilhança não-paramétrico de F*, que é a *função de distribuição empírica (FDE)*, denotada por $\hat{F}_n$, que associa massa de probabilidade 1/n em cada $x_1, x_2, \ldots, x_n$, isto é,

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(x_i \le x)} = \frac{\#\{ x_i \le x \}}{n}, \qquad (2.2.7)$$

que representa a proporção amostral de valores observados menores ou iguais a x. Logo, por exemplo, $\hat{F}_n(-2) = 0$, pois não se observou nenhum valor na amostra menor ou igual a " -2 ", $\hat{F}_n(0) = 3/10$, pois existem três pontos menores ou iguais a " 0 ", e $\hat{F}_n(6) = 10/10 = 1$. O gráfico de $\hat{F}_n$, para este conjuto de dados, é exibido na figura 2.1 juntamente com a FDA verdadeira dada em (2.12) com $\mu = 2$ e $\nu^2 = 4$. O uso da FDE $\hat{F}_n$ pode ser justificado, pelo menos do ponto de vista de grandes amostras, pelo teorema de Glivenko-Cantelli (Mood, Graybill and Boes, 1974) que mostra a convergência uniforme de $\hat{F}_n$ para F.

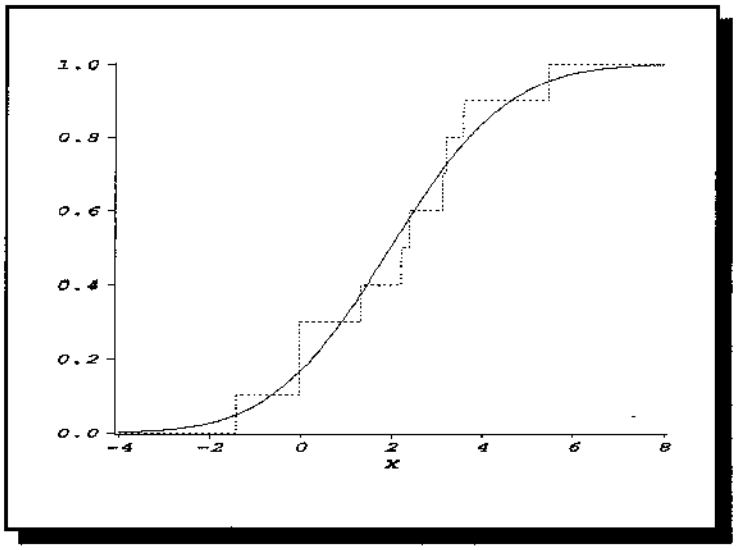

Figura 2.1. Gráfico da função de distribuição empírica (linha tracejada) e da função de distribuição acumulada dos dados $F(x) = \Phi((x-2)/2)$ (linha sólida)

O próximo passo é a obtenção da *amostra Bootstrap*. A amostra Bootstrap é um conjunto de n valores $\mathbf{X}^* = (X_1^*, X_2^*, \ldots, X_n^*)$, que são extraídos da população descrita por $\hat{F}_n$, pelo mesmo processo em que $x_1, x_2, \ldots, x_n$ foram obtidos da população com FDA F, isto é, realizando-se reamostragem *iid*. Como a FDE $\hat{F}_n$ associa massa de probabilidade 1/n sobre cada $x_1, x_2, \ldots, x_n$, então, extrair n observações *iid* da população dada por $\hat{F}_n$ é equivalente a tomar uma *amostra aleatória simples de tamanho n, com reposição, da população*

de n objetos $\{x_1, x_2,..., x_n\}$. Seja $\mathbf{X}^* = (X_1^*, X_2^*,..., X_n^*)$ e $\mathbf{x}^* = (x_1^*, x_2^*,..., x_n^*)$ a amostra Bootstrap e sua realização observada, respectivamente. Assim, a amostra Bootstrap, que é formada neste caso realizando-se reamostragem na distribuição descrita por $\hat{F}_n$, é caracterizada por *reusar* a própria amostra observada $\mathbf{x} = (x_1, x_2,..., x_n)$ e não realizando-se nova amostragem de observações da população descrita por F. Uma maneira bem simples de construir esta amostra, neste caso, é escolhendo-se aleatoriamente n inteiros $I_1, I_2,..., I_n$, com reposição, do conjunto $\{1, 2,...,n\}$. A amostra Bootstrap resultante será dada por $\mathbf{X}^* = (X_1^*, X_2^*,..., X_n^*) = (x_{I_1}, x_{I_2},...,x_{I_n})$. Assim, dado $\mathbf{X}=\mathbf{x}$, as variáveis aleatórias Bootstrap X_i^* são independentes com FDA comum $\hat{F}_n$, isto é,

$$X_i^* / \mathbf{X}=\mathbf{x} \underset{iid}{\sim} \hat{F}_n \ , \ i = 1, 2,..., n. \tag{2.2.8}$$

Com $\mathbf{X}^*$ em mão, estimativas Bootstrap são obtidas realizando-se nesta amostra, sob a distribuição $\hat{F}_n$, as mesmas operações matemáticas que se gostaria de fazer na amostra original, sob a distribuição F. Por exemplo, para o caso da média $\bar{x}$, por (2.2.4) e (2.2.5), tem-se,

$$\sigma^2(F) = \frac{v^2(F)}{n} = \frac{E_F((X_i)^2) - (E_F(X_i))^2}{n}, \tag{2.2.9}$$

onde o cálculo dos momentos $E_F(X_i)^2$ e $(E_F(X_i))^2$ são impossíveis de serem calculados sob a distribuição desconhecida F. Para os dados da tabela 2.1, o verdadeiro valor de $\sigma^2(F)$ é dado por:

$$\sigma^2(F) = \frac{v^2}{10} = \frac{2^2}{10} = 0{,}4. \tag{2.2.10}$$

A estimativa Bootstrap exata da variância de $\bar{x}$ é dada por:

$$\hat{\sigma}^2_{BOOT}(\bar{X}) = \sigma^2(\hat{F}_n) = \frac{v^2(\hat{F}_n)}{n} = \frac{E_{\hat{F}_n}((X_i^*)^2) - (E_{\hat{F}_n}(X_i^*))^2}{n}, \qquad (2.2.11)$$

isto é, substituindo-se em (2.2.9) F por $\hat{F}_n$ e X_i por X_i^*, i=1,2,...,n. Para chegar ao final do cálculo em (2.2.11), observa-se que : como X_i^* , i = 1,2,..., n, foram obtidos por um processo de seleção aleatória, com reposição, que atribue igual probabilidade de seleção a qualquer um dos valores da amostra observada $\{ x_1, x_2,..., x_n \}$, então, X_i^* pode assumir qualquer destes com a mesma probabilidade 1/n, isto é,

$$Pr_{\hat{F}_n}\{ X_i^* = x_j \} = \frac{1}{n}, \ \forall \ i,j = 1,2,...,n. \qquad (2.2.12)$$

Logo,

$$E_{\hat{F}_n}(X_i^*)^k = \sum_{i=1}^{n} x_i^k . \frac{1}{n} = \frac{\sum_{i=1}^{n} x_i^k}{n} . \qquad (2.2.13)$$

Assim,

$$\hat{\sigma}^2_{BOOT}(\bar{X}) = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2}{n} = \frac{\frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right)}{n} =$$

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n^2} = \frac{n-1}{n}\hat{\sigma}^2_{nt}. \qquad (2.2.14)$$

Como se pode ver, a estimativa Bootstrap da variância de $\bar{x}$ (2.2.14), obtida a partir da simples substituição da FDA desconhecida F, pela conhecida $\hat{F}_n$, aproxima-se da

estimativa não-tendenciosa $\hat{\sigma}^2_{nt}$, a menos do fator n/(n-1).

De uma forma geral, em situações onde se deseja estimar a variância de uma estatística $\hat{\theta} = t(\boldsymbol{x})$, dada por:

$$\sigma^2(F) = Var_F t(\boldsymbol{X}) = E_F(t(\boldsymbol{X})^2) - (E_F t(\boldsymbol{X}))^2 , \qquad (2.2.15)$$

a *estimativa Bootstrap exata (não-paramétrica) de* $\sigma^2(F)$, será dada por :

$$\hat{\sigma}^2_{BOOT}(\hat{\theta}) = \sigma^2(\hat{F}_n) = Var_{\hat{F}_n}(t(\boldsymbol{X}^*)) = E_{\hat{F}_n}(t(\boldsymbol{X}^*)^2) - (E_{\hat{F}_n} t(\boldsymbol{X}^*))^2 . \qquad (2.2.16)$$

Por exemplo, pode-se desejar avaliar propriedades da distribuição da estatística $\hat{\theta} = \bar{x}^3$ como estimador do parâmetro μ^3, logo, $t(\boldsymbol{X}) = \bar{X}^3$. Assim,

$$\sigma^2(F) = Var_F(\bar{X}^3) = E_F(\bar{X}^6) - (E_F(\bar{X}^3))^2 . \qquad (2.2.17)$$

Para os dados da tabela 2.1,

$$\sigma^2(F) = Var_F(\bar{X}^3) = 9\mu^4(\frac{\nu^2}{n}) + 36\mu^2(\frac{\nu^2}{n})^2 + 15(\frac{\nu^2}{n})^3 . \qquad (2.2.18)$$

Para n=10, μ=2 e ν=2, então

$$\sigma^2(F) = Var_F(\bar{X}^3) = 81{,}6. \qquad (2.2.19)$$

A estimativa Bootstrap exata (não-paramétrica) de $Var_F(\bar{X}^3)$ será dada pela expressão (2.2.18) com $\bar{x}$ em lugar de μ e $\hat{\sigma}^2_{BOOT}(\bar{X})$ em lugar de (ν^2/n), isto é,

$$\delta^2_{BOOT}(\bar{X}^3) = \sigma^2(\hat{F}_n) = Var_{\hat{F}_n}(\bar{X}^{*3}) = E_{\hat{F}_n}(\bar{X}^{*6}) - (E_{\hat{F}_n}(\bar{X}^{*3}))^2 =$$
$$= 9\bar{x}^4\delta^2_{BOOT}(\bar{X}) + 36\bar{x}^2[\delta^2_{BOOT}(\bar{X})]^2 + 15[\delta^2_{BOOT}(\bar{X})]^3. \quad (2.2.20)$$

A tabela 2.2 exibe os valores das estimativas Bootstrap de (2.2.14) e (2.2.20) para o conjunto de dados da tabela 2.1.

Tabela 2.2. *Estimativas Bootstrap não-paramétricas das variâncias das estatísticas* $\bar{x}$ *e* $\bar{x}^3$ *para o conjunto de dados da tabela 2.1*

Estatística $\hat{\theta}$	Característica $\sigma^2(F) = Var_F(\hat{\theta})$	$\delta^2_{BOOT}(\hat{\theta})$
$\bar{x}$	0,40	0,3778
$\bar{x}^3$	81,6	74,3021

Dependendo da forma da estatística $\hat{\theta}$, pode ser complicado efetuar o cálculo Bootstrap exato de $\delta^2_{BOOT}(\hat{\theta})$, como feito para $\bar{x}$ em (2.2.11)-(2.2.14). Um outro procedimento de cálculo exato, para $\delta^2_{BOOT}(\hat{\theta})$, pode ser obtido a partir de todas as $m = \binom{2n-1}{n}$ (que representa o número de amostras não-ordenadas de tamanho n, extraídas com reposição do conjunto $\{x_1, x_2, \ldots, x_n\}$ (Roussas, 1973)) amostras Bootstrap distintas, digamos $z^*(1), z^*(2), \ldots, z^*(m)$. Para $n = 2$, teremos $m = 3$, que são as amostras $\{x_1, x_1\}$, $\{x_2, x_2\}$ e $\{x_1, x_2\}$, pois $\{x_1, x_2\}$ e $\{x_2, x_1\}$ são as mesmas). Assim, a estimativa Bootstrap da variância de (2.2.15) poderá ser calculada pela soma finita:

$$\delta^2_{BOOT}(\hat{\theta}) = \sum_{j=1}^{m} w_j[t(z^*(j)) - t(\cdot)]^2 , \quad (2.2.21)$$

onde

$$w_j = \frac{n!}{j_1!j_2!\ldots j_n!}\prod_{i=1}^{n}\left(\frac{1}{n}\right)^{j_i} \qquad (2.2.22)$$

e

$$t(\cdot) = \sum_{j=1}^{m} w_j t(z^*(j)) \ . \qquad (2.2.23)$$

O valor w_j de (2.2.22) é a probabilidade de obtenção da amostra $z^*(j)$, com j_i representando o número de vezes em que x_i aparece na amostra $z^*(j)$. A grande dificuldade deste método é que *m* cresce muito rapidamente com o aumento do tamanho da amostra original *n*. A tabela 2.3 dá alguns valores de *m* correspondentes a pequenos valores de *n*.

Tabela 2.3. *Número de amostras distintas Bootstrap (m) referentes ao respectivo tamanho amostral (n)*

n	2	5	6	7	8	9	10	15
m	3	126	462	1.716	6.435	24.310	92.378	77.558.760

Assim, para calcular a estimativa Bootstrap exata dada em (2.2.15), com base em um tamanho amostral n = 15, seria necessário obter todas as 77.558.760 amostras distintas $z^*(\cdot)$, que pode ser uma tarefa computacional intensa, dependendo do computador disponível.

Felizmente, é sempre possível obter uma aproximação numérica para $\hat{\sigma}^2_{BOOT}(\hat{\theta})$ através de simulação de Monte Carlo, seguindo-se o algoritmo da figura 2.2.

ALGORÍTMO 1

(i) usando um gerador de números aleatórios, extraia independentemente B amostras Bootstrap $x^*(1), x^*(2),..., x^*(B)$, *isto é, B amostras aleatórias simples de tamanho n, extraídas com reposição da amostra atual* $\{x_1, x_2,..., x_n\}$;

(ii) Para cada amostra $x^*(b)$, *calcule a estatística de interesse* $\hat{\theta}^*(b) = t(x^*(b))$, $b=1,2,..., B$;

(iii) Calcule a variância amostral dos B valores $\hat{\theta}^*(1), \hat{\theta}^*(2),..., \hat{\theta}^*(B)$ *,isto é,*

$$\hat{\sigma}^2_B(\hat{\theta}) = \frac{\sum_{b=1}^{B} (\hat{\theta}^*(b) - \hat{\theta}^*(\cdot))^2}{B - 1} , \qquad (2.2.24)$$

onde $\hat{\theta}^*(\cdot) = \frac{\sum_{b=1}^{B} \hat{\theta}^*(b)}{B}$.

Figura 2.2. Procedimento de Monte Carlo para estimar a variância Bootstrap de uma estatística $\hat{\theta} = t(x)$

É fácil ver que quando $B \to \infty$, $\hat{\sigma}^2_B(\hat{\theta}) \to \hat{\sigma}^2_{BOOT}(\hat{\theta}) = \sigma^2(\hat{F}_n)$, a estimativa Bootstrap exata de $\sigma^2(F)$. É importante tomar amostras Bootstrap de mesmo tamanho que a amostra original. O algoritmo de Monte Carlo da figura 2.2 não converge para $\hat{\sigma}^2_{BOOT}(\hat{\theta})$ da expressão (2.2.16), se o tamanho da amostra Bootstrap diferir do *n* original (Efron and Tibshirani, 1986). Para fins práticos, uma pergunta que pode ser feita, a esta altura, é sobre o número de replicações Bootstrap necessárias para estimar $\hat{\sigma}^2_{BOOT}(\hat{\theta})$, via simulação, com uma boa precisão, já que o tempo computacional cresce linearmente com o aumento de *B*. Efron (1993), apresenta as seguintes regras práticas:

(i) B = 25, é usualmente informativo. B = 50 é freqüentemente suficiente para dar uma boa estimativa para $\hat{\sigma}^2_{BOOT}(\hat{\theta})$.

(ii) Raramente são necessárias mais do que B = 200 replicações para estimar uma variância ou um desvio padrão (maiores valores de B podem ser necessários para estimação de intervalos de confiança Bootstrap).

As figuras 2.3 e 2.4 exibem os resultados de um estudo de simulação aplicando-se o algorítmo 1 para observar a convergência das estimativas Monte Carlo $\hat{\sigma}_B(\hat{\theta})$ para a estimativa Bootstrap exata $\hat{\sigma}_{BOOT}(\hat{\theta})$ do desvio padrão das estatísticas $\hat{\theta} = \bar{x}$ e $\hat{\theta} = \bar{x}^3$. Neste estudo foram considerados valores de B desde 25 a 700, com incremento de 25 replicações.

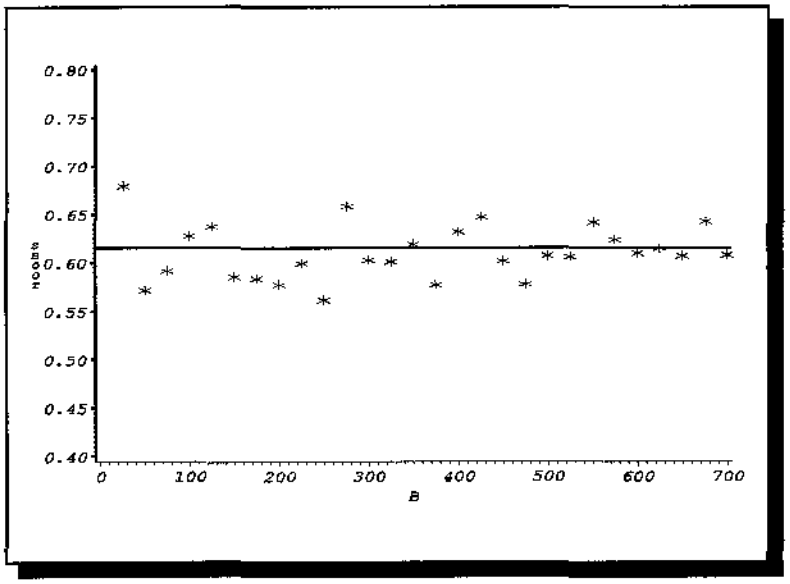

Figura 2.3. Estimativas Bootstrap Monte Carlo do desvio padrão da estatística $\hat{\theta} = \bar{x}$ (SBOOT = $\hat{\sigma}_B(\bar{X})$) em função do número de replicações B. A linha sólida representa a estimativa Bootstrap exata $\hat{\sigma}_{BOOT}(\bar{X})$ = 0,6147

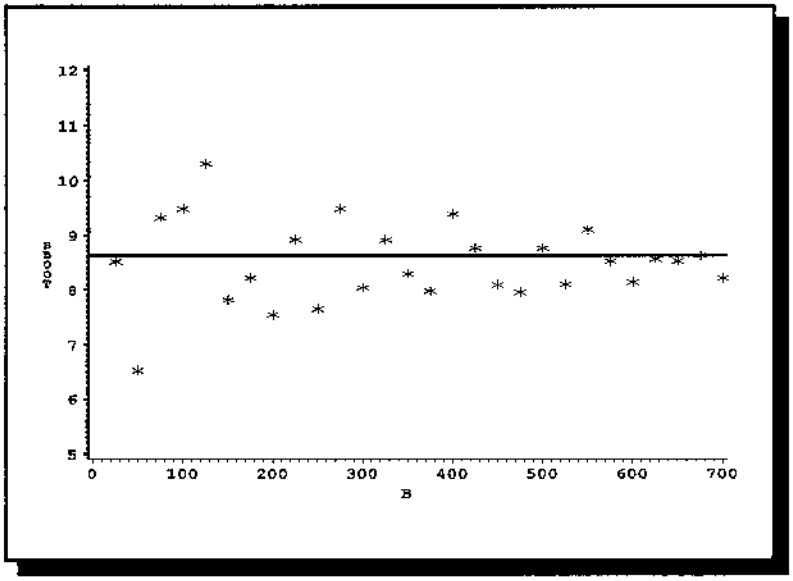

Figura 2.4. Estimativas Bootstrap Monte Carlo do desvio padrão da estatística $\hat{\theta} = \bar{x}^3$ (SBOOT = $\hat{\sigma}_B(\bar{X}^3)$) em função do número de replicações B. A linha sólida representa a estimativa Bootstrap exata $\hat{\sigma}_{BOOT}(\bar{X}^3) = 8{,}6199$

Observando-se estas duas figuras comprova-se a utilidade das regras anteriores. Em ambos os casos as estimativas de Monte Carlo, referentes ao valor B = 25, já são razoáveis caso não se requeira tanta precisão, visto que, com este número de replicações, a variabilidade inerente ao Monte Carlo é maior. Para o valor B = 200 já se observa uma diminuição desta variabilidade, e, a partir de aproximadamente B = 450, uma estabilização das estimativas. Obviamente estas duas figuras representam casos específicos. Na prática, para uma maior segurança na escolha do número de replicações em qualquer aplicação do método Bootstrap, pode-se realizar um pequeno estudo de simulação, como o anterior, variando o número de repetições e observando-se o comportamento das respectivas estimativas de Monte Carlo.

Até agora, apresentou-se o Bootstrap totalmente não-paramétrico com o uso da FDE, visto que não se está fazendo qualquer suposição sobre F, apenas a sua existência. Caso se tenha informação adicional sobre a forma da F, o Bootstrap é um método bastante versátil para incorporar este fato. Por exemplo, suponhamos que F tem a forma da FDA da distribuição normal, cuja média μ e variância v^2 são desconhecidas. Assim, F pode ser indexada por esses dois parâmetros, isto é,

$$F = F_{(\mu, v^2)}, \tag{2.2.25}$$

como no exemplo dos dados da tabela 2.1 em que, supondo-se μ e v^2 desconhecidos, a verdadeira FDA destes dados é dada por:

$$F(x) = Pr\{ X_i \leq x \} = Pr\left\{ \frac{X_i - \mu}{v} \leq \frac{x - \mu}{v} \right\} = \Phi\left(\frac{x - \mu}{v}\right), \; x \in \mathbb{R}, \tag{2.2.26}$$

onde

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{1}{2}z^2}, \; z \in \mathbb{R}, \tag{2.2.27}$$

é a função de distribuição acumulada da distribuição *normal padrão*. Portanto, um estimador paramétrico de F é dado, simplesmente, por:

$$\hat{F}_{par} = F_{(\hat{\mu}_{par}, \hat{v}^2_{par})}, \tag{2.2.28}$$

onde $\hat{\mu}_{par}$ e $\hat{v}^2_{par}$ são estimativas de μ e v^2, obtidas através de algum método paramétrico de estimação, por exemplo, máxima verossimilhança. Para os dados da tabela 2.1,

uma estimativa paramétrica de F é

$$\hat{F}_{par} = F_{(\hat{\mu}_{par}, \hat{\sigma}^2_{par})} = F_{(\bar{x}, s^2)} = \Phi((x-\bar{x})/s). \qquad (2.2.29)$$

A amostra Bootstrap é formada por amostragem *iid* da distribuição normal com média $\hat{\mu}_{par} = \bar{x}$ e variância $\hat{\sigma}^2_{par} = s^2$ (aqui, a reamostragem consiste de variáveis aleatórias com distribuição normal. A amostra observada **x** é reusada para estimar a FDA F). Isto pode ser feito através da implementação em um computador de um método de gerações de variáveis pseudo-aleatórias com distribuição normal, tal como o de Box-Müller, ou utilizando-se algum pacote estatístico apropriado. Assim,

$$X_1^*, X_2^*, \ldots, X_n^* \underset{iid}{\sim} N(\hat{\mu}_{par}, \hat{\sigma}^2_{par}) \ . \qquad (2.2.30)$$

Este processo, de conduzir o Bootstrap, substituindo-se a FDA F por uma estimativa paramétrica é conhecido como *Bootstrap paramétrico.* A *estimativa Bootstrap exata (paramétrica) da variância* de alguma estatística $\hat{\theta} = t(\mathbf{x})$ é dada por:

$$\hat{\sigma}^2_{BOOT}(\hat{\theta}) = \sigma^2(\hat{F}_{par}) = E_{\hat{F}_{par}} t(X_i^*)^2 - (E_{\hat{F}_{par}} t(X_i^*))^2 \ . \qquad (2.2.31)$$

Portanto, a estimativa Bootstrap paramétrica exata de $\sigma^2(F) = Var_F(\bar{X})$ será dada por:

$$\hat{sigma}^2_{BOOT}(\bar{X}) = \left(\frac{s^2}{n} + \mu^2\right) - \mu^2 = \frac{s^2}{n} = \hat{\sigma}^2_{nt}, \qquad (2.2.32)$$

e, a estimativa Bootstrap paramétrica de $\sigma^2(F) = Var_F(\bar{X}^3)$ será dada por:

$$\hat{\sigma}^2_{BOOT}(\bar{X}^3) = Var_{\hat{F}_{par}} \bar{X}^3 = 9\bar{x}^4\left(\frac{s^2}{n}\right) + 36\bar{x}^2\left(\frac{s^2}{n}\right)^2 + 15\left(\frac{s^2}{n}\right)^3. \qquad (2.2.33)$$

A tabela 2.4 dá os valores destas estimativas correspondentes aos dados da tabela 2.1.

Tabela 2.4. *Estimativas Bootstrap paramétricas das variâncias das estatísticas $\bar{x}$ e $\bar{x}^3$ para o conjunto de dados da tabela 2.1*

Estatística $\hat{\theta}$	Característica $\sigma^2(F) = Var_F(\hat{\theta})$	$\hat{\sigma}^2_{BOOT}(\hat{\theta})$
$\bar{x}$	0,40	0,4197
$\bar{x}^3$	81,6	85,2672

No caso de estimativas paramétricas que descrevam populações infinitas, tal como a Normal, Gama, Exponecial, Cauchy etc., torna-se impossível aplicar o método de cálculo exato de $\hat{\sigma}^2_{BOOT}(\hat{\theta})$ baseado nas extrações de todas as amostras Bootstrap distintas, isto é, todas as amostras iid de tamanho n da população que é descrita por $\hat{F}_{par}$, visto que existem infinitas amostras deste tipo. Porém, será sempre possível obter uma aproximação numérica para $\hat{\sigma}^2_{BOOT}(\hat{\theta})$, modificando-se o algorítmo 1 no passo (i) da seguinte forma : em vez de formar a amostra Bootstrap selecionando-se aleatoriamente observações com reposição, geram-se, independentemente, amostras de n obervações independentes de uma variável aleatória com distribuição normal, de média $\hat{\mu}_{par}$ e variância $\hat{\sigma}^2_{par}$, ou, de forma mais geral, da distribuição Bootstrap paramétrica. Os passos (ii) e (iii) são idênticos ao do algorítmo 1, e ao final deste processo, será produzido uma aproximação para a estimativa Bootstrap paramétrica $\hat{\sigma}^2_{BOOT}(\hat{\theta})$, de (2.2.31). Este processo foi realizado, como feito para o Bootstrap não-paramétrico, para valores de B desde 25 até 700 com incremento de 25 replicações. Os resultados estão exibidos nas figuras 2.5 e 2.6.

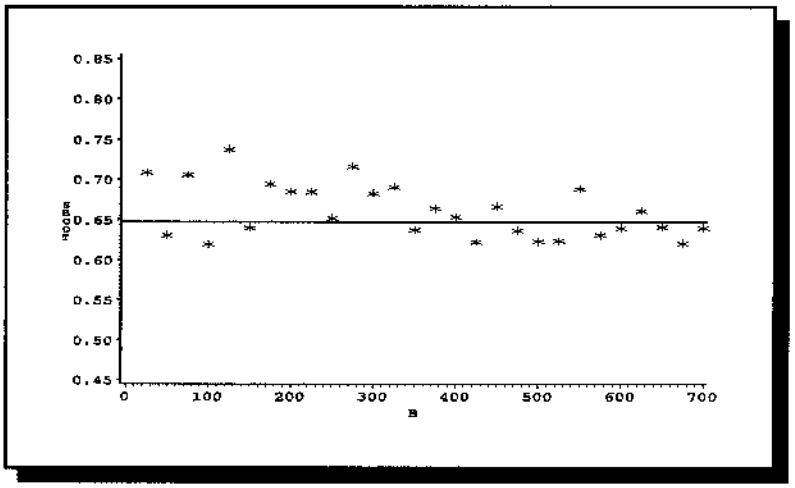

Figura 2.5. Estimativas Bootstrap Monte Carlo do desvio padrão da estatística $\hat{\theta} = \bar{x}$ (SBOOT = $\hat{\sigma}_B(\bar{X})$) em função do número de replicações B. Linha sólida representa a estimativa Bootstrap exata $\hat{\sigma}_{BOOT}(\bar{X})$ = 0,6478

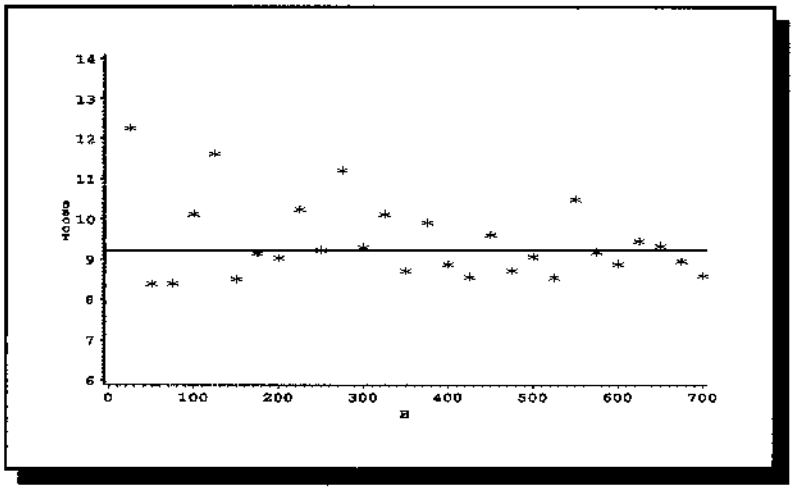

Figura 2.6. Estimativas Bootstrap Monte Carlo do desvio padrão da estatística $\hat{\theta} = \bar{x}^3$ (SBOOT = $\hat{\sigma}_B(\bar{X}^3)$) em função do número de replicações B. Linha sólida representa a estimativa Bootstrap exata $\hat{\sigma}_{BOOT}(\bar{X}^3)$ = 9,2340

Uma maneira de desenvolver o processo Bootstrap sem ser totalmente paramétrico ou não-paramétrico é usar, como estimador de F, uma FDA que represente um compromisso entre a FDE e uma FDA paramétrica, tornando o estimador resultante mais suave que $\hat{F}_n$. Isto pode ser realizado formando-se a amostra Bootstrap, em vez de selecionar aleatoriamente do conjunto $\{ x_1, x_2,..., x_n \}$, realizando-se a reamostragem da seguinte forma:

$$X_i^* = \bar{x} + c[\, x_{I_i} - \bar{x} + \hat{v}_{BOOT}(X_i)Z_i \,], \quad i = 1,2,...,n \ , \tag{2.2.34}$$

onde $I_1, I_2,..., I_n$ são escolhidos independentemente e aleatoriamente do conjunto $\{ 1, 2,..., n \}$ e $Z_1, Z_2,..., Z_n$ são variáveis aleatórias independentes com mesma distribuição fixa tendo média 0 e variância σ_Z^2. Por exemplo, a distribuição uniforme sobre $[-\frac{1}{2},\frac{1}{2}]$ ($\sigma_Z^2 = 1/12$) ou a distribuição normal etc. As quantidades $\bar{x}$, $\hat{v}_{BOOT}(X_i)$ e c da expressão (2.2.34) são a média amostral, o desvio padrão amostral $\hat{v} = (v^2(\hat{F}_n))^{1/2} = (\sum_{i=1}^{n}(x_i - \bar{x})^2/n)^{1/2}$ e c é uma constante arbitrária. Escolhendo-se $c = [\, 1 + \sigma_Z^2 \,]^{-1/2}$, então, X_i^* terá média $\bar{x}$ e variância $\hat{v}^2_{BOOT}(X_i)$ (para todo i = 1, 2, ..., n) sob o procedimento de reamostragem Bootstrap. Este processo é conhecido como *Bootstrap suavizado* (Efron, 1979a). A idéia do Bootstrap suavizado é formar amostras de uma distribuição mais suave que a FDE, que é discreta, para evitar ou, diminuir, a presença de valores repetidos da amostra observada. Existem formas mais gerais que (2.2.33), baseadas em estimadores de *Kernel*, para a formação da amostra Bootstrap, como pode ser visto em Silverman and Young (1987) e Efron (1993). Silverman and Young (1987), Hall, DiCiccio and Romano (1987) mostraram, em determinadas situações, que suavizar a FDE pode trazer benefícios ao processo Bootstrap.

Quando a FDA F é contínua e simétrica com centro de simetria θ uma outra forma não-paramétrica de condução do processo Bootstrap é a seguinte : estima-se F pela FDE simetrizada $\hat{G}_n$ que coloca massa de probabilidade 1/(2n) em cada $x_1, x_2,..., x_n, 2\hat{\theta} - x_1, 2\hat{\theta} - x_2,..., 2\hat{\theta} - x_n$, isto é,

$$\hat{G}_n(x) = \frac{\sum_{i=1}^{n} \mathbf{1}_{(x_i \le x)} + \sum_{i=1}^{n} \mathbf{1}_{(x_i \ge 2\hat{\theta}-x)}}{2n}, \qquad (2.2.35)$$

onde $\hat{\theta}$ é um estimador de θ. Sob as condições:

a) $\hat{\theta}$ é um estimador não-tendencioso para θ;
b) $\hat{\theta} - \theta$ é invariante de locação e escala;
c) $\hat{\theta}$ é assintoticamente eficiente,

Hinkley (1976) e Schuster (1975) mostraram que $\hat{G}_n$ tem melhores propriedades assintóticas que $\hat{F}_n$. A amostra Bootstrap é formada tomando-se uma amostra aleatória simples de tamanho n da população de 2n objetos $\{ x_1, x_2,..., x_n, 2\hat{\theta} - x_1, 2\hat{\theta} - x_2,..., 2\hat{\theta} - x_n \}$.

2.3 A estimativa Bootstrap da tendenciosidade

Na seção anterior, ilustramos as idéias principais do Bootstrap com o problema de estimação da variância. Naturalmente, tais idéias podem ser estendidas a outras estatísticas. Será considerado agora a estimativa da tendenciosidade de uma determinada estatística $\hat{\theta} = t(\mathbf{x})$ em relação a um parâmetro $\theta = s(F)$, seja

$$R(x,F) = \hat{\theta} - \theta = t(x) - s(F) . \qquad (2.3.1)$$

A notação $\theta = s(F)$ é para denotar que o parâmetro θ tem seu valor calculado, embora desconhecido, aplicando-se a regra dada pela função $s(\cdot)$ na distribuição F. Para exemplicar, suponha que θ seja a média populacional. Logo, $s(F) = E_F X_i = \int x dF(x)$. A tendenciosidade de $\hat{\theta}$ ao estimar θ é definida por:

$$T(F) = E_F R(\mathbf{X},F) = E_F(t(\mathbf{X})) - s(F) \ . \qquad (2.3.2)$$

No caso das estimativas $\bar{x}$ e $\bar{x}^3$, suas tendenciosidades são dadas por:

$$T_1(F) = E_F(\bar{X}) - s(F) = \mu - \mu = 0, \qquad (2.3.3)$$

e,

$$T_2(F) = E_F(\bar{X})^3 - (s(F))^3 = (\mu^3 + 3\mu\frac{v^2}{n}) - \mu^3 = 3\mu\frac{v^2}{n}, \qquad (2.3.4)$$

respectivamente. Portanto, para $\mu = 2$ $v = 2$, o verdadeiro valor da tendenciosidade de $\bar{x}^3$ é igual a $T_2(F) = 2{,}4$.

