Informes
de
Matemática

Série D-048-Agosto/92

A AUSÊNCIA DE RESPOSTA EM PESQUISAS: UMA APLICAÇÃO DE MÉTODOS DE IMPUTAÇÃO

Sonia Albieri

impa

INSTITUTO DE MATEMÁTICA PURA E APLICADA

Os artigos contidos nos INFORMES DE MATEMÁTICA DO IMPA, são "preprints", com tiragem limitada, de trabalhos distribuídos nas seguintes séries:

SÉRIE A – MATEMÁTICA PURA (artigos de pesquisa nas áreas de Matemática);
SÉRIE B – MATEMÁTICA APLICADA (artigos de pesquisa nas áreas de aplicação da Matemática, tais como: Estatística, Pesquisa Operacional, Economia Matemática, etc.);
SÉRIE C – SEMINÁRIOS DO IMPA;
SÉRIE D – DISSERTAÇÕES DE MESTRADO – Matemática Aplicada;
SÉRIE E – DISSERTAÇÕES DE MESTRADO – Matemática Pura;
SÉRIE F – TESES DE DOUTORADO;
SÉRIE G – RELATÓRIOS TÉCNICOS.

Conselho Nacional de Desenvolvimento Científico e Tecnológico
Instituto de Matemática Pura e Aplicada

INFORMES DE MATEMÁTICA

Série D-048/92

A AUSÊNCIA DE RESPOSTA EM PESQUISAS: UMA APLICAÇÃO DE MÉTODOS DE IMPUTAÇÃO

Sonia Albieri

Rio de Janeiro
Agosto/92

CONSELHO NACIONAL DE DESENVOLVIMENTO CIENTÍFICO E TECNOLÓGICO

INSTITUTO DE MATEMÁTICA PURA E APLICADA

A AUSÊNCIA DE RESPOSTA EM PESQUISAS: UMA APLICAÇÃO DE MÉTODOS DE IMPUTAÇÃO.

Sonia Albieri

Dissertação apresentada para obtenção do Grau de Mestre em Estatística

Rio de Janeiro
-1989-

A Ausência de Resposta em Pesquisas: Uma Aplicação de Métodos de Imputação

Autor: Sonia Albieri
Orientador: Wilton de Oliveira Bussab - (IME/USP)

novembro de 1989

Agradecimentos

- a José Paulo Quinhões Carneiro, chefe e amigo, por ter acreditado sempre e viabilizado a execução completa do programa de mestrado.
- a Wilton de Oliveira Bussab pela orientação firme, apesar da distância.
- ao professor Djalma Galvão Carneiro Pessoa pela confiança depositada.
- a Vandeli, companheira do IBGE e amiga, pelo apoio na obtenção não só dos dados como também das condições de trabalho, em todos os momentos.
- a Francisco, companheiro do IBGE e amigo, pelo contínuo apoio computacional.
- a Rosana, companheira do IMPA e amiga, pelo apoio computacional na utilização do OSIRIS.
- a Berta, companheira de trabalho, pela ajuda na organização das referências bibliográficas.
- a Theresinha e Denise pela boa vontade e excelente dati

5 **Uma Aplicação** 50
5.1 Introdução 50
5.2 A seleção das variáveis auxiliares 51
5.3 A geração dos dados ausentes 59
5.4 Os procedimentos computacionais 59
5.4.1 Na seleção das variáveis auxiliares 59
5.4.2 Nos procedimentos de imputação e análise 60
5.5 Os resultados 60
5.6 Os comentários 70
5.7 Conclusões 73
6 **Considerações Finais** 74
Bibliografia 76
Anexo 1 - Proposição A1 78
Anexo 2 - Proposição A2 79
Anexo 3 - Proposição A3 81
Anexo 4 - Definições do termo "hot deck" segundo alguns autores 83
Anexo 5 - Proposição A5 85
Anexo 6 - Proposição A6 86
Anexo 7 - Proposição A7 87
Anexo 8 - Proposição A8 88
Anexo 9 - Proposição A9 89
Anexo 10 - Proposição A10 90
Anexo 11 - Proposição A11 92
Anexo 12 - Os modelos MAR e MARC 95
Anexo 13 - Programa de aplicação da técnica SEARCH, do pacote OSIRIS IV..... 96
Anexo 14 - Programa de geração de dados ausentes e de aplicação de métodos de imputação 97
Anexo 15 - Resultados da aplicação na amostra original, com $\bar{m}$ = 1,46% 107
Anexo 16 - Resultados da aplicação na amostra simulada, com $\bar{m}$ = 4,90%..... 115
Anexo 17 - Resultados da aplicação na amostra simulada, com $\bar{m}$ = 9,85% 123
Anexo 18 - Resultados da aplicação na amostra simulada, com $\bar{m}$ = 18,67% 131

Apresentação

Na teoria da amostragem, a maior parte dos esquemas considera seus estimadores associados baseados no tamanho da amostra fixado a priori. Entretanto, a ausência de informações parciais ou totais de unidades indicadas para pertencerem à amostra é fato bastante comum na prática das pesquisas. Tem-se o problema já conhecido de não resposta e a questão de como contornar esse problema durante o processo de estimação das variáveis investigadas.

Este é um trabalho sobre métodos de imputação, que constituem um dentre os diversos tipos de tratamento de dados ausentes, e que vêm sendo utilizados em pesquisas realizadas pelos organismos responsáveis pelas estatísticas governamentais de vários países, inclusive o Brasil.

1. Introdução

1.1 Apresentação do problema

Grandes conjuntos de dados estão sujeitos a dados ausentes (*missing data*) por várias razões, tais como a não resposta à investigação e a eliminação de valores considerados inconsistentes durante os procedimentos de crítica e de validação das informações.

Muitas análises estatísticas multivariadas, incluindo regressão, análise fatorial e análise discriminante, são baseadas em uma redução inicial dos dados para o vetor amostral de médias e para a matriz de covariância amostral das variáveis. A questão de como estimar esses momentos com dados ausentes em uma ou mais variáveis é largamente tratada na literatura. Para o caso de variáveis categóricas, geralmente deseja-se mais do que medidas resumo da distribuição e os dados são então expressos na forma de tabelas de contingência. Muitos dos trabalhos sobre tabelas de contingência incompletas referem-se à estimação de máxima-verossimilhança supondo distribuição multinomial ou Poisson. Em ambos os casos, os trabalhos mais recentes referem-se ao algoritmo E.M. (Little & Smith (1987) e Peres (1987)).

Muitos dos problemas de dados ausentes em pesquisas por amostra são de natureza multivariada, uma vez que as pesquisas têm múltiplos propósitos e investigam uma grande quantidade de variáveis, sendo que os métodos para tratar dados incompletos multivariados têm sido muito pouco aplicados a tais problemas.

Podemos distinguir dois tipos de dados ausentes em pesquisas dessa natureza: um que se caracteriza pela não resposta de uma unidade inteira da amostra selecionada e consiste no chamado "questionário perdido" ou "não resposta da unidade"; e o outro tipo que se caracteriza pela não resposta a um item contido no questionário, ou seja, a uma das variáveis investigadas e consiste no chamado "item perdido" ou "não resposta ao item".

Uma classificação dos problemas devidos à não resposta em pesquisa encontra-se em Kalsbeek (1980). Com o objetivo de caracterizar conceitualmente a literatura existente sobre erros devidos à não resposta, Kalsbeek define seis dimensões, cada uma delas subdividida em níveis, descritas brevemente a seguir:

dimensão 1 - Visão conceitual da não resposta

. modelo determinístico - nessa visão, supõe-se que a população é constituída de dois estratos: um, cujos elementos participarão da pesquisa com certeza, se selecionados; e o outro, cujos elementos não fornecerão informações, se selecionados;

. modelo estocástico - nesse caso, a população é vista como um conjunto de elementos, cada um dos quais com uma dada probabilidade de resposta, se selecionados.

dimensão 2 - Tipo de parâmetro em estudo

. média;

. total; e

. outros.

dimensão 3 - Efeito do vício de não resposta

. impacto;

. extensão; e

. diferenças entre respondentes e não respondentes.

dimensão 4 - Nível de não resposta

. elemento ou unidade; e

. item ou quesito.

dimensão 5 - Fonte de não resposta

. não abrangência;

. não localização;

. recusa;

. incapacidade; e

. outros.

dimensão 6 - Tratamento de não resposta

1. métodos preventivos;
2. métodos corretivos ou "post hoc";
   a) estudos de identificação;
   b) métodos de imputação direta;
   c) métodos de imputação indireta;
   d) outros métodos de ajuste;
   e) métodos de substituição;
   f) métodos de subamostragem de não respondentes;
   g) métodos bayesianos; e

h) métodos de extrapolação.

Little e Rubin (1983) apresentam um histórico do problema de dados incompletos, desde o seu aparecimento na literatura referente aos experimentos planejados na agricultura, sob o enfoque das suposições acerca da distribuição dos valores, e em particular, de como as distribuições dos valores observados e dos ausentes se relacionam. Os autores classificam os métodos utilizados para tratar dados incompletos em pesquisas em uma ou mais das seguintes categorias:

i) eliminação das unidades parcialmente observadas - é um método comum que despreza as unidades com dados ausentes em algumas variáveis;

ii) procedimentos de ponderação - as inferências em pesquisas por amostra sem não resposta são comumente baseadas em pesos do desenho, os quais são inversamente proporcionais às probabilidades de seleção. Os procedimentos de ponderação modificam os pesos para considerar a não resposta, por meio de uma estimativa da probabilidade de resposta. Usualmente, a probabilidade de resposta é estimada por uma proporçao de unidades respondentes em uma subclasse da amostra, chamada de classe de ponderação ou célula de ajuste. Uma vez que os pesos são definidos para unidades, os métodos de ponderação são comumente aplicados para tratar a não resposta total (ou não resposta da unidade), onde todos os itens da pesquisa são ausentes por causa de não contato ou recusa da unidade em fornecer as informações. A única informação disponível para a formação das classes de ponderação vem das variáveis de planejamento da pesquisa;

iii) métodos de imputação - os valores ausentes são preenchidos e os dados resultantes são analisados por meio de métodos para dados completos. Em métodos distintos de imputação, é útil decompor o valor ausente y_{mi} na forma $y_{mi} = f(\underset{\sim}{z}) + e_{mi}$ onde $f(\underset{\sim}{z})$ é uma função linear de variáveis auxiliares $\underset{\sim}{z}$ e e_{mi} é um resíduo, que pode ser estimado, por exemplo, a partir dos dados das unidades respondentes;

iv) procedimentos baseados em modelos - essa classe de procedimentos caracteriza-se pela suposição de um modelo para os dados incompletos e, então, os parâmetros são estimados por procedimentos tais como a máxima verossimilhança.

As categorias consideradas por Litlle e Rubin (1983) relacionam-se com as dimensões de Kalsbeek (1980) da seguinte maneira: em função da dimensão 1, (i), (ii) e (iii) são métodos que assumem modelo determinístico e (iv) assume modelos estocásticos; com relação à dimensão 6, (i) e (ii) são métodos de imputação indireta e (iii) engloba métodos de imputação direta.

Na dimensão 3 - efeito do vício de não resposta, Kalsbeek considera o nível extensão como a quantificação da não resposta (ou da resposta), e que usualmente é medida através de taxas de resposta e taxas de não resposta. Platek e Gray (1986) tratam especificamente das definições de diversas taxas de resposta,em função da fonte de não resposta e do nível de não resposta (dimensões 5 e 4, de Kalsbeek, respectivamente).

Kalton e Kasprzyk (1982) apresentam uma revisão de vários métodos de imputação comumente usados. Na classificação de Kalsbeek, esses métodos situam-se da seguinte maneira:

dimensão 1 - modelo determinístico;
dimensão 2 - média populacional;
dimensão 3 - impacto do vício de não resposta;
dimensão 4 - não resposta a item;
dimensão 5 - recusa, incapacidade e outros;
dimensão 6 - métodos de imputação direta.

1.2 Descrição do trabalho

Este item apresenta os objetivos deste trabalho bem como os caminhos percorridos para alcançá-los.

Primeiramente, é importante ressaltar a definição de imputação a ser adotada ao longo deste trabalho. Por método de imputação entende-se todo e qualquer procedimento de tratamento de dados ausentes que atribui um valor específico individual ao dado ausente, seja ele do tipo item perdido ou unidade perdida.Ou seja, a imputação é um processo de estimação de valores individuais em um conjunto de dados.

Os objetivos deste trabalho podem ser descritos como:

1. conhecer os métodos de imputação comumente utilizados em pesquisas, suas propriedades e principais aplicações.
2. realizar uma aplicação de alguns desses métodos em uma mesma situação real de não resposta de uma pesquisa, com o objetivo de reconhecer um método que melhor se adapte ao problema específico, baseando-se nos resultados teóricos encontrados.
3. buscar evidências da influência do tamanho da não resposta no comportamento de alguns desses métodos quando efetivamente aplicados a um conjunto de dados.

Nesse sentido, o artigo de Kalton e Kasprzyk (1982) citado no item anterior abre o caminho para a consecução do primeiro objetivo, na medida em que apresenta resumidamente 10 métodos de imputação. Esses métodos estão apresentados no capítulo 2 deste trabalho.

O capítulo 3 contém, para seis métodos de imputação selecionados para análise, a apresentação teórica dos estimadores de média decorrentes do uso de imputação e dos erros associados a esses estimadores, vícios e variâncias.

O capítulo 4, ainda visando atingir o primeiro objetivo, contém uma apresentação resumida dos resultados teóricos e empíricos do estudo realizado por Santos (1981a), com os mesmos métodos de imputação considerados no capítulo 3, enfocando estimadores de variância e covariância de variáveis sujeitas a imputação.

O capítulo 5 descreve e analisa os resultados de uma aplicação de quatro métodos de imputação a uma variável com dados ausentes de uma pesquisa real, a Pesquisa Mensal de Emprego do IBGE. Nesse mesmo capítulo são apresentados três exercícios de geração de dados ausentes sobre essa mesma pesquisa, a fim de, com diferentes taxas de não resposta, atingir o 3º objetivo descrito acima. Esses exercícios devem ser considerados apenas como uma ilustração dos métodos apresentados e discutidos. O objetivo aqui não foi realizar simulações, quando então técnicas e procedimentos apropriados deveriam ser considerados.

Finalmente, o capítulo 6 contém as considerações finais sobre o trabalho realizado.

O trabalho contém ainda uma grande quantidade de anexos, são 18 ao todo, cujo objetivo é facilitar a leitura do texto, eliminando de seu corpo principal as partes mais detalhadas.

2. Métodos de Imputação

2.1 Notação

Seja $\mathcal{P}$ uma população finita de N unidades compostas de duas subpopulações: uma de respondentes $\mathcal{R}$ e uma de não respondentes $\mathcal{M}$. Sejam R e M (R + M = N) os números de unidades nas subpopulações de resposta e de não resposta, respectivamente. Suponhamos que N, R e M são grandes.

Suponhamos que uma amostra aleatória simples sem reposição A_n de tamanho n é retirada dessa população e um conjunto de variáveis é observado, digamos Y, X, e $\underset{\sim}{Z} = (Z_1, Z_2, \ldots, Z_k)$. Os dados para X e $\underset{\sim}{Z}$ são observados para todas as unidades da amostra; enquanto que os dados para a variável Y são obtidos para r entre as n unidades. Suponhamos $P(r > 1) = 1$. Então, temos m = (n-r) unidades com valores ausentes para a variável Y.

Ou seja, podemos decompor a amostra A_n em duas partes $A_n = A_m \cup A_r$, onde A_r é composta pelos indivíduos que responderam todas as variáveis e A_m é composta pelos indivíduos que responderam X e $\underset{\sim}{Z}$ e não responderam Y.

Com base nessas suposições podemos definir:

$$\bar{R} = R/N$$

$$\bar{M} = M/N$$

$$\bar{R} + \bar{M} = 1$$

$$\bar{Y} = \sum_{i=1}^{N} Y_i/N \qquad \bar{X} = \sum_{i=1}^{N} X_i/N$$

$$\bar{Y}_r = \sum_{i=1}^{R} Y_{ri}/R \qquad \bar{X}_r = \sum_{i=1}^{R} X_{ri}/R$$

$$\bar{Y}_m = \sum_{i=1}^{M} Y_{mi}/M \qquad \bar{X}_m = \sum_{i=1}^{M} X_{mi}/M$$

$$S^2_y = \sum_{i=1}^{N} (Y_i - \bar{Y})^2 / (N-1)$$

$$S^2_{ry} = \sum_{i=1}^{R} (Y_{ri} - \bar{Y}_r)^2 / (R-1)$$

$$S^2_{my} = \sum_{i=1}^{M} (Y_{mi} - \bar{Y}_m)^2 / (M-1)$$

$$S_{xy} = \sum_{i=1}^{N} (Y_i - \bar{Y})(X_i - \bar{X}) / (N-1)$$

$$S_{rxy} = \sum_{i=1}^{R} (Y_{ri} - \bar{Y}_r)(X_{ri} - \bar{X}_r) / (R-1)$$

$$S_{mxy} = \sum_{i=1}^{M} (Y_{mi} - \bar{Y}_m)(X_{mi} - \bar{X}_m) / (M-1)$$

e,

Y_i e X_i - valores das variáveis Y e X, respectivamente, da i-ésima unidade na população;

Y_{ri} e X_{ri} - valores das variáveis Y e X, respectivamente, da i-ésima unidade na subpopulação de respondentes;

Y_{mi} e X_{mi} - valores das variáveis Y e X, respectivamente. da i-ésima unidade na subpopulação de não respondentes;

Observam-se as seguintes relaçoes:

$$\bar{Y} = \bar{R}\ \bar{Y}_r + \bar{M}\ \bar{Y}_m \qquad (1)$$

$$S^2_y \doteq \bar{R}\ S^2_{ry} + \bar{M}\ S^2_{my} + \bar{R}\ \bar{M}\ (\bar{Y}_r - \bar{Y}_m)^2 \qquad (2)$$

$$S_{xy} \doteq \bar{R}\ S_{rxy} + \bar{M}\ S_{mxy} + \bar{R}\ \bar{M}\ (\bar{Y}_r - \bar{Y}_m)(\bar{X}_r - \bar{X}_m) \qquad (3)$$

Os anexos 1, 2 e 3 apresentam, respectivamente, a decomposição de $\bar{Y}$, S^2_y e S_{xy} em função das subpopulações de respondentes e de não respondentes. Analogamente, para a amostra, temos:

$$\bar{y} = \bar{r}\ \bar{y}_r + \bar{m}\ \bar{y}_r \qquad (4)$$

$$s^2_y \doteq \bar{r}\ s^2_{ry} + \bar{m}\ s^2_{my} + \bar{r}\ \bar{m}\ (\bar{y}_r - \bar{y}_m)^2 \qquad (5)$$

$$s_{xy} \doteq \bar{r}\ s_{rxy} + \bar{m}\ s_{mxy} + \bar{r}\ \bar{m}\ (\bar{y}_r - \bar{y}_m)(\bar{x}_r - \bar{x}_m) \qquad (6)$$

2.2 Breve Revisão

Quando ocorre uma não resposta a item, uma quantidade substancial de informações acerca dos não respondentes é normalmente disponível de outros itens do questionário. Ou seja, podemos enquadrar a situação de não resposta a item às suposições teóricas elaboradas em 2.1.

Muitos métodos de imputação usam um subconjunto desses itens como variáveis auxiliares para a atribuição de valores para as respostas ausentes. Em geral, o valor imputado para o i-ésimo não respondente para o item y pode ser descrito por

$$y_{mi} = f(z_{1i}, z_{2i} \ldots, z_{ki}) + e_{mi}$$

onde $f(\underset{\sim}{z})$ é uma função de variáveis auxiliares $\underset{\sim}{z}$ e e_{mi} é um resíduo estimado. Freqüentemente, $f(\underset{\sim}{z})$ pode ser expressa como uma função linear do tipo $\beta_o + \sum_{j=1}^{k} \beta_j z_{ji}$ onde os βs podem ser estimados a partir dos dados dos respondentes. Ou seja, o modelo suposto é:

$$y_i = \beta_o + \sum_{j=1}^{k} \beta_j z_{ji}$$

cujos parâmetros podem ser estimados em A_r e essas estimativas utilizadas em A_m. Então,

$$y_{ri} = b_{ro} + \sum_{j=1}^{k} b_{rj} z_{rji} \qquad \text{em } A_r$$

$$y_{mi} = b_{ro} + \sum_{j=1}^{k} b_{rj} z_{mji} \qquad \text{em } A_m$$

A consideração principal ao escolher as variáveis auxiliares é a capacidade que elas têm de predizer o valor y ausente. Dependendo da suposição feita, o resíduo pode ser determinado de várias formas tais como:

(1) supondo a característica observada normalmente distribuída e existência de homocedasticidade, um resíduo pode ser escolhido aleatoriamente de uma distribuição normal com média zero e variância igual à variância dos resíduos;

(2) se a suposição é de que os resíduos são provenientes de uma mesma distribuição não especificada, eles podem ser selecionados aleatoriamente da distribuição empírica dos resíduos dos respondentes.

Os métodos de imputação podem ser classificados de duas formas: pelo uso de variáveis auxiliares e pelo valor atribuído ao resíduo.

Alguns métodos não fazem uso de variáveis auxiliares. Outros tratam-nas como variáveis categóricas, classificando os indivíduos da amostra dentro de classes de imputação de acordo com a combinação das respostas a essas variáveis; variáveis auxiliares contínuas, tais como idade ou renda, são categorizadas para usar esse método. Outros métodos tratam todas as variáveis como contínuas, com algumas variáveis categóricas sendo tratadas como variáveis fictícias.

A segunda forma relaciona-se ao uso ou não do processo aleatório na atribuição de valores imputados. Um método de imputação é denominado estocástico quando o termo residual e_{mi} é aleatoriamente atribuído e como determinístico quando ele sempre assume o valor zero.

Os itens seguintes descrevem brevemente alguns procedimentos de imputação em ambas as classes.

(a) imputação dedutiva: esse método depende de alguma redundância dos dados, tal que uma resposta ausente possa ser deduzida de uma informação auxiliar, isto é, $y_{mi} = f(z_i)$ exatamente. Por exemplo, se um registro contém uma série de parcelas e o seu total, a ausência de informação de uma parcela pode ser deduzida por subtração. O método pode ser estendido para situações onde o valor deduzido é altamente semelhante ao valor correto ou pelo menos muito perto dele ; por exemplo, em uma pesquisa via painel com uma variável que permanece quase constante ao longo do tempo, uma resposta ausente de uma etapa da pesquisa pode ser imputada com o valor registrado para o item na etapa precedente ou posterior.

(b) imputação da média geral: esse método atribui a média geral dos respondentes, $\bar{y}_r$. para todas as respostas ausentes. É a forma determinística da função linear sem variáveis auxiliares, ou seja, $y_{mi} = b_{ro} = \bar{y}_r$.

(c) imputação aleatória geral: esse método atribui a cada não respondente o valor y de um respondente selecionado aleatoriamente da amostra total de respondentes. É a forma estocástica da função linear sem variáveis auxiliares, ou seja, $y_{mi} = \bar{y}_r + e_{mi}$, com $e_{mi} = y_{rk} - \bar{y}_r$, que se reduz a $y_{mi} = y_{rk}$, para algum k.