Seguindo-se o mesmo processo discutido na seção 2.2, calcula-se a *estimativa Bootstrap exata de T(F)* que é dada por:

$$\hat{T}_{BOOT}(\hat{\theta}) = T(\hat{F}) = E_{\hat{F}}(R(\mathbf{X}^*,\hat{F})) = E_{\hat{F}}(t(\mathbf{X}^*)) - s(\hat{F}) \ , \qquad (2.3.5)$$

onde $\mathbf{X}^* = (X_1^*, X_2^*,..., X_n^*)$ é a amostra Bootstrap.

As estimativas Bootstrap não-paramétricas de $T_1(F)$ e $T_2(F)$, para os dados da tabela 2.1, são dadas por:

$$\hat{T}_{BOOT}(\bar{X}) = T_1(\hat{F}_n) = E_{\hat{F}_n}(\bar{X}^*) - \bar{x} = \bar{x} - \bar{x} = 0 \qquad (2.3.6)$$

e

$$\hat{T}_{BOOT}(\bar{X}^3) = T_2(\hat{F}_n) = E_{\hat{F}_n}(\bar{X}^{*3}) - \bar{x}^3 = 3\bar{x}\hat{\sigma}^2_{BOOT}(\bar{X}) = 2{,}2537, \qquad (2.3.7)$$

respectivamente. Já as estimativas Bootstrap paramétricas de $T_1(F)$ e $T_2(F)$, com base no conjunto de dados da tabela 2.1, são dadas por:

$$\hat{T}_{BOOT}(\bar{X}) = T_1(\hat{F}_{par}) = E_{\hat{F}_{par}}(\bar{X}^*) - \bar{x} = \bar{x} - \bar{x} = 0 \qquad (2.3.8)$$

e

$$\hat{T}_{BOOT}(\bar{X}^3) = T_2(\hat{F}_{par}) = E_{\hat{F}_{par}}(\bar{X}^{*3}) - \bar{x}^3 = 3\bar{x}\hat{\sigma}^2_{BOOT}(\bar{X}) = 2{,}5038, \qquad (2.3.9)$$

respectivamente.

Com o uso da FDE, o método de cálculo exato da estimativa Bootstrap $\hat{T}_{BOOT}(\hat{\theta})$, pode ser utilizado retirando-se todas as $m = \binom{2n-1}{n}$ amostras Bootstrap distintas e estimando T(F) por:

$$\hat{T}_{BOOT}(\hat{\theta}) = \sum_{j=1}^{m} w_j [\, t(z^*(j)) - t(\cdot) \,], \qquad (2.3.10)$$

onde w_j e $t(\cdot)$ são dados por (2.2.22) e (2.2.23), respectivamente.

Para aproximar numericamente $\hat{T}_{BOOT}(\hat{\theta})$, pode-se usar o algoritmo 1, com as seguintes modificações nos passos (ii) e (iii):

(ii) para cada amostra Bootstrap $\mathbf{x}^*(b)$, calcule a correspondente repetição Bootstrap $R(\mathbf{x}^*(b), \hat{F}_n) = t(\mathbf{x}^*(b)) - s(\hat{F}_n)$, $b = 1, 2, \ldots, B$;

(iii) calcule a média dos B valores $R(\mathbf{x}^*(1), \hat{F}_n)$, $R(\mathbf{x}^*(2), \hat{F}_n), \ldots, R(\mathbf{x}^*(B), \hat{F}_n)$, isto é,

$$\hat{T}_B(\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B} R(x^*(b),\hat{F}) = \frac{\sum_{b=1}^{B} t(x^*(b))}{B} - s(\hat{F}) \ . \qquad (2.3.11)$$

Quando B $\rightarrow \infty$, $\hat{T}_B(\hat{\theta}) \rightarrow \hat{T}_{BOOT}(\hat{\theta})$, da expressão (2.35), tomando-se amostras Bootstrap, no passo (i) do algoritmo 1, de tamanho *n*. Naturalmente, estimativas Bootstrap da tendenciosidade podem ser calculadas parametricamente, ou através de alguma suavização. É importante ter em mente que a estimativa Bootstrap de T(F), é de fato $T(\hat{F})$, onde $\hat{F}$ é uma escolha apropriada para substituir a F.

2.4 O processo Bootstrap para estruturas de dados mais gerais

Nas duas seções anteriores, aplicou-se o método Bootstrap para a estimação da variância e da tendenciosidade de uma determinada estatística $\hat{\theta} = t(\mathbf{x})$. Estas duas situações, que serviram para introduzir as idéias básicas do método, tiveram como ponto comum a estrutura estocástica associada ao problema, em que uma única amostra $\mathbf{X} = (X_1, X_2,..., X_n)$ é gerada segundo um mecanismo probabilístico que produz, realizando-se amostragem da população descrita pela FDA F, as n observações iid $x_1, x_2,..., x_n$. Por conveniência, será denotado este mecanismo, daqui por diante, pela simbologia "$F \rightarrow \mathbf{x}$", que quer dizer : F gera **x**, segundo o processo descrito acima. O objetivo do problema consiste na obtenção de informação de alguma característica da distribuição de uma variável aleatória R, possívelmente dependendo dos dados **x** e de F, isto é,

$$R = R(x,F). \qquad (2.4.1)$$

No caso da variância, $R(\mathbf{x},F) = \hat{\theta} = t(\mathbf{x})$ e a característica estimada é $\sigma^2(F) = Var_F(R(\mathbf{X},F))$. No caso da tendenciosidade de $\hat{\theta}$ como estimador de $\theta = s(F)$, $R(\mathbf{x},F) = \hat{\theta} - \theta = t(\mathbf{x}) - s(F)$ e a característica estimada é $T(F) = E_F(R(\mathbf{X},F))$.

A figura 2.8 exibe as etapas do processo Bootstrap. A primeira parte desta figura ilustra a estrutura estocástica acima discutida. O processo Bootstrap, propriamente dito, inicia-se ao estimar a FDA F, a partir dos dados observados **x**, por uma escolha apropriada de $\hat{F}$. Esta passagem é representada na figura pela simbologia " $\mathbf{x} \Rightarrow \hat{\mathbf{F}}$ ". Este passo é crucial para o Bootstrap. Se o processo é conduzido iniciando-se com um estimador $\hat{F}$ que não aproxime a verdadeira F, então, o método provavelmente falhará.

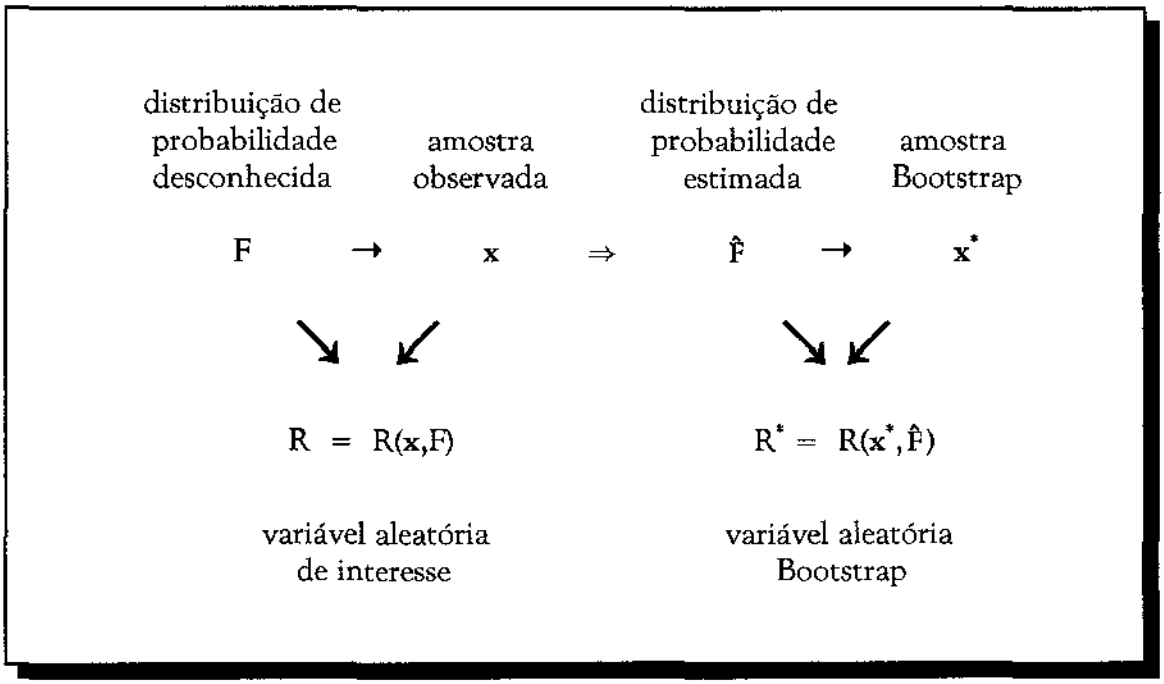

Figura 2.8. Diagrama esquemático do processo Bootstrap para uma estrutura estocástica de uma amostra (Efron and Tibshirani, 1986)

A passagem representada por " $\hat{\mathbf{F}} \rightarrow \mathbf{x}^*$ " significa que a amostra Bootstrap $\mathbf{x}^*$ é gerada de $\hat{F}$ pelas mesmas regras em que amostra original **x** foi gerada de F. A analogia Bootstrap da variável aleatória de interesse R = R(**x**,F) é, então, calculada com a amostra Bootstrap observada $\mathbf{x}^*$ e a distribuição $\hat{F}$, isto é,

$$R^* = R(\boldsymbol{x}^*, \hat{F}) \ . \qquad (2.4.2)$$

A aproximação Bootstrap para a distribuição de $R = R(\boldsymbol{X},F)$, *sob* F, *é a distribuição de* $R^* = R(\boldsymbol{X}^*,\hat{F})$, *condicional aos dados* $\boldsymbol{x}$ (esta distribuição é dita *distribuição Bootstrap de* R^*). Estimativas Bootstrap de parâmetros da distribuição verdadeira de R, tais como média, mediana, variância, tendenciosidade, quantis etc., são as correspondentes média, mediana, variância, tendenciosidade, quantis etc. da distribuição Bootstrap de R^*.

Em síntese, como vimos, o processo Bootstrap consiste num mecanismo de reprodução da estrutura real do problema, visto que os pseudos-dados da amostra Bootstrap são extraídos da distribuição descrita por $\hat{F}$, utilizando as mesmas regras em que a amostra original foi extraída da população descrita por F, e também, as estimativas Bootstrap das quantidades de interesse são calculadas na distribuição Bootstrap de R^*, pelas mesmas operações matemáticas que se desejaria realizar na verdadeira distribuição de $R = R(X,F)$.

As idéias empregadas até agora podem ser estendidas a problemas com maior número de amostras independentes. Por exemplo, pode-se desejar obter informação da distribuição da variável aleatória, função de duas amostras,

$$R = R((\boldsymbol{y},\boldsymbol{z}),(F,G)) = \hat{\theta} - \theta = t((\boldsymbol{y},\boldsymbol{z})) - s(F,G) \ , \qquad (2.4.3)$$

onde $\mathbf{y} = (y_1, y_2,..., y_m)$ é uma amostra extraída de uma população por um mecanismo probabilístico que produz as *m* observações $y_1, y_2,..., y_m$ independentes da distribuição F, e, $\mathbf{z} = (z_1, z_2,..., z_n)$ é outra amostra extraída de uma população por um mecanismo probabilístico que produz as *n* observações $z_1, z_2,..., z_n$ independentes da distribuição G. Pode-se representar o mecanismo probabilístico que produz os dados observados $\boldsymbol{x} = (\boldsymbol{y},\boldsymbol{z})$ por:

$$P = (F,G) \ , \tag{2.4.4}$$

onde " $P \rightarrow \mathbf{x}$ " (P gera $\mathbf{x}$) significa " $F \rightarrow \mathbf{y}$ " e, independentemente, " $G \rightarrow z$ ".

Dado um estimador do mecanismo da expressão (2.4.4) (passo " $\mathbf{x} \Rightarrow \hat{P}$ "), digamos,

$$\hat{P} = (\hat{F},\hat{G}) \ , \tag{2.4.5}$$

obtido a partir dos dados observados $\mathbf{x} = (\mathbf{y}, \mathbf{z})$, a amostra Bootstrap é formada pelo passo " $\hat{P} \rightarrow \mathbf{x}^*$ ", isto é, pelas mesmas regras em que " $P \rightarrow \mathbf{x}$ ", ou seja, $\mathbf{x}^* = (\mathbf{y}^*, \mathbf{z}^*)$, onde $\mathbf{y}^* = (y_1^*, y_2^*,..., y_m^*)$ são m observações extraídas independentemente da distribuição dada por $\hat{F}$ (" $\hat{F} \rightarrow \mathbf{y}^*$ "), e por outro processo independente, $z^* = (z_1^*, z_2^*,..., z_n^*)$ são n observações independentemente extraídas da distribuição dada por $\hat{G}$ (" $\hat{G} \rightarrow \mathbf{z}^*$ "). A analogia da variável aleatória

$$R = R(\mathbf{x},P) = R((\mathbf{y},\mathbf{z}),(F,G)) \ , \tag{2.4.6}$$

será dada por :

$$R^* = R(\mathbf{x}^*,\hat{P}) = R((\mathbf{y}^*,\mathbf{z}^*),(\hat{F},\hat{G})) \ . \tag{2.4.7}$$

A figura 2.9, exibe o diagrama do processo Bootstrap, para a estrutura de dados mais geral " $P \rightarrow \mathbf{x}$ ".

A aproximação Bootstrap para a distribuição de R(X,P), sob P, isto é, $\mathcal{L}_P(R)$, é a distribuição de $R^* = R(\mathbf{X}^*,\hat{P})$, sob $\hat{P}$, condicional aos dados $\mathbf{X} = \mathbf{x}$, isto é, $\mathcal{L}_{\hat{P}}(R^*/\mathbf{X}=\mathbf{x})$ (que é dita a distribuição Bootstrap de R^*). Estimativas Bootstrap de médias, variâncias,

quantis, etc. da distribuição de R, são as correspondentes médias, variâncias, quantis, etc. da distribuição Bootstrap de R^*.

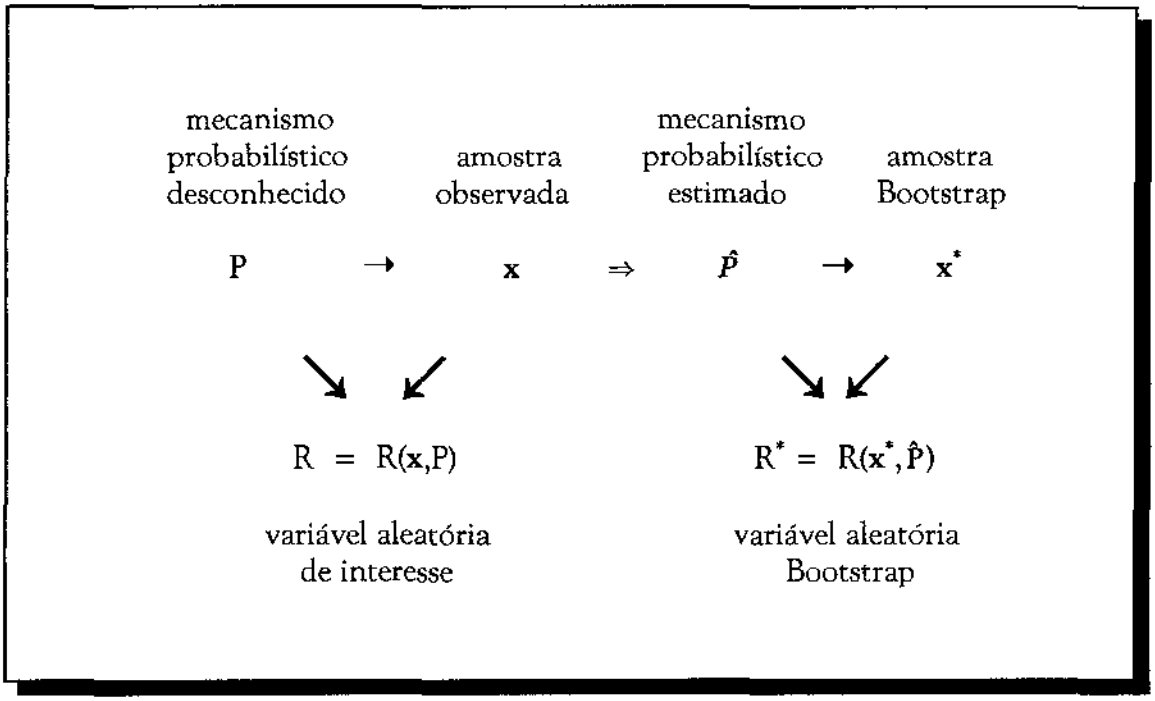

Figura 2.9. Diagrama esquemático do processo Bootstrap para uma estrutura estocástica mais geral " P → x " (Efron and Tibshirani, 1986)

Três problemas aparecem na condução do Bootstrap:

(1) estimação do mecanismo de probabilidade P dos dados atuais. Este é o passo indicado na figura 2 por " $\mathbf{x} \Rightarrow \hat{\mathbf{P}}$ " e crucial para o bom desempenho do Bootstrap;

(2) simulação dos dados de $\hat{\mathbf{P}}$, de acordo com a estrutura dos dados atual. Este é o passo " $\hat{\mathbf{P}} \rightarrow \mathbf{x}^*$" na figura 2.9. Este passo é conceituamente direto, isto é, executado da mesma forma como " $\mathbf{P} \rightarrow \mathbf{x}$ ", mas pode requerer algum cuidado se eficiência computacional é necessária;

(3) cálculo da distribuição Bootstrap de $R^* = R(\mathbf{X}^*, \hat{\mathbf{P}})$.

Com respeito à questão (1), uma das versatilidades do método Bootstrap é que ele pode ser conduzido de forma paramétrica ou não-paramétrica. Na seção 2.2 e 2.3, com a estimação da variância e da tendenciosidade, respectivamente, discutiu-se o uso da FDE $\hat{F}_n$ para estimar P = F, como também, o uso de estimativas paramétricas $\hat{F}_{par}$, ou, um compromisso entre estas duas (Bootstrap suavizado), visando tornar a FDA $\hat{F}_n$, que é discreta, mais suave. Estas escolhas podem ser adaptadas para estruturas mais gerais. Por exemplo, na situação de duas amostras " P = (F,G) $\rightarrow$ **x** = (**y**, **z**) ", o uso da FDE pode ser da seguinte forma : seja $\hat{F}_m$ a FDE que associa massa de probabilidade 1/m sobre cada $y_1, y_2, \ldots, y_m$ e $\hat{G}_n$ a FDE que associa massa de probabilidade 1/n sobre cada $z_1, z_2, \ldots, z_n$, isto é,

$$\hat{F}_m(y) = \frac{\sum_{i=1}^{m} \mathbf{1}_{(y_i \le y)}}{m} = \frac{\#\{ y_i \le y \}}{m} \qquad (2.4.8)$$

e

$$\hat{G}_n(z) = \frac{\sum_{j=1}^{n} \mathbf{1}_{(z_j \le z)}}{n} = \frac{\#\{ z_j \le z \}}{n}. \qquad (2.4.9)$$

A estimativa de P é dada por $\hat{P} = (\hat{F}_m, \hat{G}_n)$. O importante nesta passagem é que se use toda a informação disponível para obter um estimador $\hat{P}$ que melhor aproxime o mecanismo verdadeiro de P.

No que se refere à formação da amostra Bootstrap (questão (2)), foi visto que este passo é conceitualmente direto, isto é, da mesma forma em que " P $\rightarrow$ **x** ". Porém, pode-se requerer cuidados especiais se eficiência computacional na programação é necessária. Por exemplo, suponha que $\mathbf{X}^*$ será formada por seleção aleatória, com reposição, do conjunto $\{ x_1, x_2, \ldots, x_n \}$. Dependendo de onde os dados possam estar armazenados e da quantidade destes dados (*n*), o algorítmo usual baseado no acesso aleatório ao banco de dados pode

não ser tão eficiente quanto um algorítmo seqüencial que selecione as observações uma-a-uma com mesma probabilidade de seleção.

Com respeito à questão (3), o cálculo da distribuição Bootstrap de R^* é uma parte difícil do processo Bootstrap. Três métodos de cálculo são possíveis:

MÉTODO 1 : cálculo teórico direto;

MÉTODO 2 : aproximação de Monte Carlo para a distribuição Bootstrap;

MÉTODO 3 : expansões em série de Taylor podem ser usadas para obter médias variâncias aproximadas da distribuição de R^*.

O método 1 consiste em derivar a distribuição exata de R^* ou através de cálculo analítico, onde as estimativas Bootstrap das quantidades desejadas da distribuição de R são determinadas exatamente. Nas seções 2.2 e 2.3, foi visto que estas quantidades podem ser calculadas exatamente, usando a FDE, a partir da extração de todas amostras Bootstrap. De maneira geral, a estimativa Bootstrap exata da quantidade

$$E_P(g(R(\boldsymbol{X},P))) \ , \tag{2.4.10}$$

é dada por,

$$E_{\hat{P}}[g(R(\boldsymbol{X}^*,\hat{P})) \ / \ \boldsymbol{X} = \boldsymbol{x} \ ] \ . \tag{2.4.11}$$

Infelizmente, nem sempre obteremos as estimativas Bootstrap ou a distribuição Bootstrap exata de R^*. Para tanto, o método 2 é uma boa solução, e consiste realizar um

experimento de Monte Carlo, que usa o mecanismo da figura 4 e o algorítmo 1, tomando-se B amostras aleatórias independentes do modelo $\hat{P}$ (que é mantido fixo, assim como, os dados x), digamos, $x^*(1)$, $x^*(2)$,..., $x^*(B)$. A distribuição empírica das B replicações Bootstrap $r^*(1) = R(x^*(1),\hat{P})$, $r^*(2) = R(x^*(2),\hat{P})$,..., $r^*(B) = R(x^*(B),\hat{P})$, que pode ser visualizada com um histograma, é tomada como uma aproximação para a distribuição Bootstrap. Assim, a estimativa Bootstrap de (2.41) pode ser aproximada por :

$$\frac{\sum_{b=1}^{B} g(R(x^*(b),\hat{P}))}{B} \quad . \tag{2.4.12}$$

Uma vantagem deste método 2, é que quando a FDA F é estimada pela FDE dos dados observados, o processo Bootstrap é conduzido automáticamente. Naturalmente, a implementação deste método é totalmente dependente da potência computacional, visto que as aproximações tendem para as estimativas Bootstrap quando $B \rightarrow \infty$.

A exemplo do método 1, o método 3, depende também do poder analítico, e ao menos em situações específicas, perde em aplicabilidade. Em virtude disto, não será discutido aqui este método. Detalhes podem ser encontrados em Efron (1979a).

Capítulo 3

Intervalos de Confiança Bootstrap

3.1 Introdução

No capítulo 2 foi apresentado o método Bootstrap como um conjunto de técnicas para a obtenção de informações sobre características da distribuição de uma variável aleatória R(X,P), que não podem ser facilmente calculadas por métodos analíticos tradicionais ou que somente propriedades assintóticas são conhecidas. As características estudadas, até então, foram a variância, a tendenciosidade e, de uma forma geral, $E_P\, g(R(X,P))$, onde a variável aleatória R é definida, de forma conveniente, como função do estimador e do parâmetro de interesse.

Em um processo de inferência estatística, geralmente, não é suficiente somente avaliar determinadas características com um procedimento de estimação pontual. Muitas vezes, é necessário recorrer a um mecanismo de estimação por intervalo, que possibilite avaliar o erro que se comete na estimação pontual. A avaliação deste erro pode ser feita com a construção de um intervalo de confiança para o parâmetro de interesse, isto é, um intervalo do tipo

$$[\ l_1(x)\ ,\ l_2(x)\ ], \tag{3.3.1}$$

tal que, a probabilidade deste intervalo conter o verdadeiro valor do parâmetro, digamos θ, seja igual um valor pré-fixado $1 - 2\alpha$, ou seja,

$$Pr_P\{\, l_1(X) \leq \theta \leq l_2(X) \,\} = 1-2\alpha\,. \qquad (3.1.2)$$

O lado esquerdo de (3.1.2) é denominado *probabilidade de cobertura*, e, o valor $1- 2\alpha$ ($0 < \alpha < 0,5$) é o *nível de confiança* do intervalo. O valor (2α) é denominado *nível de significância* do intervalo. Intervalos de confiança satisfazendo a condição de (3.1.2), isto é, com probabilidade de cobertura igual ao nível de confiança, são ditos *exatos*. Geralmente, em aplicações práticas, busca-se construir intervalos de confiança com *caudas iguais*, ou seja, intervalos que satisfaçam a condição

$$Pr_P\{\, \theta < l_1(X) \,\} = \alpha \quad e \quad Pr_P\{\, \theta > l_2(X) \,\} = \alpha\,, \qquad (3.1.3)$$

isto é, o intervalo é construído de forma que as probabilidades das duas possibilidades do intervalo não conter o parâmetro, $\{\, \theta < l_1(X) \,\}$ e $\{\, \theta < l_2(X) \,\}$, são iguais à metade do nível de significância. É fácil ver que (3.1.3) implica (3.1.2). Porém, nem sempre o contrário é verdadeiro. Neste capítulo serão considerados intervalos de confiança com caudas iguais.

Infelizmente, nem sempre é possível construir intervalos exatos no sentido que as probabilidades em (3.1.3) são exatamente iguais ao valor α. Freqüentemente, os intervalos usados na prática são somente *aproximados*, isto é,

$$Pr_P\{\, \theta < l_1(X) \,\} \approx \alpha \quad e \quad Pr_P\{\, \theta > l_2(X) \,\} \approx \alpha\,. \qquad (3.1.4)$$

Neste caso, além do intervalo, os limites $l_1(x)$ e $l_2(x)$ são ditos também aproximados. Por exemplo, o intervalo de confiança para a média μ de uma determinada população, com variância conhecida σ^2_0, freqüentemente usado, é dado por

$$[\, l_1(x) \,,\, l_2(x) \,] = [\, \bar{x} + z_\alpha \frac{\sigma_0}{\sqrt{n}} \,,\, \bar{x} - z_\alpha \frac{\sigma_0}{\sqrt{n}} \,], \qquad (3.1.5)$$

onde, $\bar{x}$ é a média amostral de n observações independentes da população em questão, e, z_α é o α-ésimo percentil da distribuição normal padrão. Este intervalo é exato somente se as observações $X_1, X_2, \ldots, X_n$ são normalmente distribuídas com média μ e variância σ_0^2. Caso esta suposição não seja satisfeita, então, ele é somente aproximado, já que pelo teorema central do limite,

$$Pr\{ \mu < \bar{X} + z_\alpha \frac{\sigma_0}{\sqrt{n}} \} = Pr\{ \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma_0} > -z_\alpha \} \approx$$
$$\approx Pr\{ Z > -z_\alpha \} = \alpha, \tag{3.1.6}$$

e,

$$Pr\{ \mu < \bar{X} - z_\alpha \frac{\sigma_0}{\sqrt{n}} \} = Pr\{ \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma_0} < z_\alpha \} \approx$$
$$\approx Pr\{ Z < z_\alpha \} = \alpha, \tag{3.1.7}$$

onde Z tem distribuição normal padrão.

Uma primeira questão que pode aparecer quando se trabalha com intervalos de confiança aproximados é, por exemplo, com respeito a proximidade das probabilidades em (3.1.4), já que quanto mais próximas, intuitivamente pode-se dizer que, mais exato será o intervalo. Outra questão que pode ser colocada é quão próximos estão os limites aproximados $l_1(\mathbf{x})$ e $l_2(\mathbf{x})$ de limites exatos, como os definidos em (3.1.3). Estas duas questões definem, respectivamente, propriedades importantes para a avaliação e comparação de intervalos de confiança aproximados, que são a *acurácia* e a *corretibilidade* do intervalo. Estes conceitos serão definidos na seção 3.2 juntamente com outras propriedades desejáveis para os intervalos de confiança aproximados que são a *invariância a transformações monótonas* e a *preservação de amplitude.*

No decorrer deste capítulo, será considerada a estrutura probabilística de uma única amostra para os dados observados, conforme a figura 2.3, isto é, $P = F \rightarrow \mathbf{x}$, onde $\mathbf{x} = (x_1, x_2,..., x_n)$ é uma realização da amostra $\mathbf{X} = (X_1, X_2,..., X_n)$, cujas componentes são variáveis aleatórias *iid* com FDA F. A característica de interesse será representada por $\theta = s(F)$ que será estimada por $\hat{\theta} = t(\mathbf{X})$. A FDA de $\hat{\theta}$ será denotada e definida por:

$$G_F(s) = Pr_F\{ \hat{\theta} \leq s \}, \tag{3.1.8}$$

e, a FDA da estatística Bootstrap $\hat{\theta}^* = t(\mathbf{X}^*)$, será denotada e definida por:

$$\hat{G}_F(s) = Pr_{\hat{F}}\{ \hat{\theta}^* \leq s \}. \tag{3.1.9}$$

O objetivo deste capítulo é descrever como o método Bootstrap pode ser usado para a produção de intervalos de confiança para o parâmetro de interesse θ. Quase sempre estes intervalos serão aproximados. Porém, como será visto, esses intervalos poderão representar melhores aproximações, no sentido das propriedades definidas na seção 3.2, que outros intervalos aproximados baseados em métodos clássicos. Intervalos de confiança Bootstrap podem requerer um gasto computacional maior que os intervalos aproximados usuais, quando eles existirem. Caso estes não existam ou, não sejam fáceis de construir, então os intervalos de confiança Bootstrap representam pelos menos uma alternativa implementável para quem dispõe de um computador. Nas seções 3.3-3.5 serão apresentados os intervalos Bootstrap padrão, t-Student Bootstrap e t-Bootstrap. O primeiro desses intervalos é construído sob a suposição de normalidade aproximada da estatística $(\hat{\theta} - \theta)/\hat{\sigma}_{BOOT}(\hat{\theta})$. Já o segundo é construído com base na aproximação pela distribuição t-Student com n-1 g.l. em vez da normal padrão. O terceiro intervalo é construído sob a aproximação Bootstrap para a distribuição de $(\hat{\theta} - \theta)/\hat{\sigma}_{BOOT}(\hat{\theta})$, de forma análoga à construção do intervalo clássico que usa a a aproximação t-Student para esta distribuição. Finalmente, nas seções 3.6-3.8 serão discutidos os intervalos Bootstrap baseados nos percentis da distribuição Bootstrap da estatística $\hat{\theta}^*$. Estes intervalos são os então denominados percentil, percentil com correção para tendência e percentil com correção para tendência

e aceleração.

3.2 Definições básicas e propriedades

Nesta seção serão discutidas propriedades desejáveis para comparação e avaliação dos intervalos de confiança que serão apresentados no decorrer do capítulo. Será usada a notação $[\ \hat{\theta}[\alpha]\ ,\ \hat{\theta}[1-\alpha]\ ]$ para representar um intervalo de confiança aproximado, onde $\hat{\theta}[\alpha]$ e $\hat{\theta}[1-\alpha]$ são limites de confiança aproximados inferior e superior, respectivamente. Estes limites aproximados são tais que,

$$Pr_F\{\ \theta < \hat{\theta}[\alpha]\ \} \approx \alpha \quad e \quad Pr_F\{\ \theta > \hat{\theta}[1-\alpha]\ \} \approx \alpha . \tag{3.2.1}$$

Considere os seguintes tipos de erros:

$$ERRO_I(\hat{\theta}[\alpha]) = Pr_F\{\ \theta < \hat{\theta}[\alpha]\ \} - \alpha , \tag{3.2.2}$$

e,

$$ERRO_I(\hat{\theta}[1-\alpha]) = Pr_F\{\ \theta > \hat{\theta}[1-\alpha]\ \} - \alpha . \tag{3.2.3}$$

Propriedade 3.1 (*acurácia de ordem 1*) : *Um intervalo de confiança aproximado para o parâmetro θ, com probabilidade de cobertura aproximadamente igual a $1-2\alpha$, é denominado acurado de ordem 1, se,*

$$ERRO_I(\hat{\theta}[\alpha]) \doteq \frac{c_1}{\sqrt{n}} \quad e \quad ERRO_I(\hat{\theta}[1-\alpha]) \doteq \frac{c_2}{\sqrt{n}}, \tag{3.2.4}$$

para duas constantes c_1 e c_2.

Propriedade 3.2 (*acurácia de ordem 2*) : *Um intervalo de confiança aproximado para o parâmetro θ, com probabilidade de cobertura de aproximadamente igual a 1-2α, é denominado acurado de ordem 2, se,*

$$ERRO_I(\hat{\theta}[\alpha]) \doteq \frac{c_1}{n} \quad e \quad ERRO_I(\hat{\theta}[1-\alpha]) \doteq \frac{c_2}{n}, \tag{3.2.5}$$

para duas constantes c_1 e c_2.

Como se pode observar, a noção de acurácia está relacionada com a proximidade entre as probabilidades de cobertura dos intervalos $(-\infty,\hat{\theta}[\alpha])$ e $(\hat{\theta}[\alpha],+\infty)$, e, o nível desejado α. Portanto, quanto menores forem os erros definido em (3.2.2) e (3.2.3), mais acurado será o intervalo de confiança aproximado. A propriedade 3.1 estabelece que os erros dos intervalos acurado de ordem 1 tendem a zero, a medida que o tamanho da amostra n tende para o infinito, a uma taxa (ou velocidade) de $1/\sqrt{n}$. A propriedade 3.2 estabelece que os erros dos intervalos acurados de ordem 2 tendem a zero, a medida que n tende para o infinito, a uma taxa de $1/n$. A comparação destas duas taxas fornece não somente uma magnitude do erro que se comete nos intervalos mas, também, um critério de escolha entre eles, visto que a segunda taxa converge para zero bem mais rapidamente que a primeira.

Seja $\hat{\sigma}$ uma estimativa do desvio padrão de $\hat{\theta}$ e $\hat{\theta}_{exato}[\alpha]$ um limite de confiança exato para θ de nível α que satisfaz

$$Pr_F\{\ \theta \leq \hat{\theta}_{exato}[\alpha]\ \} = \alpha. \tag{3.2.6}$$

Considere agora o estudo do seguinte tipo de erro:

$$ERRO_{II}(\hat{\theta}[\alpha]) = \hat{\theta}[\alpha] - \hat{\theta}_{exato}[\alpha]. \tag{3.2.7}$$

Propriedade 3.3 (*corretibilidade de ordem 1*) : *Um limite de confiança aproximado* $\hat{\theta}[\alpha]$ *é denominado correto de ordem 1, se,*

$$ERRO_{II}(\hat{\theta}[\alpha]) = O_p(n^{-1}), \tag{3.2.8}$$

ou, equivalentemente,

$$ERRO_{II}(\hat{\theta}[\alpha]) = O_p(n^{-1/2})\hat{\sigma}, \tag{3.2.9}$$

desde que $\hat{\sigma}$ *é usualmente de ordem* $n^{-1/2}$.

Propriedade 3.4 (*corretibilidade de ordem 2*) : *Um limite de confiança aproximado* $\hat{\theta}[\alpha]$ *é denominado correto de ordem 2, se,*

$$ERRO_{II}(\hat{\theta}[\alpha]) = O_p(n^{-3/2}), \tag{3.2.10}$$

ou, equivalentemente,

$$ERRO_{II}(\hat{\theta}[\alpha]) = O_p(n^{-1})\hat{\sigma}. \tag{3.2.11}$$

Ao contrário da noção de acurácia, o conceito de corretibilidade se refere à proximidade de um limite de confiança aproximado para um limite de confiança exato. Corretibilidade de uma dada ordem, implica em acurácia da mesma ordem.

Propriedade 3.5 (*invariância a transformações monótonas*) : *Um intervalo de confiança para um parâmetro* θ *é dito possuir a propriedade de invariância a transformações monótonas se o intervalo obtido para um novo parâmetro* $\phi = g(\theta)$, *onde* $g(\cdot)$ *é uma transformação monótona, corresponde ao intervalo para* θ, *mapeado por* $g(\theta)$. *Se* $g(\cdot)$ *é monótona crescente então o intervalo para o novo parâmetro é dado por:*

$$[\hat{\phi}[\alpha],\hat{\phi}[1-\alpha]] = [g(\hat{\theta}[\alpha]),g(\hat{\theta}[1-\alpha])], \tag{3.2.12}$$

e, no caso de $g(\cdot)$ *ser monótona decrescente,*

$$[\hat{\phi}[\alpha],\hat{\phi}[1-\alpha]] = [g(\hat{\theta}[1-\alpha]),g(\hat{\theta}[\alpha])]. \qquad (3.2.13)$$

A propriedade de invariância a transformações monótonas fornece um resultado prático bastante útil. Por exemplo, se é desejável construir um intervalo para a um parâmetro $\phi = \log\theta$, dado um intervalo para θ com esta propriedade, então, basta aplicar o logarítmo nos limites deste intervalo para formar o intervalo para ϕ.

Propriedade 3.6 (*preservação de amplitude*) : *Um intervalo de confiança* $[\ \hat{\theta}[\alpha]\ ,\ \hat{\theta}[1-\alpha]\ ]$ *para um parâmetro* θ, *que assume valores em um conjunto* Θ, *tal que* $\Theta = [a,b] \subset \mathbb{R}$, $a \leq b$, *é dito possuir a propriedade de preservação de amplitude, se,* $[\ \hat{\theta}[\alpha]\ ,\ \hat{\theta}[1-\alpha]\ ] \subseteq \Theta$.

Esta propriedade de preservação de amplitude nem sempre se cumpre em muitos intervalos de confiança usados na prática. Por exemplo, considere a seguinte situação : Sejam $X_1, X_2,..., X_n$ n variáveis aleatórias independentes que assumem o valor " 1 " com probabilidade p ($p \in [0,1]$) e o valor " 0 " com probabilidade $1 - p$. Um intervalo de confiança bastante usado, com probabilidade de cobertura de aproximadamente $1-2\alpha$, é o seguinte:

$$[\ \hat{p} + z_{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\ ,\ \hat{p} - z_{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\ ], \qquad (3.2.14)$$

onde

$$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n}. \qquad (3.2.15)$$

Entretanto, se em uma amostra de tamanho $n = 15$ observou-se uma proporção amostral $\hat{p} = 0,1$, então o intervalo (3.2.14), para uma probabilidade de cobertura de aproximadamente $1 - 2\alpha = 0,9$ ($\alpha = 0,05$), será dado por:

$$[0,1-1,645(0,0775);0,1+1,645(0,0775)] \approx [-0,027;0,227] \not\subset [0,1]. \qquad (3.2.16)$$

Portanto este intervalo não possui a propriedade de preservação de amplitude.

3.3 Intervalo Bootstrap padrão

Seja $\hat{\sigma}_{BOOT}(\hat{\theta})$ a estimativa Bootstrap de $\sigma = \{ Var_F \hat{\theta} \}^{1/2}$, definida na seção 2.2. O intervalo de confiança *Bootstrap padrão (BOOTPAD)* para o parâmetro θ, com probabilidade de cobertura de aproximadamente $1-2\alpha$, é dado por:

$$\begin{aligned} &[\hat{\theta}_{BOOTPAD}[\alpha], \hat{\theta}_{BOOTPAD}[1-\alpha]] = \\ &= [\hat{\theta} + z_\alpha \hat{\sigma}_{BOOT}(\hat{\theta}), \hat{\theta} - z_\alpha \hat{\sigma}_{BOOT}(\hat{\theta})], \end{aligned} \qquad (3.3.1)$$

onde z_α é o α-ésimo percentil da distribuição normal padrão, isto é, $z_\alpha = \Phi^{-1}(\alpha)$. A construção deste intervalo baseia-se na aproximação assintótica,

$$T = \frac{\hat{\theta} - \theta}{\hat{\sigma}_{BOOT}(\hat{\theta})} \sim N(0,1). \qquad (3.3.2)$$

Caso a distribuição de T seja perfeitamente normal padrão, então, o intervalo definido em (3.3.1) é exato. Caso contrário, o intervalo é somente aproximado. Por exemplo, usando os dados da tabela 2.1, se o interesse é construir o intervalo de confiança BOOTPAD para o parâmetro μ com base na estatística $\hat{\theta} = \bar{x}$, com uma probabilidade de cobertura de aproximadamente 0,90 ($=1-2\alpha \Rightarrow \alpha = 0,05$), procede-se da seguinte forma : de acordo com a tabela 2.2, a estimativa Bootstrap (não-paramétrica) $\hat{\sigma}_{BOOT}(\bar{X})$ é dada por:

$$\hat{\sigma}_{BOOT}(\bar{X}) = \{ \hat{\sigma}^2_{BOOT}(\bar{X}) \}^{1/2} = (0,3778)^{1/2} = 0,6147. \qquad (3.3.3)$$

Portanto, os limites do intervalo BOOTPAD ($z_\alpha = z_{0,05} = -1,6449$) serão dados por:

$$\hat{\theta}_{BOOTPAD}\,[0,05] = \bar{x} + z_{0,05}\,\hat{\sigma}_{BOOT}\,(\bar{X}) = 1,9886 - 1,6449 \,.\, 0,6147 = 0,9775 \tag{3.3.4}$$

e

$$\hat{\theta}_{BOOTPAD}\,[0,95] = \bar{x} - z_{0,05}\,\hat{\sigma}_{BOOT}\,(\bar{X}) = 1,9886 + 1,6449 \,.\, 0,6147 = 2,9997, \tag{3.3.5}$$

respectivamente. De forma análoga, pode-se construir o intervalo BOOTPAD para o parâmetro μ, utilizando-se a estimativa Bootstrap paramétrica (a partir da estimativa da variância da tabela 2.4)

$$\hat{\sigma}_{BOOT}\,(\bar{X}) = \{\ \hat{\sigma}^2_{BOOT}\,(\bar{X})\ \}^{1/2} = (0,4197)^{1/2} = 0,6478. \tag{3.3.6}$$

Os intervalos BOOTPAD resultantes, destas duas formas de condução do processo Bootstrap, encontram-se exibidos na tabela 3.1.

Tabela 3.1. *Intervalo de confiança BOOTPAD para o parâmetro μ (= 2) com probabilidade de cobertura de aproximadamente 0,90, para os dados da tabela 2.1 (n = 10)*

BOOTSTRAP	$\hat{\theta}_{BOOTPAD}\,[0.05]$	$\hat{\theta}_{BOOTPAD}\,[0.95]$
Não-paramétrico	0,9775	2,9997
Paramétrico	0,9230	3,0542

Este exemplo corresponde a uma situação em que se dispõe de um intervalo de confiança exato, levando-se em conta a normalidade dos dados e o conhecimento da variância v^2 (= 4), que é dado por:

$$[\ \bar{x} + z_\alpha (\frac{v}{\sqrt{n}})\ ,\ \bar{x} - z_\alpha \frac{v}{\sqrt{n}}\ ] = [0{,}9483\ ;\ 3{,}0289\ ]. \qquad (3.3.6)$$

Assim, comparando-se os limites dos intervalos da tabela 3.1, vê-se que, mesmo para uma amostra de n=10, ambos os intervalos fornecem boas aproximações. As figuras 3.1 e 3.2 exibem a distribuição empírica da estatística T^* com $\hat{\theta} = \bar{x}$ (nos casos paramétrico e não-paramétrico), gerada pela simulação de 10000 amostras de $n = 10$ observações da distribuição N(2,4).

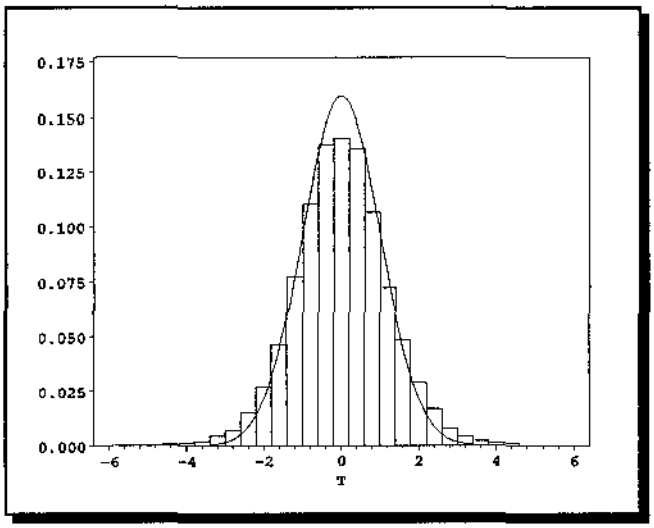

Figura 3.1. Distribuição empírica da estatística $T = (\bar{X} - 2)/\hat{\sigma}_{BOOT}(\bar{X})$ entre as 10000 amostras simuladas da distribuição N(2,4) (curva ajustada sobre o histograma N(0,1)) com a estimativa Bootstrap não-paramétrica $\hat{\sigma}_{BOOT}(\bar{X})$ calculada a partir da fórmula (2.2.14)

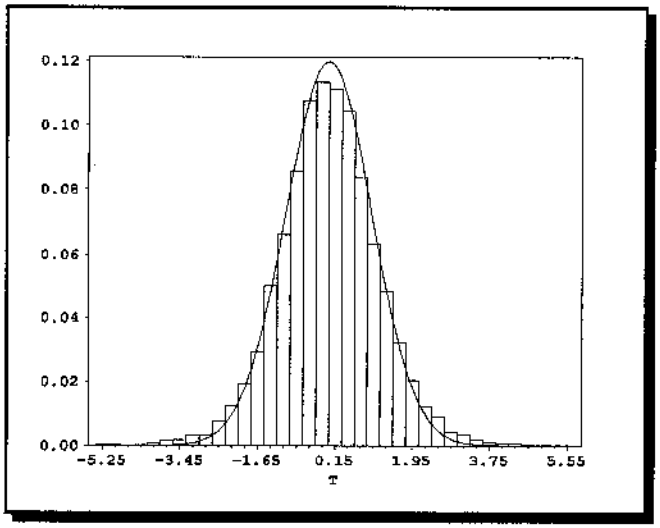

Figura 3.2. Distribuição empírica da estatística $T = (\bar{X} - 2)/\hat{\sigma}_{BOOT}(\bar{X})$ entre as 10000 amostras simuladas da distribuição N(2,4) (curva ajustada sobre o histograma N(0,1)) com a estimativa Bootstrap paramétrica $\hat{\sigma}_{BOOT}(\bar{X})$ calculada a partir da fórmula (2.2.31)

A principal vantagem do intervalo BOOTPAD é a sua facilidade de construção. Mesmo quando a estimativa Bootstrap $\hat{\sigma}_{BOOT}(\hat{\theta})$ for complicada de ser calculada analíticamente, conforme descrito na seção 2.2, pode-se utilizar em seu lugar, sua aproximação de Monte Carlo $\hat{\sigma}_B(\hat{\theta})$, que é obtida aplicando-se o algorítmo da figura 2.2. Neste caso, com a ajuda de um computador o intervalo BOOTPAD torna-se completamente automático, sendo apenas necessário para um usuário obter um intervalo de confiança aproximado para o parâmetro θ, fornecer a forma da estatística $\hat{\theta}$.

A grande desvantagem do intervalo BOOTPAD é com respeito a sua acurácia. Para tamanhos amostrais que não sejam suficientemente grandes para a validade da aproximação de (3.3.2), ou, que, características como assimetria, tendenciosidades, etc. que podem estar presentes na distribuição de T, podem prejudicar o desempenho do intervalo BOOTPAD no sentido da sua acurácia. No exemplo discutido acima, a aproximação para

a distribuição de T poderá ser razoável, visto que os dados foram gerados da distribuição normal. Agora, suponhamos que, ainda neste exemplo, deseja-se construir um intervalo de confiança para o parâmetro μ^3 com base na estatística $\hat{\theta} = \bar{x}^3$. O intervalo BOOTPAD (paramétrico e não-paramétrico) para este novo parâmetro , com probabilidade de cobertura de aproximadamente 0,90, encontra-se na tabela 3.2.

Tabela 3.2. *Intervalo de confiança BOOTPAD para o parâmetro μ^3 (= 8) com probabilidade de cobertura de aproximadamente 0,90, para os dados da tabela 2.1 (n = 10)*

BOOTSTRAP	$\hat{\theta}_{BOOTPAD}$ [0.05]	$\hat{\theta}_{BOOTPAD}$ [0.95]
Não-paramétrico	-6,3149	22,0429
Paramétrico	-7,3250	23,0530

Para analisar a acurácia deste intervalo foi conduzido um estudo de Monte Carlo com 10.000 amostras de $n = 10$ observações independentes, simuladas da distribuição normal com média $\mu = 2$ e variância $v^2 = 4$. Para cada uma destas amostras foi calculado o correspondente intervalo BOOTPAD (paramétrico e não-paramétrico) e observado as percentagens:

$$\hat{p}_1 = \frac{\#\{\ \mu^3 < \hat{\theta}_{BOOTPAD}\ [0.05]\ \}}{10000} \tag{3.3.7}$$

e

$$\hat{p}_2 = \frac{\#\{\ \mu^3 > \hat{\theta}_{BOOTPAD}\ [0.95]\ \}}{10000}. \tag{3.3.8}$$

Os resultados encontram-se na tabela 3.3.

Tabela 3.3. *Percentagens de não-cobertura, à esquerda ($\hat{p}_1$) e à direita ($\hat{p}_2$), em 10.000 realizações do intervalo BOOTPAD para o parâmetro μ^3 (= 8), com uma probabilidade de cobertura de aproximadamente 0,90*

BOOTSTRAP	$\hat{p}_1$(%)	$\hat{p}_2$(%)
Não-paramétrico	2,14	11,65
Paramétrico	1,69	10,39

Como se pode observar, as percentagens de não-cobertura, em ambos os casos, diferem das probabilidades nominais iguais a 5%. As figuras 3.3 e 3.4 exibem a distribuição empírica da estatística T das 10.000 amostras da distribuição N(2,4), para os casos paramétrico e não-paramétrico, respectivamente, e, a função de densidades da distribuição normal padrão. Observas-se que a aproximação não é satisfatória, pois a distribuição empírica de T é assimetrica à esquerda. Portanto, como os limites do intervalo BOOTPAD são sempre simétricos em torno da estimativa $\hat{\theta}$, então, a assimetria não é levada em conta nos cálculos dos limites que poderão ser bastante inacurados como o caso dos limites da tabela 3.2. Nas situações em que limites de confiança exatos podem ser definidos, os limites do intervalo BOOTPAD são apenas corretos de ordem 1 (Efron, 1987), e portanto, acurados de ordem 1. Uma outra desvantagem deste intervalo é que, em geral, ele não possue as propriedades 3.5 (invariância a transformações monótonas) e 3.6 (preservação de amplitude).

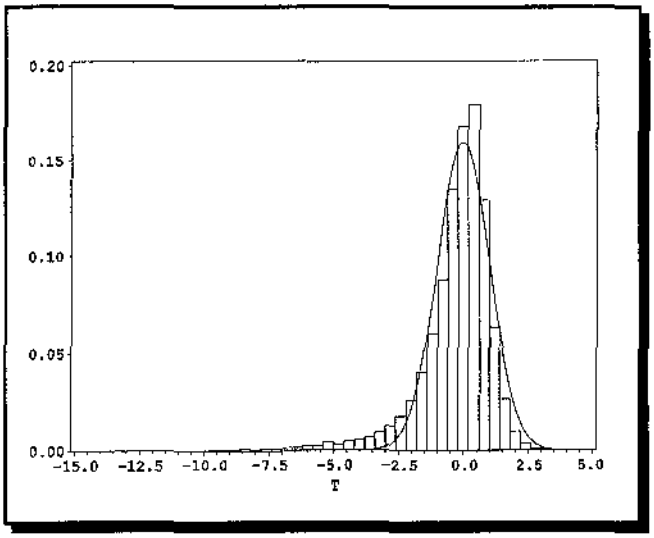

Figura 3.3. Distribuição empírica da estatística $T = (\bar{X}^3 - 8)/\hat{\sigma}_{BOOT}(\bar{X}^3)$ entre as 10.000 amostras simuladas da distribuição N(2,4) (curva ajustada sobre o histograma N(0,1)) com a estimativa Bootstrap não-paramétrica $\hat{\sigma}_{BOOT}(\bar{X}^3)$ calculada a partir da fórmula (2.2.20)

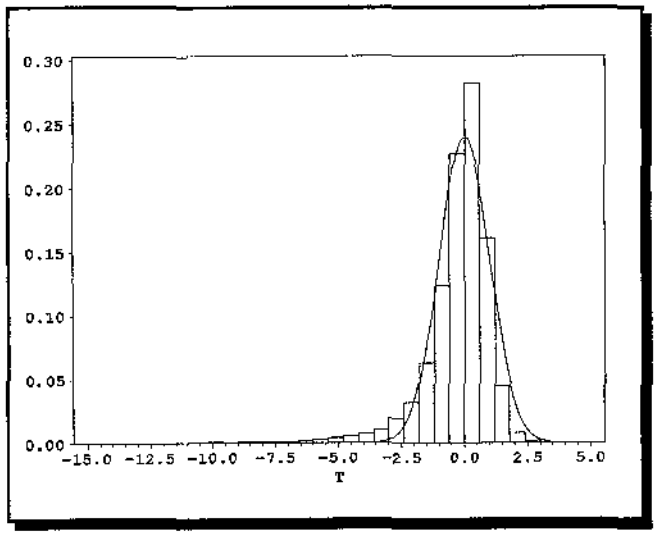

Figura 3.4. Distribuição empírica da estatística $T = (\bar{X}^3 - 8)/\hat{\sigma}_{BOOT}(\bar{X}^3)$ entre as 10.000 amostras simuladas da distribuição N(2,4) (curva ajustada sobre o histograma N(0,1)) com a estimativa Bootstrap paramétrica $\hat{\sigma}_{BOOT}(\bar{X}^3)$ calculada a partir da fórmula (2.2.33).

Um procedimento bastante usado na estatística clássica, que pode fornecer um intervalo mais acurado que o intervalo BOOTPAD é o seguinte: dado que existe uma transformação monótona g(·) tal que, para $\phi = g(\theta)$ e $\hat{\phi} = g(\hat{\theta})$,

$$\frac{\hat{\phi} - \phi}{DP(\hat{\phi})} \approx N(0,1), \tag{3.3.9}$$

onde $DP(\hat{\phi})$ é o desvio padrão (ou uma estimativa) de $\hat{\phi}$, ou seja, a distribuição de $\hat{\phi}$ está mais próxima à distribuição normal do que a distribuição de $\hat{\theta}$, então, o intervalo de confiança central para ϕ, com probabilidade de cobertura de aproximadamente $1-2\alpha$, é dado por :

$$[\hat{\phi}[\alpha],\hat{\phi}[1-\alpha]] = [\hat{\phi} + z_{\alpha}DP(\hat{\phi}),\hat{\phi} - z_{\alpha}DP(\hat{\phi})]. \tag{3.3.10}$$

O intervalo desejado é obtido convertendo-se à escala θ, os limites do intervalo anterior através da transformação inversa $g^{-1}(\cdot)$. Por exemplo, se g(·) é monótona crescente, então,

$$[\hat{\theta}[\alpha],\hat{\theta}[1-\alpha]] = [g^{-1}(\hat{\phi} + z_{\alpha}DP(\hat{\phi})),g^{-1}(\hat{\phi} - z_{\alpha}DP(\hat{\phi}))], \tag{3.3.11}$$

e, o inverso dos limites do intervalo anterior caso g(·) seja monótona decrescente. Por exemplo, no caso da construção do intervalo de confiança para o parâmetro μ^3, com base na estatística $\hat{\theta} = \bar{x}^3$, foi visto que os limites do intervalo BOOTPAD eram bastantes inacurados. Porém, uma transformação monótona que normaliza a distribuição de $\hat{\theta}$ é dada por $g(y) = y^{(1/3)}$, pois

$$\hat{\phi} = g(\hat{\theta}) = \hat{\theta}^{1/3} = (\bar{X}^3)^{1/3} = \bar{X} \approx N(\phi,\frac{\nu^2}{n}), \tag{3.3.12}$$

onde

$$\phi = g(\theta) = \theta^{1/3} = (\mu^3)^{1/3} = \mu. \tag{3.3.13}$$

Logo, um intervalo de confiança aproximado para ϕ, com base em $\hat{\phi}$, pode ser dado pelos

limites do intervalo BOOTPAD para o parâmetro μ descritos na tabela 3.1, ou seja,

$$[\hat{\phi}[0,05], \hat{\phi}[0,95]] = [\bar{x} + z_{\alpha}\hat{\sigma}_{BOOT}(\bar{X}), \bar{x} - z_{\alpha}\hat{\sigma}_{BOOT}(\bar{X})]. \qquad (3.3.14)$$

Assim, os limites do intervalo de confiança desejado são obtidos convertendo-se os limites da tabela 3.1 à escala μ^3, pela transformação inversa $g^{-1}(y) = y^3$, ou seja,

$$\begin{aligned}[\hat{\theta}[0,05], \hat{\theta}[0,95]] &= [(\hat{\phi}[0,05])^3, (\hat{\phi}[0,95])^3] = \\ &= [(\bar{x} + z_{\alpha}\hat{\sigma}_{BOOT}(\bar{X}))^3, (\bar{x} - z_{\alpha}\hat{\sigma}_{BOOT}(\bar{X}))^3].\end{aligned} \qquad (3.3.15)$$

Os resultados do intervalo resultante deste procedimento estão descritos na tabela 3.4.

Tabela 3.4. *Intervalo de confiança para o parâmetro μ^3 (= 8), com probabilidade de cobertura de aproximadamente 0,90, obtido elevando-se ao cubo os limites do intervalo BOOTPAD da tabela 3.1*

BOOTSTRAP	$\hat{\theta}$[0.05]	$\hat{\theta}$[0.95]
Não-paramétrico	0,9340	26,9919
Paramétrico	0,7863	28,4900

Para constatar a melhora na acurácia destes limites, produziu-se as correspondentes realizações deste intervalo com base nas 10.000 amostras simulada da distribuição N(2,4) e calculou-se as percentagens de não-cobertura à direita e à esquerda, conforme a tabela 3.5. Como se pode observar, pelos resultados encontrados, os valores de $\hat{p}_1$ e $\hat{p}_2$ estão mais próximos do valor nominal 5%, do que anteriormente. Portanto, o intervalo resultante deste procedimento de transformação é sem dúvida mais acurado que o intervalo BOOTPAD aplicado direto na escala de μ^3.

Tabela 3.5. *Percentagens de não-cobertura, à esquerda ($\hat{p}_1$) e à direita ($\hat{p}_2$), em 10000 realizações do intervalo de confiança para o parâmetro μ^3 (= 8) com uma probabilidade de cobertura de aproximadamente 0,90, cujos limites são os cubos dos limites do intervalo BOOTPAD para o parâmetro μ*

BOOTSTRAP	$\hat{p}_1$(%)	$\hat{p}_2$(%)
Não-paramétrico	7,89	7,44
Paramétrico	6,97	6,50

Um outro exemplo de aplicação deste procedimento é a transformação de Fisher do coeficiente de correlação, $\hat{\phi} = \tanh^{-1}\hat{\rho}$, cuja distribuição está mais próxima à distribuição normal do que a distribuição de $\hat{\rho}$. A desvantagem desta estratégia é a necessidade do analista de dados deduzir, para cada aplicação, a transformação que torna a distribuição da estatística transformada mais próxima à normal que a distribuição da estatística de interesse. Mais adiante será mostrado que alguns intervalos de confiança Bootstrap incorporam em seus cálculos, automaticamente, transformações como a anterior sem requerer o seu conhecimento.

3.4 Intervalo t-Student Bootstrap

O intervalo de confiança BOOTPAD, discutido na seção anterior, foi derivado sob a suposição que a distribuição de T pode ser aproximada pela distribuição normal padrão. Em determinadas situações esta suposição torna-se válida quando o tamanho da amostra *n* tende para infinito. Em pequenas amostras, pelo menos para o caso em que $\hat{\theta} = \bar{x}$, uma aproximação melhor para a distribuição de T é dada pela distribuição t-Student com *(n-1)* graus de liberdade, isto é,

$$T = \frac{\hat{\theta} - \theta}{\hat{\sigma}_{BOOT}(\hat{\theta})} \sim t_{n-1}. \tag{3.4.1}$$

Se $\hat{\theta} = \bar{x}$, $\hat{\sigma}_{BOOT}(\hat{\theta}) = \sqrt{\frac{s^2}{n}} = \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2/(n(n-1))}$ e as observações são normalmente distribuídas, então esta aproximação é exata.

Seja $t_{n-1}^{(\alpha)}$ o α-ésimo percentil da distribuição *t* com *n-1* graus de liberdade, ou seja, o valor desta distribuição que satisfaz à equação $\alpha = Pr_F\{ t_{n-1} \leq t_{n-1}^{(\alpha)} \}$. Assim, sob (3.4.1),

$$\begin{aligned} 1 - 2\alpha &\approx Pr_F \{ t_{n-1}^{(\alpha)} \leq T \leq t_{n-1}^{(1-\alpha)} \} = Pr_F \{ t_{n-1}^{(\alpha)} \leq T \leq -t_{n-1}^{(\alpha)} \} = \\ &= Pr_F \{ \hat{\theta} + t_{n-1}^{(\alpha)} \hat{\sigma}_{BOOT}(\hat{\theta}) \leq \theta \leq \hat{\theta} - t_{n-1}^{(\alpha)} \hat{\sigma}_{BOOT}(\hat{\theta}) \}. \end{aligned} \tag{3.4.2}$$

Portanto, o intervalo de confiança *t-Student Bootstrap (t-SBOOT)* com probabilidade de cobertura aproximadamente igual a 1-2α será dado por:

$$\begin{aligned} &[\hat{\theta}_{t\text{-}SBOOT}[\alpha], \hat{\theta}_{t\text{-}SBOOT}[1-\alpha]] = \\ &= [\hat{\theta} + t_{n-1}^{(\alpha)} \hat{\sigma}_{BOOT}(\hat{\theta}), \hat{\theta} - t_{n-1}^{(\alpha)} \hat{\sigma}_{BOOT}(\hat{\theta})]. \end{aligned} \tag{3.4.3}$$

Como se pode observar este intervalo tem forma similar ao intervalo BOOTPAD, exceto que os percentis $t_{n-1}^{(\alpha)}$ e $-t_{n-1}^{(\alpha)}$ são utilizados no lugar de z_α e $-z_\alpha$, respectivamente. Na tabela 3.6 é apresentado o intervalo t-SBOOT para os parâmetros μ e μ^3, obtidos a partir dos dados da tabela 2.1.

Tabela 3.6. Intervalo de confiança t-SBOOT para os parâmetros μ (= 2) e μ^3 (= 8) com probabilidade de cobertura de aproximadamente 0,90, para os dados da tabela 2.1

Parâmetro	BOOTSTRAP	$\hat{\theta}_{t\text{-}SBOOT}$ [0.05]	$\hat{\theta}_{t\text{-}SBOOT}$ [0.95]
μ	Não-paramétrico	0,8618	3,1154
μ	Paramétrico	0,8011	3,1761
μ^3	Não-paramétrico	-7,9375	23,6650
μ^3	Paramétrico	-9,0632	24,7907

Analisando-se os intervalos da tabela 3.6, para o parâmetro μ, percebe-se uma diferença entre os não-paramétricos e paramétricos. Em (2.2.31) vimos para o Bootstrap paramétrico que $\hat{\sigma}_{BOOT}(\bar{X}) = s/\sqrt{n}$. Logo, o intervalo t-SBOOT paramétrico para o parâmetro μ (segunda linha da tabela 3.6) é exato, pois, com a suposição de normalidade dos dados da tabela 2.1, a estatística T é exatamente distribuída segundo uma *t*-Student com *(n-1)* g.l. Assim, este intervalo dá a probabilidade de cobertura correta. Para o parâmetro μ^3, a exemplo do intervalo BOOTPAD, os limites do intervalo t-SBOOT são bastantes inacurados, como se pode ver na tabela 3.7 que exibe os resultados com base nas 10.000 amostras simuladas da distribuição N(2,4).

Tabela 3.7. *Percentagens de não-cobertura, à esquerda ($\hat{p}_1$) e à direita ($\hat{p}_2$), em 10000 realizações do intervalo t-SBOOT para o parâmetro μ^3 (= 8), com uma probabilidade de cobertura de aproximadamente 0,90*

BOOTSTRAP	$\hat{p}_1$(%)	$\hat{p}_2$(%)
Não-paramétrico	1,26	10,01
Paramétrico	0,93	8,86

Isto acontece, pois, a exemplo do intervalo BOOTPAD, o intervalo t-SBOOT é simétrico em torno de $\hat{\theta}$, em virtude da simetria em torno do valor zero dos percentis da distribuição t de Student. Assim, o intervalo de (3.4.3) não leva em conta assimetria na distribuição de $\hat{\theta}$, ou outras características quando $\hat{\theta}$ não é a média amostral. Um melhor intervalo pode ser obtido aplicando-se o procedimento de transformação para alcançar uma simetria da distribuição de $\hat{\theta}$. Como realizado para o intervalo BOOTPAD , os limites do intervalo t-SBOOT para o parâmetro μ (linhas 1 e 2 da tabela 3.6) foram transformados pela transformação cúbica para à escala μ^3 e suas percentagens de não-cobertura foram calculadas. Os resultados encontram-se nas tabelas 3.8 e 3.9.

Tabela 3.8. *Intervalo de confiança para o parâmetro μ^3 (= 8)-, com probabilidade de cobertura de aproximadamente 0,90, obtido elevando-se ao cubo os limites do intervalo t-SBOOT da tabela 3.6 (linhas 1 e 2)*

BOOTSTRAP	$\hat{\theta}[0.05]$	$\hat{\theta}[0.95]$
Não-paramétrico	0,6401	30,2372
Paramétrico	0,5141	32,0393

Tabela 3.9. *Percentagens de não-cobertura, à esquerda ($\hat{p}_1$) e à direita ($\hat{p}_2$), em 10000 realizações do intervalo de confiança para o parâmetro μ^3 (= 8) com uma probabilidade de cobertura de aproximadamente 0,90, cujos limites são os cubos dos limites do intervalo t-SBOOT para o parâmetro μ .*

BOOTSTRAP	$\hat{p}_1$(%)	$\hat{p}_2$(%)
Não-paramétrico	5,99	5,52
Paramétrico	5,10	4,72

De acordo com a tabela 3.9, comprova-se a necessidade do conhecimento de uma transformação para situações mais gerais, que a da média amostral, para um bom desempenho do intervalo t-SBOOT. Nas situações onde limites de confiança podem ser definidos, o intervalo t-SBOOT é, em geral, correto de ordem 1 (Efron, 1987). Portanto este intervalo é acurado de ordem 1. Também, este intervalo não possue, em geral, as propriedades de preservação de amplitude e invariância a transformações monótonas. Na próxima seção será apresentado o intervalo de confiança t-Bootstrap, cuja construção

consiste do uso do Bootstrap não somente para estimar $\sigma = \{ Var_F \hat{\theta} \}^{1/2}$, mas também, para obter uma aproximação para a distribuição de T, com base nos dados observados **x**.

3.5 Intervalo t-Bootstrap

Foram apresentados nas duas últimas seções os intervalos BOOTPAD e t-SBOOT, cujas construções baseiam-se em suposições sobre a aproximação da distribuição da estatística T à distribuição normal padrão e a *t*-Student com *(n-1)* g.l., respectivamente. Porém, nem sempre esta suposição é válida. Será descrito, a seguir, uma outra metodologia Bootstrap para a construção de um intervalo de confiança para o parâmetro $\theta = s(F)$, usando uma aproximação para a distribuição de T, fornecida pelo Bootstrap. A idéia é, com base nesta aproximação, obter estimativas Bootstrap dos percentis da distribuição de T, usando-os para formar o intervalo de confiança de forma análoga à seção 3.3 e 3.4.

De acordo com a seção 2.4, a aproximação Bootstrap para a distribuição de

$$T = T(X,F) = \frac{\hat{\theta} - \theta}{\hat{\sigma}_{BOOT}(\hat{\theta})} = \frac{t(X) - s(F)}{\hat{\sigma}_{BOOT}(\hat{\theta})} \qquad (3.5.1)$$

é a distribuição de

$$T^* = T(X^*,\hat{F}) = \frac{\hat{\theta}^* - s(\hat{F})}{\hat{\sigma}_{BOOT}(\hat{\theta})} = \frac{t(X^*) - s(\hat{F})}{\{Var_{\hat{F}}(t(X^*))\}^{1/2}}, \qquad (3.5.2)$$

condicional aos dados observados **x**. A construção do intervalo de confiança t-Bootstrap (Efron, 1981) é feita da seguinte forma : sejam t^*_{α} e $t^*_{1-\alpha}$ os α-ésimo e $(1-\alpha)$-ésimo percentis da distribuição Bootstrap de T^*, tais que,

$$Pr_*\{ T^* \leq t_\alpha^* \} = \alpha \quad e \quad Pr_*\{ T^* \leq t_{1-\alpha}^* \} = 1-\alpha, \tag{3.5.3}$$

onde, " Pr_* " indica que as probabilidades acima acima são calculadas sob a distribuição Bootstrap de T^*. Assim,

$$Pr_*\{ t_\alpha^* \leq T^* \leq t_{1-\alpha}^* \} = 1-2\alpha. \tag{3.5.4}$$

Se a distribuição Bootstrap de T^* fornece uma boa aproximação para a distribuição de T, então

$$Pr_*\{ t_\alpha^* \leq T^* \leq t_{1-\alpha}^* \} \approx Pr_F\{ t_\alpha^* \leq T \leq t_{1-\alpha}^* \}. \tag{3.5.5}$$

Agora, combinando os resultados de (3.5.4) e (3.5.5), teremos que,

$$\begin{aligned} 1 - 2\alpha &\approx Pr_F\{ t_\alpha^* \leq T \leq t_{1-\alpha}^* \} = \\ &= Pr_F\{ \hat{\theta} - t_{1-\alpha}^* \hat{\sigma}_{BOOT}(\hat{\theta}) \leq \theta \leq \hat{\theta} - t_\alpha^* \hat{\sigma}_{BOOT}(\hat{\theta}) \}. \end{aligned} \tag{3.5.6}$$

Logo, desta última relação, temos que o intervalo de confiança t-Bootstrap (t-BOOT) para o parâmetro θ, com probabilidade de cobertura de aproximadamente $1 - 2\alpha$, será dado por:

$$\begin{aligned} &[\hat{\theta}_{t\text{-}BOOT}[\alpha] , \hat{\theta}_{t\text{-}BOOT}[1-\alpha]] = \\ &= [\hat{\theta} - t_{1-\alpha}^* \hat{\sigma}_{BOOT}(\hat{\theta}) , \hat{\theta} - t_\alpha^* \hat{\sigma}_{BOOT}(\hat{\theta})]. \end{aligned} \tag{3.5.7}$$

Por exemplo, para construir o intervalo de confiança t-BOOT para o parâmetro μ, usando o Bootstrap paramétrico, temos que:

$$\mu = s(F) = \int x dF(x), \tag{3.5.8}$$

$$s(\hat{F}) = s(\hat{F}_{par}) = \int x d\hat{F}_{par}(x) = \bar{x} \tag{3.5.9}$$

e

$$T^* = \frac{\bar{X}^* - \bar{x}}{\hat{\sigma}_{BOOT}(\bar{X})} = \frac{\bar{X}^* - \bar{x}}{s/\sqrt{n}}. \tag{3.5.10}$$

Como neste caso do Bootstrap paramétrico,

$$\bar{X}^* \underset{*}{\sim} N(\bar{x}, s^2/n), \tag{3.5.11}$$

então a distribuição Bootstrap da estatística T^* da expressão (3.5.10) é a normal padrão. Portanto,

$$t^*_{\alpha} = z_{\alpha} \quad e \quad t^*_{1-\alpha} = -z_{\alpha}. \tag{3.5.12}$$

Logo, o intervalo t-BOOT coincide com o intervalo BOOTPAD para o parâmetro μ (linha 2 da tabela 3.1) neste caso paramétrico.