(d) imputação da média dentro de classes: esse método considera a amostra total dividida em classes de imputação de acordo com valores das variáveis auxiliares. Dentro de cada classe, a média dos respondentes para a variável y é atribuída a todos os não respondentes dessa classe, ou seja, $y_{mhi} = \bar{y}_{rh}$ para o i-ésimo não respondente na classe h (h=1,...,H) . As classes podem ser definidas como todas as celulas de uma tabulação cruzada de variáveis categóricas auxiliares; ao invés disso, algumas variáveis auxiliares podem ser usadas para uma parte da amostra, enquanto outras são usadas para outra parte.Se todas as células de uma tabela são usadas, a função linear pode ser expressa como um modelo com o efeito principal e todos os níveis de interação para as variáveis auxiliares. Em geral o modelo pode ser representado por $y_{mi} = b_{ro} + \sum_{j=1}^{k} b_{rj} z_{jmi}$, onde z_{ji} são variáveis fictícias. $z_{jmi} = 1$ se o i-ésimo não respondente esta na classe j e $z_{jmi} = 0$ caso contrario. Uma vez que $e_{mi} = 0$, o método é determinístico.

(e) imputação aleatória dentro de classes: esse método corresponde ao da imputação aleatória geral, exceto pelo fato que ele é aplicado dentro de classes. A cada não respondente é atribuído o valor y de um respondente selecionado aleatoriamente da mesma classe de imputação. O método é o estocástico equivalente ao método da média dentro de classes, com $Y_{mhi} = \bar{y}_{rh} + e_{mhi}$ e $e_{mhi} = y_{rhk} - \bar{y}_{rh}$, reduzindo-se a $y_{mhi} = y_{rhk}$, para algum k pertencente à classe h. Uma expressão alternativa é $y_{mji} = b_{ro} + \sum_{j=1}^{k} b_{rj} z_{jmi} + e_{mji}$ onde e_{mji} é um resíduo de respondente selecionado aleatoriamente dentro da classe de imputação j na qual o não respondente i é classificado. É também denominado de "hot deck" aleatório(1).

(f) imputação da média da regressão: esse método usa os dados dos respondentes para regredir y sobre as variáveis auxiliares. Os valores ausentes de y são então imputados pelo valor predito da equação de regressão $y_{mi} = b_{ro} + \sum_{j=1}^{k} b_{rj} z_{jmi}$. Esse é o método determinístico, com $e_{mi} = 0$. As variáveis auxiliares podem ser quantitativas ou qualitativas, as últimas sendo incorporadas por meio de variáveis fictícias. Se a variável y é qualitativa, modelos log-lineares ou logísticos podem ser usados. Como em qualquer análise de regressão, termos específicos de interação podem ser incluídos na equação de regressão e transformações nas variáveis podem ser úteis. Um caso especial de modelo de regressão é o modelo da razão $y_{mi} = b_r z_{mi}$ com uma única variável auxiliar e intercepto zero.

(g) imputação por regressão mais efeito de aleatoriedade: esse método é a versão estocástica do anterior; os valores imputados são os valores preditos da equação de regressão mais um resíduo e_{mi}.

---

(1) Outras definições do termo "hot deck" encontram-se no anexo 4.

(h) imputação "hot deck" seqüencial: o procedimento considera a amostra total dividida em classes de imputação, e para cada classe atribui um único valor para a variável y para dar início ao processo. Esse valor inicial pode ser obtido por meio de um valor de um respondente de cada classe, ou pelo valor médio na classe de uma pesquisa prévia. Os registros da pesquisa são então tratados seqüencialmente. Se o registro tem o valor para a variável y, este valor substitui o inicial para a classe a que pertence. Se o valor de y é ausente, ele é imputado pelo valor inicial da classe. A principal atração desse procedimento é a sua economia computacional, uma vez que todas as imputações são feitas com uma única leitura do arquivo de dados. Se a ordem dos registros no arquivo de dados for aleatória, o método é semelhante ao método aleatório dentro de classes, exceto pelo início do processo. O procedimento "hot deck" seqüencial geralmente se beneficia da ordem não aleatória do arquivo de dados, uma vez que o uso de um doador precedente na classe de imputação propicia um grau de associação adicional, que é vantajoso se a ordem do arquivo cria autocorrelação positiva. Esse benefício dilui-se, entretanto, quando as classes de imputação são pequenas e as unidades estão espalhadas por todo o arquivo. A desvantagem desse método é que pode facilmente gerar o uso múltiplo de um doador,o que leva a uma perda de precisão para os estimadores da pesquisa. Isso ocorre quando, dentro de uma classe de imputação, um registro com valor ausente é seguido de outros registros com a mesma resposta ausente; então,todos esses registros terão o mesmo valor imputado proveniente do último respondente da classe.

(i) imputação por associação flexível: esse termo é usado para o procedimento "hot deck" modificado que vem sendo usado desde 1976 pelo Bureau of the Census dos Estados Unidos no Suplemento de Renda da Current Population Survey - CPS, aplicado em todos os meses de março. O procedimento classifica respondentes e não respondentes em um grande número de classes de imputação, construídas de uma detalhada categorização de um conjunto relativamente grande de variáveis auxiliares. Não respondentes são então associados com respondentes de uma forma hierárquica, no sentido que se um não respondente não pode ser associado com um respondente na classe inicial (mais detalhada), as classes são agrupadas e a associação é feita em um outro nível maior. Três níveis são usados na CPS, de tal forma que o maior nível (o menos detalhado) é tal que a associação sempre ocorre. Ele também evita o uso múltiplo de doadores em classes onde o número de respondentes excede o de não respondentes.

(j) associação por função distância: esse método atribui o valor y do respondente mais próximo a cada não respondente, com o "mais próximo" definido por uma função de distância das variáveis auxiliares. O método refere-se principalmente a variáveis quantitativas; entretanto, variáveis qualitativas podem ser incluídas seja pelo uso de função de distância dentro de classes de imputação formadas por essas variáveis qualitativas, seja pela incorporação dessas variáveis na função de distância. Com uma unica variavel auxiliar, a amostra pode ser ordenada pela variável e o respondente vizinho mais próximo a cada não respondente é tomado para ser o doador, onde o mais próximo pode ser definido como a diferença absoluta mínima entre os valores do doador e do não respondente, da

variável auxiliar ou de alguma transformação dela. Quando várias variáveis auxiliares são usadas, a questão das transformações torna-se mais crítica; uma técnica é transformar todas as variáveis auxiliares em seus postos. Então, uma função distância proposta é dada por $D(i,k) = Sup_h W_h |R_{hi} - R_{hk}|$, onde R_{hi} e R_{hk} são os postos do não respondente i e do doador em potencial k, para a variável h, e W_h é o peso representando a importância da variável h na função distância (Sande, I (1983)). Outra técnica é a baseada na distância de Mahalanobis.

3. Métodos de Imputação Selecionados para Estudo

Para o estudo em questão, foram escolhidos os seguintes métodos:

(i) imputação da média geral;
(ii) imputação aleatoria geral;
(iii) imputação da média dentro de classes;
(iv) imputação aleatória dentro de classes;
(v) imputação da média da regressão;
(vi) imputação por regressão mais um efeito de aleatoriedade gerado pela própria distribuição empírica dos resíduos.

Os métodos(i), (iii) e (v) são determinísticos e os métodos(ii), (iv) e (vi) são estocásticos, no sentido definido no item 2.2.

Os itens 3.1, 3.2 e 3.3, a seguir, apresentam, respectivamente, os estimadores de média, seus vícios e suas variâncias, obtidos dos dados completados por meio dos métodos de imputação considerados.

3.1. Os estimadores de média

Os estimadores de média decorrentes da utilização dos diversos métodos de imputação são apresentados a seguir, em função da expressão (4) da parte 2.1, qual seja,

$$\bar{y} = \bar{r}\bar{y}_r + \bar{m}\bar{y}_m \qquad (4)$$

onde $\bar{y}_m$ é a parte desconhecida.

3.1.1. Imputação da média geral-$\bar{y}_{MO}$

Conforme o item(b) da parte 2.2, temos:

$$y_{miMO} = \bar{y}_r \qquad \forall mi \in A_m$$

então, $\bar{y}_{mMO} = \bar{y}_r$

e substituindo em (4),

$$\bar{y}_{MO} = \bar{r}\bar{y}_r + \bar{m}\bar{y}_r = \bar{y}_r$$

3.1.2. Imputação aleatória geral-$\bar{y}_{RO}$

Conforme o item (c) da parte 2.2, temos:

$$y_{miMO} = \bar{y}_r + e_{miRO} \qquad \forall\, mi \in A_m$$

onde $e_{miRO} = e_{rk} = y_{rk} - \bar{y}_r$ para algum $rk \in A_r$

então, $\bar{y}_{mRO} = \bar{y}_r + \bar{e}_{mRO}$ onde $\bar{e}_{mRO} = \sum_{i=1}^{m} e_{miRO}/m$

e substituindo em (4),

$\bar{y}_{RO} = \bar{r}\bar{y}_r + \bar{m}\bar{y}_{mRO} = \bar{y}_r + \bar{m}\bar{e}_{mRO}$

Neste método supõe-se:

$e_{miRO} \sim f_{ry}\ (0, \frac{r-1}{r}\ s^2_{ry})$ onde

f_{ry} é a distribuição na amostra da variável y para os respondentes, ou seja, associada a A_r.

3.1.3. Imputação da média dentro de classes-$\bar{y}_{MC}$

Seja $\mathcal{H}$ uma partição de A_n de tamanho H construída através das variáveis auxiliares $\underset{\sim}{Z}$, ou seja, $\mathcal{H} = \{ I_1\ I_2 \ldots I_H \}$

Conforme o item (d) da parte 2.2, temos:

$y_{mhiMC} = \bar{y}_{rh}$ $\forall\, mhi \in (I_h \cap A_m), I_h \in \mathcal{H}$

então, $\bar{y}_{mhMC} = \bar{y}_{rh}$ $\forall\ I_h \in \mathcal{H}$

e, $\bar{y}_{mMC} = \sum_{h=1}^{H} w_{mh}\bar{y}_{rh}$ onde

$w_{mh} = \frac{m_h}{\bar{m}}$ e m_h é o tamanho de $(I_h \cap A_m)$

substituindo em (4),

$\bar{y}_{MC} = \bar{r}\bar{y}_r + \bar{m}\bar{y}_{mMC} = \bar{r}\bar{y}_r + \bar{m} \sum_{h=1}^{H} w_{mh}\bar{y}_{rh}$

3.1.4. Imputação aleatória dentro de classes-$\bar{y}_{RC}$

Conforme o item (e) da parte 2.2, temos:

$y_{mhiRC} = \bar{y}_{rh} + e_{mhiRC}$ $\forall\, mhi \in (I_h \cap A_m)$, $I_h \in \mathcal{H}$ tal como definido em 3.1.3

onde, $e_{mhiRC} = e_{rhk} = y_{rhk} - \bar{y}_{rh}$ para algum $rhk \in (I_h \cap A_r)$

então, $\bar{y}_{mhRC} = \bar{y}_{rh} + \bar{e}_{mhRC}$

onde $\bar{e}_{mhRC} = \sum_{i=1}^{m_h} e_{mhiRC}/m_h$ e m_h é o tamanho de $(I_h \cap A_m)$

e, $\bar{y}_{mRC} = \sum_{h=1}^{H} w_{mh}\ \bar{y}_{mhRC}$

substituindo em (4),

$$\bar{y}_{RC} = \bar{r}\bar{y}_r + \bar{m}\bar{y}_{mRC} = \bar{r}\bar{y}_r + \bar{m} \sum_{h=1}^{H} w_{mh}\ \bar{y}_{mhRC}$$

Neste método supõe-se:

$$e_{mhi} \sim f_{rhy}\ (0, \frac{r_h - 1}{r_h}\ s^2_{rhy})$$

onde f_{rhy} é a distribuição na amostra de y associada a $(I_h \cap A_r)$.

s^2_{rhy} é a variância na amostra de y calculada em $(I_h \cap A_r)$;

r_h é o tamanho de $(I_h \cap A_r)$.