Na maioria das vezes, como já mencionado no capítulo 2, poderá ser difícil calcular analíticamente a distribuição Bootstrap exata de T^*, como foi feito acima. Neste caso, uma saída é usar a aproximação de Monte Carlo, descrita na seção 2.4, para estimar a distribuição Bootstrap de T^* e, por conseguinte, os percentis desta distribuição. Este procedimento pode ser feito da seguinte forma : sejam $\mathbf{x}^*(1)$, $\mathbf{x}^*(2)$,..., $\mathbf{x}^*(B)$, onde

$$\mathbf{x}^*(b) = (x^*_{b_1}, x^*_{b_2}, ..., x^*_{b_n}), \quad b = 1, 2, ..., B, \tag{3.5.13}$$

B amostras Bootstrap geradas reproduzindo-se o mecanismo $P = F \rightarrow x$, conforme a figura 2.9. Para cada uma destas amostras, seja

$$T^*(b) = T(x^*(b), \hat{F}) = \frac{t(x^*(b)) - s(\hat{F})}{\{Var_{\hat{F}}\,(t(x^*(b)))\}^{1/2}}, \quad b=1,2,...,B, \tag{3.5.14}$$

onde $\{\ var_{\hat{F}}\, t(x^*(b))\ \}^{1/2}$ representa a estimativa Bootstrap do desvio padrão de $\hat{\theta}$ em cada amostra Bootstrap. O α-ésimo percentil da distribuição de T^* é estimado pelo valor $\hat{t}^*_\alpha$, tal que,

$$\frac{\#\{\ T^*(b) \le \hat{t}^*_\alpha\ \}}{B} = \alpha. \tag{3.5.15}$$

Este valor pode ser calculado da seguinte forma : ordenando-se os B valores $T^*(b)$, digamos $T^*_{(1)}, T^*_{(2)},..., T^*_{(B)}$, então, $\hat{t}^*_\alpha$ corresponde ao $B.\alpha$ maior valor, destes ordenados, isto é, $T^*_{(B\alpha)}$. Por exemplo, se B = 1000, então, o percentil 5% da distribuição Bootstrap de T^* é estimado pelo quinquagésimo maior valor entre os $T^*_{(b)}$, ou seja, $T^*_{(50)}$. Analogamente, o percentil 95% é estimado por $T^*_{(950)}$. Caso $B.\alpha$ não seja inteiro um procedimento indicado para o cálculo de $\hat{t}^*_\alpha$ e $\hat{t}^*_{1-\alpha}$, segundo Efron(1993), se $\alpha \le .5$, consiste em tomar $k = [(B+1)\alpha]$, o maior inteiro menor ou igual a $(B+1)\alpha$, e, os quantis empíricos α e $1 - \alpha$ são os k-ésimo e (B+1-k)-ésimo valores ordenados de $T^*(b)$, b=1, 2,..., B, isto é, $T^*_{(k)}$ e $T^*_{(B+1-k)}$, respectivamente. Dessa forma, o intervalo t-Bootstrap será dado por (3.5.7), com $\hat{t}^*_\alpha$ e $\hat{t}^*_{1-\alpha}$ em lugar de t^*_α e $t^*_{1-\alpha}$, respectivamente. Por exemplo, para construir o intervalo de confiança t-BOOT para o parâmetro $\mu = s(F)$ utilizando o estimador $\hat{F} = \hat{F}_n$, ou seja, um Bootstrap não-paramétrico, então

$$s(\hat{F}) = s(\hat{F}_n) = \int x d\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x} \tag{3.5.16}$$

e

$$T^* = \frac{\bar{X}^* - \bar{x}}{\hat{\sigma}_{BOOT}(\bar{X})} = \frac{(\bar{X}^* - \bar{x})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2/n^2}}. \qquad (3.5.17)$$

Como não se conhece a distribuição de T^*, neste caso não-paramétrico, então, com base em B amostras Bootstrap da FDE $\hat{F}_n$, calcula-se a estatística

$$T^*(b) = \frac{(\bar{x}_b^* - \bar{x})}{\hat{\sigma}_{BOOT}(\bar{x}_b^*)} = \frac{(\bar{x}_b^* - \bar{x})}{\sqrt{\sum_{i=1}^{n} (x_{b_i}^* - \bar{x}_b^*)^2/n^2}}. \qquad (3.5.18)$$

Dessa forma, a distribuição Bootstrap de T^*, pode ser aproximada pela distribuição empírica dos B valores de $T^*(b)$. Para os dados da tabela 2.1 foi realizado este processo com B = 1000 replicações Bootstrap. A figura 3.5 exibe um histograma dos valores calculados de $T^*(b)$.

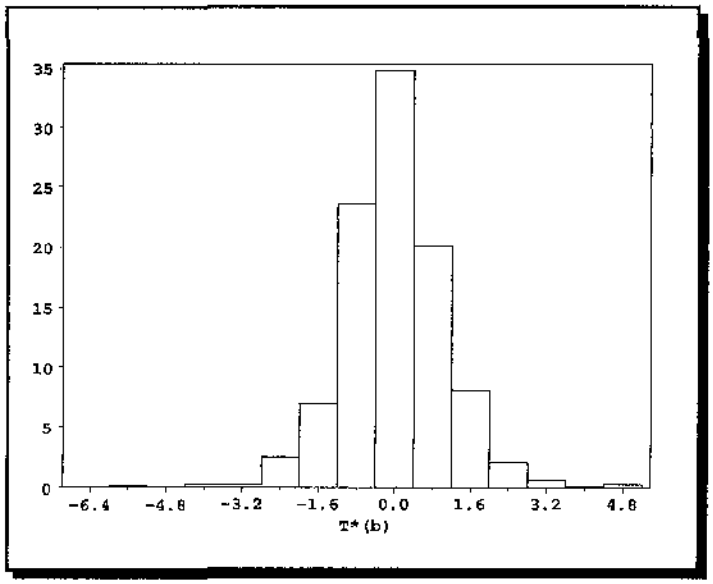

Figura 3.5. Distribuicão empírica das B = 1.000 replicações da estatística $T^*(b)$

Os percentis 0,05 e 0,95 desta distribuição empírica foram

$$\hat{t}^*_{0,05} = -1,7074 \quad e \quad \hat{t}^*_{0,95} = 1,7727. \tag{3.5.19}$$

Portanto, o intervalo t-BOOT para o parâmetro μ, usando o Bootstrap não-paramétrico, será dado por:

$$\hat{\theta}_{t\text{-}BOOT}\,[0,05] = 1,9886 - 1,7727 \,.\, 0,6147 = 0,8990 \tag{3.5.20}$$

e

$$\hat{\theta}_{t\text{-}BOOT}\,[0,95] = 1,9886 - (-1,7074) \,.\, 0,6147 = 3,0381\,. \tag{3.5.21}$$

A tabela abaixo descreve os limites do intervalo t-BOOT para os parâmetros μ e μ^3.

Tabela 3.10. *Intervalo de confiança t-BOOT para os parâmetros μ (= 2) e μ^3 (= 8) com probabilidade de cobertura de aproximadamente 0,90, para os dados da tabela 2.1 (n = 10)*

Parâmetro	BOOTSTRAP	$\hat{\theta}_{t\text{-}BOOT}$ [0.05]	$\hat{\theta}_{t\text{-}BOOT}$ [0.95]
μ	Não-paramétrico*	0,8990	3,0381
μ	Paramétrico	0,9230	3,0542
μ^3	Não-paramétrico*	-7,4169	22,5818
μ^3	Paramétrico*	-6,8697	26,2288

* Intervalos calculados por simulação de Monte Carlo com B = 1000 replicações.

Analisando a tabela 3.10, observa-se que não há muita diferença entre os os intervalos t-BOOT para este caso com respeito aos intervalos das seções anteriores. Um

fato importante é que os intervalos para o parâmetro μ^3 são tão inacurados quanto as versões não-transformadas dos intervalos BOOTPAD e t-SBOOT. Efron (1993) sugere que este intervalo seja aplicado a estatísticas de locação.

Existe um problema computacional e interpretativo com o intervalo t-BOOT. No denominador da estatística $T^*(b)$ em (3.5.14), é requerido o cálculo da variância Bootstrap de $\hat{\theta}$, para cada amostra Bootstrap $\mathbf{x}^*(b)$, ou seja, $Var_{\hat{F}}(t(x^*(b)))$. No simples caso em que $\hat{\theta} = \bar{x}$, com o uso da FDE $\hat{F}_n$,

$$Var_{\hat{F}_n}(t(x^*(b))) = \frac{\sum_{i=1}^{n}(x_{b_i}^* - \bar{x}_b^*)^2}{n^2}, \tag{3.5.22}$$

como visto em (2.2.14). Neste caso, também, poderia ser usada, a estimativa não-tendenciosa da variância, que é dada por,

$$\hat{\sigma}_{nt}^{2*}(b) = \frac{\sum_{i=1}^{n}(x_{b_i} - \bar{x}_b^*)^2}{n(n-1)}. \tag{3.5.23}$$

A dificuldade aparece quando $\hat{\theta} = t(x)$ é um estimador mais complicado do que $\bar{x}$, cuja expressão analítica para $Var_{\hat{F}}(t(x^*(b)))$ não seja fácil de deduzir. Uma saída para este problema, que não prejudica a automacidade do método, é usar a aproximação de Monte Carlo para estimar a variância Bootstrap $Var_{\hat{F}}(t(x^*(b)))$ conforme o algorítmo 1. Este procedimento implica em dois níveis embutidos de reamostragem Bootstrap, ou seja, sobre cada amostra Bootstrap $\mathbf{x}^*(b)$, são extraídas B_1 amostras Bootstrap de tamanho n $\mathbf{x}^*(b,s) = (x_{b,s,1}^*, x_{b,s,2}, \ldots, x_{b,s,n}^*)$, $s = 1, 2, \ldots, B_1$. Assim, na primeira etapa de reamostragem são obtidas B amostras Bootstrap da amostra original $\mathbf{x}$, e, na segunda, B_1 amostras Bootstrap da cada amostra da primeira etapa $\mathbf{x}^*(b)$. Com este processo, $Var_{\hat{F}}(t(x^*(b)))$ será estimada por:

$$\frac{\sum_{s=1}^{B_1} \left(t(x^*(b,s)) - t(x^*(b),.) \right)^2}{B_1 - 1}, \tag{3.5.24}$$

onde,

$$t(x^*(b),.) = \frac{\sum_{s=1}^{B_1} t(x^*(b,s))}{B_1}. \tag{3.5.25}$$

O preço que se paga com esta abordagem é o grande número de replicações Bootstrap, que é dado por $B.B_1$ replicações.

O intervalo t-Bootstrap não tem em geral as propriedades 3.5 e 3.6. Em termos de acurácia, este intervalo pode ser acurado de ordem 2 (Efron, 1993).

3.6 Intervalo percentil

A partir desta seção será discutida uma outra metodologia de construção de intervalos de confiança Bootstrap diferente das apresentadas até o momento. Como será visto, esta metodologia caracteriza-se pela formação dos intervalos de confiança utilizando-se percentis da distribuição Bootstrap da estatística $\hat{\theta}^*$.

O intervalo de confiança *percentil (PERC)* (Efron, 1981) para o parâmetro θ, com probabilidade de cobertura de aproximadamente $1 - 2\alpha$ é definido pelos percentis α e $1 - \alpha$, da distribuição Bootstrap de $\hat{\theta}^*$, isto é,

$$[\hat{\theta}_{PERC}[\alpha], \hat{\theta}_{PERC}[1-\alpha]] = [\hat{G}_{\hat{F}}^{-1}(\alpha), \hat{G}_{\hat{F}}^{-1}(1-\alpha)]. \qquad (3.6.1)$$

De acordo com (3.6.1), os limites do intervalo PERC são calculados na distribuição Bootstrap de $\hat{\theta}^*$, cuja FDA é $\hat{G}_{\hat{F}}$. Este intervalo é bastante simples de ser construído. Por exemplo, considere a construção do intervalo de confiança para o parâmetro μ, com os dados da tabela 2.1, com uma probabilidade de cobertura de aproximadamente 0,90. Utilizando-se o Bootstrap paramétrico com $\hat{F} = \hat{F}_{par}$ da expressão (2.2.29), então, a estatística Bootstrap $\bar{X}^*$ tem distribuição Bootstrap $N(\bar{x}, s^2/n)$. Portanto,

$$\hat{G}_{\hat{F}_{par}}(x) = Pr_{\hat{F}_{par}}\{\bar{X}^* \leq x\} = \Phi(\sqrt{n}(x - \bar{x})/s). \qquad (3.6.2)$$

Como a função inversa de (3.6.2) é dada por:

$$\hat{G}_{\hat{F}_{par}}^{-1}(x) = \bar{x} + \Phi^{-1}(x)\frac{s}{\sqrt{n}}, \qquad (3.6.3)$$

então, os limites do intervalo percentil serão dados, para $0 < \alpha < 0,5$, por:

$$\hat{\theta}_{PERC}[\alpha] = \hat{G}^{-1}_{\hat{F}_{par}}(\alpha) = \bar{x} + \Phi^{-1}(\alpha)\frac{s}{\sqrt{n}} = \bar{x} + z_{\alpha}\frac{s}{\sqrt{n}} \qquad (3.6.4)$$

e

$$\hat{\theta}_{PERC}[1-\alpha] = \hat{G}^{-1}_{\hat{F}_{par}}(1-\alpha) = \bar{x} - \Phi^{-1}(\alpha)\frac{s}{\sqrt{n}} = \bar{x} - z_{\alpha}\frac{s}{\sqrt{n}}, \qquad (3.6.5)$$

que correspondem aos limites do intervalo BOOTPAD paramétrico. Assim, da tabela 3.1,

$$\hat{\theta}_{PERC}[0,05] = \hat{\theta}_{BOOTPAD}[0,05] = 0,9230, \qquad (3.6.6)$$

e,

$$\hat{\theta}_{PERC}[0,95] = \hat{\theta}_{BOOTPAD}[0,95] = 3,0542. \qquad (3.6.7)$$

A igualdade deste dois intervalos, neste caso, não foi por coincidência. Será mostrado no teorema 3.1 quando ela ocorre.

Teorema 3.1 : *Se a distribuição Bootstrap de* $\hat{\theta}^*$ *é normal, com média* $\hat{\theta}$ *e variância* $\hat{\sigma}^2_{BOOT}(\hat{\theta})$, *isto é,*

$$\hat{\theta}^* \underset{\cdot}{\sim} N(\hat{\theta}, \hat{\sigma}^2_{BOOT}(\hat{\theta})), \tag{3.6.8}$$

então, o intervalo PERC é igual ao intervalo BOOTPAD.

PROVA :

$$\hat{G}_{\hat{F}}\ (\ \hat{\theta} + t\hat{\sigma}_{BOOT}(\hat{\theta})\ ) = Pr_{\hat{F}}\{\ \hat{\theta}^* \le \hat{\theta} + t\hat{\sigma}_{BOOT}(\hat{\theta})\ \} =$$

$$= Pr_{\hat{F}}\{\ \frac{\hat{\theta}^* - \hat{\theta}}{\hat{\sigma}_{BOOT}(\hat{\theta})} \le t\ \} = (\ por\ (3.6.8)\ ) = Pr\{\ Z \le t\ \} = \tag{3.6.9}$$

$$= \Phi(t),\ \forall t,$$

onde

$$Z \sim N(0,1). \tag{3.6.10}$$

Logo,

$$\hat{\theta} + t\hat{\sigma}_{BOOT}(\hat{\theta}) = \hat{G}^{-1}_{\hat{F}}\Phi(t),\ \forall t. \tag{3.6.11}$$

Para $t = z_\alpha$ e $t = -z_\alpha$, teremos que:

$$\hat{\theta} + z_\alpha\hat{\sigma}_{BOOT} = \hat{\theta}_{BOOTPAD}[\alpha] = \hat{G}^{-1}_{\hat{F}}\Phi(z_\alpha) = \hat{G}^{-1}_{\hat{F}}(\alpha) = \hat{\theta}_{PERC}[\alpha] \tag{3.6.12}$$

e

$$\hat{\theta} - z_{\alpha}\hat{\sigma}_{BOOT}(\hat{\theta}) = \hat{\theta}_{PADNOR}[1-\alpha] = \hat{G}_{\hat{F}}^{-1}\Phi(-z_{\alpha}) = \hat{G}_{\hat{F}}^{-1}\Phi(z_{1-\alpha}) =$$
$$= \hat{G}_{\hat{F}}^{-1}(1-\alpha) = \hat{\theta}_{PERC}[1-\alpha]. \quad \blacksquare \tag{3.6.13}$$

Quando a suposição (3.6.8) não é válida, então, estes dois intervalos podem diferir.

Um pequeno problema prático que pode aparecer na aplicação do intervalo percentil é o cálculo analítico da FDA $\hat{G}_{\hat{F}}$, como realizado em (3.6.2), para fornecer os percentis α e 1-α. Este cálculo poderá ser bastante díficil nas aplicações, dependendo da distribuição da estatística $\hat{\theta}^*$. Nestes casos, uma saída é usar a aproximação de Monte Carlo para esta distribuição, que é obtida, como descrito para o intervalo t-BOOT, da seguinte forma: Sejam $\mathbf{x}^*(1)$, $\mathbf{x}^*(2)$,..., $\mathbf{x}^*(B)$, B amostras Bootstrap com FDA comum $\hat{F}$. A aproximação para $\hat{G}_{\hat{F}}$ será dada por:

$$\hat{G}_{\hat{F}}^{*}(t) = \frac{\#\{\ \hat{\theta}^*(b) \leq t\ \}}{B} = \frac{\#\{\ t(\mathbf{x}^*(b)) \leq t\ \}}{B}, \tag{3.6.14}$$

de forma que o p-ésimo percentil da distribuição de $\hat{\theta}^*$ é estimado por $\hat{G}_{\hat{F}}^{*-1}(p)$, que é um valor nesta distribuição satisfazendo:

$$\frac{\#\{\ \hat{\theta}^*(b) \leq \hat{G}_{\hat{F}}^{*-1}(p)\ \}}{B} = p. \tag{3.6.15}$$

Para calcular os valores $\hat{G}_{\hat{F}}^{*-1}(\alpha)$ e $\hat{G}_{\hat{F}}^{*-1}(1-\alpha)$, pode-se proceder da seguinte forma: (i) ordena-se os valores da estatística $\hat{\theta}^*$ de cada amostra Bootstrap, obtendo-se os novos valores $\hat{\theta}^*_{(1)}$, $\hat{\theta}^*_{(2)}$,..., $\hat{\theta}^*_{(B)}$; (ii) estima-se os percentis empíricos da distribuição Bootstrap de $\hat{\theta}^*$ pelos valores $\hat{G}_{\hat{F}}^{*-1}(\alpha) = \hat{\theta}^*_{(B.\alpha)}$ e $\hat{G}_{\hat{F}}^{*-1}(1-\alpha) = \hat{\theta}^*_{(B.(1-\alpha))}$, se B.$\alpha$ é inteiro. Caso B.α não seja inteiro, pode-se usar o procedimento descrito na seção 3.5 para o intervalo t-BOOT. Para mostrar a eficácia deste procedimento, realizou-se um estudo de simulação que consistiu em calcular os limites do intervalo percentil para os dados da tabela 2.1, com este procedimento,

variando o número de replicações B desde 500 até 10000 (com incremento de 50). Os resultados encontram-se nas figuras 3.6 e 3.7.

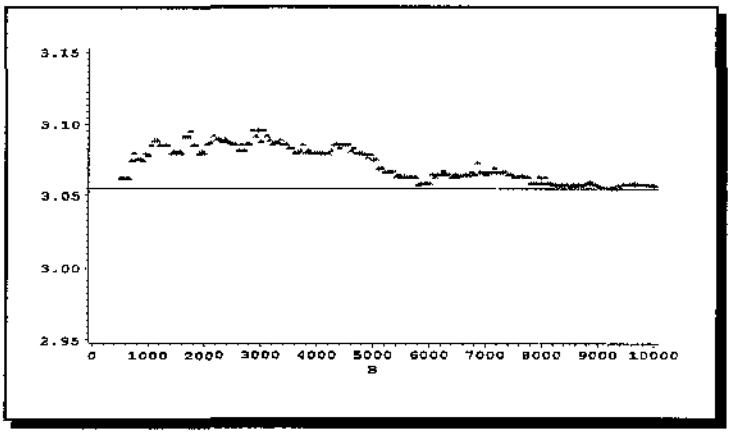

Figura 3.6. Limites superiores do intervalo percentil para o parâmetro μ, calculados por simulacão, em função do número de replicações B. Linha sólida é o verdadeiro valor do limite superior $\hat{\theta}_{PERC}$ [0,95] = 3,0542

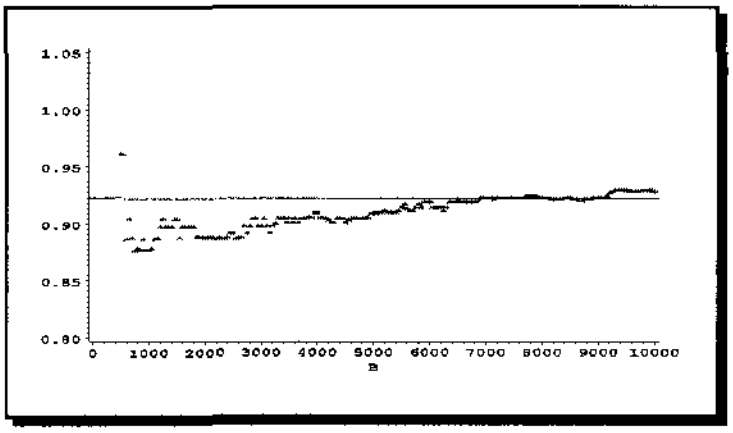

Figura 3.7. Limites inferiores do intervalo percentil para o parâmetro μ, calculados por simulacão, em função do número de replicações B. Linha sólida é o verdadeiro valor do limite inferior $\hat{\theta}_{PERC}$ [0,05] = 0,9230

Pode-se observar em ambas figuras anteriores que a convergência é excelente para um número de replicações B = 6.000. Porém, B = 1.000, que se trata apenas de uma recomendação, já produz um intervalo satisfatório, em ambos os casos, pois as diferenças são irrelevantes para fins práticos. Se precisão for necessária para os limites estimados por Monte Carlo, seja qual for a aplicação, pode-se realizar um estudo de simulação, como o anterior, variando-se o número de replicações e observando-se a estabilização dos respectivos limites estimados.

O intervalo PERC, ao contrário dos intervalos BOOTPAD, t-SBOOT e t-BOOT, possui a propriedade 3.5 (invariância a transformações monótonas), como será mostrado no teorema 3.2, e a propriedade 3.6, que trata da preservação de amplitude pelo intervalo. Quanto a propriedade de corretibilidade, o intervalo PERC é, geralmente, correto de ordem 1 (DiCiccio and Romano, 1988). A grande vantagem do intervalo PERC sobre os anteriores é nas situações em que a distribuição da estatística $\hat{\theta}$ pode ser assimétrica, o que compromete o desempenho dos intervalos BOOTPAD, t-SBOOT e t-BOOT. Se existe uma transformação monótona g(·), tal que as quantidades transformadas

$$\phi = g(\theta),\ \hat{\phi} = g(\hat{\theta})\ e\ \hat{\phi}^* = g(\hat{\theta}^*), \tag{3.6.16}$$

satisfazem

$$\tau\{\ \hat{\phi} - \phi\ \} \underset{\sim}{} W \tag{3.6.17}$$

e

$$\tau\{\ \hat{\phi}^* - \hat{\phi}\ \} \underset{*}{\sim} W, \tag{3.6.18}$$

onde τ é uma constante e W é uma variável aleatória simétricamente distribuída em torno do ponto zero, com variância constante (" $\underset{*}{\sim}$ " indica distribuição Bootstrap), então, o intervalo PERC, cujos limites podem ser calculados diretamente na distribuição Bootstrap de $\hat{\theta}^*$, coincide com o intervalo de confiança exato para o parâmetro ϕ, convertido à escala

θ pela transformação $g^{-1}(\cdot)$. Isto quer dizer que, neste caso, o intervalo PERC é também exato. Este resultado será provado no teorema 3.3. As suposições (3.6.17) e (3.6.18) significam que $\hat{\phi} - \phi$ e $\hat{\phi}^* - \hat{\phi}$ devem ter a mesma distribuição simétrica em torno da origem, e, em particular, a mesma variância. Na prática, ambas podem não ocorrer exatamente, mas aproximadamente, como é o caso do exemplo considerado até agora em que para $\theta = \mu^3$, $\hat{\theta} = \bar{x}^3$, $g(x) = x^{1/3}$ e $\tau = 1$, então

$$\hat{\phi} - \phi = \bar{X} - \mu \sim N(0, v^2/n) \tag{3.6.19}$$

e

$$\hat{\phi}^* - \hat{\phi} = \bar{X}^* - \bar{x} \underset{*}{\sim} N(0, s^2/n), \tag{3.6.20}$$

e, neste caso, a probabilidade de cobertura não pode ser mais igual a $1\text{-}2\alpha$. Porém, quanto mais próxima s^2 estiver de v^2 maior será a proximidade entre estas duas distribuições e, conseqüentemente, mais próxima estará a probabilidade de cobertura de $1\text{-}2\alpha$.

Tabela 3.11. *Intervalo de confiança PERC para os parâmetros μ (= 2) e μ^3 (= 8) com probabilidade de cobertura de aproximadamente 0,90, para os dados da tabela 2.1 (n = 10)*

Parâmetro	BOOTSTRAP	$\hat{\theta}_{PERC}$ [0.05]	$\hat{\theta}_{PERC}$ [0,95]
μ	Não-paramétrico*	1,0179	2,9873
μ	Paramétrico	0,9230	3,0542
μ^3	Não-paramétrico*	1,0546	26,6585
μ^3	Paramétrico	0,7863	28,4900

* Intervalos calculados por simulação de Monte Carlo com B = 1000 replicações.

A tabela 3.11 exibe o intervalo PERC para os parâmetros μ e μ^3, com os dados da tabela 3.1. Numericamente, pode-se observar nesta tabela que os limites dos intervalos para μ^3 são o cubo dos limites dos intervalos para μ. Outro fato importante para ser observado na tabela 3.11 é com respeito a acurácia atingida dos limites encontrados para o parâmetro μ^3 que, automaticamente, correspondem aos limites do intervalo para o parâmetro μ, elevados ao cubo, sem ter que necessariamente usar a transformação que normaliza a distribuição das estatísticas $\hat{\theta} = \bar{x}^3$ e $\hat{\theta}^* = \bar{x}^{*3}$. Este desempenho, que é uma grande vantagem do intervalo percentil com respeito aos intervalos anteriores, é devido ao resultado que será provado no teorema 3.3.

Teorema 3.2 : *O intervalo percentil para o parâmetro $\phi = g(\theta)$, com base na estatística transformada $\hat{\phi} = g(\hat{\theta})$, onde $g(\cdot)$ é uma função monótona, crescente ou decrescente, é dado por:*

$$[\hat{\phi}_{PERC}[\alpha], \hat{\phi}_{PERC}[1-\alpha]] = [g(\hat{\theta}_{PERC}[\alpha]), g(\hat{\theta}_{PERC}[1-\alpha])], \qquad (3.6.21)$$

se $g(\cdot)$ é monótona crescente, e,

$$[\hat{\phi}_{PERC}[\alpha], \hat{\phi}_{PERC}[1-\alpha]] = [g(\hat{\theta}_{PERC}[1-\alpha]), g(\hat{\theta}_{PERC}[\alpha])], \qquad (3.6.22)$$

se $g(\cdot)$ é monótona decrescente.

PROVA : Seja $\hat{\phi}^* = g(\hat{\theta}^*)$ a estatística Bootstrap transformada com FDA dada por:

$$\hat{f}(t) = Pr_*\{\hat{\phi}^* \leq t\} = Pr_{\hat{F}}\{\hat{\phi}^* \leq t\}, \qquad (3.6.23)$$

onde " * " indica que a probabilidade em (3.6.23) é calculada sob a distribuição Bootstrap de $\hat{\phi}^*$. Considere o caso em que $g(\cdot)$ é monótona crescente. Portanto,

$$\hat{G}_{\hat{F}}\,(g^{-1}(t)) = Pr_{\hat{F}}\{\,\hat{\theta}^* \le g^{-1}(t)\,\} = Pr_{\hat{F}}\{\,g(\hat{\theta}^*) \le t\,\} = \\ = Pr_*\{\,\hat{\phi}^* \le t\,\} = \hat{f}(t)\,,\ \forall\ t. \tag{3.6.24}$$

Logo,

$$\hat{G}_{\hat{F}}\,g^{-1} = \hat{f}. \tag{3.6.25}$$

O α-ésimo percentil da distribuição Bootstrap de $\hat{\phi}^*$ é dado por :

$$\hat{\phi}_{PERC}\,[\alpha] = \hat{f}^{-1}(\alpha) = (\ por\ (3.6.25)\ ) = (\hat{G}_{\hat{F}}\,g^{-1})^{-1}(\alpha) = \\ = (g^{-1})^{-1}\hat{G}_{\hat{F}}^{-1}(\alpha) = g(\hat{G}_{\hat{F}}^{-1}(\alpha)) = g(\hat{\theta}_{PERC}\,[\alpha])\,. \tag{3.6.26}$$

Analogamente,

$$\hat{\phi}_{PERC}\,[1-\alpha] = g(\hat{G}^{-1}(1-\alpha)) = g(\hat{\theta}_{PERC}\,[1-\alpha])\,. \tag{3.6.27}$$

A prova para o caso em que g(·) é monótona decrescente é feita de forma similar. ■

Teorema 3.3 : *Suponha que existe uma transformação monótona g(·), crescente ou decrescente, e uma constante τ satisfazendo (3.6.16)-(3.6.18). Então, o intervalo PERC para o parâmetro θ, coincide com o intervalo de confiança, baseado na distribuição de W (3.6.17 3.6.18), convertido à escala θ pela transformação inversa $g^{-1}(\cdot)$, isto é,*

$$[\hat{\theta}_{PERC}\,[\alpha]\,,\hat{\theta}_{PERC}\,[1-\alpha]] = [g^{-1}(\hat{\phi} + \frac{w_\alpha}{\tau}),g^{-1}(\hat{\phi} - \frac{w_\alpha}{\tau})]\,, \tag{3.6.28}$$

se g(·) é monótona crescente, e,

$$[\hat{\theta}_{PERC}[\alpha], \hat{\theta}_{PERC}[1-\alpha]] = [g^{-1}(\hat{\phi} - \frac{w_\alpha}{\tau}), g^{-1}(\hat{\phi} + \frac{w_\alpha}{\tau})], \qquad (3.6.29)$$

se g(·) é monótona decrescente. Neste caso, a probabilidade de cobertura do intervalo percentil será igual a 1-2α. O valor w_α é o α-ésimo percentil da distribuição de W, isto é, $w_\alpha = H^{-1}(\alpha)$, onde H é a FDA de W.

PROVA : Da relação

$$Pr\{ w_\alpha \leq W \leq w_{1-\alpha} \} = Pr\{ w_\alpha \leq W \leq -w_\alpha \} = 1 - 2\alpha, \qquad (3.6.30)$$

temos por (3.6.17), que

$$\begin{aligned} 1 - 2\alpha &= Pr_F\{ w_\alpha \leq \tau[g(\hat{\theta}) - g(\theta)] \leq -w_\alpha \} = \\ &= Pr_F\{ g(\hat{\theta}) + \frac{w_\alpha}{\tau} \leq g(\theta) \leq g(\hat{\theta}) - \frac{w_\alpha}{\tau} \}. \end{aligned} \qquad (3.6.31)$$

Considere o caso em que g(·) é monótona crescente. Portanto, o intervalo de confiança central para $\phi = g(\theta)$, com probabilidade de cobertura 1-2α, será dado por:

$$[\, g(\hat{\theta}) + \frac{w_\alpha}{\tau} \,,\, g(\hat{\theta}) - \frac{w_\alpha}{\tau} \,]. \qquad (3.6.32)$$

Como g(·) é uma função monótona crescente, então, da relação (3.6.31), temos que,

$$Pr_F\{ g^{-1}(g(\hat{\theta}) + \frac{w_\alpha}{\tau}) \leq \theta \leq g^{-1}(g(\hat{\theta}) - \frac{w_\alpha}{\tau}) \} = 1 - 2\alpha. \qquad (3.6.33)$$

Logo, o intervalo de confiança para o parâmetro θ, com probabilidade de cobertura 1 - 2α, convertido à escala θ pela transformação inversa $g^{-1}(\cdot)$, será dado por:

$$[\ g^{-1}(g(\hat{\theta}) + \frac{w_\alpha}{\tau})\ ,\ g^{-1}(g(\hat{\theta}) - \frac{w_\alpha}{\tau})\ ] =$$

$$= [\ g^{-1}(\hat{\phi} + \frac{w_\alpha}{\tau})\ ,\ g^{-1}(\hat{\phi} - \frac{w_\alpha}{\tau})\ ]. \qquad (3.6.34)$$

Agora, seja $\hat{f}$ a FDA da distribuição Bootstrap de $\hat{\phi}^* = g(\hat{\theta}^*)$ como em (3.6.23). Assim,

$$\hat{f}(\ \hat{\phi} + \frac{w_\alpha}{\tau}\ ) = Pr_{\hat{F}}\{\ \hat{\phi}^* \le \hat{\phi} + \frac{w_\alpha}{\tau}\ \} =$$

$$= Pr_{\hat{F}}\{\ \tau[\hat{\phi}^* - \hat{\phi}] \le w_\alpha\ \} = (\ por\ (3.6.18)\ ) = \qquad (3.6.35)$$

$$= Pr\{\ W \le w_\alpha\ \} = H(w_\alpha) = \alpha,$$

e, analogamente,

$$\hat{f}(\ \hat{\phi} - \frac{w_\alpha}{\tau}\ ) = H(-w_\alpha) = H(w_{1-\alpha}) = 1-\alpha. \qquad (3.6.36)$$

Portanto,

$$\hat{\phi} + \frac{w_\alpha}{\tau} = \hat{f}^{-1}(\alpha) = \hat{\phi}_{PERC}[\alpha] \qquad (3.6.37)$$

e

$$\hat{\phi} - \frac{w_\alpha}{\tau} = \hat{f}^{-1}(1-\alpha) = \hat{\phi}_{PERC}[1-\alpha], \qquad (3.6.38)$$

ou seja, $\hat{\phi} + \frac{w_\alpha}{\tau}$ e $\hat{\phi} - \frac{w_\alpha}{\tau}$ são os α-ésimo e $(1-\alpha)$-ésimo percentis da distribuição Bootstrap

de $\hat{\phi}^*$ que são, por definição, os limites do intervalo de confiança percentil para o parâmetro $\phi = g(\theta)$, respectivamente.

Assim, como g(·) é monótona crescente, pelo teorema 3.2, temos que:

$$[\ \hat{\phi} + \frac{w_\alpha}{\tau}\ ,\ \hat{\phi} - \frac{w_\alpha}{\tau}\ ] = [\ \hat{\phi}_{PERC}\ [\alpha]\ ,\ \hat{\phi}_{PERC}\ [1-\alpha]\ ] = \\ = [\ g(\hat{\theta}_{PERC}\ [\alpha])\ ,\ g(\hat{\theta}_{PERC}\ [1-\alpha])\ ]. \qquad (3.6.39)$$

Logo,

$$[\ \hat{\theta}_{PERC}\ [\alpha]\ ,\ \hat{\theta}_{PERC}\ [1-\alpha]\ ] = [\ g^{-1}(\hat{\phi} + \frac{h_\alpha}{\tau})\ ,\ g^{-1}(\hat{\phi} - \frac{h_\alpha}{\tau})\ ], \qquad (3.6.40)$$

que corresponde ao intervalo de (3.6.34), que tem probabilidade de cobertura igual a $1-2\alpha$. A prova para o caso em que g(·) é monótona decrescente é feita de forma análoga. ■

3.7 Intervalo percentil com correção para tendência

Foi apresentado na seção anterior o intervalo percentil que, entre boas propriedades, pode fornecer probabilidade de cobertura exata em situações de assimetria, desde que exista uma transformação monótona g(·) satisfazendo (3.6.16)-(3.6.18). Uma das características destas suposições é que as estatísticas $\hat{\phi}$ e $\hat{\phi}^*$ devam ser estimadores não-tendenciosos dos seus respectivos parâmetros ϕ e $\hat{\phi}$. Caso este fato não ocorra, então, dependendo da tendenciosidade, a probabilidade de cobertura do intervalo PERC poderá ser bastante afetada. Será apresentado a seguir um outro intervalo de confiança Bootstrap, cujos limites representam modificações nos limites do intervalo PERC para levar em conta a presença de tendências nas estimativas $\hat{\phi}$ e $\hat{\phi}^*$. Este intervalo será denominado percentil com correção para tendência (Efron, 1981, 1982).

O intervalo de confiança *percentil com correção para tendência (PCT)* para o parâmetro θ, com probabilidade de cobertura de aproximadamente $1 - 2\alpha$, é definido pelos percentis

$$\alpha_1 = \Phi(2z_0 + z_\alpha) \tag{3.7.1}$$

e

$$\alpha_2 = \Phi(2z_0 - z_\alpha), \tag{3.7.2}$$

da distribuição Bootstrap de $\hat{\theta}^*$, isto é,

$$\begin{aligned} [\ \hat{\theta}_{PCT}[\alpha]\ ,\ \hat{\theta}_{PCT}[1-\alpha]\ ] &= [\ \hat{G}_{\hat{F}}^{-1}(\alpha_1)\ ,\ \hat{G}_{\hat{F}}^{-1}(\alpha_2)\ ] = \\ &= [\ \hat{G}_{\hat{F}}^{-1}(\Phi(2z_0 + z_\alpha))\ ,\ \hat{G}_{\hat{F}}^{-1}(\Phi(2z_0 - z_\alpha))\ ]. \end{aligned} \tag{3.7.3}$$

Como se pode observar, este intervalo depende de uma constante denotada por z_0. Esta constante é conhecida como *constante de correção de tendência* e é dada por:

$$z_0 = \Phi^{-1}(\hat{G}_{\hat{F}}(\hat{\theta})). \tag{3.7.4}$$

Se a estatística Bootstrap $\hat{\theta}^*$ tem esperança Bootstrap igual a $\hat{\theta}$, então, $G_{\hat{F}}(\hat{\theta}) = Pr_{\hat{F}}\{\ \hat{\theta}^* \leq \hat{\theta}\ \} = 1/2$. Logo, $z_0 = \Phi^{-1}(1/2) = 0$ e, neste caso, os limites do intervalo PCT serão iguais aos do intervalo PERC. Por exemplo, para os dados da tabela 2.1,

$$z_0 = \Phi^{-1}(\ \hat{G}_{\hat{F}_{par}}(\bar{x})\ ) = \Phi^{-1}(\ \Phi(\ \sqrt{n}(\bar{x} - \bar{x})/s\ )\ ) = \Phi^{-1}(\Phi(0)) = 0, \tag{3.7.5}$$

Portanto, o intervalo PCT para o parâmetro μ será igual ao intervalo PERC. Também, para o parâmetro μ^3 esta igualdade também acontecerá, pois, neste caso, a constante z_0 será

dada por:

$$z_0 = \Phi^{-1}(Pr_{\hat{F}_{par}} \{ \bar{x}^{*3} \leq \bar{x}^3 \})) = \\ = \Phi^{-1}(Pr_{\hat{F}_{par}} \{ \bar{x}^{*} \leq \bar{x} \})) = \Phi^{-1}(G_{\hat{F}_{par}} (\bar{x})) = 0. \tag{3.7.6}$$

Porém, se $G_{\hat{F}}(\hat{\theta}) \neq 1/2$, então, $z_0 \neq 0$ e os limites do intervalo PCT diferiram dos limites do intervalo PERC.

Caso seja difícil obter $\hat{G}_{\hat{F}}$ analíticamente, o intervalo PCT de (3.7.3) pode ser aproximado pelo intervalo:

$$[\hat{\theta}^*_{(B\alpha_1)} , \hat{\theta}^*_{(B\alpha_2)}], \tag{3.7.7}$$

onde $\hat{\theta}^*_{(B\alpha_1)}$ e $\hat{\theta}^*_{(B\alpha_2)}$ são os $B.\alpha_1$ e $B.\alpha_2$, respectivamente, valores ordenados de B replicações Bootstrap da estatística $\hat{\theta}^*$, isto é, $\hat{\theta}^*(b)$, $b = 1,2,...,B$. Caso $B.\alpha_1$ e $B.\alpha_2$ não sejam inteiros, então pode-se adaptar o algorítmo discutido para o intervalo t-BOOT. Neste procedimento, a constante z_0 é estimada por:

$$\hat{z}_0 = \Phi^{-1}(\#\{ \hat{\theta}^*(b) \leq \hat{\theta} \} / B). \tag{3.7.8}$$

Portanto, se metade das replicações Bootstrap $\hat{\theta}^*(b)$ são menores ou iguais ao valor observado da estatísitca $\hat{\theta}$, então, $\hat{z}_0$ será igual a zero.

O intervalo PCT, apesar da correção de tendência, ainda conserva as propriedades do intervalo PERC de invariância a transformações monótonas, como será mostrado no teorema 3.4, e, preservação de amplitude. Quanto a acurácia, este intervalo é, usualmente, acuarado de ordem 2. Porém, será mostrado no teorema 3.5 que se existe uma

transformação monótona g(·), tal que as quantidades tranformadas

$$\phi = g(\theta),\ \hat{\phi} = g(\hat{\theta}) \ e\ \hat{\phi}^* = g(\hat{\theta}^*), \tag{3.7.9}$$

satisfaçam

$$\tau\{\ \hat{\phi} - \phi\ \} + z_0 \ \sim\ N(0,1) \tag{3.7.10}$$

e

$$\tau\{\ \hat{\phi}^* - \hat{\phi}\ \} + z_0 \ \underset{*}{\sim}\ N(0,1), \tag{3.7.11}$$

onde τ é uma constante, então o intervalo PCT para o parâmetro θ, cujos limites podem ser obtidos da distribuição Bootstrap de $\hat{\theta}^*$, coincide com o intervalo de confiança exato para o parâmetro ϕ, convertido à escala θ pela transformação inversa $g^{-1}(\cdot)$. Isto quer dizer que sob (3.7.9)-(3.7.11) o intervalo PCT fornece probabilidade de cobertura igual a $1\text{-}2\alpha$. Um exemplo clássico deste tipo de situação é o da transformação de Fisher do coeficiente de correlação, em que, para $g(\cdot) = \tanh^{-1}(\cdot)$,

$$\hat{\phi} \approx N(\ \phi + \frac{\rho}{2(n-1)}\ ,\ \frac{1}{n-3}\ ), \tag{3.7.12}$$

ou seja,

$$\hat{\phi} - \phi + (-\rho/(2(n-1))) \approx N(\ 0\ ,\ \frac{1}{n-3}\ ). \tag{3.7.13}$$

Este caso é um exemplo onde o intervalo PERC pode falhar.

Teorema 3.4 : *O intervalo PCT para o parâmetro $\phi = g(\theta)$, onde $g(\cdot)$ é uma função monótona crescente ou decrescente, será dado por:*

$$[\hat{\phi}_{PCT}[\alpha_1], \hat{\phi}_{PCT}[\alpha_2]] = [g(\hat{\theta}_{PCT}[\alpha_1]), g(\hat{\theta}_{PCT}[\alpha_2])], \qquad (3.7.14)$$

se $g(\cdot)$ é monótona crescente, e,

$$[\hat{\phi}_{PCT}[\alpha_1], \hat{\phi}_{PCT}[\alpha_2]] = [g(\hat{\theta}_{PCT}[\alpha_2]), g(\hat{\theta}_{PCT}[\alpha_1])], \qquad (3.7.15)$$

se $g(\cdot)$ é monótona decrescente.

PROVA : Similar à prova do teorema 3.2. ■

Teorema 3.5 : *Suponha que existe uma transformação monótona $g(\cdot)$, crescente ou decrescente, e constantes τ e z_0, satisfazendo (3.7.9)-(3.7.11). Então, o intervalo PCT para o parâmetro θ, corresponde ao intervalo de confiança baseado na distribuição normal, convertido à escala θ pela transformação inversa $g^{-1}(\cdot)$, isto é,*

$$[\hat{\theta}_{PCT}[\alpha_1], \hat{\theta}_{PCT}[\alpha_2]] = [g^{-1}(\hat{\phi} + \frac{z_0}{\tau} + \frac{z_\alpha}{\tau}), g^{-1}(\hat{\phi} + \frac{z_0}{\tau} - \frac{z_\alpha}{\tau})], \qquad (3.7.16)$$

se $g(\cdot)$ é monótona crescente, e,

$$[\hat{\theta}_{PCT}[\alpha_1], \hat{\theta}_{PCT}[\alpha_2]] = [g^{-1}(\hat{\phi} + \frac{z_0}{\tau} - \frac{z_\alpha}{\tau}), g^{-1}(\hat{\phi} + \frac{z_0}{\tau} + \frac{z_\alpha}{\tau})], \qquad (3.7.17)$$

se $g(\cdot)$ é monótona decrescente. Neste caso, a probabilidade de cobertura do intervalo PCT será igual a $1-2\alpha$.

PROVA : Da relação

$$Pr\{ z_\alpha \le Z \le z_{1-\alpha} \} = Pr\{ z_\alpha \le Z \le -z_\alpha \} = 1 - 2\alpha, \tag{3.7.18}$$

temos por (3.7.10), que

$$\begin{aligned} 1 - 2\alpha &= Pr_F \{ z_\alpha \le \tau[g(\hat\theta) - g(\theta)] + z_0 \le -z_\alpha \} = \\ &= Pr_F \{ g(\hat\theta) + \frac{z_0}{\tau} + \frac{z_\alpha}{\tau} \le g(\theta) \le g(\hat\theta) + \frac{z_0}{\tau} - \frac{z_\alpha}{\tau} \}. \end{aligned} \tag{3.7.19}$$

Considere o caso em que g(·) é monótona crescente. Portanto, o intervalo de confiança central para $\phi = g(\theta)$, com probabilidade de cobertura $1- 2\alpha$, será dado por:

$$[g(\hat\theta) + \frac{z_0}{\tau} + \frac{z_\alpha}{\tau} , g(\hat\theta) + \frac{z_0}{\tau} - \frac{z_\alpha}{\tau}]. \tag{3.7.20}$$

Como g(·) é uma função monótona crescente, então, da relação (3.7.19), temos que,

$$\begin{aligned} Pr_F \{ g^{-1}(g(\hat\theta) + \frac{z_0}{\tau} + \frac{z_\alpha}{\tau}) \le \theta \le g^{-1}(g(\hat\theta) + \frac{z_0}{\tau} - \frac{z_\alpha}{\tau}) \} = \\ = 1 - 2\alpha. \end{aligned} \tag{3.7.21}$$

Logo, o intervalo de confiança para o parâmetro θ, com probabilidade de cobertura $1 - 2\alpha$, convertido à escala θ pela transformação inversa $g^{-1}(\cdot)$, será dado por:

$$[g^{-1}(\hat\phi + \frac{z_0}{\tau} + \frac{z_\alpha}{\tau}) , g^{-1}(\hat\phi + \frac{z_0}{\tau} - \frac{z_\alpha}{\tau})]. \tag{3.7.22}$$

Agora, seja $\hat f$ a FDA da distribuição Bootstrap de $\hat\phi^* = g(\hat\theta^*)$como em (3.6.23). Assim,

$$\hat{f}\left(\hat{\phi} + \frac{z_0}{\tau} + \frac{z_\alpha}{\tau} \right) = Pr_{\hat{F}}\left\{ \hat{\phi}^* \le \hat{\phi} + \frac{z_0}{\tau} + \frac{z_\alpha}{\tau} \right\} =$$

$$= Pr_{\hat{F}}\left\{ \tau[\hat{\phi}^* - \hat{\phi}] + z_0 \le 2z_0 + z_\alpha \right\} = \tag{3.7.23}$$

$$= (\ por\ (3.7.11)\ ) = Pr\{ Z \le 2z_0 + z_\alpha \} = \Phi(2z_0 + z_\alpha) = \alpha_1,$$

e, analogamente,

$$\hat{f}\left(\hat{\phi} + \frac{z_0}{\tau} - \frac{z_\alpha}{\tau} \right) = Pr\{ Z \le 2z_0 - z_\alpha \} = \Phi(2z_0 - z_\alpha) = \alpha_2. \tag{3.7.24}$$

Portanto,

$$\hat{\phi} + \frac{z_0}{\tau} + \frac{z_\alpha}{\tau} = \hat{f}^{-1}(\Phi(2z_0 + z_\alpha)) = \hat{f}^{-1}(\alpha_1) = \hat{\phi}_{PCT}[\alpha_1] \tag{3.7.25}$$

e

$$\hat{\phi} + \frac{z_0}{\tau} - \frac{z_\alpha}{\tau} = \hat{f}^{-1}(\Phi(2z_0 - z_\alpha)) = \hat{f}^{-1}(\alpha_2) = \hat{\phi}_{PCT}[\alpha_2]. \tag{3.7.26}$$

ou seja, $\hat{\phi} + \frac{z_0}{\tau} + \frac{z_\alpha}{\tau}$ e $\hat{\phi} + \frac{z_0}{\tau} - \frac{z_\alpha}{\tau}$ correspondem aos percentis $\alpha_1 = \Phi(2z_0 + z_\alpha)$ e $\alpha_2 = \Phi(2z_0 - z_\alpha)$ da distribuição Bootstrap de $\hat{\phi}^*$ que são, por definição, os limites do intervalo de confiança PCT para o parâmetro $\phi = g(\theta)$.

Assim, como g(·) é monótona crescente, pelo teorema 3.4, temos que:

$$\left[\hat{\phi} + \frac{z_0}{\tau} + \frac{z_\alpha}{\tau} , \hat{\phi} + \frac{z_0}{\tau} - \frac{z_\alpha}{\tau} \right] = \left[\hat{\phi}_{PERC}[\alpha_1] , \hat{\phi}_{PERC}[\alpha_2] \right] =$$

$$= \left[g(\hat{\theta}_{PERC}[\alpha_1]) , g(\hat{\theta}_{PERC}[\alpha_2]) \right]. \tag{3.7.27}$$

Logo,

$$[\ \hat{\theta}_{PERC}[\alpha_1]\ ,\ \hat{\theta}_{PERC}[\alpha_2]\ ] = \\ = [\ g^{-1}(\hat{\phi} + \frac{z_0}{\tau} + \frac{z_\alpha}{\tau})\ ,\ g^{-1}(\hat{\phi} + \frac{z_0}{\tau} - \frac{z_\alpha}{\tau})\ ], \tag{3.7.28}$$

que corresponde ao intervalo de (3.7.22) que tem probabildiade de cobertura igual a 1-2α. A prova para o caso em que g(·) é monótona decrescente é feita de forma análoga. ■

Um fato importante do intervalo PCT é que o teorema 3.5 ainda vale para suposições mais gerais que (3.7.10) e (3.7.11), que são:

$$\tau\{\ \hat{\phi} - \phi\ \} + z_0 \ \sim\ W \tag{3.7.29}$$

e

$$\tau\{\ \hat{\phi}^* - \hat{\phi}\ \} + z_0 \ \underset{*}{\sim}\ W, \tag{3.7.30}$$

onde τ é uma constante e W é uma variável aleatória simetricamente distribuída sobre o ponto zero com variância constante, com as seguintes modificações:

$$\alpha_1 = H(2z_0 + w_\alpha), \tag{3.7.31}$$

$$\alpha_2 = H(2z_0 - w_\alpha) \tag{3.7.32}$$

e

$$z_0 = H^{-1}(\hat{G}_{\hat{F}}(\hat{\theta})), \tag{3.7.33}$$

onde H é a FDA da variável aleatória W e w_α é o α-ésimo percentil da distribuição de W, isto é, $w_\alpha = H^{-1}(\alpha)$.

A diferença das suposições (3.7.29) e (3.7.30), comparadas com (3.6.17) e (3.6.18), respectivamente, é a generalidade da constante z_0. Se esta constante é zero, então, elas serão iguais, assim como os intervalos PERC e PCT.

De uma forma geral, a constante z_0 mede a discrepância entre a a mediana de $\hat{\theta}^*$ e $\hat{\theta}$, em unidades da distribuição de W. A fórmula para seu cálculo, aparece das seguintes relações:

$$\begin{aligned} \hat{f}(\hat{\phi}) &= Pr_{\hat{F}}\{\ \hat{\phi}^* \leq \hat{\phi}\ \} = Pr_{\hat{F}}\{\ \tau[\hat{\phi}^* - \hat{\phi}] \leq 0\ \} = \\ &= Pr_{\hat{F}}\{\ \tau[\hat{\phi}^* - \hat{\phi}] <= 0\ \} = (\ por\ (3.7.30)\ ) = \\ &= Pr\ \{\ W \leq z_0\ \} = H(z_0). \end{aligned} \tag{3.7.34}$$

Por outro lado,

$$J(\hat{\phi}) = \hat{G}_{\hat{F}}(\hat{\theta}). \tag{3.7.35}$$

Logo,

$$H(z_0) = \hat{G}_{\hat{F}}(\hat{\theta}) \quad \Rightarrow \quad z_0 = H^{-1}(\hat{G}_{\hat{F}}(\hat{\theta})). \tag{3.7.36}$$

Para $H = \Phi$, isto é, W tem distribuição normal padrão, então

$$z_0 = \Phi^{-1}(\hat{G}_{\hat{F}}(\hat{\theta})). \tag{3.7.37}$$

3.8 Intervalo percentil com correção para tendência e aceleração

Na seção anterior foi apresentado o intervalo de confiança PCT cujos limites representam modificações dos limites do intervalo PERC para fornecer probabilidade de cobertura exata, também, nas situações em que existe uma transformação g(·) satisfazendo as mesmas suposições sobre a transformação referente ao intervalo PERC, porém, admitindo-se tendenciosidade nas estatísticas transformadas $\hat{\phi} = g(\hat{\theta})$ e $\hat{\phi}^* = g(\hat{\theta}^*)$. Foi discutido que, caso esta tendenciosidade seja nula, então, os dois intervalos são iguais. Por isso, pode-se entender o intervalo PCT como sendo mais geral que o intervalo PERC. Nesta seção, será apresentado um outro intervalo de confiança Bootstrap, cujos limites são modificações dos limites do intervalo PCT para fornecer probabilidade de cobertura exata nas situações em que existe uma transformação g(·) que satisfaz as suposições referentes a transformação do intervalo PCT, porém, admitindo-se que o desvio das estatísticas transformadas $\hat{\phi} = g(\hat{\theta})$ e $\hat{\phi}^* = g(\hat{\theta}^*)$ possa variar com seus respectivos parâmetros ϕ e $\hat{\phi}$, respectivamente, que se trata de uma situação de heterocedasticidade pela qual os intervalos PERC e PCT não garantem, em geral, probabilidade de cobertura exata. Este intervalo é denominado *intervalo percentil com correção para tendência e aceleração* (Efron, 1984 , 1987), cujos limites serão definidos a seguir.

O intervalo de confiança percentil com correção de tendência-aceleração (PCTa) para o parâmetro θ, com probabilidade de cobertura de aproximadamente $1 - 2\alpha$, é definido pelos percentis:

$$\alpha_1' = \Phi\left(z_0 + \frac{(z_0 + z_\alpha)}{1 - a(z_0 + z_\alpha)} \right) \qquad (3.8.1)$$

e

$$\alpha_2' = \Phi\left(z_0 + \frac{(z_0 - z_\alpha)}{1 - a(z_0 - z_\alpha)} \right), \tag{3.8.2}$$

da distribuição Bootstrap de $\hat{\theta}^*$, isto é,

$$[\hat{\theta}_{PCTa}[\alpha_1'] , \hat{\theta}_{PCTa}[\alpha_2']] = [\hat{G}_{\hat{F}}^{-1}(\alpha_1') , \hat{G}_{\hat{F}}^{-1}(\alpha_2')], \tag{3.8.3}$$

Como se pode observar, este intervalo depende de duas constantes a serem determinadas: z_0, que é a constante de correção de tendência discutida na seção anterior, e, a que é denominada *constante de aceleração*, cujos cálculos e interpretação serão discutidos mais adiante. Se $a = 0$, então, o intervalo PCTa será igual ao intervalo PCT, e, se $a = 0$ e $z_0 = 0$, então o intervalo PCTa será igual ao intervalo PERC. Caso seja difícil obter $\hat{G}_{\hat{F}}$ analíticamente, o intervalo PCTa de (3.8.3) poderá ser aproximado, com procedimento análogo ao da seção anterior, pelo intervalo:

$$[\hat{\theta}^*_{(B\alpha_1')} , \hat{\theta}^*_{(B\alpha_2')}]. \tag{3.8.4}$$

O intervalo PCTa conserva as propriedades de invariância a transformações monótonas, como será mostrado no teorema 3.6, e, preservação de amplitude. Uma grande propriedade deste intervalo é com respeito a sua corretibilidade. Efron (1987) mostrou em determinadas classes de modelos paramétricos que o intervalo PCTa é correto de ordem 2. Porém, será mostrado no teorema 3.7 que, se existe uma transformação monótona g(·) tal que as quantidades transformadas

$$\phi = g(\theta), \ \hat{\phi} = g(\hat{\theta}) \ e \ \hat{\phi}^* = g(\hat{\theta}^*), \tag{3.8.5}$$

satisfaçam

$$\tau\{\hat{\phi} - \phi\} + z_0 \sim N(0, \sigma_\phi^2), \quad \sigma_\phi = 1 + a\tau\phi, \tag{3.8.6}$$

e

$$\tau\{\hat{\phi}^* - \hat{\phi}\} + z_0 \sim N(0, \sigma_{\hat{\phi}}^2), \quad \sigma_{\hat{\phi}} = 1 + a\tau\hat{\phi}, \tag{3.8.7}$$

então o intervalo PCTa para o parâmetro θ, cujos limites podem ser obtidos da distribuição Bootstrap de $\hat{\theta}^*$, coincide com o intervalo de confiança exato para o parâmetro ϕ, convertido à escala θ pela transformação inversa $g^{-1}(\cdot)$. Isto quer dizer que sob (3.8.5)-(3.8.7) o intervalo PCTa fornece probabilidade de cobertura igual a $1-2\alpha$. As suposições (3.8.6) e (3.8.7) generalizam as suposições correspondentes dos intervalos PERC, visto que, permite-se o desvio padrão das estatísticas $\hat{\phi}$ e $\hat{\phi}^*$ possa variar linearmente com seus parâmetros ϕ e $\hat{\phi}$, respectivamente, a uma taxa que depende de a. Esta é uma das razões para a denominação de " constante de aceleração ". De um forma geral, esta constante *mede a taxa de mudança do desvio padrão de* $\hat{\phi}$ *e* $\hat{\phi}^*$ *sobre a escala da distribuição normal padrão.* Ao contrário da constante de correção de tendência z_0, seu cálculo com base na distribuição Bootstrap de $\hat{\theta}^*$ não é fácilmente determinado. Em famílias uni-paramétricas, onde a estatística $\hat{\theta}$ tem uma função de densidade ou função de probabilidades dada por $g_\theta(t)$, Efron (1987) sugere que uma boa aproximação para a é dada por:

$$\hat{a} = \frac{1}{6} ASSIMT(U_\theta)\big|_{\theta=\hat{\theta}} = \left\{ \frac{1}{6} \frac{E_G\,(U_\theta - E_G U_\theta)^3}{[E_G\,(U_\theta - E_G U_\theta)^2]^{3/2}} \right\}\Bigg|_{\theta=\hat{\theta}}, \tag{3.8.8}$$

onde, $ASSIMT(X)|_{\theta=\hat{\theta}}$ é o *coeficiente de assimetria* da distribuição de X, calculado no valor do parâmetro $\theta = \hat{\theta}$, U_θ é a *função score*, isto é,

$$U_\theta(t) = \frac{\partial}{\partial\theta}\log g_\theta(t). \qquad (3.8.9)$$

Em famílias multi-paramétricas, caso em que as variáveis aleatórias independentes X_1, X_2,..., X_n são identicamentes distribuídas com uma função de densidades dada por g_η, onde,

$$\eta = (\eta_1, \eta_2, ..., \eta_p)', \qquad (3.8.10)$$

o parâmetro e a estatística de interesse são dados por:

$$\theta = t(\eta) \qquad (3.8.11)$$

e

$$\hat{\theta} = t(\hat{\eta}), \qquad (3.8.12)$$

onde

$$\hat{\eta} = (\hat{\eta}_1, \hat{\eta}_2, ..., \hat{\eta}_p)', \qquad (3.8.13)$$

é o estimador de maxima verossimilhança de η. A distribuição Bootstrap (paramétrica) da estatística $\hat{\theta}$ é dada por $g_{\hat{\eta}}$. Seja $\ddot{l}$ uma matriz *pxp* cujo i-ésimo e j-ésimo elemento é dado por $(\partial^2/\partial\eta_i\partial\eta_j)\log g_\eta(x)|_{\eta=\hat{\eta}}$ e $\hat{\nabla} = (\hat{\nabla}_1, \hat{\nabla}_2, ..., \hat{\nabla}_n)$, onde $\hat{\nabla}_i = (\partial t(\eta)/\partial\eta_i)|_{\eta=\hat{\eta}}$. Então, uma aproximação para a constante a (Efron, 1987) é dada por:

$$\hat{a} = \frac{1}{6}ASSIMT\ (\dot{l}(X))\ |_{\lambda=0}, \qquad (3.8.14)$$

onde

$$\dot{l}(x) = (\partial/\partial\lambda)\log g_{\delta}(x) \tag{3.8.15}$$

e

$$\hat{\delta} = \hat{\eta} - \lambda(\ddot{l}(X))^{-1}\hat{\nabla}. \tag{3.8.16}$$

Em modelos não-paramétricos, uma expressão simples para a, é dada por:

$$\hat{a} = \frac{\sum_{i=1}^{n}(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)})^3}{6\{\sum_{i=1}^{n}(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)})^2\}^{3/2}}, \tag{3.8.17}$$

onde $\hat{\theta}_{(i)}$ é o i-ésimo valor jackknife, isto é,

$$\hat{\theta}_{(i)} = t(x_{(i)}) = t(x_1,\ldots, x_{i-1}, x_{i+1},\ldots, x_n), \quad \forall \quad i = 1, 2,\ldots, n, \tag{3.8.18}$$

e

$$\hat{\theta}_{(\cdot)} = \frac{\sum_{i=1}^{n}\hat{\theta}_{(i)}}{n}. \tag{3.8.19}$$

Observa-se, portanto, que a constante de aceleração a é essencialmente uma estimativa de assimetria. Segundo Hall (1992), esta constante é um ingrediente chave para que o intervalo de confiança seja construído com correção da assimetria da distribuição da estatística de interesse.

Teorema 3.6 : *O intervalo PCTa para o parâmetro $\phi = g(\theta)$, onde $g(\cdot)$ é uma função monótona, crescente ou decrescente, será dado por:*

$$[\hat{\phi}_{PCTa}[\alpha_1'], \hat{\phi}_{PCTa}[\alpha_2']] = [g(\hat{\theta}_{PCTa}[\alpha_1']), g(\hat{\theta}_{PCTa}[\alpha_2'])], \tag{3.8.20}$$

se $g(\cdot)$ é monótona crescente e,

$$[\hat{\phi}_{PCTa}[\alpha_1'], \hat{\phi}_{PCTa}[\alpha_2']] = [g(\hat{\theta}_{PCTa}[\alpha_2']), g(\hat{\theta}_{PCTa}[\alpha_1'])], \tag{3.8.21}$$

se $g(\cdot)$ é monótona decrescente.

PROVA : Similar à prova do teorema 3.2. ■

Teorema 3.7 : *Suponha que existe uma transformação monótona $g(\cdot)$, crescente ou decrescente, e constantes τ e z_0 e a, satisfazendo (3.8.5)-(3.8.7). Então, o intervalo PCTa para o parâmetro θ, corresponde ao intervalo de confiança para ϕ, baseado na distribuição normal, convertido à escala θ pela transformação inversa $g^{-1}(\cdot)$, isto é,*

$$[\hat{\theta}_{PCTa}[\alpha_1'], \hat{\theta}_{PCTa}[\alpha_2']] =$$

$$= \left[g^{-1}\left(\hat{\phi} + \frac{(z_0 + z_\alpha)(1 + a\tau\hat{\phi})}{\tau(1 - a(z_0 + z_\alpha))}\right), g^{-1}\left(\hat{\phi} + \frac{(z_0 - z_\alpha)(1 + a\tau\hat{\phi})}{\tau(1 - a(z_0 - z_\alpha))}\right)\right], \tag{3.8.22}$$

se $g(\cdot)$ é monótona crescente, e,

$$[\hat{\theta}_{PCTa}[\alpha_1'], \hat{\theta}_{PCTa}[\alpha_2']] =$$

$$= \left[g^{-1}\left(\hat{\phi} + \frac{(z_0 - z_\alpha)(1 + a\tau\hat{\phi})}{\tau\{ 1 - a(z_0 - z_\alpha) \}} \right), g^{-1}\left(\hat{\phi} + \frac{(z_0 + z_\alpha)(1 + a\tau\hat{\phi})}{\tau\{ 1 - a(z_0 + z_\alpha) \}} \right) \right], \tag{3.8.23}$$

se g(·) é monótona decrescente. Neste caso, a probabilidade de cobertura do intervalo PCTa será igual a 1-2α.

PROVA : Da relação

$$Pr\{ z_\alpha \le Z \le z_{1-\alpha} \} = Pr\{ z_\alpha \le Z \le -z_\alpha \} = 1 - 2\alpha, \tag{3.8.24}$$

temos, por (3.8.6), que

$$1 - 2\alpha = Pr_F \{ z_\alpha \le \tau[g(\hat{\theta}) - g(\theta)] + z_0 \le -z_\alpha \} =$$

$$= Pr_F \left\{ z_\alpha \le \tau\left\{ \frac{\hat{\phi} - \phi}{1 + a\tau\phi} \right\} + z_0 \le -z_\alpha \right\} = \tag{3.8.25}$$

$$= Pr_F \left\{ \hat{\phi} + \frac{(z_0 + z_\alpha)(1 + \tau\hat{\phi})}{\tau\{ 1 - a(z_0 + z_\alpha) \}} \le \phi \le \hat{\phi} + \frac{(z_0 - z_\alpha)(1 + \tau\hat{\phi})}{\tau\{ 1 - a(z_0 - z_\alpha) \}} \right\}.$$

Consideremos o caso em que g(·) é monótona crescente. Portanto, o intervalo de confiança central para $\phi = g(\theta)$, com probabilidade de cobertura 1- 2α, será dado por:

$$\left[\hat{\phi} + \frac{(z_0 + z_\alpha)(1 + \tau\hat{\phi})}{\tau\{ 1 - a(z_0 + z_\alpha) \}} , \hat{\phi} + \frac{(z_0 - z_\alpha)(1 + \tau\hat{\phi})}{\tau\{ 1 - a(z_0 - z_\alpha) \}} \right]. \qquad (3.8.26)$$

Como g(·) é uma função monótona crescente, então, da relação (3.8.25), temos que:

$$Pr_F\left\{ g^{-1}\left(\hat{\phi} + \frac{(z_0 + z_\alpha)(1 + \tau\hat{\phi})}{\tau\{ 1 - a(z_0 + z_\alpha) \}} \right) \le \theta \le g^{-1}\left(\hat{\phi} + \frac{(z_0 - z_\alpha)(1 + \tau\hat{\phi})}{\tau\{ 1 - a(z_0 - z_\alpha) \}} \right) \right\} = \qquad (3.8.27)$$

$$= 1 - 2\alpha$$

Logo, o intervalo de confiança para o parâmetro θ, com probabilidade de cobertura $1 - 2\alpha$, convertido à escala θ pela transformação inversa $g^{-1}(\cdot)$, será dado por:

$$\left[g^{-1}\left(\hat{\phi} + \frac{(z_0 + z_\alpha)(1 + \tau\hat{\phi})}{\tau\{ 1 - a(z_0 + z_\alpha) \}} \right) , g^{-1}\left(\hat{\phi} + \frac{(z_0 - z_\alpha)(1 + \tau\hat{\phi})}{\tau\{ 1 - a(z_0 - z_\alpha) \}} \right) \right]. \qquad (3.8.28)$$

Agora, seja $\hat{f}$ a FDA da distribuição Bootstrap de $\hat{\phi}^* = g(\hat{\theta}^*)$ como em (3.6.23). Assim,

$$\hat{f}\left(\hat{\phi} + \frac{(z_0 + z_\alpha)(1 + a\tau\hat{\phi})}{\tau\{ 1-a(z_0 + z_\alpha) \}} \right) = Pr_{\hat{F}}\left\{ \hat{\phi}^* \leq \hat{\phi} + \frac{(z_0 + z_\alpha)(1 + a\tau\hat{\phi})}{\tau\{ 1 - a(z_0 + z_\alpha) \}} \right\} =$$

$$= Pr_{\hat{F}}\left\{ \tau\left(\frac{\hat{\phi} - \phi}{1 + a(z_0 + z_\alpha)} \leq z_0 + \frac{(z_0 + z_\alpha)}{1 - a(z_0 + z_\alpha)} \right) = \right. \tag{3.8.29}$$

$$= (\ \textit{por}\ (3.43)\ ) = \Phi\left(z_0 + \frac{(z_0 + z_\alpha)}{1 - a(z_0 + z_\alpha)} \right) = \alpha_1'.$$

e, analogamente,

$$\hat{f}\left(\hat{\phi} + \frac{(z_0 - z_\alpha)(1 + a\tau\hat{\phi})}{\tau\{ 1-a(z_0 - z_\alpha) \}} \right) = \Phi\left(z_0 + \frac{(z_0 - z_\alpha)}{1 - a(z_0 - z_\alpha)} \right) = \alpha_2'. \tag{3.8.30}$$

Portanto,

$$\hat{\phi} + \frac{(z_0 + z_\alpha)(1 + a\tau\hat{\phi})}{\tau\{ 1 - a(z_0 + z_\alpha) \}} = \hat{f}^{-1}(\alpha_1') \tag{3.8.31}$$

e

$$\hat{\phi} + \frac{(z_0 - z_\alpha)(1 + a\tau\hat{\phi})}{\tau\{ 1 - a(z_0 - z_\alpha) \}} = \hat{f}^{-1}(\alpha_2'), \tag{3.8.32}$$

ou seja, $\hat{\phi} + \frac{(z_0 + z_\alpha)(1 + a\tau\hat{\phi})}{\tau\{ 1 - a(z_0 + z_\alpha) \}}$ e $\hat{\phi} + \frac{(z_0 - z_\alpha)(1 + a\tau\hat{\phi})}{\tau\{ 1 - a(z_0 - z_\alpha) \}}$ correspondem aos percentis α_1' e α_2' da distribuição Bootstrap de $\hat{\phi}^*$ que são, por definição, os limites do intervalo de confiança PCT para o parâmetro $\phi = g(\theta)$.