3.1.5. Imputação da média da regressão-$\bar{y}_{MR}$

Conforme o item (f) da parte 2.2, temos:

$$y_{miMR} = b_{ro} + \sum_{j=1}^{k} b_{rj}\ z_{jmi} \qquad \forall\ mi \in A_m$$

onde z_{jmi} é o valor da variável Z_j para o indivíduo mi $mi \in A_m$; e

b_{rj}, j= 0.1.....k são os parâmetros estimados pelo ajuste do modelo de regressão múltipla efetuado com os elementos de A_r.

Então, $\bar{y}_{mMR} = b_{ro} + \sum_{j=1}^{k} b_{rj}\ \bar{z}_{jm}$ onde

$$\bar{z}_{jm} = \sum_{i=1}^{m} z_{jmi}/m \qquad , \; mi \in A_m$$

substituindo em (4),

$$\bar{y}_{MR} = \bar{r}\bar{y}_r + \bar{m}\bar{y}_{mMR} = \bar{r}\bar{y}_r + \bar{m}\left(b_{ro} + \sum_{j=1}^{k} b_{rj}\, \bar{z}_{jm} \right)$$

3.1.6. Imputação por regressão mais efeito de aleatoriedade- $\bar{y}_{RR}$

Conforme o item (g) da parte 2.2, temos:

$$y_{miRR} = b_{ro} + \sum_{j=1}^{k} b_{rj}\, z_{jmi} + e_{miRR} \qquad \forall \; mi \in A_m$$

onde $e_{miRR} = e_{rk} = y_{rk} - \hat{y}_{rk}$

e $\hat{y}_{rk} = b_{ro} + \sum_{j=1}^{k} b_{rj}\, z_{jrk}$ para algum $rk \in A_r$

Então, $\bar{y}_{mRR} = b_{ro} + \sum_{j=1}^{k} b_{rj}\, \bar{z}_{jm} + \bar{e}_{mRR}$

onde $\bar{e}_{mRR} = \sum_{i=1}^{m} e_{miRR}/m$

substituindo em (4),

$$\bar{y}_{RR} = \bar{r}\bar{y}_r + \bar{m}\bar{y}_{mRR} = \bar{r}\bar{y}_r + \bar{m}\left(b_{ro} + \sum_{j=1}^{k} b_{rj}\, \bar{z}_{jm} + \bar{e}_{mRR} \right)$$

Neste método supõe-se:

$$e_{mi} \sim f_{ry}\left(0, \frac{r-1}{r}\, s^2_{ry.\underset{\sim}{z}} \right)$$

onde f_{ry} é a distribuição amostral de y

associada a A_r.

e $s^2_{ry.\underset{\sim}{z}}$ é a variância residual de Y dadas as variáveis $\underset{\sim}{Z}$, calculada em A_r.

3.2. Os vícios dos estimadores de média

Primeiramente, podemos verificar que os estimadores de média baseados em métodos estocásticos de imputação têm o mesmo vício que aqueles baseados nos correspondentes métodos determinísticos.

Proposição 3.2.1

$$B(\bar{y}_{MO}) = B(\bar{y}_{RO})$$

$$B(\bar{y}_{MC}) = B(\bar{y}_{RC})$$

$$B(\bar{y}_{MR}) = B(\bar{y}_{RR})$$

onde $B(t) = E(t)-\bar{Y}$ é o vício do estimador de média,t.

Demonstração: sejam:

y_{mis} - o valor imputado para a unidade $mi \in A_m$, por meio de um método estocástico;

y_{mid} - o valor imputado para a unidade $mi \in A_m$, por meio do método determinístico correspondente ao estocástico;

e_{mi} - o resíduo estimado para a unidade $mi \in A_m$, selecionado aleatoriamente dentre os resíduos dos respondentes tal como definido em 3.1.2, 3.1.4 e 3.1.6.

$$\bar{y}_{ms} = \sum_{i=1}^{m} y_{mis}/m$$

$$\bar{y}_{md} = \sum_{i=1}^{m} y_{mid}/m$$

$\bar{y}_s$ - o estimador de média decorrente da utilização de um método estocástico;

$\bar{y}_d$ - o estimador de média decorrente da utilização do método determinístico correspondente ao estocástico.

Da expressão (4) do item 3.1 podemos escrever:

$$\bar{y}_s = \bar{r}\bar{y}_r + \bar{m}\bar{y}_{ms}$$

$$y_d = ry_r + my_{md}$$

e das definições acima podemos escrever:

$$y_{mis} = y_{mid} + e_{mis}$$

Se E_2 é a esperança condicional sobre a amostra dos resíduos, dada a amostra inicial e E_1 é a esperança sobre a amostra inicial, temos que $E = E_1\ E_2$. Uma apresentação formal dos operadores E_1 e E_2 encontra-se no anexo 5 e é devida a Raj, D. (1968).

Mas, $E_2(y_{mis}) = E_2(y_{mid} + e_{mi}) = y_{mid}$

pois $E_2\ (e_{mi}) = 0$

então, $E\ (y_{mis}) = E_1\ E_2\ (y_{mis}) = E_1(y_{mid}) = E(y_{mid})$

ou seja, $E(\bar{y}_{ms}) = E(\bar{y}_{md}) \Rightarrow E(\bar{y}_s) = E(\bar{y}_d)$

Concluindo : $E(\bar{y}_{MO}) = E(\bar{y}_{RO})$

$E(\bar{y}_{MC}) = E(\bar{y}_{RC})$

$E(\bar{y}_{MR}) = E(\bar{y}_{RR})$ c.q.d.

Proposição 3.2.2 $B(\bar{y}_{MO}) = B(\bar{y}_{RO}) = \bar{M}\ (\bar{Y}_r - \bar{Y}_m)$

Demonstração: do item 3.1.1 temos que

$\bar{y}_{MO} = \bar{y}_r$. Então,

$E\ (\bar{y}_{MO}) = E(\bar{y}_r) = \bar{Y}_r$.

O cálculo dessa última esperança encontra-se no anexo 6.

Usando a decomposição apresentada em (1) do item 2.1, temos:

$B(\bar{y}_{MO}) = E(\bar{y}_{MO}) - \bar{Y} = \bar{Y}_r - (\bar{R}\bar{Y}_r + \bar{M}\ \bar{Y}_m) =$

$= (1-\bar{R})\ \bar{Y}_r - \bar{M}\ \bar{Y}_m = \bar{M}\ (\bar{Y}_r - \bar{Y}_m)$.

c.q.d.

Analisando a expressão, vemos que o vício do estimador da média pelo método de imputação da média geral é diretamente proporcional à taxa de não resposta esperada e à diferença entre as médias das duas subpopulações.

Uma vez que a imputação é feita pela média dos respondentes, a suposição implícita nesse método é que as médias de ambas as subpopulações são iguais. Se a suposição for verdadeira, o estimador é não viciado qualquer que seja a taxa esperada de não resposta.

Proposição 3.2.3 : $B(\bar{y}_{MC}) = B(\bar{y}_{RC}) = \sum_{h=1}^{H} \frac{M_h}{N}(\bar{Y}_{rh} - \bar{Y}_{mh})$

Demonstração: Para uma classe h fixa, podemos decompor $\bar{y}_h$, seguindo a expressão (4) da parte 2.1,em:

$\bar{y}_h = \bar{r}_h \bar{y}_{rh} + \bar{m}_h \bar{y}_{mh}$ onde

$\bar{r}_h = r_h/n_h$

$\bar{m}_h = m_h/n_h$

$n_h = r_h + m_h$

e r_h é o tamanho de $(I_h \cap A_r)$, ou seja, o número de respondentes na classe h, na amostra;

m_h é o tamanho de $(I_h \cap A_m)$, ou seja, o número de não respondentes na classe h, na amostra;

n_h é o tamanho de $(I_h \cap A_n)$, ou seja, o número total de elementos na classe h, na amostra.

Do item 3.1.3 temos que $\bar{y}_{mh} = \bar{y}_{rh}$, $\forall$ h.

Então, $\bar{y}_h = \bar{r}_h \bar{y}_{rh} + \bar{m}_h \bar{y}_{rh} = \bar{y}_{rh}$

Como podemos definir

$\bar{y}_{MC} = \sum_{h=1}^{H} w_h \bar{y}_h$, onde $w_h = \frac{n_h}{n}$

temos: $\bar{y}_{MC} = \sum_{h=1}^{H} w_h \bar{y}_{rh}$

Condicionando em $\underset{\sim}{n} = (n_1,\ldots,n_h,\ldots,n_H)$ e

em $\underset{\sim}{r} = (r_1,\ldots,r_h,\ldots r_H)$, temos:

$$E(\bar{y}_{MC}|\underset{\sim}{n},\underset{\sim}{r}) = \sum_{h=1}^{H} w_h E(\bar{y}_{rh}|\underset{\sim}{n},\underset{\sim}{r}) = \sum_{h=1}^{H} w_h \bar{Y}_{rh}$$

onde $E(\bar{y}_{rh}|\underset{\sim}{n},\underset{\sim}{r}) = \bar{Y}_{rh}$ justifica-se pela demonstração

apresentada no anexo 6, considerando-se uma classe h fixa.

Como $\sum_{h=1}^{H} w_h \bar{Y}_{rh}$ independe de $\underset{\sim}{r}$, temos que

$$E(\bar{y}_{rh}|\underset{\sim}{n},\underset{\sim}{r}) = E(\bar{y}_{rh}|\underset{\sim}{n})$$

Então,

$$E(\bar{y}_{MC}) = E[E(\bar{y}_{MC}|\underset{\sim}{n},\underset{\sim}{r})] = E\left[\sum_{h=1}^{H} w_h \bar{Y}_{rh}\right] =$$

$$= \sum_{h=1}^{H} \bar{Y}_{rh} E(w_h) = \sum_{h=1}^{H} W_h \bar{Y}_{rh} \quad \text{pois}$$

$E(w_h) = E\left(\frac{n_h}{n}\right) = \frac{1}{n}.\frac{n}{N}.Nh = W_h$ e

$E(n_h) = \frac{n}{N}.N_h$ vem do fato de $\underset{\sim}{n}$ seguir uma distribuição multinomial.

Usando a expressão (1) do item 2.1, para uma classe h fixa, qual seja, $\bar{Y}_h \approx \bar{R}_h \bar{Y}_{rh} + \bar{M}_h \bar{Y}_{mh}$,

temos:

$$B(\bar{y}_{MC}) = E(\bar{y}_{MC}) - \bar{Y} = \sum_{h=1}^{H} W_h \bar{Y}_{rh} - \bar{Y} =$$

$$= \sum_{h=1}^{H} W_h \bar{Y}_{rh} - \sum_{h=1}^{H} W_h \bar{Y}_h =$$

$$= \sum_{h=1}^{H} W_h (\bar{Y}_{rh} - \bar{R}_h \bar{Y}_{rh} - \bar{M}_h \bar{Y}_{mh}) =$$

$$= \sum_{h=1}^{H} W_h (1 - \bar{R}_h) \bar{Y}_{rh} - \bar{M}_h \bar{Y}_{mh} =$$

$$= \sum_{h=1}^{H} W_h \bar{M}_h (\bar{Y}_{rh} - \bar{Y}_{mh}) =$$

$$= \sum_{h=1}^{H} \frac{M_h}{N} (\bar{Y}_{rh} - \bar{Y}_{mh})$$

c.q.d.

*Verificamos que o vício do estimador de média pe*lo método de imputação da média dentro de classes depende das taxas de não resposta esperadas e das diferenças entre as mé*dias das duas subpopulações, dentro das classes. Como a imputa*ção é feita pela média dos respondentes dentro das classes, a suposição implícita desse método de imputação é que,dentro das classes, as duas subpopulações têm a mesma média. Se a suposição for verdadeira, o vício se anula, quaisquer que sejam as taxas de não resposta esperadas dentro das classes.

Proposição 3.2.4 : $B(\bar{y}_{MC}) = \bar{M}(\bar{Y}_r - \bar{Y}_m) - \sum_{h=1}^{H} \frac{W_h}{\bar{R}} (\bar{Y}_{rh} - \bar{Y}_r)(\bar{R}_h - \bar{R})$

$= B(\bar{y}_{MO}) - \sum_{h=1}^{H} \frac{W_h}{\bar{R}} (\bar{Y}_{rh} - \bar{Y}_r)(\bar{R}_h - \bar{R})$

Demonstração : encontra-se no anexo 7.

Da proposição acima vemos que o 2º termo da expressão à direita pode ser interpretado, a menos de uma constante, como uma covariância entre classes (h = 1,...H), no sentido de covariância entre e dentro da análise de variância, entre as variáveis Y- variável de interesse e R- número de respondentes. O que nos diz que quando não houver covariação das médias da variável de interesse e das taxas de resposta esperadas entre as classes, não há ganho, em termos de vício, em se utilizar o método de imputação da média dentro de classes ao invés de usar o método de imputação da média geral.

Proposição 3.2.5 : $B(\bar{y}_{MR}) = B(\bar{y}_{RR}) = \bar{M}\,[(\bar{Y}_r - \sum_{j=1}^{k} \beta_{rj}\,\bar{z}_{jr}) - (\bar{Y}_m - \sum_{j=1}^{k} \beta_{rj}\,\bar{z}_{jm})]$

Condicional a z_j , $j = 1,\ldots,k$.

Demonstração: do item 3.1.5 temos que:

$$\bar{y}_{MR} = \bar{r}\,\bar{y}_r + \bar{m}\,\bar{y}_{mMR} \qquad \text{onde}$$

$$\bar{y}_{mMR} = b_{ro} + \sum_{j=1}^{k} b_{rj}\,\bar{z}_{jm}.$$

Mas, das equações normais do modelo de regressão múltipla, temos que:

$$b_{ro} = \bar{y}_r - \sum_{j=1}^{k} b_{rj}\,\bar{z}_{jr} \qquad \text{onde}$$

$\bar{z}_{jr}$ é a média para a variável Z_j entre os respondentes da amostra, ou seja, em A_r.

Então, $\bar{y}_{MR} = \bar{r}\bar{y}_r + \bar{m}\,(\,\bar{y}_r - \sum_{j=1}^{k} b_{rj}\,\bar{z}_{jr} + \sum_{j=1}^{k} b_{rj}\,\bar{z}_{jm})$

$$= \bar{y}_r + \bar{m}\,(\sum_{j=1}^{k} b_{rj}\,\bar{z}_{jm} - \sum_{j=1}^{k} b_{rj}\,\bar{z}_{jr})$$

Considerando z_j fixo, $\forall j = 1,\ldots,k$, na amostra total temos:

$$E\,(\bar{y}_{MR}|r) = \bar{Y}_r + \bar{m} \sum_{j=1}^{k} \beta_{rj}\,\bar{z}_{jm} - \bar{m} \sum_{j=1}^{k} \beta_{rj}\,\bar{z}_{jr}$$

$$E\,(\bar{y}_{MR}) = E[E(\bar{y}_{MR}|r)] = \bar{Y}_r + \bar{M} \sum_{j=1}^{k} \beta_{rj}\,\bar{z}_{jm} - \bar{M} \sum_{j=1}^{k} \beta_{rj}\,\bar{z}_{jr}$$

Usando a decomposição dada pela expressão (1) do item 2.1, temos:

$$B\,(\bar{y}_{MR}) = E\,(\bar{y}_{MR}) - \bar{Y} = E\,(\bar{y}_{MR}) - \bar{R}\,\bar{Y}_r - \bar{M}\,\bar{Y}_m =$$

$$= \bar{Y}_r + \bar{M} \sum_{j=1}^{k} \beta_{rj}\,\bar{z}_{jm} - \bar{M} \sum_{j=1}^{k} \beta_{rj}\,\bar{z}_{jr} - \bar{R}\,\bar{Y}_r - \bar{M}\,\bar{Y}_m =$$

$$= \bar{M}\,(\,\bar{Y}_r - \sum_{j=1}^{k} \beta_{rj}\,\bar{z}_{jr}) - \bar{M}\,(\,\bar{Y}_m - \sum_{j=1}^{k} \beta_{rj}\,\bar{z}_{jm})$$

c.q.d.

A expressão do vício do estimador de média com o método de imputação da regressão múltipla sobre as variáveis auxiliares nos mostra que ele depende da taxa de não resposta esperada e de uma medida da diferença, entre as duas subpopulações, do modelo de regressão considerado. Como o método imputa valo

res para a subpopulação de não respondentes através do modelo de regressão ajustado com os valores da subpopulação de respondentes, a suposição implícita é que os dois modelos são iguais. Se a suposição for verdadeira, o vício se anula. Caso contrário, o vício absoluto é tanto maior quanto maior a taxa de não resposta esperada e maior a distância entre a realidade e a suposição de validade do mesmo modelo nas duas subpopulações.

Proposição 3.2.6:

$$B(\bar{y}_{MR}) = B(\bar{y}_{RR}) = \bar{M}(\bar{Y}_r - \bar{Y}_m) - \bar{M} \sum_{j=1}^{k} \beta_{rj}(\bar{z}_{jr} - \bar{z}_{jm})$$

$$\doteq B(\bar{y}_{MO}) - \bar{M} \sum_{j=1}^{k} \beta_{rj}(\bar{z}_{jr} - \bar{z}_{jm})$$

para $Z_j = z_j$ fixo, $j = 1,\ldots,k$.

Demonstração : é imediata a partir da proposição anterior.

3.3. As variâncias dos estimadores de média

Proposição 3.3.1: $\mathrm{Var}(\bar{y}_{MO}) \doteq \dfrac{S^2_{ry}}{n\bar{R}}$

Demonstração: sabemos que (Raj(1968)):

$$\mathrm{Var}(\bar{y}_{MO}) = E[\mathrm{Var}(\bar{y}_{MO}|r)] + \mathrm{Var}[E(\bar{y}_{MO}|r)]$$

Como já visto no anexo 6, $E(\bar{y}_{MO}|r) = \bar{Y}_r$.

então, $\mathrm{Var}[E(\bar{y}_{MO}|r)] = \mathrm{Var}(\bar{Y}_r) = 0$

mas, $\mathrm{Var}(\bar{y}_{MO}|r) = \mathrm{Var}(\bar{y}_r|r) = \dfrac{S^2_{ry}}{r}$

então $E[\mathrm{Var}(\bar{y}_{MO}|r)] = E(\dfrac{S^2_{ry}}{r}) = S^2_{ry}\ E(\dfrac{1}{r})$

$$\doteq S^2_{ry}\ \frac{1}{E(r)} = \frac{S^2_{ry}}{n\bar{R}}$$

c.q.d.

A aproximação da E(1/r) encontra-se no anexo 8 e a $E(r) \simeq n\bar{R}$ justifica-se pelo fato de r seguir uma distribuição hipergeométrica com parâmetros N, R e n.

Essa expressão da variância tem a mesma forma da expressão da variância de um estimador de média na ausência de não resposta, sendo que no caso, a população de referência é a subpopulação de respondentes e o tamanho da amostra é na realidade o tamanho esperado da amostra. Podemos observar que ela é inversamente proporcional à taxa de resposta esperada, ou seja, quanto maior a taxa de resposta, menor a variância do estimador, e que só depende da variância da variável na subpopulação de respondentes, o que é bastante razoável na medida em que o estimador considera valores imputados provenientes da amostra de respondentes.

Proposição 3.3.2: $\mathrm{Var}(\bar{y}_{RO}) \doteq \frac{S^2_{ry}}{n\bar{R}} + \frac{\bar{M}}{n} S^2_{ry}$

Demonstração : como anteriormente, temos:

$$\mathrm{Var}(\bar{y}_{RO}) = E[\mathrm{Var}(\bar{y}_{RO}|r)] + \mathrm{Var}[E(\bar{y}_{RO}|r)]$$

Mas $E(\bar{y}_{RO}|r) = E(\bar{y}_{MO}|r) = \bar{Y}_r$

então $\mathrm{Var}[E(\bar{y}_{RO}|r)] = 0$ como no caso anterior.

Utilizando os operadores E_1, V_1, E_2 e V_2 anteriormente definidos, temos:

$$\mathrm{Var}(\bar{y}_{RO}|r) = V_1\, E_2(\bar{y}_{RO}|r) + E_1\, V_2(\bar{y}_{RO}|r) =$$

$$= V_1(\bar{y}_{MO}|r) + E_1\, V_2(\bar{y}_{RO}|r)$$

pois, como já visto no item 3:2 $E_2(\bar{y}_{RO}|r) = \bar{y}_{MO}$

Ainda, temos que: $V_1(\bar{y}_{MO}|r) = \frac{S^2_{ry}}{r}$

Cálculo de $V_2(\bar{y}_{RO}|r)$:

$$V_2(\bar{y}_{RO}|r) = V_2((\bar{r}\bar{y}_r + \bar{m}\bar{y}_{mRO})|r) =$$

$$= \bar{r}^2 V_2(\bar{y}_r|r) + \bar{m}^2 V_2(\bar{y}_{mRO}|r) =$$

$$= \bar{m}^2 V_2(\bar{y}_{mRO}|r) = \frac{(r-1)s^2_{ry}\bar{m}^2}{rm}$$

pois $V_2(\bar{y}_r|r) = 0$ e $\bar{y}_{mRO}$ é a média amostral de uma amostra aleatoria simples com reposiçao de tamanho m, retirada de uma população de tamanho r, ou seja,

$$V_2(\bar{y}_{mRO}|r) = \sum_{i=1}^{r} \frac{(y_{ri} - \bar{y}_r)^2}{rm} = \frac{(r-1)\ s^2_{ry}}{rm}$$

Então, $E_1\ V_2(\bar{y}_{RO}|r) = E_1[\bar{m}^2 \frac{(r-1)\ s^2_{ry}}{rm}|r] =$

$$= \frac{\bar{m}^2(r-1)}{rm} S^2_{ry}$$

pois, pelas suposições em 2.1, $E_1(s^2_{ry}|r) = S^2_{ry}$

Voltando à expressão inicial:

$$Var(\bar{y}_{RO}) = E[Var(\bar{y}_{RO}|r)] = E[\frac{S^2_{ry}}{r} + \frac{\bar{m}^2(r-1)}{rm} S^2_{ry}]$$

$$\doteq E(\frac{S^2_{ry}}{r}) + E(\frac{\bar{m}}{n} S^2_{ry})$$

$$\doteq \frac{S^2_{ry}}{n\bar{R}} + \frac{\bar{M}}{n} S^2_{ry}$$ c.q.d.

As aproximações consideram $(r-1) \doteq r$, $E(1/r) \doteq 1/E(r)$ e $E(\bar{m}) = \bar{M}$. Esta última justifica-se pelo fato de m seguir uma distribuição hipergeométrica com parâmetros N, M, n.