Assim, como $g(\cdot)$ é monótona crescente, pelo teorema 3.6, temos que :

$$\left[\hat{\phi} + \frac{(z_0 + z_\alpha)(1 + \tau\hat{\phi})}{\tau\{ 1 - a(z_0 + z_\alpha) \}} , \hat{\phi} + \frac{(z_0 - z_\alpha)(1 + \tau\hat{\phi})}{\tau\{ 1 - a(z_0 - z_\alpha) \}} \right] = \tag{3.8.33}$$

$$= [\hat{\phi}_{PCTa}[\alpha_1'] , \hat{\phi}_{PCTa}[\alpha_2']] = [g(\hat{\theta}_{PCTa}[\alpha_1']) , g(\hat{\theta}_{PCTa}[\alpha_2'])].$$

Logo,

$$[\hat{\theta}_{PCTa}[\alpha_1'] , \hat{\theta}_{PCTa}[\alpha_2']] =$$

$$= \left[g^{-1}\left(\hat{\phi} + \frac{(z_0 + z_\alpha)(1 + \tau\hat{\phi})}{\tau\{ 1 - a(z_0 + z_\alpha) \}} \right) , g^{-1}\left(\hat{\phi} + \frac{(z_0 - z_\alpha)(1 + \tau\hat{\phi})}{\tau\{ 1 - a(z_0 - z_\alpha) \}} \right) \right], \tag{3.8.34}$$

que corresponde ao intervalo de (3.8.28) que tem probabilidade de cobertura $1\text{-}2\alpha$. A prova para o caso em que $g(\cdot)$ é monótona decrescente é feita de forma similar. ■

Um fato importante do intervalo PCTa é que o teorema 3.7 vale para suposições mais gerais que (3.8.6) e (3.8.7), que são as seguintes:

$$\tau\left\{ \frac{\hat{\phi} - \phi}{\sigma_{\phi}} \right\} + z_0 \sim W, \quad \sigma_{\phi} = 1 + a\tau\phi, \tag{3.8.35}$$

e

$$\tau\left\{ \frac{\hat{\phi}^* - \hat{\phi}}{\sigma_{\hat{\phi}}} \right\} + z_0 \sim W, \quad \sigma_{\hat{\phi}} = 1 + a\tau\hat{\phi}, \tag{3.8.36}$$

onde τ é uma constante e W é uma variável aleatória simetricamente distribuída sobre o ponto zero, com variância constante e FDA H. Assim, o intervalo PCTa com as seguintes modificações

$$[\ \hat{\theta}_{PCTa}\ [\alpha_1']\ ,\ \hat{\theta}_{PCTa}\ [\alpha_2']\ ] = [\ \hat{G}_{\hat{F}}^{-1}(\alpha_1')\ ,\ \hat{G}_{\hat{F}}^{-1}(\alpha_2')\ ], \tag{3.8.37}$$

onde

$$\alpha_1' = H\left(z_0 + \frac{z_0 + w_{\alpha}}{1 - a(z_0 + w_{\alpha})} \right), \tag{3.8.38}$$

$$\alpha_2' = H\left(z_0 + \frac{z_0 - w_{\alpha}}{1 - a(z_0 - w_{\alpha})} \right), \tag{3.8.39}$$

$$w_\alpha = H^{-1}(\alpha) \tag{3.8.40}$$

e

$$z_0 = H^{-1}(\hat{G}_t(\hat{\theta})), \tag{3.8.41}$$

fornece probabilidade de cobertura igual a $1\text{-}2\alpha$.

3.9 Discussões

Para finalizar este capítulo, será descrita nesta seção um breve resumo dos intervalos de confiança Bootstrap apresentados nas seções 3.3-3.8, segundo as propriedades da seção 3.2 de acurácia, corretibilidade, invariância a transformações monótonas e preservação de amplitude.

• Os intervalos de confiança BOOTPAD e t-SBOOT tem a grande vantagem de serem facilmente construídos, dada uma estimativa Bootstrap do erro padrão da estatística $\hat{\theta}$. Porém, estes intervalos só devem ser aplicados a situações em que as aproximações (3.3.2) e (3.4.1) sejam válidas, visto que características como assimetria e tendenciosidade podem comprometer o desempenho de ambos. Estes intervalos não possuem, em geral, as propriedades de invariância a transformações monótonas e, também, são apenas corretos de ordem 1 e, portanto, acurados de ordem 1.

• O intervalo t-BOOT tem duas grandes vantagens sobre os intervalos BOOTPAD e t-SBOOT : (i) evita aproximar a distribuição da estatística T pela distribuição normal padrão ou, t-Student com *n-1* g.l., etc., utilizando-se a aproximação Bootstrap que pode

ser obtida usando a informação dos dados observados **x** (ii) são acurados de ordem 2. Porém, este intervalo tem a desvantagem de não possuir, em geral, as propriedades de invariância a transformações monótonas e preservação de amplitude. Efron (1993) sugere a aplicação deste intervalo a estatísticas de locação.

• Os intervalos PERC, PCT e PCTa possuem, sempre, as propriedades de invariância a transformações monótonas e preservação de amplitude. Estes são, sem dúvida, dois pontos a favor destes intervalos com respeito aos intervalos BOOTPAD, t-SBOOT e t-BOOT. Porém, em termos de acurácia, os intervalos PERC e PCT, que são acurados de ordem 1, perdem para o intervalo t-BOOT. O intervalo PCTa foi mostrado em algumas famílias paramétricas ser correto de ordem 2 e, portanto, acurado de ordem 2. Este intervalo generaliza os intervalos PERC e PCT por levar em conta correção de tendenciosidade e assimetria, portanto, é recomendado para situações mais gerais. Se a constante a estimada por alguma das fórmulas apresentadas na seção anterior for próxima de zero, então os limites do intervalo PCTa não devem diferir muito dos limites do intervalo PCT que, por sua vez, se a a constante z_0 for próxima de zero , então seus limites estarão próximos dos limites do intervalo PERC.

Capítulo 4

O Método Bootstrap em Regressão de Mínimos Quadrados

4.1 Introdução

Nos capítulos 2 e 3 foram desenvolvidos aspectos básicos da metodologia Bootstrap, para o estudo de propriedades da distribuição de uma variável aleatória, com base em procedimentos de estimação pontual e por intervalo para uma estrutura de dados $P \rightarrow \mathbf{x}$, como definida na seção 2.4. Será mostrada neste capítulo a aplicação destes métodos para uma estrutura um pouco mais complicada definida em análise de regressão linear múltipla, como será visto a seguir.

A regressão linear múltipla é uma importante área da estatística de grande utilização em problemas cujo interesse é estudar relações entre uma variável resposta Y e um conjunto de variáveis preditoras $X_1, X_2, \ldots, X_p$, com base em n observações de Y,

$$\boldsymbol{y} = (y_1, y_2, \ldots, y_n)', \tag{4.1.1}$$

a partir de n valores *fixados* das variáveis preditoras,

$$\boldsymbol{c}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})', \quad i = 1, 2, \ldots, n. \tag{4.1.2}$$

Assim, o conjunto de dados necessário para uma análise de regressão pode ser representado por $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$, onde

$$\boldsymbol{x}_i = (\boldsymbol{y}_i , \boldsymbol{c}_i'), \quad i = 1, 2, \ldots, n. \tag{4.1.3}$$

O modelo de regressão linear múltipla para estes dados é especificado, em termos matriciais, por :

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{4.1.4}$$

onde,

$$X = \begin{pmatrix} 1 & X_{11} & \ldots & X_{1p} \\ 1 & X_{21} & \ldots & X_{2p} \\ \vdots & \vdots & \ldots & \vdots \\ 1 & X_{n1} & \ldots & X_{np} \end{pmatrix} = \begin{pmatrix} 1 & \boldsymbol{c}_1' \\ 1 & \boldsymbol{c}_2' \\ \vdots & \vdots \\ 1 & \boldsymbol{c}_n' \end{pmatrix}, \tag{4.1.5}$$

é uma matriz (fixa) $nx(p+1)$, de posto igual a $(p+1)$,

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \tag{4.1.6}$$

é um vetor $(p+1)x1$ de parâmetros desconhecidos (β_0 é denominado *parâmetro de intercepto*) , e,

$$\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}, \tag{4.1.7}$$

é um vetor *nx1* de erros aleatórios não-observáveis. Faremos as seguintes suposições sobre este vetor de erros aleatórios :

- *Suas componentes são variáveis aleatórias não observáveis independentes com FDA comum F;* (4.1.8)
- $E(\epsilon) = \mathbf{0}$; (4.1.9)
- $\Sigma(\epsilon) = Var(\epsilon) = \sigma^2 I_n$, $(\sigma^2 > 0)$. (4.1.10)

onde $E(\epsilon)$ e $\Sigma(\epsilon)$ denotam a esperança e a matriz de variância-covariância do vetor ϵ, respectivamente.

Outra representação que será útil para o modelo (4.1.4) e as suposições (4.1.8)-(4.1.10) é a seguinte:

$$y_i = \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j + \epsilon_i, \qquad i = 1, 2, ..., n, \tag{4.1.11}$$

onde,

- $\epsilon_1, \epsilon_2, ..., \epsilon_n \underset{iid}{\sim} F$; (4.1.12)
- $E_F(\epsilon_i) = 0, \quad i = 1, 2, ..., n$; (4.1.12)
- $Var_F(\epsilon_i) = \sigma^2, \quad i = 1, 2, ..., n$ (4.1.12)

A estrutura probabilística para o modelo de regressão linear múltipla, que gera os dados **x**, pode ser identificada como tendo duas componentes desconhecidas, F e β, ou seja,

$$P = (F, \beta) \rightarrow x = ((y_1, c_1'), (y_2, c_2'), \ldots, (y_n, c_n')), \tag{4.1.15}$$

de acordo com as seguinte regras:

$$F \rightarrow \epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n), \tag{4.1.16}$$

segundo a suposição de (4.1.12), e

$$y_i = (1 \quad c_i')\beta + \epsilon_i, \qquad i = 1, 2, \ldots, n, \tag{4.1.17}$$

por (4.1.11).

O estudo das relações entre a variável Y e as preditoras X_1, X_2,..., X_n é feito com base em inferências sobre o vetor de parâmetros β, face os dados observados **x**, a partir de estimação pontual, intervalos de confiança e testes de significância sobre β ou funções lineares de β,

$$\lambda'\beta \quad (\lambda \in \mathbb{R}^{p+1}). \tag{4.1.18}$$

Um estimador comumente usado do vetor β é obtido pelo método de mínimos quadrados ordinários (MQO),

$$\hat{\beta} = (X'X)^{-1}X'y, \tag{4.1.19}$$

que minimiza a expressão

$$Q(\beta) = (\, y - X\beta \,)'(\, y - X\beta \,), \tag{4.1.20}$$

isto é,

$$\min_{\beta \in \mathbb{R}^p} Q(\beta) = Q(\hat{\beta}). \tag{4.1.21}$$

Sob as suposições (4.1.12)-(4.1.14), $\hat{\beta}$ é o *melhor estimador linear não-tendencioso (BLUE)* de β, com matriz de variância-covariância dada por :

$$\Sigma_p(\hat{\beta}) = Var_p(\hat{\beta}) = E_p(\, \hat{\beta} - \beta \,)(\, \hat{\beta} - \beta \,)' = \sigma^2(X'X)^{-1}, \tag{4.1.22}$$

que pode ser estimada não-tendenciosamente por :

$$\hat{\Sigma}(\hat{\beta}) = s^2(X'X)^{-1}, \tag{4.1.23}$$

onde

$$s^2 = \frac{\sum_{i=1}^{n} (\, y_i - (1 \;\; c_i')\hat{\beta} \,)^2}{(\, n - p - 1 \,)}. \tag{4.1.24}$$

Portanto, se o interesse consiste em apenas avaliar uma estimativa do desvio padrão de cada componente do vetor β então, toma-se, simplesmente, a raiz quadrada dos elementos diagonais da matriz da expressão (4.1.23). Porém, em aplicações práticas, deseja-se a partir do conhecimento da variabilidade do estimador $\hat{\beta}$, fazer inferências através de procedimentos como intervalos de confiança e testes de significância. Estes procedimentos estatísticos são realizados na análise de regressão clássica supondo-se que os erros aleatórios são normalmente distribuídos, ou seja, que F é a função de distribuição acumulada da

distribuição normal com média zero e variância constante σ^2. Sendo assim,

$$\hat{\beta} \sim N_p(\beta, \sigma^2(X'X)^{-1}), \tag{4.1.25}$$

e substituindo-se σ^2 por s^2 da expressão (4.1.24), constrói-se facilmente pelos métodos tradicionais intervalos de confiança e testes de significância. Mesmo quando não se conhece nada sobre a distribuição dos erros, se

$$\frac{1}{n}(X'X) \underset{n\to\infty}{\to} V \tag{4.1.26}$$

e

os elementos de X são uniformemente pequenos comparados com $\sqrt{n}$, (4.1.27)

então o estimador de mínimos quadrados ordinários tem a propriedade assintótica (Freedman, 1981),

$$\sqrt{n}(\hat{\beta} - \beta) \approx N_p(0, \sigma^2 V^{-1}). \tag{4.1.28}$$

Assim, a aproximação de (4.1.28) possibilita que as inferências desejadas sejam válidas pelo menos para n suficientemente grande. O problema é saber, em termos práticos, quão grande n deva ser.

O objetivo geral deste capítulo é mostrar a metodologia Bootstrap como método alternativo para responder questões de interesse em análise de regressão linear múltipla, quando o vetor de parâmetros é estimado por mínimos quadrados ordinários, sem fazer o uso da suposição de normalidade dos erros aleatórios, ou a aproximação assintótica de (4.1.28). Na seções 4.2 e 4.3, serão discutidos dois métodos básicos de reamostragem Bootstrap para formação da amostra Bootstrap. O primeiro é o método de reamostragem dos resíduos que consiste em formar a amostra Bootstrap reproduzindo-se a estrutura do

modelo de regressão definida em (4.1.15)-(4.1.17), usando a informação dos resíduos do ajuste de MQO do modelo (4.1.11). O segundo é o método de reamostragem das observações que caracteriza-se por reproduzir a estrutura estocástica de um modelo de correlação, mas que pode ser usado em problemas de regressão em uma estrutura de erros heterocedástica. Nestas seções, serão discutidas a formação da amostra Bootstrap, suas propriedades, estimação da matriz de variância e covariância de $\hat{\boldsymbol{\beta}}$ e resultados assintóticos. Na seção 4.4, será mostrado como adaptar a metodologia apresentada nas seções 3.3 a 3.8, para construção de intervalos de confiança para o vetor $\boldsymbol{\beta}$, ou funcões lineares deste vetor e, na seção 4.5, para a realização de testes de significância.

4.2 O método Bootstrap com reamostragem de resíduos

4.2.1 Reamostragem dos resíduos

Em (4.1.15)-(4.1.17) foi definido o mecanismo probabilístico P gerador dos dados para a análise de regressão linear múltipla, cujas componentes desconhecidas eram, F e $\boldsymbol{\beta}$. Portanto, para formar uma amostra Bootstrap reproduzindo-se este mecanismo, conforme descrito na seção 2.4, é necessário, primeiramente, fornecer um estimador de P, que será denotado por:

$$\hat{P} = (\hat{F}, \hat{\boldsymbol{\beta}}). \tag{4.2.1}$$

A escolha a ser utilizada para $\hat{\boldsymbol{\beta}}$, aqui neste capítulo, será o estimador de mínimos quadrados ordinários definido em (4.1.19). Sejam

$$\hat{e}_i = y_i - (1 \; \boldsymbol{c}_i')\hat{\boldsymbol{\beta}}, \quad i = 1, 2, \ldots, n, \tag{4.2.2}$$

os resíduos resultantes do ajuste de mínimos quadrados ordinários ao modelo (4.1.11). O método de reamostragem dos resíduos ajustados (Efron, 1979), consiste em formar a amostra Bootstrap utilizando-se a informação destes resíduos para construir uma amostra Bootstrap de pseudos erros aleatórios e, depois, com estes erros estimados, construir os pseudos dados Bootstrap $x^* = (x_1^*, x_2^*,..., x_n^*)$. Este mecanismo pode ser implementado utilizando-se o processo Bootstrap não-paramétrico realizando-se, simplesmente, a reamostragem da funcão de distribuição empírica que associa massa de probabilidade 1/n sobre cada resíduo $\hat{\epsilon}_i$, ou seja,

$$\hat{F}_n(x) = \frac{\#\{ \hat{\epsilon}_i <= x \}}{n}. \tag{4.2.3}$$

Assim, a componente F será estimada por $\hat{F}_n$, e o passo $x \rightarrow \hat{P}$ produzirá o estimador do mecanismo probabilístico P,

$$\hat{P} = (\hat{F}_n, \hat{\beta}). \tag{4.2.4}$$

O passo

$$\hat{P} = (\hat{F}_n, \hat{\beta}) \rightarrow x^* = ((y_1^*, c_1'), (y_2^*, c_2'),..., (y_n^*, c_n')), \tag{4.2.5}$$

é realizado analogamente às regras (4.1.16) e (4.1.17), isto é,

$$\hat{F}_n \rightarrow \epsilon^* = (\epsilon_1^*, \epsilon_2^*,..., \epsilon_n^*) \tag{4.2.6}$$

e

$$y_i^* = (1 \quad c_i')\hat{\beta} + \epsilon_i^*, \qquad i = 1, 2,..., n, \tag{4.2.7}$$

onde (4.2.6) quer dizer que (ϵ_1^*, ϵ_2^*,..., ϵ_n^*) representa uma amostra de n observações independentes da população descrita por $\hat{F}_n$, que pode ser obtida, como já mencionado na seção 2.2, como uma amostra aletória simples extraída com reposição do conjunto de valores

$$A = \{ \hat{\epsilon}_1, \hat{\epsilon}_2,..., \hat{\epsilon}_n \}. \tag{4.2.8}$$

Logo, por construção,

$$Pr_{\hat{F}_n}\{ \epsilon_i^* = \hat{\epsilon}_j \} = \frac{1}{n}, \quad i,j = 1,2,...,n. \tag{4.2.9}$$

Os pseudos erros aleatórios Bootstrap (ϵ_1^*, ϵ_2^*,..., ϵ_n^*) são, condicionalmente aos dados observados **x**, variáveis aleatórias independentes (observáveis), isto é,

$$\epsilon_1^*/x, \ \epsilon_2^*/x, \ \epsilon_n^*/x \underset{iid}{\sim} \hat{F}_n, \tag{4.2.10}$$

tem esperança zero, na distribuição Bootstrap, já que o termo de intercepto está presente no modelo, isto é,

$$E_{\hat{F}_n}(\epsilon_i^*/x) = \int_A z d\hat{F}_n(z) = \frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i = 0, \tag{4.2.11}$$

e variância

$$\hat{\sigma}^2_{BOOT}(\epsilon_i^*) = Var_{\hat{F}_n}(\epsilon_i^*/x) = E_{\hat{F}_n}(\epsilon_i^{*2}/x) = \int_A z^2 d\hat{F}_n(z) = \frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i^2, \tag{4.2.12}$$

que é a estimativa Bootstrap de σ^2. É importante observar que

$$\hat{\sigma}^2_{BOOT}(\epsilon_i^*) = \frac{(n-p-1)}{n} s^2, \qquad (4.2.13)$$

que é igual à estimativa não-tendenciosa de σ^2, a menos do fator $(n-p-1)/n$. Da expressão (4.2.13), temos que:

$$E_F(\hat{\sigma}^2_{BOOT}(\epsilon_i^*)) = E_F\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i^2\right) = \frac{(n-p-1)}{n}\sigma^2 = \sigma^2 - \frac{(p+1)}{n}\sigma^2. \qquad (4.2.14)$$

Logo, a estimativa Bootstrap $\hat{\sigma}^2_{BOOT}(\epsilon_i^*)$ tem tendenciosidade negativa dada por $-((p+1)/n)\sigma^2$, portanto subestima a variância σ^2. Mas, para p fixo, e n crescendo para o infinito, esta tendenciosidade converge para zero. Existe uma forma de corrigir esta tendenciosidade que é pré-multiplicando, antes da reamostragem, cada resíduo em A, pelo fator $f_1 = \sqrt{n/(n-p-1)}$, obtendo-se o novo conjunto

$$A' = \{ f_1\hat{\epsilon}_1, f_1\hat{\epsilon}_2, \ldots, f_1\hat{\epsilon}_n \}. \qquad (4.2.15)$$

Realizando-se a reamostragem de A', então, por construção,

$$Pr_{F_n}\{ \epsilon_i^* = f_1\hat{\epsilon}_j \} = Pr_{F_n}\{ \epsilon_i^* = \sqrt{\frac{n}{(n-p-1)}}\hat{\epsilon}_j \} = \frac{1}{n}, \quad i,j = 1,2,\ldots,n, \qquad (4.2.16)$$

e, as propriedades (4.2.10)-(4.2.12) valem, exceto que,

$$\hat{\sigma}^2_{BOOT}(\epsilon_i^*) = Var_{\hat{F}_n}(\epsilon_i^*/x) = E_{\hat{F}_n}(\epsilon_i^{*2}/x) = \int_{A'} z^2 d\hat{F}_n(z) =$$

$$= \frac{1}{n}\sum_{i=1}^{n}(f_1\hat{\epsilon}_i)^2 = \frac{n}{(n-p-1)}\frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i^2 = \frac{\sum_{i=1}^{n}\hat{\epsilon}_i^2}{(n-p-1)} = s^2 \qquad (4.2.17)$$

,

ou seja, com a simples modificação na etapa de reamostragem, iguala-se a estimativa Bootstrap da variância dos erros aleatórios à estimativa não-tendenciosa clássica.

Com base nos pseudo dados

$$x_i^* = (y_i^*, c_i') , \quad i = 1, 2, ..., n , \qquad (4.2.18)$$

ajusta-se pelo mesmo método que produziu a estimativa original $\hat{\beta}$, isto é, neste caso, mínimos quadrados ordinários, o modelo

$$y_i^* = (1 \quad c_i')\beta + \epsilon_i , \qquad i = 1, 2, ..., n , \qquad (4.2.19)$$

ou, em forma matricial,

$$y^* = X\beta + \epsilon , \qquad (4.2.20)$$

cujo resultado é dado por,

$$\hat{\beta}^* = (X'X)^{-1}X'y^* , \qquad (4.2.21)$$

que minimiza a expressão

$$Q^*(\beta) = (y^* - X\beta)'(y^* - X\beta). \qquad (4.2.22)$$

A aproximação Bootstrap para a distribuição de uma variável aleatória R(x,P) (ou um vetor aleatório R(x,P)), sob P, função de $\hat{\beta}$ e β, é a distribuição de $R(x^*,\hat{P})$, condicional aos dados observados x, que é definida como a distribuição Bootstrap de $R(x^*,\hat{P})$. Estimativas Bootstrap de características de interesse da distribuição da variável aleatória R(x,P), são as respectivas características da distribuição Bootstrap de $R(x^*,\hat{P})$. Por exemplo, para estimação da matriz de variância-covariância de $\hat{\beta}$, de (4.1.22), seja

$$R(x,P) = \hat{\beta} - \beta. \qquad (4.2.23)$$

Então,

$$\Sigma_P (\hat{\beta}) = Var_P(\hat{\beta}) = E_P R(x,P)R(x,P)' \qquad (4.2.24)$$

e

$$R(x^*,\hat{P}) = \hat{\beta}^* - \hat{\beta}. \qquad (4.2.25)$$

A estimativa Bootstrap de $\Sigma_P (\hat{\beta})$ será dada por

$$\hat{\Sigma}_{BOOT} (\hat{\beta}) = Var_{\hat{P}}(\hat{\beta}/x) = E_{\hat{P}} \{ R(x^*,\hat{P})R(x^*,\hat{P})' / x \} =$$

$$= E_{\hat{P}} \{ (\hat{\beta}^* - \hat{\beta})(\hat{\beta}^* - \hat{\beta})' / x \} = \qquad (4.2.26)$$

$$= \hat{\sigma}^2_{BOOT} (\epsilon_i^*) (X'X)^{-1} = s^2 (X'X)^{-1}.$$

Desde que,

$$s^2 \xrightarrow{p} \sigma^2, \tag{4.2.27}$$

onde " $\xrightarrow{p}$ " significa *convergência em probabilidade*, então,

$$\hat{\Sigma}_{BOOT}(\hat{\beta}) \xrightarrow{p} \Sigma_P(\hat{\beta}). \tag{4.2.28}$$

Destes resultados segue que o processo Bootstrap anterior produz uma estimativa não-tendenciosa de $\Sigma_P(\hat{\beta})$, que também é *consistente*. Na realidade, esta é uma situação em que não é necessário aplicar o método Bootstrap, devido ao conhecimento de (4.1.22) e (4.1.23). Porém, ela serve para ilustrar a implementação do método para a estrutura de dados de um problema de regressão que pode ser facilmente generalizada para uma característica mais complicada da distribuição do estimador de mínimos quadrados, ou, um estimador mais complicado que (4.1.25), como será discutido no próximo capítulo.

4.2.2 Reamostragem dos resíduos centralizados

Quando for desejável ajustar um modelo como (4.1.17) sem o intercepto , ou seja,

$$y_i = \sum_{j=1}^{p} X_{ij}\beta_j + \epsilon_i, \qquad i = 1, 2, ..., n, \tag{4.2.29}$$

ou em termos matriciais,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{4.2.30}$$

onde, a matriz X de (4.2.30) não contém o vetor $\mathbf{1}_n = (1, 1, ..., 1)'$ na primeira coluna, e, assume-se que *posto*$(X) = p$, alguns cuidados devem ser tomados na reamostragem, pois, a menos que $\mathbf{1}_n \in C(X)$, então, $\sum_{i=1}^{n} \hat{e}_i \neq 0$, e,

$$E_{\hat{F}_n}(\epsilon_i^*/\boldsymbol{x}) = \int_{A'} z d\hat{F}_n(z) = \frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i \neq 0, \quad \forall\, i = 1,2,...,n. \tag{4.2.31}$$

Como,

$$\begin{aligned}\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}} &= (X'X)^{-1}X'\boldsymbol{y}^* - \hat{\boldsymbol{\beta}} = (X'X)^{-1}X'(X\hat{\boldsymbol{\beta}} + \boldsymbol{\epsilon}^*) - \hat{\boldsymbol{\beta}} = \\ &= \hat{\boldsymbol{\beta}} + (X'X)^{-1}X'\boldsymbol{\epsilon}^* - \hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\boldsymbol{\epsilon}^*,\end{aligned} \tag{4.2.32}$$

então

$$\begin{aligned}E_{\hat{F}}\{\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}} / \boldsymbol{x}\} &= (X'X)^{-1}X'E_{\hat{F}}\{\boldsymbol{\epsilon}^* / \boldsymbol{x}\} = \\ &= (\frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i)(X'X)^{-1}X'\mathbf{1}_n \neq \mathbf{0}_n,\end{aligned} \tag{4.2.33}$$

ou seja, a distribuição Bootstrap de $\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}$ incorporará uma tendenciosidade aleatória, que poderá afetar o desempenho do método. Porém, pode-se corrigir esta tendenciosidade realizando-se a reamostragem, de forma mais geral, do conjunto,

$$A'' = \{ f_2(\hat{\epsilon}_1 - \frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i),\ f_2(\hat{\epsilon}_2 - \frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i),...,\ f_2(\hat{\epsilon}_n - \frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i) \}, \tag{4.2.34}$$

onde

$$f_2 = \left(\frac{n - posto(X) - 1}{n} + \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} h_{ij}}{n^2} \right)^{-1/2} \tag{4.2.35}$$

e h_{ij} é o elemento *ij* da matriz

$$H = X(X'X)^{-1}X'. \tag{4.2.36}$$

Logo, por construção,

$$Pr_{\hat{F}_n}\{ \epsilon_i^* = f_2(\hat{\epsilon}_i - \tfrac{1}{n}\Sigma\hat{\epsilon}_j) \} = \frac{1}{n}, \quad i = 1,2,...,n. \tag{4.2.37}$$

Então,

$$\begin{aligned} E_{\hat{F}_n}(\epsilon_i^*/\boldsymbol{x}) &= \int_{A''} z d\hat{F}_n(z) = \frac{1}{n}\sum_{i=1}^{n} f_2(\hat{\epsilon}_i - \tfrac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i) = \\ &= \frac{1}{n} f_2 \sum_{i=1}^{n} (\hat{\epsilon}_i - \tfrac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i) = 0 \end{aligned} \tag{4.2.38}$$

e

$$\begin{aligned} \hat{\sigma}^2_{BOOT}(\epsilon_i^*) &= Var_{\hat{F}_n}(\epsilon_i^*/\boldsymbol{x}) = E_{\hat{F}_n}(\epsilon_i^{*2}/\boldsymbol{x}) = \int_{A''} z^2 d\hat{F}_n(z) = \\ &= \frac{1}{n}\sum_{i=1}^{n} (f_2(\hat{\epsilon}_i - \tfrac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i))^2 = f_2^2 \frac{1}{n}\sum_{i=1}^{n} (\hat{\epsilon}_i - \tfrac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i)^2, \end{aligned} \tag{4.2.39}$$

que no modelo sem intercepto (4.2.29) é um estimador não-tendencioso de σ^2, assim como, no modelo com intercepto (4.1.11), pois, neste caso, $\sum_{i=1}^{n}\hat{\epsilon}_i = 0$ e

$$\mathbf{1}_n \in C(X) \Rightarrow \sum_{i=1}^{n}\sum_{j=1}^{n} h_{ij} = n \Rightarrow f_2 = ((n - posto(X))/n)^{-1/2} = f_1, \tag{4.2.40}$$

e, a fórmula da expressão (4.2.40) será igual a da expressão (4.2.17).

4.2.3 Cálculos Bootstrap-Monte Carlo

A exemplo de outras situações discutidas nos capítulos anteriores, pode também acontecer no contexto de regressão que a distribuição Bootstrap exata de $R(x^*,\hat{P})$ ou características desta distribuição, possa ser difícil de ser calculada analiticamente. Porém, pode-se obter uma aproximação para a distribuição Bootstrap ou as características desta distribuição, adaptando-se o algorítmo de Monte Carlo da figura 2.2, como descrito na figura 4.1. A distribuição empírica das B replicações $r^*(1), r^*(2), ..., r^*(B)$ é tomada como uma aproximação para a distribuição Bootstrap de $R(x^*,\hat{P})$. Por exemplo, considere a estimação da matriz de variância-covariância de $\hat{\beta}$. Com base nas B replicações do vetor aleatório Bootstrap

$$R(x^*,\hat{P}) = \hat{\beta}^* - \hat{\beta}, \tag{4.2.41}$$

isto é,

$$r^*(b) = R(x^*(b),\hat{P}) = \hat{\beta}^*(b) - \hat{\beta}, \tag{4.2.42}$$

então, a aproximação de Monte Carlo para (4.2.26) será dada por:

$$\hat{\Sigma}_B(\hat{\beta}) = \frac{\sum_{b=1}^{B} (r^*(b) - r^*(\cdot))(r^*(b) - r^*(\cdot))'}{B - 1} = \frac{\sum_{b=1}^{B} (\hat{\beta}^*(b) - \hat{\beta}^*(\cdot))(\hat{\beta}^*(b) - \hat{\beta}^*(\cdot))'}{B - 1}, \tag{4.2.43}$$

onde,

$$r^*(\cdot) = \frac{1}{B}\sum_{b=1}^{B} r^*(b), \tag{4.2.44}$$

e,

$$\hat{\beta}^*(\cdot) = \frac{1}{B}\sum_{b=1}^{B} \hat{\beta}^*(b). \tag{4.2.45}$$

Quando $B \to \infty$, então, $\hat{\Sigma}_B \to \hat{\Sigma}_{BOOT}(\hat{\beta})$.