A expressão da variância de $\bar{y}_{RO}$ tem duas componentes. A primeira é idêntica à do estimador que considera a imputação da média geral. A segunda pode ser atribuída ao processo

de aleatorização que o método considera na imputação dos valores ausentes. Ou seja, a aleatorização é responsável por um acréscimo na variância do estimador de média, cuja ordem de grandeza é dada pela segunda parcela. Mas, é fácil verificar que:

$$\frac{Var(\bar{y}_{RO}) - Var(\bar{y}_{MO})}{Var(\bar{y}_{MO})} = \bar{M}(1-\bar{M})$$

Ou, seja, o acrescimo relativo da variancia do estimador de média pela introdução da aleatoriedade é $\bar{M}(1-\bar{M})$, que tem valor máximo 0,25 quando $\bar{M} = 0,5$. O máximo que se aumenta a variância do estimador pelo uso de aleatorização é em 25% do valor obtido pelo método determinístico da média geral, e isso ocorre quando a taxa de não resposta esperada é igual a 0,5.

Proposição 3.3.3:

$$Var(\bar{y}_{MC}) \doteq \sum_{h=1}^{H} \frac{W_h S^2_{rhy}}{n\bar{R}_h} + \sum_{h=1}^{H} \frac{W_h(\bar{Y}_{rh} - \bar{Y}_r)^2}{n}$$

Demonstração : podemos definir:

$\bar{y}_{MC} = \sum_{h=1}^{H} w_h \bar{y}_{hMO}$ onde $\bar{y}_{hMO}$ é o estimador pelo método MO na classe h, $\forall$ h = 1,...,H.

Mas, para uma classe h fixa, já foram calculados:

$E(\bar{y}_{hMO}|n_h) = \bar{Y}_{rh}$

$Var(\bar{y}_{hMO}|n_h) \doteq \frac{S^2_{rhy}}{n_h\bar{R}_h}$ com $\bar{R}_h = \frac{R_h}{N_h}$

Temos que,

$Var(\bar{y}_{MC}) = E[Var(\bar{y}_{MC}|n_h)] + Var[E(\bar{y}_{MC}|n_h)]$

e, $E(\bar{y}_{MC}|n_h) = E\left[\sum_{h=1}^{H} w_h \bar{y}_{hMO}|n_h\right] = \sum_{h=1}^{H} w_h \bar{Y}_{rh}$

e, $Var(\bar{y}_{MC}|n_h) = Var\left[\sum_{h=1}^{H} w_h \bar{y}_{hMO}|n_h\right] \doteq \sum_{h=1}^{H} w_h^2 \frac{S^2_{rhy}}{n_h \bar{R}_h}$

então, $E[Var(\bar{y}_{MC}|n_h)] = E\left[\sum_{h=1}^{H} w_h^2 \frac{S^2_{rhy}}{n_h \bar{R}_h}\right] =$

$$= \sum_{h=1}^{H} \frac{S^2_{rhy}}{\bar{R}_h} E\left(\frac{w_h^2}{n_h}\right) = \sum_{h=1}^{H} \frac{S^2_{rhy}}{\bar{R}_h} \frac{W_h}{n}$$

pois $E\left(\frac{w_h^2}{n_h}\right) = \frac{1}{n^2} E(n_h) = \frac{W_h}{n}$

e $E(n_h) = \frac{n N_h}{N}$ vem do fato de $\underline{n}$ seguir uma distribuição multinomial.

E, $Var[E(\bar{y}_{MC}|n_h)] = Var\left[\sum_{h=1}^{H} w_h \bar{Y}_{rh}\right] =$

$$= \sum_{h=1}^{H} \bar{Y}^2_{rh} Var(w_h) + 2 \sum_{h<k}^{H} \bar{Y}_{rh} \bar{Y}_{rk} Cov(w_h, w_k)$$

$$= \sum_{h=1}^{H} \bar{Y}^2_{rh} \frac{W_h(1-W_h)}{n} - 2 \sum_{h<k} \frac{\bar{Y}_{rh} \bar{Y}_{rk} W_h W_k}{n} =$$

$$= \frac{1}{n} \sum_{h=1}^{H} W_h \bar{Y}^2_{rh} - \frac{1}{n}\left[\sum_{h=1}^{H} W_h^2 \bar{Y}^2_{rh} - 2 \sum_{h<k} W_h W_k \bar{Y}_{rh} \bar{Y}_{rk}\right]$$

$$= \frac{1}{n} \sum_{h=1}^{H} W_h \bar{Y}^2_{rh} - \frac{1}{n}\left(\sum_{h=1}^{H} W_h \bar{Y}_{rh}\right)^2$$

$$= \frac{1}{n} \sum_{h=1}^{H} W_h (\bar{Y}_{rh} - \bar{Y}_r)^2$$

Então, $Var(\bar{y}_{MC}) = \sum_{h=1}^{H} \frac{W_h S^2_{rhy}}{n \bar{R}_h} + \frac{1}{n} \sum_{h=1}^{H} W_h (\bar{Y}_{rh} - \bar{Y}_r)^2$

onde $Var(w_h) = \frac{W_h(1-W_h)}{n}$ e $Cov(w_h, w_k) = -\frac{W_h W_k}{n}$

justificam-se pelo fato de $\underline{n}$ seguir uma distribuição multinomial.

Vemos que a variância do estimador de média obtido pelo método de imputação da média dentro de classes possui duas parcelas: uma, que depende da taxa de resposta esperada sendo inversamente porporcional a ela, e que lembra a variância de um estimador de média de uma amostra estratificada com alocação proporcional, só que envolvendo apenas a variabilidade na subpopulação de respondentes; e uma segunda parcela, refletindo a variabilidade entre classes, que surge em função do fato que os pesos amostrais das classes são variáveis, uma vez que a classificação é feita após a seleção da amostra.

Proposição 3.3.4:

$$Var(\bar{y}_{MO}) - Var(\bar{y}_{MC}) = \sum_{h=1}^{H} \frac{W_h(\bar{Y}_{rh} - \bar{Y}_r)^2}{n}\left(\frac{1}{\bar{R}} - 1\right) +$$

$$\sum_{h=1}^{H} \frac{W_h \, S^2_{rhy}}{n}\left(\frac{1}{\bar{R}} - \frac{1}{\bar{R}_h}\right)$$

Demonstração : encontra-se no anexo 9.

Da proposição acima, o que interessa analisar é quando $Var(\bar{y}_{MO}) - Var(\bar{y}_{MC}) > 0$, ou seja, quando é que $\bar{y}_{MC}$ é melhor que $\bar{y}_{MO}$ em termos de variância.Se $\bar{R}_h = \bar{R}$ $\forall$ h, então a segunda parcela da expressão à direita se anula,$Var(\bar{y}_{MO}) > Var(\bar{y}_{MC})$ e então $\bar{y}_{MC}$ será melhor que $\bar{y}_{MO}$ quanto maior for a variabilidade entre classes. Como não podemos ter $\bar{R}_h > \bar{R}, \forall$ h simultaneamente, naquelas classes onde $\bar{R}_h$ for menor $\bar{R}$, quanto menor a variância dentro, melhor será $\bar{y}_{MC}$.

Proposiçao 3.3.5:

$$Var(\bar{y}_{RC}) \doteq \sum_{h=1}^{H} \frac{W_h \, S^2_{rhy}}{n\bar{R}_h} + \sum_{h=1}^{H} \frac{W_h}{n}(\bar{Y}_{rh} - \bar{Y}_r)^2 + \sum_{h=1}^{H} \frac{W_h \, \bar{M}_h}{n} S^2_{rhy}$$

Demonstração : podemos definir

$$\bar{y}_{RC} = \sum_{h=1}^{H} w_h \, \bar{y}_{hRO}$$

onde $\bar{y}_{hRO}$ é o estimador RO na classe h, h = 1,...,H.

Mas, para uma classe h fixa, já foram calculados

$$E(\bar{y}_{hRO}|n_h) \doteq \bar{Y}_{rh}$$

$$Var(\bar{y}_{hRO}|n_h) \doteq \frac{S^2_{rhy}}{n_h \bar{R}_h} + \frac{\bar{M}_h \; S^2_{rhy}}{n_h}$$

Temos que : $Var(\bar{y}_{RC}) = E[Var(\bar{y}_{RC}|n_h)] + Var[E(\bar{y}_{RC}|n_h)]$

e, $E(\bar{y}_{RC}|n_h) \doteq \sum_{h=1}^{H} w_h \bar{Y}_{rh}$

e, $Var(\bar{y}_{RC}|n_h) \doteq \sum_{h=1}^{H} w_h^2 [\frac{S^2_{rhy}}{n_h \bar{R}_h} + \frac{\bar{M}_h \; S^2_{rhy}}{n_h}]$

então, $E\ [Var(\bar{y}_{RC}|n_h)] = \sum_{h=1}^{H} [\frac{S^2_{rhy}}{\bar{R}_h} + \bar{M}_h \; S^2_{rhy}] E(\frac{w_h^2}{n_h}) =$

$$= \sum_{h=1}^{H} \frac{W_h \; S^2_{rhy}}{n \bar{R}_h} + \sum_{h=1}^{H} \frac{W_h \; \bar{M}_h \; S^2_{rhy}}{n}$$

pois $E(\frac{w_h^2}{n_h}) = \frac{W_h}{n}$ já calculado anteriormente,

e, $Var[E(\bar{y}_{RC}|n_h)] = Var[\sum_{h=1}^{H} w_h \; \bar{Y}_{rh}] = \frac{1}{n} \sum_{h=1}^{H} W_h (\bar{Y}_{rh} - \bar{Y}_r)$

já calculada anteriormente também.

Então, voltando à expressão inicial:

$$Var(\bar{y}_{RC}) \doteq \sum_{h=1}^{H} \frac{W_h \; S^2_{rhy}}{n \bar{R}_h} + \sum_{h=1}^{H} \frac{W_h (\bar{Y}_{rh} - \bar{Y}_r)^2}{n} + \sum_{h=1}^{H} \frac{W_h \; \bar{M}_h \; S^2_{rhy}}{n}$$

c.q.d.

Das expressões de $Var(\bar{y}_{RC})$ e $Var(\bar{y}_{MC})$ temos que

$$Var(\bar{y}_{RC}) = Var(\bar{y}_{MC}) + \sum_{h=1}^{H} \frac{W_h \bar{M}_h S^2_{rhy}}{n}$$

, ou seja, tal como no caso geral, há um acréscimo na variância do estimador pelo fato de se introduzir o processo de aleatoriedade na atribuição de valores aos dados ausentes. Ou seja, as três parcelas componentes da $Var(\bar{y}_{RC})$ podem ter a seguinte interpretação: a primeira, pelo fato do método considerar uma classificação; a segunda, pelo fato dessa classificação ter sido efetuada após a seleção; e a terceira devido ao fato do método ser estocástico.

Proposição 3.3.6 Se k = 1 ,

$$Var(\bar{y}_{MR}) \doteq \frac{S^2_{ry.z}}{n\bar{R}} + \left[\frac{\bar{M}(1-\bar{M})}{n} + \bar{M}^2 \right] \frac{(\bar{z}_m - \bar{z}_r)^2}{\sum_{i=1}^{r} (z_{ri} - \bar{z}_r)^2} S^2_{ry.z} +$$

$$+ \beta^2_{r1} (\bar{z}_m - \bar{z}_r)^2 \frac{\bar{M}(1 - \bar{M})}{n}$$

Condicional a Z

Demonstração : encontra-se no anexo 10.

Proposição 3.3.7 : Se k = 1

$$Var(\bar{y}_{RR}) \doteq Var(\bar{y}_{MR}) + \bar{M} \frac{S^2_{ry.z}}{n}$$

Condicional a Z

Demonstração : encontra-se no anexo 11.

4. As Variâncias e Covariâncias Amostrais

4.1. - Introdução

Santos (1981a) estudou teoricamente os efeitos de diversos *métodos de imputação sobre algumas* estatísticas complexas consi derando dois modelos diferentes para os dados ausentes. Para ca da estatística, cada método de imputação e cada modelo de dados ausentes foi analisado o vício da estatística amostral como um estimador natural da correspondente estatística populacional.

As estatísticas estudadas foram covariância (s_{xy}), variân cia (s_y^2), coeficiente de correlação linear ($s_{xy} / s_x s_y$) e co eficiente de regressão (s_{xy} / s_x^2), onde o índice x representa uma variável não sujeita a dados ausentes e o y representa a variá vel que passou pelo processo de imputação.

Os métodos estudados foram: imputação da média geral(MO), imputação aleatória geral(RO), imputação da média dentro de clas ses(MC), imputação aleatória dentro de classes(RC), imputação da média da regressão(MR) e imputação da média da regressão mais um efeito de aleatoriedade(RR). Para os dois últimos métodos foram considerados dois casos: regressão simples e regressão múltipla. A apresentação desses métodos encontra-se no item 3.1.

Foi também considerada a situação de rejeitar todos os ca sos incompletos e só utilizar os respondentes, ou seja, não fa zer imputação. A esse caso o autor denominou de médodo DN (do nothing).

Os métodos adotados para os dados ausentes foram:dados au sentes ao acaso (designado por MAR - missing at random) e dados ausentes ao acaso dentro de classes (designado por MARC- missing at random within cells). As definições dos modelos MAR e MARC utilizados por Santos encontram-se no anexo 12.

As demais suposições teóricas efetuadas pelo autor estão de acordo com o apresentado em 2.1.

Para as estatísticas coeficiente de regressão e coeficien te de correlação, Santos verificou que os vícios relativos depen dem dos vícios relativos das variâncias e covariâncias. Ou seja, não há necessidade de promover uma investigação detalhada do vício

de imputação para essas estatísticas, pois vícios relativos pequenos para variâncias e covariâncias garantem vícios relativos pequenos para elas. Por essa razão, os itens 4.2, 4.3 e 4.4 a seguir só consideram as estatísticas s_{xy} e s^2_y.

4.2. - Os resultados teóricos.

A partir das relações (4),(5) e (6) do item 2.1, para um método de imputação genericamente designado por I, podemos escrever:

$$\bar{y}_I = r\bar{y}_r + m\bar{y}_{mI}$$

$$s^2_{\bar{y}I} \doteq \bar{r}s^2_{ry} + \bar{m}s^2_{my} + \bar{r}\bar{m}\,(\bar{y}_r - \bar{y}_{mI})^2$$

$$s_{xy} \doteq \bar{r}s_{rxy} + \bar{m}s_{mxyI} + \bar{r}\bar{m}\,(\bar{x}_r - \bar{x}_m)(\bar{y}_r - \bar{y}_{mI})$$

onde:

$\bar{y}_I$ é o estimador de média decorrente do método de imputação I, e encontra-se apresentado no item 3.1 para os diversos métodos considerados;

s^2_{yI} é a variância na amostra da variável y que sofreu imputação pelo método I;

s_{xyI} é a covariância na amostra entre as variáveis x e y, sendo que y sofreu imputação pelo método I.

As expressões de s^2_{yI} e s_{xyI} podem ser reescritas em função das definições de y_{mI}, $\bar{y}_{mI}$ e $\bar{y}_I$, para cada I, e que encontram-se no item 3.1

A partir dessas novas expressões foram então calculadas as esperanças e os vícios de cada uma dessas estatísticas, para cada um dos métodos considerados.

As expressões obtidas para os vícios consideram o modelo teórico apresentado em 2.1, as aproximações decorrentes da consideração de n, r e m grandes, de 4.1, e desprezam os termos de $O(n^{-1})$ ou menores.

A seguir, no quadro 4.1 são apresentados os vícios das estatísticas s_{xy} e s^2_y, quando consideradas estimadores para as

correspondentes estatísticas populacionais sob o modelo MAR, para os métodos estudados por Santos. O quadro 4.2 apresenta os mesmos vícios só que relativos aos valores dos parâmetros populacionais. Os quadros 4.3 e 4.4 apresentam os vícios sob o modelo MARC para as estatísticas s^2_y e s_{xy}, respectivamente.

Quadro 4.1 - Vícios da variância e covariância amostral, para diversos métodos de imputação, sob o modelo MAR.

Métodos de Imputação	Vício (s^2_y)	Vício (s_{xy})
DN	0	0
MO	$-\bar{M}\ S^2_y$	$-\bar{M}\ S_{xy}$
RO	0	$-\bar{M}\ S_{xy}$
MC	$-\bar{M} \sum_{h=1}^{H} W_h\ S^2_{yh}$	$-\bar{M} \sum_{h=1}^{H} W_h S_{xyh}$
RC	0	$-\bar{M} \sum_{h=1}^{H} W_h S_{xyh}$
MR k=1	$-\bar{M}\ S^2_{y.z}$	$-\bar{M}\ S_{xy.z}$
RR k=1	0	$-\bar{M}\ S_{xy.z}$
(*) MR k=2	$-\bar{M}\ S^2_{y.xz}$	0
(*) RR k=2	0	0

(*) o modelo de regressão contém X como variável independente

$S_{xy.z}$ - covariância parcial de X e Y dado Z

$S^2_{y.z}$ - variância de Y dado Z

$S^2_{y.xz}$ - variância de Y dado Z e X

Quadro 4.2 - Vício relativo da variância e covariância amostral, para diversos métodos de imputação, sob o modelo MAR.

Métodos de Imputação	Vício Relativo (s^2_y)	Vício Relativo (s_{xy})
DN	0	0
MO	$-\bar{M}$	$-\bar{M}$
RO	0	$-\bar{M}$
MC	$-\bar{M}\ (1-\eta^2)$	$-\bar{M} \sum_{h=1}^{H} W_h\ S_{xyh} / S_{xy}$
RC	0	$-\bar{M} \sum_{h=1}^{H} W_h\ S_{xyh} / S_{xy}$
MR k=1	$-\bar{M}\ (1-R^2)$	$-\bar{M}\ S_{xy.z} / S_{xy}$
RR k=1	0	$-\bar{M}\ S_{xy.z} / S_{xy}$
(*) MR k=2	$-\bar{M}\ (1-R^2)$	0
(*) RR k=2	0	0

(*) o modelo de regressão contém X como variável independente

η^2 - proporção da variância explicada pelas classes de imputação

R^2 - proporção da variância explicada pela regressão

$S_{xy.z}$ - covariância parcial de X e Y dado Z

Quadro 4.3 - Vício da variância amostral, para diversos métodos de imputação, sob o modelo MARC.

	Métodos de Imputação		Vício (s^2_y)
	DN		$-\bar{M} \sum_{h=1}^{H} (W_{mh}-W_{rh})[S^2_{yh}+(\bar{Y}_h-\bar{Y})^2]-\bar{M}^2(\bar{Y}_r-\bar{Y}_m)^2$
	MO		$-\bar{M} \sum_{h=1}^{H} W_{mh}[S^2_{yh}+(\bar{Y}_h-\bar{Y})^2]-\bar{R}\ \bar{M}^2(\bar{Y}_r-\bar{Y}_m)^2$
	RO		$-\bar{M} \sum_{h=1}^{H} (W_{mh}-W_{rh})[S^2_{yh}+(Y_h-Y)^2]-\bar{M}^2(Y_r-Y_m)^2$
	MC		$-\bar{M} \sum_{h=1}^{H} W_{mh}\ S^2_{yh}$.
	RC		0
(**)	MR	k=1	$-\bar{M}\ S^2_{ry.z}+(S^2_{ry}-S^2_y)-B^2_{r1}(S^2_{rz}-S^2_z)$
(**)	RR	k=1	$(S^2_{ry}-S^2_y)-B^2_{r1}(S^2_{rz}-S^2_z)$
(*)	MR	k=2	$-\bar{M}\ S^2_{ry.xz}+(S^2_{ry}-S^2_y)-B^{*2}_{r1}(S^2_{rx}-S^2_x)$ $-B^{*2}_{r2}(S^2_{rz}-S^2_z)-2B^*_{r1}B^*_{r2}(S_{rxz}-S_{xz})$
(*)	RR	k=2	$(S^2_{ry}-S^2_y)-B^{*2}_{r1}(S^2_{rx}-S^2_x)$ $-B^{*2}_{r2}(S^2_{rz}-S^2_z)-2B^*_{r1}B^*_{r2}(S_{rxy}-S_{xy})$

(*) o modelo de regressão contém X como variável independente

B^*_{r1} e B^*_{r2} são os coeficientes da regressão múltipla na subpopulação de respondentes

(**) B_{r1} é o coeficiente da regressão simples na subpopulação de respondentes

Quadro 4.4 - Vício da covariância amostral, para diversos métodos de imputação, sob o modelo MARC.

Método de Imputação	Vício (s_{xy})
DN	$-\bar{M}\sum_{h=1}^{H}(W_{mh}-W_{rh})[S_{xyh}+(\bar{X}_h-\bar{X})(\bar{Y}_h-\bar{Y})]$
	$-\bar{M}^2(\bar{X}_r-X_m)(\bar{Y}_r-\bar{Y}_m)$
MO e RO	$-\bar{M}\sum_{h=1}^{H}W_{mh}[S_{xyh}+(\bar{X}_h-\bar{X})(\bar{Y}_h-\bar{Y})]$
	$-\bar{R}\,\bar{M}^2(\bar{X}_r-\bar{X}_m)(\bar{Y}_r-\bar{Y}_m)$
MC e RC	$-\bar{M}\sum_{h=1}^{H}W_{mh}\,S_{xyh}$
(**) MR e RR (k=1)	$S_{xz}(B_{r1}-B_1)+(\bar{R}\,S_{rxy.z}-S_{xy.z})$
(*) MR e RR (k=2)	$S^2_x(B^{*2}_{r1}-B^*_1)+S_{xy}(B^*_{r2}-B^*_2)$

(*) o modelo de regressão contém X como variável independente B^*_{r1}, B^*_{r2}, B^*_1 e B^*_2 são os coeficientes da regressão múltipla na subpopulação de respondentes e na população total.

(**) B_{r1} e B_1 são os coeficientes da regressão simples na subpopulação de respondentes e na população total.

4.3 - Os estudos empíricos

Santos (1981a,b) realizou, além de estudos teóricos, um estudo empírico, no qual, por meio de simulação e a partir de um conjunto de dados completos, foram criados conjuntos de dados com valores ausentes segundo o modelo MARC.

Os dados utilizados no estudo foram retirados de uma pesquisa nacional americana que faz parte do ISDP - Income Survey Development Program do Departament of Health and Human Services - DHSS. As variáveis estudadas foram: salário recebido por hora (hourly rate of pay), por emprego, para aquelas pessoas que recebiam somente por hora, investigada na pesquisa do mês de julho de 1978; e rendimento trimestral (quarterly earnings), por emprego, investigada no questionário do mês de abril de 1978.