ALGORÍTMO 2

(i) forma-se B vetores de pseudo erros Bootstrap, retirando-se independentemente B amostras aleatória de n observações, $\epsilon^*(b) = (\epsilon_{b1}^*, \epsilon_{b2}^*, \ldots, \epsilon_{bn}^*)'$, $b = 1, 2, \ldots, B$, *com reposição, do conjunto* A'' ;
(ii) para cada um dos B vetores ϵ_b^*, *calcula-se os pseudos dados Bootstrap*

$$y^*(b) = X\hat{\beta} + \epsilon^*(b), \tag{4.2.46}$$

onde, $y^*(b) = (y_{b1}^*, y_{b2}^*, \ldots, y_{bn}^*)$;
(iii) para cada um dos B conjunto de dados Bootstrap $x^*(b) = (x_{b1}^*, x_{b2}^*, \ldots, x_{bn}^*)$, *onde* $x_{bi}^* = (y_{bi}^*, c_i')$, $i = 1, 2, \ldots, n$, *calcula-se a estimativa de* MQO

$$\hat{\beta}^*(b) = (X'X)^{-1}X'y^*(b), \tag{4.2.47}$$

e, a b-ésima replicação da variável aleatória Bootstrap $R(x^*, \hat{P})$, *que será dada por*

$$r^*(b) = R(x^*(b), \hat{P}), \quad b = 1, 2, \ldots, B. \tag{4.2.48}$$

Figura 4.1. Algorítmo de Monte Carlo para construir B replicações da variável aleatória Bootstrap $R(x^*, \hat{P})$ com reamostragem de resíduos

4.2.4 Resultados assintóticos

O comportamento assintótico da distribuição Bootstrap de

$$\sqrt{n}(\ \hat{\beta}^* - \hat{\beta}\ ), \tag{4.2.49}$$

no modelo de regressão, foi estudado em Freedman (1981), com a reamostragem feita sobre os resíduos ajustados pela média, sem o fator de correção que converge para 1 quando n tende para infinito, ou seja,

$$\hat{F}_n(x) = \frac{\#\{\ (\ \hat{e}_i - \frac{1}{n}\sum_{i=1}^{n}\hat{e}_i\ ) \le x\ \}}{n}. \tag{4.2.50}$$

Seja $\Psi_n(F)$ a verdadeira distribuição de

$$\sqrt{n}(\ \hat{\beta} - \beta\ ). \tag{4.2.51}$$

Logo, a distribuição Bootstrap de (4.2.49) será dada por $\Psi_n(\hat{F}_n)$. Freedman (1981) mostrou que, para n suficientemente grande, a distribuição Bootstrap $\Psi_n(\hat{F}_n)$ estará próxima da distribuição $\Psi_n(F)$, desde que $\sigma^2 p.traço\ (X'X)^{-1}$ seja pequeno. Logo, se as condições (4.1.26) e (4.1.27) são satisfeitas, então, a distribuição Bootstrap $\Psi_n(\hat{F}_n)$ também, aproxima-se-á da distribuição normal p-variada com vetor de médias $\mathbf{0}$ e matriz de variância-covariância $\sigma^2 V^{-1}$, isto é,

$$\sqrt{n}(\ \hat{\beta}^* - \hat{\beta}\ ) \underset{*}{\sim} N_p\ (\mathbf{0}, \sigma^2 V^{-1}). \tag{4.2.52}$$

4.2.5 Problemas com modelagem incorreta

Quando o modelo estocástico P que se supõe gerar os dados **x** é incorreto, problemas sérios podem acontecer com as estimativas Bootstrap produzidas pelo método de reamostragem de resíduos, já que este caracteriza-se por reproduzir fielmente a estrutura estocástica de P. Por exemplo, suponha que os erros aleatórios são heterocedásticos em vez de homocedásticos como postulado por P, ou seja,

$$\Sigma_{P'}(\epsilon) = diag(\sigma_1^2, \sigma_2^2, ..., \sigma_n^2) = D, \tag{4.2.53}$$

onde P′ é o modelo estocástico correto. Portanto,

$$\Sigma_{P'}(\hat{\epsilon}) = Var_{P'}(\hat{\epsilon}) = Var_{P'}((I-H)\epsilon) = (I-H)D(I-H) \tag{4.2.54}$$

e

$$\begin{aligned}\Sigma_{P'}(\sqrt{n}(\hat{\beta}-\beta)) = Var_{P'}(\sqrt{n}(\hat{\beta}-\beta)) = n(X'X)^{-1}X'DX(X'X)^{-1} = \\ = (\tfrac{1}{n}X'X)^{-1}\tfrac{1}{n}X'DX(\tfrac{1}{n}X'X)^{-1}.\end{aligned} \tag{4.2.55}$$

Sob a condição (4.1.26) e

$$\frac{1}{n}X'DX \underset{n\to\infty}{\to} \Delta, \tag{4.2.56}$$

onde Δ é uma matriz positiva definida, então,

$$\Sigma_{P'}(\sqrt{n}(\hat{\beta}-\beta)) \underset{n\to\infty}{\to} V^{-1}\Delta V^{-1}, \tag{4.2.57}$$

ou seja, a matriz de variância e covariância assintótica da distribuição de $\sqrt{n}(\hat{\beta}-\beta)$ é dada

por $V^{-1}\Delta V^{-1}$. Conduzindo-se o processo Bootstrap segundo o modelo P, que assume erros homocedásticos, da expressão (4.2.17) (para um modelo com intercepto), temos,

$$\hat{\sigma}^2_{BOOT}(\epsilon_i^*) = \frac{\sum_{i=1}^{n} \hat{\epsilon}_i^2}{n} = \frac{\hat{\epsilon}'\hat{\epsilon}}{n-p-1}. \tag{4.2.58}$$

Porém, supondo-se a estrutura de erros heterocedásticos,

$$E_{p'}(\hat{\sigma}^2_{BOOT}(\epsilon_i^*)) = \frac{E_{p'}(\hat{\epsilon}'\hat{\epsilon})}{n-p-1} = \frac{E_{p'}(\epsilon'(I-H)\epsilon)}{n-p-1} = \frac{traço((I-H)D)}{n-p-1} =$$

$$= \frac{traço(D) - traço(HD)}{n-p-1} = \frac{\sum_{i=1}^{n} \sigma_i^2 - traço(X(X'X)^{-1}X'D)}{n-p-1}. \tag{4.2.59}$$

Portanto, se

$$\frac{\sum_{i=1}^{n} \sigma_i^2}{n-p-1} \underset{n\to\infty}{\to} \bar{\sigma}^2, \tag{4.2.60}$$

onde $\bar{\sigma}^2$ é uma constante, então, prova-se, pela desigualdade de Markov, que

$$\hat{\sigma}^2_{BOOT}(\epsilon_i^*) \xrightarrow{p} \bar{\sigma}^2, \tag{4.2.61}$$

e, conseqüentemente,

$$\hat{\Sigma}_{BOOT}(\sqrt{n}(\hat{\beta} - \beta)) \xrightarrow{p} \bar{\sigma}^2 V^{-1} \neq V^{-1}\Delta V^{-1}. \tag{4.2.62}$$

Este resultado significa que a estimativa Bootstrap da matriz de variância e covariância de

$\sqrt{n}(\hat{\beta} - \beta)$, não corresponde a sua matriz de variância e covariância assintótica. Em outras palavras, a estimativa Bootstrap não é *consistente*.

4.2.6 O método Bootstrap ponderado

Uma outra situação que pode comprometer o desempenho do método de reamostragem dos resíduos é quando um ou mais pontos $\mathbf{x}_i = (y_i, \mathbf{c}_i')$ são discrepantes dos restantes, visto que estes podem afetar o ajuste de mínimos quadrados. O *Bootstrap ponderado* (Weber, 1984 e 1986) é uma tentativa de tornar o Bootstrap, ou, as estimativas Bootstrap, mais robustos à presença de observações discrepantes. Considere-se os resíduos *studentizados* ajustados pela média

$$\hat{e}_i = \hat{\varepsilon}_i(1 - h_{ii})^{-1/2} - \hat{e}_{\cdot}, \quad i = 1, 2, \ldots, n, \tag{4.2.63}$$

onde,

$$\hat{e}_{\cdot} = \frac{1}{n}\sum_{i=1}^{n} \hat{e}_i, \tag{4.2.64}$$

e o conjunto destes resíduos

$$A''' = \{ \hat{e}_1, \hat{e}_2, \ldots, \hat{e}_n \}. \tag{4.2.65}$$

A idéia do Bootstrap ponderado é ajustar as probabilidades de reamostragem deste conjunto para que resíduos $\hat{e}_i$ não muito grandes sejam igualmentes prováveis de serem selecionados, e, ao contrário, resíduos grandes sejam menos prováveis de serem selecionados. A sugestão apresentada em Weber (1984) consiste em reamostrar da FDE $\hat{F}_n^{w}$ que associa massa de probabilidade $p_i = (1 - h_{ii})/(n - posto(X))$ em cada resíduo do conjunto A''', isto é,

$$\hat{F}_n^w = \textit{massa de probabilidade } p_i = \frac{1 - h_{ii}}{n - \textit{posto}(X)} \textit{ em cada } \hat{e}_i. \qquad (4.2.66)$$

Assim, tomando-se $\epsilon^* = (\epsilon_1^*, \epsilon_2^*, ..., \epsilon_n^*)$ como uma amostra aleatória de tamanho n do conjunto A''', segundo as probabilidades de seleção p_i, então, por construção,

$$Pr_{\hat{F}_n^w}\{ \epsilon_i^* = \hat{e}_j \} = p_j = \frac{1 - h_{jj}}{n - \textit{posto}(X)}, \quad i,j = 1,2,...,n. \qquad (4.2.67)$$

Com estes pseudo erros aleatórios, formam-se os dados Bootstrap

$$x_i = (y_i^{*w}, c_i'), \qquad (4.2.68)$$

onde,

$$y_i^{*w} = (1 \; c_i')\hat{\beta} + \epsilon_i^{*w}, \quad i = 1,2,...,n. \qquad (4.2.69)$$

Com base nestes dados, calcula-se a estimativa Bootstrap

$$\hat{\beta}^{*w} = (X'X)^{-1}X'y^{*w}. \qquad (4.2.70)$$

Weber (1984) mostrou que esse processo não afeta o comportamento assintótico do Bootstrap e, sob a (4.1.17), para uma funcão f com segunda derivada limitada a distribuição Bootstrap de

$$\sqrt{n}(f(\hat{\beta}^{*w}) - f(\hat{\beta})) \qquad (4.2.71)$$

aproximar-se-á, para n suficientemente grande, da distribuição normal com vetor de médias zero e matriz de variância-covariância $\sigma^2 f^{(1)}(\beta)'V^{-1}f^{(1)}(\beta)$, onde $f^{(1)}(\beta)$ é o vetor de

derivada parciais de *f* com respeito as componentes do vetor β. O Bootstrap ponderado foi proposto com o intuito de melhorar o desempenho do método Bootstrap para valores moderados de *n*, na presença de pontos influentes.

4.3 O método Bootstrap com reamostragem das observações

Outro método Bootstrap para obtenção de informação da distribuição de um estimador do vetor de parâmetros β, é o *método de reamostragem das observações* (Efron, 1983 e 1986). Este método consiste em formar a amostra Bootstrap reamostrando-se a função de distribuição empírica que atribue massa de probabilidade *1/n* a cada um *n* dos vetores de observações $\mathbf{x}_i$, i =1, 2,..., n, ou seja,

$$\hat{F}_n : \textit{massa de probabilidade } \frac{1}{n} \textit{ sobre cada } x_i = (y_i, c_i'), \quad i = 1,2,...,n. \quad (4.3.1)$$

Assim, a construção da amostra Bootstrap é feita retirando-se uma amostra aleatória simples de tamanho n, $x^* = (x_1^*, x_2^*, ..., x_n^*)$, com reposição, do conjunto de pares

$$A = \{ (y_1, c_1'), (y_2, c_2'), ..., (y_n, c_n') \}. \quad (4.3.2)$$

Deste modo,

$$x_i^* = (y_i^*, c_i^{*\prime}), \quad i = 1,2,...,n, \quad (4.3.3)$$

tal que,

$$Pr_{\hat{F}_n}(x_i^* = (y_j, c_j')) = \frac{1}{n}, \quad i,j = 1,2,...,n. \quad (4.3.4)$$

Portanto, ambas as componentes do vetor $\mathbf{x}_i^*$ serão aleatórias. A matriz X de (4.1.5) será substituída por :

$$X^*_{n\times(p+1)} = \begin{pmatrix} 1 & c_1^{*\prime} \\ 1 & c_2^{*\prime} \\ \vdots & \vdots \\ 1 & c_n^{*\prime} \end{pmatrix}, \tag{4.3.5}$$

que é uma matriz aleatória. Então, a estimativa Bootstrap do vetor $\boldsymbol{\beta}$ será dada por :

$$\hat{\boldsymbol{\beta}}^* = (X^{*\prime}X^*)^{-1}X^{*\prime}\mathbf{y}^*. \tag{4.3.6}$$

Estimativas da distribuição Bootstrap $\hat{\boldsymbol{\beta}}^*$ (ou de uma outra variável aleatória função de $\hat{\boldsymbol{\beta}}^*$ como a matriz de variância e covariância de $\hat{\boldsymbol{\beta}}$) não são facilmente obtidas através de cálculo analítico direto, como realizado no método de reamostragem de resíduos. Mas, com a possibilidade de poder gerar várias amostras Bootstrap do conjunto de pares A, pode-se adaptar o algorítmo de Monte Carlo da figura 4.1 com o mesmo propósito de aproximar a distribuição Bootstrap da variável aleatória de interesse, conforme está descrito no algorítmo da figura 4.2. Assim, características de interesse da distribuição Bootstrap de uma variável aleatória $R(\mathbf{x}^*,\hat{P})$ são aproximadas pelas características correspondentes da distribuição empírica das replicações $r^*(b)$, $b = 1,2,...,B$. Por exemplo, a estimativa Bootstrap de $\boldsymbol{\Sigma}_P(\hat{\boldsymbol{\beta}})$ é aproximada pela mesma quantidade da expressão (4.2.43), com $\hat{\boldsymbol{\beta}}^*(b)$ calculado pela fórmula da expressão (4.3.7).

ALGORÍTMO 3

(i) Forma-se B vetores de pseudo dados Bootstrap, retirando-se independentemente B amostras aleatória de n pares de observações, $x^*(b) = (x^*_{b1}, x^*_{b2}, ..., x^*_{bn}), b = 1, 2, ..., B,$ *com reposição, do conjunto A;*
(ii) Para cada um dos B conjunto de dados Bootstrap $x^*(b) = (x^*_{b1}, x^*_{b2}, ..., x^*_{bn})$, *onde* $x^*_{bi} = (y^*_{bi}, c^{*\prime}_i)$, $i = 1, 2, ..., n$, *calcula-se a estimativa de* MQO

$$\hat{\beta}^*(b) = (X^{*\prime}X^*)^{-1}X^{*\prime}y^*(b), \qquad (4.3.7)$$

e, a b-ésima replicação da variável aleatória Bootstrap $R(x^*, \hat{P})$ *será dada por*

$$r^*(b) = R(x^*(b), \hat{P}), \quad b = 1, 2, ..., B. \qquad (4.3.8)$$

Figura 4.2. Algorítmo de Monte Carlo para construir B replicações da variável aleatória Bootstrap $R(x^*, \hat{P})$ com de reamostragem de pares de observações

O método de reamostragem dos pares de observações reproduz uma estrutura estocástica de um modelo de correlação (Freedman, 1981) que é diferente da estrutura de um modelo de regressão. A estrutura estocástica do modelo de correlação supõe que os n vetores de dados observados representam uma amostra de uma distribuição $(p+1)$-dimensional, ou seja, o passo $P \rightarrow x$ é representado por:

$$F \rightarrow x = ((y_1, c_1'), (y_2, c_2'), ..., (y_n, c_n')), \qquad (4.3.9)$$

onde a FDA F é definida sobre $\mathbb{R}^{(p+1)}$. O comportamento assintótico do método Bootstrap, neste método de reamostragem, foi estudado em Freedman (1981) com uma modificação na etapa de reamostragem, selecionando aleatoriamente m pares de observações em vez de n como descrito originalmente em Efron (1983, 1986). Assim, a matriz X^* será dada por:

$$X^*_{m\times(p+1)} = \begin{pmatrix} 1 & c_1^{*\prime} \\ 1 & c_2^{*\prime} \\ \vdots & \vdots \\ 1 & c_m^{*\prime} \end{pmatrix}, \qquad (4.3.10)$$

e a estimativa Bootstrap $\hat{\beta}^*$ será (4.3.6) com X^* igual a (4.3.10). Os principais resultados assintóticos mostrados foram os seguintes :

• A estimativa Bootstrap $\hat{\beta}^*$ converge *quase certamente para* $\hat{\beta}$ quando m $\to \infty$, isto é,

$$\hat{\beta}^* \underset{m\to\infty}{\overset{a.s}{\to}} \hat{\beta},$$

que quer dizer que a menos de pontos y^* com probabilidade nula, $\hat{\beta}^*$ tende para $\hat{\beta}$ com probabilidade igual a 1, quando m $\to \infty$.

• Quando *m* e *n* tendem para infinito, sob (4.1.17) e (4.2.57), a estimativa Bootstrap da matriz de variância-covariância converge para $V^{-1}MV^{-1}$, isto é,

$$\hat{\Sigma}_{BOOT}\left(\sqrt{n}\left(\hat{\beta}-\beta\right)\right) \underset{m,n\to\infty}{\to} V^{-1}MV^{-1},$$

onde

$$M = \lim_{n\to\infty}\left(\frac{1}{n}(X'\Sigma(\epsilon)X\right).$$

Logo, se $\Sigma(\epsilon) = \sigma^2 I_n$, isto é, os erros são homocesdásticos, então, sob (4.1.17), $M = \sigma^2 V$ e $V^{-1}MV^{-1} = \sigma^2 V^{-1}$. Se $\Sigma(\epsilon) = D$, isto é, os erros são heterocedásticos, então, sob (4.2.56),

$M = \Delta$, e, $V^{-1}MV^{-1} = V^{-1}\Delta V^{-1}$. Estes resultados mostram que a estimativa Bootstrap da matriz de variância e covariância da distribuição de $\sqrt{n}(\hat{\beta}^* - \hat{\beta})$, baseada na reamostragem das observações, é consitente tanto, no caso de homocedasticidade, quanto, no caso de heterocedasticidade, ao contrário do método de reamostragem de resíduos que é válido apenas no primeiro caso. Este método de reamostragem é mais simples que o método de reamostragem de resíduos no sentido que não é necessário ajustá-los ou corrigí-los antes da reamostragem.

• *Sob determinadas condições, a distribuição Bootstrap de* $(\sqrt{n}(\hat{\beta}^* - \hat{\beta}))$ *é assintóticamente (quando m e n tendem para o infinito) normal p-variada com vetor de médias zero e matriz de variância-covariância* $V^{-1}MV^{-1}$.

4.4 Intervalos de confiança

Nesta seção será implementada a metodologia apresentada no capítulo 3, para construir intervalos de confiança Bootstrap para funções lineares do vetor de parâmetros desconhecidos do modelo (4.1.4), isto é,

$$\theta = \lambda'\beta, \tag{4.4.1}$$

onde λ é um vetor do $\mathbb{R}^{(p+1)}$ que será denotado por:

$$\lambda = (\lambda_0, \lambda_1, ..., \lambda_p)'. \tag{4.4.2}$$

Neste caso, a estatística de interesse será dada por:

$$\hat{\theta} = \lambda'\hat{\beta}, \tag{4.4.3}$$

onde $\hat{\beta}$ é o estimador de MQO de β, que será denotado por

$$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p)'. \tag{4.4.4}$$

A estatística Bootstrap será dada por:

$$\hat{\theta}^* = \lambda'\hat{\beta}^*, \tag{4.4.5}$$

onde $\hat{\beta}^*$ é o estimador de MQO calculado na amostra Bootstrap $\mathbf{y}^*$, que pode ser obtida pelo método de reamostragem de resíduos ou pelo método de reamostragem das observações. Nesta seção serão descritos os intervalos diretamente com o uso do método de reamostragem de resíduos. A estimativa Bootstrap da variância de $\hat{\theta}$ é dada por

$$\hat{\sigma}^2_{BOOT}(\lambda'\hat{\beta}) = \lambda'\hat{\Sigma}_{BOOT}(\hat{\beta})\lambda. \tag{4.4.6}$$

4.4.1 Intervalo Bootstrap padrão

O intervalo de confiança Bootstrap padrão (BOOTPAD) para o parâmetro $\theta = \lambda'\beta$, com probabilidade de cobertura de aproximadamente $1-2\alpha$, é dado por:

$$\begin{aligned} &[\hat{\theta}_{BOOTPAD}[\alpha], \hat{\theta}_{BOOTPAD}[1-\alpha]] = \\ &= [\lambda'\hat{\beta} + z_\alpha \hat{\sigma}_{BOOT}(\lambda'\hat{\beta}), \lambda'\hat{\beta} - z_\alpha \hat{\sigma}_{BOOT}(\lambda'\hat{\beta})] = \\ &= [\lambda'\hat{\beta} + z_\alpha \lambda'\hat{\Sigma}_{BOOT}(\hat{\beta})\lambda, \lambda'\hat{\beta} - z_\alpha \lambda'\hat{\Sigma}_{BOOT}(\hat{\beta})\lambda]. \end{aligned} \tag{4.4.7}$$

4.4.2 Intervalo t-Student Bootstrap

O intervalo de confiança t-Student Bootstrap (t-SBOOT) para o parâmetro $\theta = \lambda'\beta$, com probabilidade de cobertura aproximadamente igual a $1-2\alpha$, é dado por:

$$[\hat{\theta}_{t\text{-}SBOOT}[\alpha], \hat{\theta}_{t\text{-}SBOOT}[1-\alpha]] =$$

$$= [\lambda'\hat{\beta} + t^{(\alpha)}_{n-p-1}\hat{\sigma}_{BOOT}(\lambda'\hat{\beta}), \lambda'\hat{\beta} - t^{(\alpha)}_{n-p-1}\hat{\sigma}_{BOOT}(\lambda'\hat{\beta})] = \quad (4.4.8)$$

$$[\lambda'\hat{\beta} + t^{(\alpha)}_{n-p-1}\lambda'\hat{\Sigma}_{BOOT}(\hat{\beta})\lambda, \lambda'\hat{\beta} - t^{(\alpha)}_{n-p-1}\lambda'\hat{\Sigma}_{BOOT}(\hat{\beta})\lambda].$$

4.4.3 Intervalo t-Bootstrap

Para construir o intervalo t-Bootstrap para o parâmetro θ, foi visto na seção 3.5 que é necessário primeiramente calcular os percentis α e $1-\alpha$ da distribuição Bootstrap da estatística

$$T^* = \frac{\lambda'\hat{\beta}^* - \lambda'\hat{\beta}}{\hat{\sigma}_{BOOT}(\lambda'\hat{\beta})} = \frac{\lambda'(\hat{\beta}^* - \hat{\beta})}{\{\lambda'\hat{\Sigma}_{BOOT}(\hat{\beta})\lambda\}^{1/2}}, \quad (4.4.9)$$

isto é, t^*_{α} e $t^*_{1-\alpha}$. Porém, devido aos esquemas de reamostragem apresentados neste capítulo serem de natureza não-paramétrica, com o uso da FDE, poderá ser díficil o cálculo destes percentis. Uma saída para este problema é utilizar o procedimento descrito na seção 3.5, seguindo-se os seguintes passos do processo Bootstrap não-paramétrico com reamostragem de resíduos :

(i) geram-se B amostras Bootstrap de pseudo erros

$$\epsilon^*(b) = (\epsilon^*_{b1}, \epsilon^*_{b2}, ..., \epsilon^*_{bn})', \quad b = 1, 2, ..., B, \tag{4.4.10}$$

selecionando-se com reposição do conjunto de resíduos A'';

(ii) formam-se B amostras Bootstrap de pseudo dados

$$y^*(b) = X\hat{\beta} + \epsilon^*(b), \quad b = 1, 2, ..., B, \tag{4.4.11}$$

(iii) calculam-se as estimativas de MQO Bootstrap

$$\hat{\beta}^*(b) = (X'X)^{-1}X'y^*(b), \quad b = 1, 2, ..., B, \tag{4.4.12}$$

para cada uma das B amostras Bootstrap do passo anterior;

(iv) Com base em $\epsilon^*(b)$ e $\hat{\beta}^*(b)$, b=1, 2,..., B, calculam-se as correspondentes replicações da estatística T^*,

$$T^*(b) = \frac{\lambda'\hat{\beta}^*(b) - \lambda'\hat{\beta}}{\hat{\sigma}_{BOOT}(\lambda'\hat{\beta}(b))} = \frac{\lambda'(\hat{\beta}^*(b) - \hat{\beta})}{\{\lambda'\hat{\Sigma}_{BOOT}(\hat{\beta}(b))\lambda\}^{1/2}}, \quad b = 1, 2, ..., B, \tag{4.4.13}$$

onde

$$\hat{\Sigma}_{BOOT}(\hat{\beta}(b)) = s^{*2}(X'X)^{-1}, \tag{4.4.14}$$

$$s^{*2}(b) = f_2^2 \frac{\sum_{i=1}^{n} (\epsilon^*_{bi} - \bar{\epsilon}^*_b)^2}{n} \tag{4.4.15}$$

e

$$\bar{\epsilon}_b^* = \frac{1}{n}\sum_{i=1}^{n} \epsilon_{bi}^* . \qquad (4.4.16)$$

(v) estima-se os quantis α e 1-α da distribuição Bootstrap de T^* pelos quantis empíricos $\hat{t}_{\alpha}^*$ e $\hat{t}_{1-\alpha}^*$, que satisfazem a

$$\frac{\#\{\ T^*(b)\ <=\ \hat{t}_{\alpha}^*\ \}}{B} = \alpha \qquad (4.4.17)$$

e

$$\frac{\#\{\ T^*(b)\ <=\ \hat{t}_{1-\alpha}^*\ \}}{B} = 1-\alpha . \qquad (4.4.18)$$

Estes valores podem calculados ordenando-se os B valores de $T^*(b)$, obtendo-se $T^*_{(1)}$, $T^*_{(2)}$,..., $T^*_{(B)}$, e, tomando-se $\hat{t}_{\alpha}^* = T^*_{(B.\alpha)}$ e $\hat{t}_{1-\alpha}^* = T^*_{B.(1-\alpha)}$, se $B.\alpha$ é inteiro. Caso $B.\alpha$ não seja inteiro, então, pode-se seguir o procedimento indicado para o intervalo t-Bootstrap na seção 3.5. Portanto, o intervalo de confiança t-Bootstrap (t-BOOT) para o parâmetro θ, utilizando-se os percentis aproximados da distribuição Bootstrap de T^*, com probabilidade de cobertura de aproximadamente 1 - 2α, será dado por:

$$[\ \hat{\theta}_{t\text{-}BOOT}[\alpha]\ ,\ \hat{\theta}_{t\text{-}BOOT}[1-\alpha]\ ] =$$

$$= [\ \lambda'\hat{\beta} - \hat{t}_{1-\alpha}^*\hat{\sigma}_{BOOT}(\lambda'\hat{\beta})\ ,\ \lambda'\hat{\beta} - \hat{t}_{\alpha}^*\hat{\sigma}_{BOOT}(\lambda'\hat{\beta})\ ] = \qquad (4.4.19)$$

$$= [\ \lambda'\hat{\beta} - \hat{t}_{1-\alpha}^*\lambda'\hat{\Sigma}_{BOOT}(\hat{\beta})\lambda\ ,\ \lambda'\hat{\beta} - \hat{t}_{\alpha}^*\lambda'\hat{\Sigma}_{BOOT}(\hat{\beta})\lambda\ ].$$

4.4.4 Intervalo percentil

Sejam

$$\hat{\theta}^*(1),\ \hat{\theta}^*(2),...,\ \hat{\theta}^*(B) \tag{4.4.20}$$

B replicações da estatística $\hat{\theta}^*$ baseada em B estimativas Bootstrap $\hat{\beta}^*(b)$, b = 1, 2,..., B, isto é,

$$\hat{\theta}^*(b) = \lambda'\hat{\beta}^*(b), \quad b = 1\ 2,..., B. \tag{4.4.21}$$

Considere os B valores ordenados destas B replicações dados por $\hat{\theta}^*_{(1)}$, $\hat{\theta}^*_{(2)}$,..., $\hat{\theta}^*_{(B)}$. O intervalo de confiança percentil (PERC) para o parâmetro θ, com probabilidade de cobertura de aproximadamente $1 - 2\alpha$ será dado por:

$$[\hat{\theta}_{PERC}[\alpha], \hat{\theta}_{PERC}[1-\alpha]] = [\hat{\theta}_{(B.\alpha)}, \hat{\theta}_{(B.(1-\alpha))}], \tag{4.4.22}$$

se $B.\alpha$ é inteiro. Caso $B.\alpha$ não seja inteiro, então, como vém sendo recomendado, estima-se os percentis seguindo-se o procedimento descrito na seção 3.5.

4.4.5 Intervalo percentil com correção para tendência

Sejam

$$\alpha_1 = \Phi(2\hat{z}_0 + z_\alpha) \tag{4.4.23}$$

e

$$\alpha_2 = \Phi(2\hat{z}_0 - z_\alpha), \tag{4.4.24}$$

onde

$$\hat{z}_0 = \Phi^{-1}(\#\{ \hat{\theta}^*(b) \leq \hat{\theta} \} / B) = \Phi^{-1}(\#\{ \lambda'\hat{\beta}^*(b) \leq \lambda'\hat{\beta} \} / B), \tag{4.4.25}$$

é a estimativa da constante de correção de tendência z_0, baseada nas B replicações de (4.4.21). Assim, o intervalo de confiança percentil com correção para tendência (PCT) para o parâmetro θ, com probabilidade de cobertura de aproximadamente $1 - 2\alpha$, será dado por:

$$[\hat{\theta}_{PCT}[\alpha] , \hat{\theta}_{PCT}[1-\alpha]] = [\hat{\theta}^*_{(B\alpha_1)} , \hat{\theta}^*_{(B\alpha_2)}]. \tag{4.4.26}$$

4.4.6 Intervalo percentil com correção para tendência e aceleração

Sejam

$$\alpha_1' = \Phi\left(\hat{z}_0 + \frac{(\hat{z}_0 + z_\alpha)}{1 - \hat{a}(\hat{z}_0 + z_\alpha)} \right) \tag{4.4.27}$$

e

$$\alpha_2' = \Phi\left(\hat{z}_0 + \frac{(\hat{z}_0 - z_\alpha)}{1 - \hat{a}(\hat{z}_0 - z_\alpha)} \right), \tag{4.4.28}$$

onde $\hat{z}_0$ é dada por (4.4.25) e $\hat{a}$ é uma estimativa da constante de aceleração a que, pela facilidade de cálculo, será usada a expressão (3.8.17) baseada nos valores *jackknife*, isto é,

$$\hat{a} = \frac{\sum_{i=1}^{n} (\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)})^3}{6\{\sum_{i=1}^{n} (\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)})^2\}^{3/2}}, \tag{4.4.29}$$

onde

$$\hat{\theta}_{(i)} = \lambda' \hat{\beta}_{(i)}, \quad i = 1, 2, \ldots, n, \tag{4.4.30}$$

$$\hat{\beta}_{(i)} = (X'_{(i)} X_{(i)})^{-1} X'_{(i)} y_{(i)}, \quad b = 1, 2, \ldots, B, \tag{4.4.31}$$

$X_{(i)}$ e $y_{(i)}$ representam a matriz X e o vetor de dados observados **y**, ambos sem a i-ésima linha ($X_{(i)}$ tem dimensão (n-1)x(p+1) e $y_{(i)}$ tem dimensão (n-1)x1). Portanto, o intervalo de confiança percentil com correção de tendência-aceleração (PCTa) para o parâmetro θ, com probabilidade de cobertura de aproximadamente $1 - 2\alpha$, será dado por:

$$[\ \hat{\theta}_{PCTa}[\alpha'_1]\ ,\ \hat{\theta}_{PCTa}[\alpha'_2]\ ] = [\ \hat{\theta}^*_{(B\alpha'_1)}\ ,\ \hat{\theta}^*_{(B\alpha'_2)}\ ]. \tag{4.4.32}$$

se $B.\alpha_1$ e $B.\alpha_2$ não for inteiro pode-se usar o procedimento descrito na seção 3.5.

4.5 Testes de hipóteses e significância

Nesta seção será demonstrado como se utilizar métodos Bootstrap para a realização de testes de hipóteses e de significância. Para ambos, serão consideradas a *hipótese linear geral*, nula e alternativa, respectivamente,

$$H_0 : C\beta = \boldsymbol{d} \tag{4.5.1}$$

e

$$H_A : C\beta \neq \boldsymbol{d}, \tag{4.5.2}$$

onde C é um matriz de dimensão $mx(p+1)$, de posto igual a $m < (p+1)$, e $\boldsymbol{d}$ é um vetor pertencente ao $\mathbb{R}^m$. A estatística comumente usada para a realização destes testes é dada por:

$$\mathcal{F} = \frac{\boldsymbol{m}^{-1}(C\hat{\beta} - \boldsymbol{d})'[C(X'X)^{-1}C']^{-1}(C\hat{\beta} - \boldsymbol{d})}{S^2}. \tag{4.5.3}$$

Se

$$\epsilon \sim N(\mathbf{0}, \sigma^2 I_n), \tag{4.5.4}$$

então, sob H_0, $\mathcal{F}$ tem distribuição *F de Snedecor* com m e $n-p-1$ graus de liberdade no numerador e denominador respectivamente, ou seja,

$$\mathcal{F} \sim F_{m,n-p-1}. \tag{4.5.5}$$

Porém, sem as suposições distribuicionais sobre os erros aleatórios de (4.5.4), não se garante que (4.5.5) seja verdadeira e a validade de testes de hipóteses e significância poderá ficar comprometida. Este é um bom exemplo de onde se pode aplicar o método Bootstrap como será visto a seguir.

Uma forma de usar o método Bootstrap para testes de hipóteses e de significância consiste em utilizar a distribuição Bootstrap da estatística

$$\mathcal{F}^* = \frac{m^{-1}(C\hat{\beta}^* - d)'[C(X'X)^{-1}C']^{-1}(C\hat{\beta}^* - d)}{Var_{F^*}(\epsilon_i^*)}. \qquad (4.5.6)$$

para aproximar a distribuição da estatística $\mathcal{F}$. Isto pode ser feito, utilizando-se o método de reamostragem dos resíduos, da seguinte forma:

• Para cada uma de B amostras Bootstrap de pseudos erros aleatórios e replicações do estimador de MQO para cada uma destas amostras, digamos $\epsilon^*(b)$ e $\hat{\beta}^*(b)$, b=1, 2,..., B como em (4.4.10) e (4.4.12), calculam-se as correspondentes replicações da estatística $\mathcal{F}^*$, isto é,

$$\mathcal{F}^*(b) = \frac{m^{-1}(C\hat{\beta}^*(b) - d)'[C(X'X)^{-1}C']^{-1}(C\hat{\beta}^*(b) - d)}{s^{*2}(b)}, \quad b = 1, 2,..., B, \qquad (4.5.7)$$

onde $s^{*2}(b)$ é dada por (4.4.15);

• ordena-se estes B valores de $\mathcal{F}^*$(b), obtendo-se $\mathcal{F}^*_{(1)}, \mathcal{F}^*_{(2)},..., \mathcal{F}^*_{(B)}$;

• estima-se o *valor p* que é igual a $p = Pr\{ \mathcal{F}^* > \mathcal{F}_0 \}$, onde $\mathcal{F}_0$ é o valor observado da estatística $\mathcal{F}$, por:

$$\hat{p} = \frac{\#\{\ \mathcal{F}^*(b) > \mathcal{F}_0\ \}}{B}. \qquad (4.5.8)$$

Dessa forma, um teste de hipótese para a hipótese H_0, a um nível aproximado α, será descrito por: *rejeite* H_0 *se* $\mathcal{F}_0$ *for superior a* $\mathcal{F}_{(B.(1-\alpha))}$. Já o teste de significância é realizado analisando-se a magnitude de $\hat{p}$.

Uma outra forma de utilização do Bootstrap para testar a hipótese $H_0 : C\boldsymbol{\beta} = \mathbf{d}$ é invertendo-se um intervalo de confiança Bootstrap para o parâmetro $\theta = C\boldsymbol{\beta}$, isto é, dado um intervalo de confiança Bootstrap para θ, $[\hat{\theta}_{BOOT}[\alpha], \hat{\theta}_{BOOT}[1-\alpha]]$, com probabilidade de cobertura aproximadamente igual a $1-2\alpha$, rejeita-se H_0 ao nível 2α se $\mathbf{d}$ não pertence a este intervalo.

Capítulo 5

O Método Bootstrap em Regressão Robusta

5.1 Introdução

No capítulo anterior foi discutida a aplicação do método Bootstrap em modelos de análise de regressão linear múltipla, utilizando-se o método de mínimos quadrados ordinários para estimar o vetor de parâmetros β. O estimador de mínimos quadrados é bastante utilizado na prática para se fazer inferências sobre β devido às suas propriedades ótimas quando os erros são normais (o que o torna com variância mínima entre todos os estimadores de não-tendenciosos de β, lineares e não-lineares) e por sua facilidade de cálculo matemático e computacional.

Quando a distribuição dos erros não é normal o desempenho do estimador de MQO pode não ser mais o mesmo. Sob as suposições (4.1.12)-(4.1.14), $\hat{\beta}$ continua possuindo a propriedade de variância mínima, porém em uma classe mais restrita que é a dos estimadores lineares não-tendenciosos. Isto quer dizer que se pode encontrar um estimador não-linear das observações $y_1, y_2,..., y_n$, por algum outro critério diferente do MQO, porém, com variância menor que a de $\hat{\beta}$. Em geral, os estimadores de MQO não são resistentes à presença de observações discrepantes na amostra resultantes, por exemplo, de erros grosseiros nos dados y_i ou X_{ij}. Também, não são considerados robustos no sentido de manter uma alta eficiência, como na situação de normalidade dos erros, quando estes

tiverem uma distribuição com cauda mais alongada que a distribuição normal. Estas questões levantadas agora motivaram o desenvolvimento do que se denomina *regressão robusta*.

A regressão robusta refere-se a aplicação de métodos de estimação robusta para modelos de regressão, isto é, critérios alternativos ao de mínimos quadrados que não só sejam eficientes nas situações em que a distribuição dos erros tenha cauda mais longa que a normal, mas, também, na presença de normalidade. Uma família de estimadores, que contém o de MQO, mas que pode fornecer alternativas mais robustas, é dos *M-estimadores*[1] (Huber, 1981), que será definida a seguir.

Um M-estimador do vetor de parâmetros β é definido como sendo um valor $\hat{\beta}_M$ que minimiza a expressão

$$\sum_{i=1}^{n} \rho(\epsilon_i) = \sum_{i=1}^{n} \rho(y_i - (1\ c_i')\beta), \tag{5.1.1}$$

isto é,

$$\min_{\beta \in \mathbb{R}^{p-1}} \sum_{i=1}^{n} \rho(\epsilon_i) = \sum_{i=1}^{n} \rho(y_i - (1\ c_i')\hat{\beta}_M), \tag{5.1.2}$$

onde ρ é uma função convexa e simétrica sobre o ponto zero. Geralmente, um valor que minimiza (5.1.1) pode ser obtido como solução da equação

$$\sum_{i=1}^{n} \psi(y_i - (1\ c_i')\beta)(1\ c_i')' = 0, \tag{5.1.3}$$

onde $\psi(u) = (\partial/\partial u)\rho(u)$. Como se pode observar, a classe dos M-estimadores é bastante

[1] O termo *M-estimadores* é devido ao fato que eles correspondem aos estimadores de máxima verossimilhança quando a distribuição dos erros tém densidade proporcional a $\exp(-\rho(u))$.

rica. Dois casos importantes são o próprio estimador de mínimos quadrados ordinários $\hat{\beta}$ que é obtido com $\rho(u) = \frac{1}{2}u^2$ ($\rightarrow \psi(u) = u$) e o estimador baseado na *norma* L_1, $\hat{\beta}_{L_1}$, que minimiza a soma dos desvio absolutos (MDA),

$$\min_{\beta \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} |y_i - (1 \; c_i')\beta| = \sum_{i=1}^{n} |y_i - (1 \; c_i')\hat{\beta}_{L_1}|, \qquad (5.1.5)$$

ou seja, este estimador é obtido com $\rho(u) = |u|$ ($\rightarrow \psi(u) = sinal(u)$).

Até bem pouco tempo, a falta de programas computacionais para o cálculo dos M-estimadores era um dos problemas que dificultava o uso da regressão robusta. Hoje, já se pode encontrar em pacotes estatísticos, como SAS® (no caso da regressão L_1) e S-PLUS, programas para estimar parâmetros do modelo de regressão com critérios distintos ao de mínimos quadrados.

Porém, uma dificuldade que ainda persiste é a falta de conhecimento das propriedades estatísticas dos estimadores robustos. Todos os resultados analíticos dísponíveis são de caráter assintótico, sem muito tratamento sobre as respectivas taxas de convergência. No caso dos M-estimadores, Huber (1981) mostrou que sob condições suaves de regularidade eles são consistentes e assintoticamente normal com matriz de variância e covariância assintótica proporcional à obtida pelo critério de mínimos quadrados. Propriedades de pequenas amostras vem sendo estudadas com auxílio de simulação de Monte Carlo para constatar quando, ou não, as fórmulas assintóticas são válidas supondo determinada distribuição para os erros com cauda mais alongada que a normal.

Porém, sem fazer uso das aproximações assintóticas, para um determinado tamanho amostral finito, como avaliar alguma medida de dispersão de uma particular estimativa $\hat{\beta}_M$ ou construir um intervalo de confiança (ou um teste de significância) para $\lambda'\beta$ ($\lambda \in \mathbb{R}^{p+1}$), com base em $\lambda'\hat{\beta}_M$? Uma resposta para estas questões pode ser dada pelo método

Bootstrap. Embora também aproximados, os resultados fornecidos aplicando-se a metodologia Bootstrap apresentada até agora podem ser bastantes úteis dado que, geralmente, devido à complexidade dos estimadores robustos não se conhecem soluções exatas. Do ponto de vista assintótico, Shorack (1982) mostrou que o método Bootstrap é um método válido. A grande vantagem do Bootstrap é a substituição das complexidades analíticas pela potência computacional, que torna-se cada vez mais mais rápida a custos mais acessíveis.

5.2 O método de reamostragem de resíduos para M-estimadores

O método que será descrito a seguir, para formação da amostra Bootstrap, é o baseado na reamostragem de resíduos. O método da reamostragem das observações não será considerado devido a falta de conhecimento, em geral, das propriedades das estimativas Bootstrap para M-estimadores, como no caso do ajuste por mínimos quadrados em que a estimativa Bootstrap da matriz de variância-covariância de $\hat{\boldsymbol{\beta}}$ é consistente mesmo que os erros sejam heterocedásticos, e, também, por não se tratar de um método que reproduza a estrutura estocástica do modelo de regressão.

Sejam

$$\hat{\epsilon}_i = y_i - (1 \; c_i')\hat{\boldsymbol{\beta}}_M, \quad i = 1, 2, ..., n, \tag{5.2.1}$$

os resíduos resultantes de um ajuste ao modelo de regressão (4.1.11) com um critério da classe dos M-estimadores, que foram definidos em (5.1.2), isto é, com alguma escolha particular da função ρ. Como foi visto no capítulo anterior, a formação de uma amostra Bootstrap, pelo método de reamostragem de resíduos, consiste em usar a informação dos resíduos correspondentes ao ajuste do modelo de regressão (no caso, os valores de (5.2.1)), para formar primeiro uma amostra de erros Bootstrap $\boldsymbol{\epsilon}^* = (\epsilon_1^*, \epsilon_2^*, ..., \epsilon_n^*)'$ e, em seguida, construir os pseudos dados $\boldsymbol{y}^* = (y_1^*, y_2^*, ..., y_n^*)' = X\hat{\boldsymbol{\beta}}_M + \boldsymbol{\epsilon}^*$. Com base em

$x^* = ((y_1^*, c_1'), (y_2^*, c_2'), \ldots, (y_n^*, c_n'))'$, calcula-se a estimativa Bootstrap $\hat{\beta}^*_M$ pelo mesmo critério que forneceu a estimativa $\hat{\beta}_M$ dos dados originais $x = ((y_1, c_1'), (y_2, c_2'), \ldots, (y_n, c_n'))'$, isto é, $\hat{\beta}^*_M$ é o valor tal que,

$$\min_{\beta \in \mathbb{R}^{p-1}} \sum_{i=1}^{n} \rho(\epsilon_i) = \sum_{i=1}^{n} \rho(y_i^* - (1 \; c_i')\hat{\beta}^*_M), \qquad (5.2.2)$$

onde

$$\epsilon_i = y_i^* - (1 \; c_i')\beta, \quad i = 1, 2, \ldots, n. \qquad (5.2.3)$$

Como já foi discutido, inferências sobre β são realizadas com base na distribuição Bootstrap de $\hat{\beta}^*_M$.

Como se pode ver o processo é idêntico ao visto na seção 4.2. Porém, alguns cuidados com respeito às propriedades do critério de estimação utilizado deve ser levado em conta na formação da amostra de pseudos erros Bootstrap, que no caso do critério de mínimos quadrados toma-se uma amostra aleatória, com reposição, de tamanho n, do conjunto de resíduo puros, ou ajustados, com ou sem fatores de correção. Por exemplo, no caso da critério baseado na norma L_1 ($\hat{\beta}_M = \hat{\beta}_{L_1}$, $\rho(u) = |u|$), sabe-se que o ajuste resultante passa pelo menos por $(p+1)$ das n observações $\mathbf{x}_i = (y_i, \mathbf{c}_i')$, i = 1, ,2,..., n. Isto quer dizer que ao menos $(p+1)$ dos resíduos deste ajuste são iguais a zero. Schrader & McKean (1987) sugerem que é melhor eliminar os resíduos com valor igual a zero para a reamostragem. Supondo que $(p+1)$ são nulos, seja $m = n - (p+1)$ e $\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_m$ os resíduos não-nulos, então, toma-se as componentes de ϵ^*, $\epsilon_1^*, \epsilon_2^*, \ldots, \epsilon_n^*$, como uma amostra aleatória simples de tamanho n, com reposição, do conjunto de valores $\{\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_m\}$. Para assegurar que cada pseudo erro ϵ_i^* tenha esperança Bootstrap zero pode-se ajustar, antes da reamostragem, cada um dos m resíduos pela sua média $\hat{e}_\cdot = (1/n)\sum_{i=1}^{m} \hat{e}_i$. Em geral, devido a complexidade do estimador baseado na norma L_1, ou qualquer outro M-estimador que não

seja o de mínimos quadrados, poderá ser díficil determinar analiticamente fatores de correção para tornar a estimativa Bootstrap da variância de cada componente ϵ_i, não-tendenciosa para σ^2. Por esse motivo é que a reamostragem é feita apenas sobre os resíduos puros ou ajustados pela média.

ALGORÍTMO 4

(i) forma-se B vetores de pseudo erros Bootstrap, retirando-se independentemente B amostras aleatória de n observações, $\epsilon^*(b) = (\epsilon^*_{b1}, \epsilon^*_{b2},..., \epsilon^*_{bn})'$, $b = 1, 2, ..., B$, *com reposição, do conjunto de m resíduos não-nulos* $\{\hat{e}_1, \hat{e}_2,..., \hat{e}_m\}$;
(ii) para cada um dos B vetores $\epsilon^*(b)$, *calcula-se os pseudos dados Bootstrap*

$$y^*(b) = X\hat{\beta}_{L_1} + \epsilon^*(b), \qquad (5.2.4)$$

onde, $y^*(b) = (y^*_{b1}, y^*_{b2},..., y^*_{bn})'$, $b = 1, 2,..., B$;
(iii) para cada um dos B conjuntos de dados Bootstrap $x^*(b) = (x^*_{b1}, x^*_{b2},..., x^*_{bn})$, *onde* $x^*_{bi} = (y^*_{bi}, c_i')$, $i = 1, 2,..., n$, *calcula-se a estimativa de* β *pelo critério da norma* L_1, $\hat{\beta}^*_{L_1}(b)$, *que minimiza* $\sum_{i=1} |y^*_{bi} - (1\ c_i')\beta|$, *ou seja,*

$$\min_{\beta \in \mathbb{R}^{p+1}} \sum_{i=1}^{n} |y^*_{bi} - (1\ c_i')\beta| = \sum_{i=1}^{n} |y^*_{bi} - (1\ c_i')\hat{\beta}^*_{L_1}(b)|. \qquad (5.2.5)$$

Figura 5.1. Algorítmo de Monte Carlo para construir B replicações Bootstrap da estimativa $\hat{\beta}_{L_1}$, com reamostragem de resíduos, para a regressão L_1

Para continuar o processo Bootstrap, nesta situação de regressão robusta, é necessário calcular a distribuição Bootstrap da estatística $\hat{\beta}^*_M$ ou características de interesse desta. Mas, ao contrário do que foi feito para o estimador de mínimos quadrado no capítulo anterior, a complexidade agora é muito maior. Como calcular, por exemplo, a

esperança Bootstrap do vetor $\hat{\beta}^*_M$, pelo qual nenhuma expressão analítica relacionando as observações y_i é conhecida ? Uma saída para esta questão, que vem sendo sugerida ao longo deste trabalho, é usar simulação de Monte Carlo, que se torna fundamental neste caso. A figura 5.1 exibe este processo para o caso particular da regressão L_1. A distribuição empírica das B replicações $\hat{\beta}^*_{L_1}(1), \hat{\beta}^*_{L_1}(2),..., \hat{\beta}^*_{L_1}(B)$ fornece uma aproximação para a distribuição Bootstrap de $\hat{\beta}^*_{L_1}$. Logo, uma aproximação de Monte Carlo para a estimativa Bootstrap da matriz de variância-covariância do estimador baseado na norma L_1, $\hat{\beta}_{L_1}$, é dada por:

$$\hat{\Sigma}_B(\hat{\beta}_{L_1}) = \frac{\sum_{b=1}^{B} (\hat{\beta}^*_{L_1}(b) - \hat{\beta}^*_{L_1}(\cdot))(\hat{\beta}^*_{L_1}(b) - \hat{\beta}^*_{L_1}(\cdot))'}{B - 1}, \qquad (5.2.6)$$

onde

$$\hat{\beta}^*_{L_1}(\cdot) = \frac{1}{B}\sum_{b=1}^{B} \hat{\beta}^*_{L_1}(b). \qquad (5.2.7)$$

5.3 Estimação da tendenciosidade de $\hat{\beta}_{L_1}$

Na seção anterior foi mostrado como se utilizar a metodologia Bootstrap para obter uma estimativa da matriz de variância-covariância do estimador $\hat{\beta}_{L_1}$, assim como foi realizado no capítulo 4 para o estimador de mínimos quadrados ordinários $\hat{\beta}$. Uma outra característica que se pode desejar avaliar da distribuição de $\hat{\beta}_{L_1}$ é a sua tendenciosidade como estimador de β, isto é,

$$T(F) = E_F(\hat{\beta}_{L_1} - \beta) = E_F\hat{\beta}_{L_1} - \beta. \qquad (5.3.1)$$

Esta característica não foi estudada no capítulo 4, pois, sob as condições (4.1.8)-(4.1.10) $\hat{\beta}$ é um estimador não-tendencioso para β, isto é, T(F) = 0. Neste caso da regressão L_1, ou outro estimador robusto, a avaliação da tendenciosidade pode ser importante e o Bootstrap poderá ser utilizado para uma estimativa.

De acordo com a seção 2.3, a estimativa Bootstrap de T(F) será dada por:

$$\hat{T}_{BOOT}(\hat{\beta}_{L_1}) = T(\hat{F}_n) = E_{\hat{F}_n}(\hat{\beta}^*_{L_1} - \hat{\beta}_{L_1}) = E_{\hat{F}_n}\hat{\beta}^*_{L_1} - \hat{\beta}_{L_1}. \qquad (5.3.2)$$

Mas, como é complicado calcular analiticamente a quantidade $E_{\hat{F}_n}\hat{\beta}^*_{L_1}$, mais uma vez, é necessário recorrer ao procedimento de Monte Carlo para obter uma aproximação para $\hat{T}_{BOOT}(\hat{\beta}_{L_1})$. Isto é feito calculando-se B replicações Bootstrap do estimador $\hat{\beta}^*_{L_1}$, $\hat{\beta}^*_{L_1}(1), \hat{\beta}^*_{L_1}(2),..., \hat{\beta}^*_{L_1}(B)$, utilizando-se o algoritmo da figura 5.1. Assim, uma aproximação de Monte Carlo para (5.3.2) será dada por:

$$\hat{T}_B(\hat{\beta}_{L_1}) = \frac{\sum_{b=1}^{B}\hat{\beta}^*_{L_1}(b)}{B} - \hat{\beta}_{L_1}. \qquad (5.3.3)$$

5.4 Resultados sobre intervalos de confiança e testes de significância

A construção de intervalos de confiança para os M-estimadores ou realização de um teste de significância são procedimentos de inferência estatística que podem ser elaborados facilmente com aplicação do Bootstrap, evitando o uso de aproximações assintóticas ou métodos analíticos complicados. A abordagem Bootstrap para estes procedimentos, neste caso dos M-estimadores é feita de forma análoga a realizada nas seções 4.4 e 4.5 para o caso do estimador de mínimos quadrados.

Com respeito a trabalhos de utilização do Bootstrap para estudar propriedades de

um estimador mais complicado que o de mínimos quadrados e de avaliação do seu desempenho, pode-se destacar : Dielman & Pfaffenberger (1988) e Stangenhaus, Narula and Filho (1993). Este último trabalho consistiu de um estudo de simulação para avaliar o desempenho dos intervalos BOOTPAD e PERC para as componentes do vetor de parâmetros de um modelo de regressão linear simples (p =1), β_0 e β_1, utilizando-se o critério de estimação baseado na norma L_1 . Foram utilizados tamanhos amostrais (n) iguais a 10, 20, 30 50, 75 e 100, e, as seguintes distribuições para os erros aleatórios : normal, normal contaminada, Laplace e Cauchy. As conclusões obtidas para estes dois intervalos, sugerem o uso do intervalo PERC para n <= 50, e, para n > 50, o intervalo BOOTPAD que requer um menor número de simulações que o PERC, já que ambos obtiveram desempenhos equivalentes.

Capítulo 6

Aplicação

6.1 Descrição do problema

Uma tarefa de grande importância na química é a interpretação e predição da reatividade que, em geral, pode ser expressa em termos da constante de equilíbrio, K, ou a constante de intensidade, k. Sabe-se que ambas quantidades dependem fortemente da temperatura, T.

Em uma série de reações químicas, se a lei de Arrhenius for satisfeita, a relação entre o logaritmo natural da constante de intensidade, $\ln k$, e o inverso da temperatura, $1/T$, é linear para cada reação. Isto quer dizer que um gráfico de vários valores $\ln k$ e $1/T$ deverá indicar uma reta para cada reação. Se, além disso, uma relação isocinética, cujo estudo é de grande importância em química estrutural, orgânica e inorgânica, for também satisfeita, então, as retas correspondentes a cada reação deverão interceptar-se em um mesmo ponto. A recíproca deste ponto tem interpretação química e é denominada *temperatura isocinética* (β). É fácil mostrar que quando as retas se interceptam em mesmo ponto, então, a correlação entre os interceptos e os coeficientes angulares de cada reta é igual a 1 ou a -1. Deseja-se, a partir de dados de um experimento químico, estudar a relação isocinética em reações com 19 bases de Schiff, estimando o ponto de interseção das retas, como também construir uma elipsóide de concentração para este ponto.

A estimação do ponto de interseção trata-se de um problema não-linear como será

visto na próxima seção. A construção da elipsóide de concentração pode ser realizada pelos métodos tradicionais, utilizando-se aproximações assintóticas. Porém, como será mostrado na seção 6.3, pode-se facilmente implementar o método Bootstrap para construí-lo, sem a necessidade de usar tais aproximações.

6.2 Estimação do ponto de interseção das retas

Os dados do experimento químico foram obtidos da seguinte forma : para cada um dos 19 reagentes (bases de Schiff) foram medidos valores da constante de intensidade a diferentes temperaturas (°K). Estas quantidades serão denotadas por k_{ij} e T_{ij}, respectivamente, i = 1, 2,..., 19 e j = 1, 2,..., n_i, onde n_i representa o número de temperaturas em cada reagente.

Sejam $y_{ij} = \ln k_{ij}$ e $x_{ij} = 1/T_{ij}$. As retas estimadas foram obtidas ajustando-se, para cada reagente, o modelo de regressão linear simples

$$y_{ij} = a_i + b_i x_{ij} + \epsilon_{ij}, \quad i = 1, 2, \ldots, 19, \; j = 1, 2, \ldots, n_i. \tag{6.2.1}$$

O coeficente de corrrelação das estimativas dos parâmetros a_i e b_i foi aproximadamente -0,99519 indicando que a lei de Arrhenius e a relação isocinética são praticamente satisfeitas.

Seja (x_0 , y_0) o verdadeiro valor do ponto de interseção das retas. A estimação deste ponto pode ser feita ajustando-se o modelo (6.2.1), com a restrição que, para $x_{ij} = x_0$, então, $y_{ij} = y_0$. Isto pode ser feito ajustando-se o modelo de regressão não-linear

$$y_{ij} = y_0 + b_i(x_{ij} - x_0) + \epsilon_{ij}, \quad i = 1, 2, \ldots, 19, \; j = 1, 2, \ldots, n_i. \tag{6.2.2}$$

As estimativas dos parâmetros x_0, y_0 e b_i podem ser obtidas pelo método de mínimos quadrados, utilizando-se um método iterativo para a resolução das equações normais não-lineares. Com ajuda do procedimento NLIN do sistema SAS®, versão 6.04, usando o método iterativo de *Marquardt*, obteve-se as seguintes estimativas de x_0 e y_0, respectivamente : $\hat{x}_0 = 0.00345$ e $\hat{y}_0 = -6.50$. Isto quer dizer que a estimativa da temperatura isocinética é $\hat{\beta} = \hat{x}_0^{-1} = 289{,}86°K$.

6.3 Construção da elipsóide de concentração

O processo Bootstrap para construir a elipsóide de concentração é o seguinte :

1) para cada uma das 19 retas estimadas ($\hat{y}_{ij} = \hat{a}_i + \hat{b}_i x_{ij}$), sejam

$$\hat{\epsilon}_{ij}, \quad i = 1, 2, ..., 19 \ , \ j = 1, 2, ..., n_i, \tag{6.3.1}$$

os resíduos resultantes. Retira-se de cada um destes 19 conjuntos de n_i resíduos, uma amostra aleatória de n_i observações com reposição, para formar os pseudo erros bootstrap

$$\epsilon^*_{ij}, \quad i = 1, 2, ..., 19 \ , \ j = 1, 2, ..., n_i; \tag{6.3.2}$$

2) os pseudo dados bootstrap são dados por:

$$y^*_{ij} = \hat{a}_i + \hat{b}_i x_{ij} + \epsilon^*_{ij}, \quad i = 1, 2, ..., 19 \ , \ j = 1, 2, ..., n_i; \tag{6.3.3}$$

3) repetindo-se independentemente os passos 1) e 2), B vezes, obtém-se B conjuntos de pseudo dados bootstrap $(y^*_{ij}(b), x_{ij})$, $b = 1, 2, ..., B$.

4) para cada um destes B conjuntos dados, obtidos no passo 3, ajusta-se o modelo não-

linear

$$y_{ij}^{*} = y_0 + \hat{b}_i(x_{ij} - x_0) + \epsilon_{ij}, \quad i = 1, 2, ..., 19 \ , \ j = 1, 2, ..., n_i, \tag{6.3.4}$$

obtendo-se as estimativas bootstrap de x_0 e y_0, $\hat{x}_0^*(b)$ e $\hat{y}_0^*(b)$, b = 1, 2,..., B.

5) a estimativa bootstrap da matriz de variância-covariância das estimativas $\hat{x}_0$ e $\hat{y}_0$, será dada por:

$$\hat{\Sigma}_B = \begin{pmatrix} \dfrac{\sum_{b=1}^{B} (\hat{x}_0^*(b) - \hat{x}_0^*(.))^2}{B-1} & \dfrac{\sum_{b=1}^{B} (\hat{x}_0^*(b) - \hat{x}_0^*(.))(\hat{y}_0^*(b) - \hat{y}_0^*(.))}{B-1} \\ \dfrac{\sum_{b=1}^{B} (\hat{x}_0^*(b) - \hat{x}_0^*(.))(\hat{y}_0^*(b) - \hat{y}_0^*(.))}{B-1} & \dfrac{\sum_{b=1}^{B} (\hat{y}_0^*(b) - \hat{y}_0^*(.))^2}{B-1} \end{pmatrix} \tag{6.3.5}$$

6) a elipsóide de concentração para o ponto (x_0, y_0), com 100(1-α)% de significância, será dada pela região:

$$R_\alpha = \{(x,y) \in \mathbb{R}^2 : [(x - \hat{x}_0, y - \hat{y}_0)\hat{\Sigma}_B^{-1}(x - \hat{x}_0, y - \hat{y}_0)']^{1/2} \leq (\chi^2_{2,1-\alpha})^{1/2}\}. \tag{6.3.6}$$

A estimativa Bootstrap-Monte Carlo $\hat{\Sigma}_B$ para os dados do experimento químiuco, com B = 250, foi a seguinte:

$$\hat{\Sigma}_B = \begin{pmatrix} 3{,}86.10^{-10} & -7{,}14.10^{-7} \\ -7{,}14.10^{-7} & 2{,}69.10^{-3} \end{pmatrix} \tag{6.3.7}$$

A figura 6.1 exibe o gráfico com as 19 retas estimadas e a elipsóide de concentração, de 95% de significância, para o ponto de interseção (x_0, y_0).

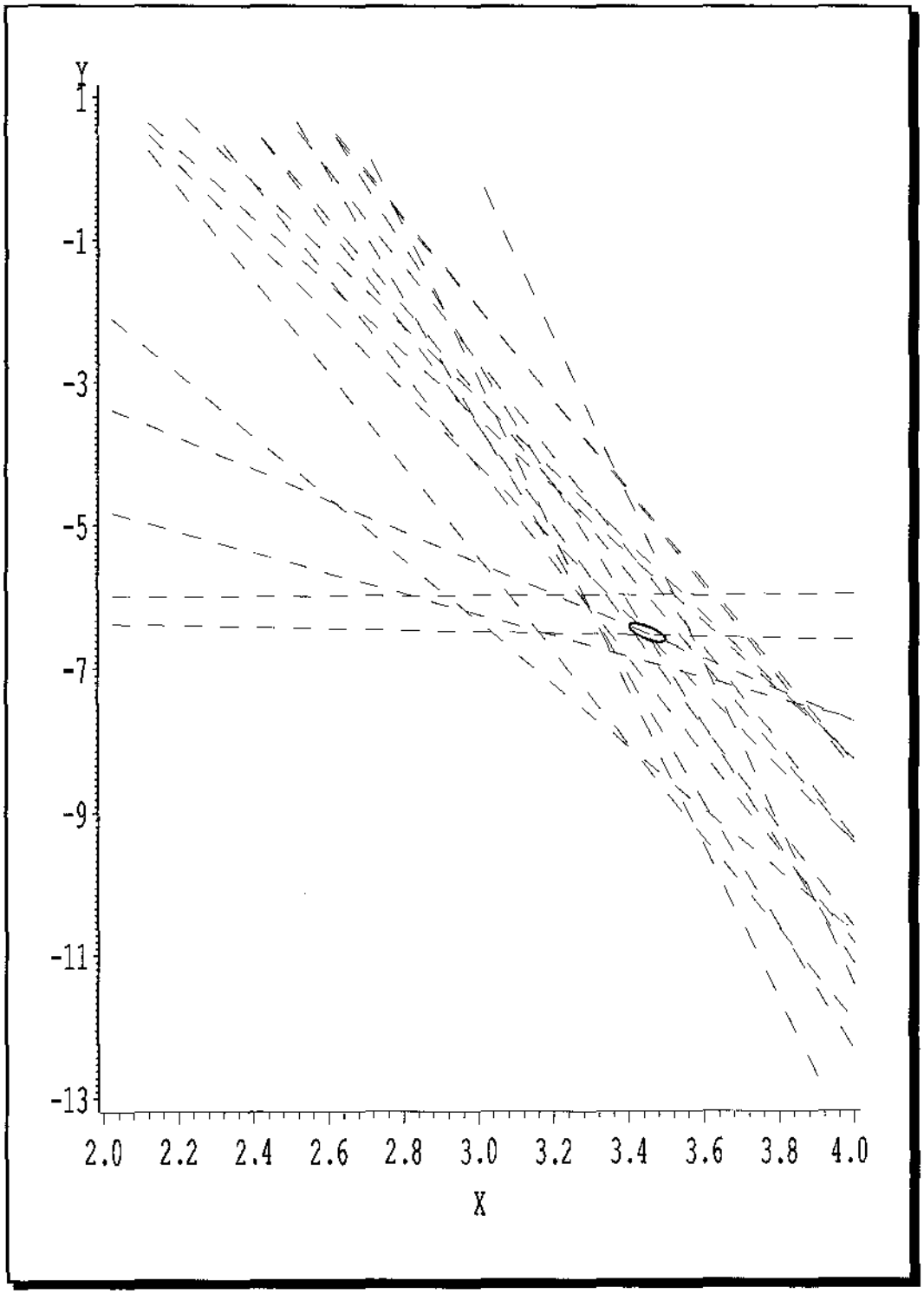

Figura 6.1. Retas estimadas e a elipsóide de concentração de 95% de significância para o ponto de interseção. Y = lnk e X = (1/T)x1.000

Referências Bibliográficas

ALEONG, J. & HOWARD, D. *Bootstrapping regression equations.* **American Statistical Association (ASA) - Proceedings of Statistics Computing Section.** 287-292, 1986.

BICKEL, P. J. & FREEDMAN, D. A. *Some asymptotic theory for the bootstrap.* **The Annals of Statistics.** 9:1196-1217, 1981.

BROWNSTONE, D. *Bootstrapping improved estimators for linear regression models.* **Journal of Econometrics.** 44:171-187, 1990.

BUCKLAND, S. T. *Monte carlo confidence intervals.* **Biometrics.** 40:811-817, 1984.

CARSON, R. T. *SAS® macros for bootstrapping and cross-validation regression equations.* **Proceedings of the SAS® users Groups International Conference.** 10:1064-1069, 1985.

DIACONIS, P. & EFRON, B. *Computer-intensive methods in statistics.* **Scientific American.** 248:116-130, 1983.

DICICCIO, T & ROMANO, J. P. *A review of bootstrap confidence intervals (with discussion).* **Journal of the Royal Statistical Society, B.** 50:338-370, 1988.

DIELMAN, T. E. & PFAFFENBERGER, R. C. *Bootstrapping in least absolute value regression : an application to hypothesis testing.* **Communications in Statistics-Simulation and Computation.** 17:843-856, 1988.

DRAPER, N. R. & SMITH, H. **Applied regression analysis.** 2nd Ed., New York, John Wiley & Sons, 1981.

EFRON, B. *Bootstrap methods : another look at the jackknife.* **The Annals of Statistics.** 21-460-480, 1979a.

EFRON, B. *Computers and theory of statistics : thinking the unthinkable.* **SIAM Review.** 21:460-480, 1979b.

EFRON, B. *Nonparametric standard errors and confidence intervals.* **The Canadian Journal of Statistics.** 9:139-172, 1981a.

EFRON, B. *Nonparametric estimates of Standard Error : The Jacknife, the bootstrap and other methods.* **Biometrika.** 68:589-599, 1981b.

EFRON, B. *The jackknife, the bootstrap and other resampling plans.* **Volume 38 of CBMS-NSF Regional Conference Series in Applied Mathematics.** SIAM, 1982.

EFRON, B. *Better bootstrap confidence intervals.* **Tech. Rep Stanford Univ. Dept. Statist,** 1984.

EFRON, B. *Bootstrap confidence intervals for a class of parametric problems.* **Biometrika.** 72:421-45-58, 1985.

EFRON, B. *Better bootstrap confidence intervals.* **Journal of the American Statistical Association.** 82:171-200, 1987.

EFRON, B. *Computer intensive methods in statistical regression.* **SIAM Review.** 30:421-449, 1988.

EFRON, B. & GONG, G. *A leisurely look at the bootstrap, the jackknife and cross-validation.* **The American Statistician.** 37:36-48, 1983.

EFRON, B. & TIBSHIRANI, R. *Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy (with discussion).* **Statistical Science.** 1:54-57, 1986.

EFRON, B. & TIBSHIRANI, R. **An introduction to the bootstrap.** New York, London, Chapman & Hall, 1993.

FREEDMAN, D. A. *Bootstrapping regression models.* **The Annals of Statistics.** 9:1218-1228, 1981.

FREEDMAN, D. A. & PETERS, S. C. *Bootstrapping an econometric model : some empirical results.* **Journal of the Business & Economic Statistics.** 2:150-158, 1984a.

FREEDMAN, D. A. & PETERS, S. C. *Bootstrapping a regression equation : some empirical results.* **Journal of the American Statistical Association.** 81:97-106, 1984b.

HALL, P. **The bootstrap and Edgeworth expansion.** New York, Berlin, heidelberg, London, Paris, Tokyo, Hong Kong, Barcelona, Budapest, Springer-Verlag, 1992.

HALL, P.; DICICCIO, T. J. AND ROMANO, J. P. *On smoothing and the bootstrap.* **The Annals of Statistics.** 17:692-704, 1989.

HAND, M. L. *Bootstrap resampling for regression models under non-least-squares estimation criteria.* **American Statistical Association (ASA) - Proceedings of Statistics Computing Section.** 172-176, 1986.

HINKLEY, D. *On estimating a symmetric distribution.* **Biometrika**. 63:680-681, 1976.

HINKLEY, D *Bootstrap methods (with discussion).* **Journal of the Royal Statistical Society, B**. 50:321-337, 1988.

HUBER, P. J. **Robust statistics**. New York, John Wiley & Sons, 1981.

JOHNSON, N. L. & KOTZ, S. **Continuous univariate distributions 2**. Boston, Houghton Mifflin, 1970.

LÉGER, C.; POLITIS, D. N. AND ROMANO, J. P. *Bootstrap technology and applications.* **Technometrics**. 34:378-398, 1992.

LEPAGE, R. & BILLARD, L. (Eds.) **Exploring limits of Bootstrap**. New York, Chichester, Brisbane, Toronto, John Wiley and Sons, 1992.

MAMMEN, E. *When does bootstrap work ? asymptotic results and simulations.* 77, Springer-Verlag, 1992.

MILLER, R. G. *The jackknife - a review.* **Biometrika**. 61:1-17, 1974.

MONTGOMERY, D. C. & PECK, E. A. **Introduction to linear regression analysis**. New York, Chichester, Brisbane, Toronto, Singapore, John Wiley & Sons, 1982.

MOOD, A. M.; GRAYBILL, F. A. AND BOES, D. C. **Introduction to the theory of Statistics**. Third Edition. Auckland, Singapore, McGRAW-HILL, 1974.

PETERS, S. C. & FREEDMAN, D. A. *Some notes on the bootstrap in regression problems.* **Journal of the Business & Economic Statistics**. 2:406-409, 1984.

TIBSHIRANI, R. *Bootstrap computations.* **Proceedings of the SAS® users Groups International Conference**. 10:1059-1063, 1985.

SAS Institute Inc. *SAS® User's Guide : Statistics, version 5 Edition*, Cary, NC : SAS Institute Inc., 1986.

SCHRADER, R. M. & MCKEAN, J. W. *Small sample properties of least absolute errors analysis of variance.* **Statistical Data Analysis Based on the L_1-norm and Related methods (Y. Dodge, Ed.)**. North-Holland, Netherland, 307-321, 1987.

SCHUSTER, E. F. *Estimating the distribution function of a symmetric distribution.* **Biometrika**. 62:631-635, 1975.

SEN, A. & SRIVASTAVA, M. **Regression analysis : theory, methods, and aplications**. New York, Berlin, heidelberg, London, Paris, Tokyo, Hong Kong, Springer-Verlag, 1990.

SILVERMAN, B. W. & YOUNG, G. A. *The bootstrap : to smooth or not smooth ?.* **Biometrika**. 74:469-479, 1987.

SHORACK, G. R. *Bootstrapping robust regression.* **Communications in Statistics-Theory and Methods**. 11:961-972, 1982.

STANGENHAUS, G. & NARULA, S. C. *Inference procedures for the L_1 regression.* **Computational Statistics and Data Analysis**. 12:79-85, 1991.

STANGENHAUS, G.; NARULA, S. C. AND FILHO, P. F. *Bootstrap confidence intervals for the minimum sum of absolute errors regression.* **Journal of Statistical Computation and Simulation**. 48:127-133, 1993.

STINE, R. *An Introduction to Bootstrap Methods.* **Sociological Methods & Research.** 18:243-291, 1989.

WEBER, N. C. *On Resampling Techniques For Regression Models.* **Statistics & Probability Lettters.** 2:275-278, 1984.

WEBER, N. C. *On The Jackknife and Bootstrap Techniques For Regression Models.* **Proceedings of the Pacific Statistical Congress.** Auckland (Francis, I. S., Manly, B. F. J. and Lam, F. C., eds), North-Holland, Groningen, pp. 51-55, 1986.

WU, F. J. *Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis.* **The Annals of Statistics.** 4:1261-1295, 1986.