Para simular o estrato de não respondentes, a partir do conjunto de respondentes, Santos utilizou os resultados de um estudo realizado por Heeringa em 1980 com os dados da mesma pesquisa. Nesse estudo foram detectadas taxas de não resposta diferenciais segundo algumas variáveis auxiliares. Ou seja, foi analisado o padrão de não resposta das variáveis da pesquisa. Para a 1ª variável, o salário por hora, as variáveis auxiliares escolhidas para a construção do modelo de simulação foram sexo e informante, onde essa última tem duas categorias: numa delas a própria pessoa investigada na pesquisa é o informante; na outra, o informante é qualquer pessoa da unidade domiciliar relacionada com a investigada. Para a 2ª variável, o rendimento trimestral, as variáveis associadas para a construção do modelo de simulação foram o informante e a renda mensal domiciliar, onde essa última possui duas categorias que discriminam as faixas de renda. As tabelas 4.1 e 4.2 apresentam, para as variáveis de interesse salário por hora e rendimento trimestral, respectivamente, a distribuição da amostra e dos dados ausentes ao longo das classes consideradas.

Primeiramente, as informações de alguns indivíduos foram desprezadas de modo a criar para os respondentes um arquivo com a mesma distribuição amostral total nas classes definidas pelas variáveis associadas.

Então, com base no padrão de não resposta detectado para as duas variáveis de interesse ao longo das classes formadas pelas respectivas variáveis associadas, foram realizadas 10 simulações de dados ausentes, gerando 10 repetições de um conjunto de dados segundo o modelo MARC definido no anexo 12.

Os procedimentos de imputação aplicados às duas variáveis de interesse foram os descritos em 4.1, exceto a imputação aleatória geral (RO).

Para a definição das variáveis auxiliares utilizadas nos procedimentos de imputação, Santos baseou-se nos estudos de Hendrix (1980) sobre os melhores preditores para as duas variáveis dependentes em questão. Hendrix (1980), por sua vez, utilizou a técnica de análise denominada SEARCH, sucessora da técnica AID - Automatic Interaction Detector. Essa técnica determina os principais preditores de uma variável através de uma série de divisões binárias da amostra, sendo que a cada divisão a amostra é partida em duas, segundo uma variável preditora, de tal forma que a proporção da variância explicada seja máxima. O item 5.2 volta a discutir essa técnica, porém mais detalhes podem ser encontrados em Sonquist et al (1973).

Uma vez realizadas as imputações, foram então calculadas as estatísticas amostrais e os seus vícios relativos. Esses vícios foram calculados em relação às estatísticas amostrais para a amostra total, ou seja, a amostra disponível antes de serem realizadas as simulações de dados ausentes.

Os resultados obtidos da análise dos vícios apresentaram-se de acordo com os resultados teóricos encontrados por Santos para o modelo MAR (quadros 4.1 e 4.2), embora o modelo construído tenha sido o MARC.

Tabela 4.1 - Distribuição da amostra total e taxa de não resposta para a variável salário recebido por hora, segundo as classes associadas

Classes		Distribuição da amostra total(%)	Taxas de não-resposta (%)
Sexo	Informante		
homem	a própria pessoa	27,0	3,5
homem	outra pessoa	27,5	17,4
mulher	a própria pessoa	34,4	7,4
mulher	outra pessoa	11,1	16,5
TOTAL		100,0	10,1

Tabela 4.2 - Distribuição da amostra total e taxa de não resposta para a variável rendimento trimestral, segundo as classes associadas.

Classes		Distribuição da amostra total (%)	Taxa de não-resposta (%)
Renda média domiciliar	Informante		
abaixo de 900$	a própria pessoa	21,7	53,5
abaixo de 900$	outra pessoa	9,5	78,4
900$ ou mais	a própria pessoa	41,5	45,0
900$ ou mais	outra pessoa	27,3	62,3
TOTAL		100,0	54,7

A título de ilustração, a tabela 4.3 apresenta, para as variáveis salários por hora e rendimento trimestral, os vícios relativos da variância na amostra como estimador da variância na população, para os diversos métodos de imputação estudados por Santos.

Tabela 4.3 - Vícios relativos de s^2_{y1} e s^2_{y2} onde y_1 é a variável salário por hora e y_2 é a variável rendimento trimestral, segundo os métodos de imputação.

Métodos de Imputação	Vício relativo(s^2_{y1})	Vício relativo(s^2_{y2})
DN	- 2,5%	- 0,7%
MO	-12,4%	-55,4%
MC	- 6,7%	-26,5%
RC	- 1,2%	- 3,6%
MR	- 6,7%	-25,3%
RR	- 2,1%	- 1,2%

Obs: $\bar{m}_1$ = 10,1% (taxa de não resposta para y_1)

$\bar{m}_2$ = 54,7% (taxa de não resposta para y_2)

4.4 - As conclusões

O autor apresenta uma análise comparativa entre métodos determinísticos e estocásticos. Enquanto ambos reduzem o vício de estimativas de médias e totais, os métodos determinísticos são mais precisos que os correspondentes estocásticos. Entretanto, os métodos determinísticos produzem distribuições com vários picos, enquanto os estocásticos praticamente retêm as propriedades da distribuição original. Ainda mais, nos métodos determinísticos a variância na amostra tende a ser subestimada e a covariância na amostra tende a ser atenuada, especialmente entre a variável que sofreu imputação e uma variável qualquer não utilizada como variável auxiliar no processo de imputação.

É interessante ressaltar que o método denominado DN, ou seja, não fazer imputação foi o que apresentou a melhor performance para todas as estatísticas consideradas no estudo. O autor, entretanto, adverte para a falha do experimento na construção do conjunto de dados sob o modelo MARC e sugere novas avaliações antes de se concluir que não fazer imputação é melhor que qualquer dos métodos estudados.

5. Uma aplicação

5.1. Introdução

Com o objetivo de realizar uma aplicação de alguns dos métodos de imputação estudados foi escolhida a Pesquisa Mensal de Emprego realizada pelo IBGE [IBGE (1983)].

A Pesquisa Mensal de Emprego - PME tem por objetivo principal obter indicadores de emprego e desemprego. É realizada independentemente em seis regiões metropolitanas do país e utiliza-se de uma amostra probabilística cujo desenho envolve conglomeração em dois estágios e resulta autoponderado. O primeiro estágio é denominado setor, que se constitui em uma delimitação geográfica para fins de pesquisa. O segundo estágio é a unidade domiciliar (domicílio ou unidade de habitação em domicílio coletivo). Dentro de uma unidade domiciliar selecionada para a amostra, todas as pessoas de 10 anos ou mais são investigadas. A pesquisa investiga uma série de itens que visam classificar as pessoas em inativas ou ativas e, dentro dessas últimas, em ocupadas ou desocupadas, e, além disso, caracterizar cada um desses tipos segundo algumas variáveis sócio-econômico-demográficas.

A região metropolitana do Rio de Janeiro e o mês de setembro de 1987 delimitam o conjunto de dados em estudo.

A variável escolhida para sofrer imputação (Y) foi o rendimento mensal efetivamente recebido, no trabalho principal, pelas pessoas ocupadas de 15 anos ou mais com remuneração, no mês de referência da pesquisa. Todas as demais variáveis investigadas para essas pessoas foram consideradas como variáveis auxiliares em potencial ($\underset{\sim}{Z}$).

Dentre as variáveis investigadas para as pessoas ocupadas com remuneração, apenas 5 são do tipo quantitativo, quais sejam, o próprio rendimento mensal efetivamente recebido no trabalho principal, o número de horas trabalhadas no trabalho principal, o rendimento mensal efetivamente recebido nos outros trabalhos, o número de horas trabalhadas nos outros trabalhos e a idade.As demais variáveis são do tipo categóricas.

Pelo fato de a incidência de pessoas com mais de um trabalho ser baixa, de a variável número de horas trabalhadas por se

mana *no trabalho principal não* variar muito, e de a variável idade comportar-se menos errática quando tratada em classes do que pontualmente, decidiu-se não construir o modelo usual de regressão para a aplicação dos métodos da média da regressão e do seu correspondente estocástico.

Dessa forma, os métodos de imputação aplicados foram: média geral (MG), aleatória geral (RG), média dentro de classes (MC) e aleatória dentro de classes (RC).

Com o objetivo de estudar o efeito das taxas de não resposta sobre os diferentes métodos de imputação e devido ao fato de a taxa de não resposta observada para a variável considerada ser apenas de 1,46% na amostra original da PME, foram realizados três exercícios de geração de dados ausentes, com base nessa amostra, para a aplicação dos métodos e a análise de seus resultados.

O objetivo dessa aplicação foi o de observar o comportamento dos métodos, avaliando a magnitude dos vícios sobre a variável rendimento mensal, e, para os métodos que se utilizam de variáveis auxiliares, selecionar um conjunto delas, considerado como o melhor para os processos de imputação em questão, no conjunto de dados em estudo.

Vale repetir que esses exercícios devem ser considerados apenas como uma ilustração dos métodos apresentados, e que o objetivo aqui não foi realizar simulações, quando então técnicas e procedimentos apropriados deveriam ser considerados.

Este capítulo descreve os procedimentos adotados para a aplicação, bem como seus resultados.

5.2. - A seleção das variáveis auxiliares

Para definir quais variáveis auxiliares seriam utilizadas para a formação das classes de imputação nos métodos MC e RC, foi analisado o comportamento dos dados da subpopulação de respondentes para variáveis investigadas na mesma pesquisa, cujas informações estão disponíveis para os elementos da subpopulação de não respondentes também.

A análise consistiu na aplicação da técnica SEARCH (ver Sonquist et al(1973)) para a variável dependente valor do rendimento, em cruzados, considerando como independentes as variáveis auxiliares apresentadas no quadro 5.1.

Quadro 5.1 - Variáveis investigadasna pesquisa e suas categori as.

Variáveis	Categorias
A_1 - Sexo	1 - Masculino 3 - Feminino
A_2 - Condição no domicílio	1 - Chefe 2 - Cônjuge 3 - Filho 4 - Outro parente 5 - Agregado 6 - Pensionista 7 - Empregado doméstico 8 - Parente de empregado do méstico
A_3 - Ramo de atividade	1 - Indústria de transforma ção 3 - Construção Civil 5 - Comércio 7 - Serviços 8 - Outro
A_4 - Posição na ocupação	2 - Empregado 4 - Conta - própria 6 - Empregador 8 - Não - remunerado
A_5 - Ganhava por	1 - Mês 3 - Quinzena 5 - Semana 7 - Outra forma
A_6 - Tinha carteira de trabalho assinada	2 - Sim 4 - Não

Quadro 5.1 - (continuação)

Variáveis	Categorias
A_7 - Faixa de idade	1 - Menos de 10 anos 2 - 10 a 17 anos 3 - 18 a 24 anos 4 - 25 a 29 anos 5 - 30 a 34 anos 6 - 35 a 39 anos 7 - 40 a 49 anos 8 - 50 anos ou mais 9 - idade ignorada
A_8 - Grau de instrução	0 - Sem instrução 1 - Elementar incompleto 2 - Elementar completo 3 - 1º grau incompleto 4 - 1º grau completo 5 - 2º grau incompleto 6 - 2º grau completo 7 - Superior incompleto 8 - Superior completo ou mais 9 - Grau ignorado

Essa técnica pretende responder à seguinte questão: "quais categorias de quais variáveis preditoras simples proporcionam uma melhoria máxima na capacidade de prever valores para a variável dependente?". Para tanto, o procedimento usa a técnica de análise de variância a um fator iterativamente, para explicar a variância da variável dependente tanto quanto possível. Para cada variável preditora especificada, é procurada a melhor divisão simples entre as suas categorias, onde "melhor" significa maior redução no erro (soma de quadrados dos resíduos). Ou seja, a amostra é dividida através de uma série de categorias binárias, em série de subgrupos mutuamente exclusivos, escolhidos tal que, a cada passo, as duas novas médias da variável dependente contribuam mais para o total da soma de quadrados que as médias de qualquer outro par de subgrupos. O processo de divisão termina quando a redução na soma de quadrados dos resíduos for inferior a um valor especificado. No caso, foi utilizado o valor "default" do programa, que é 0,8%.

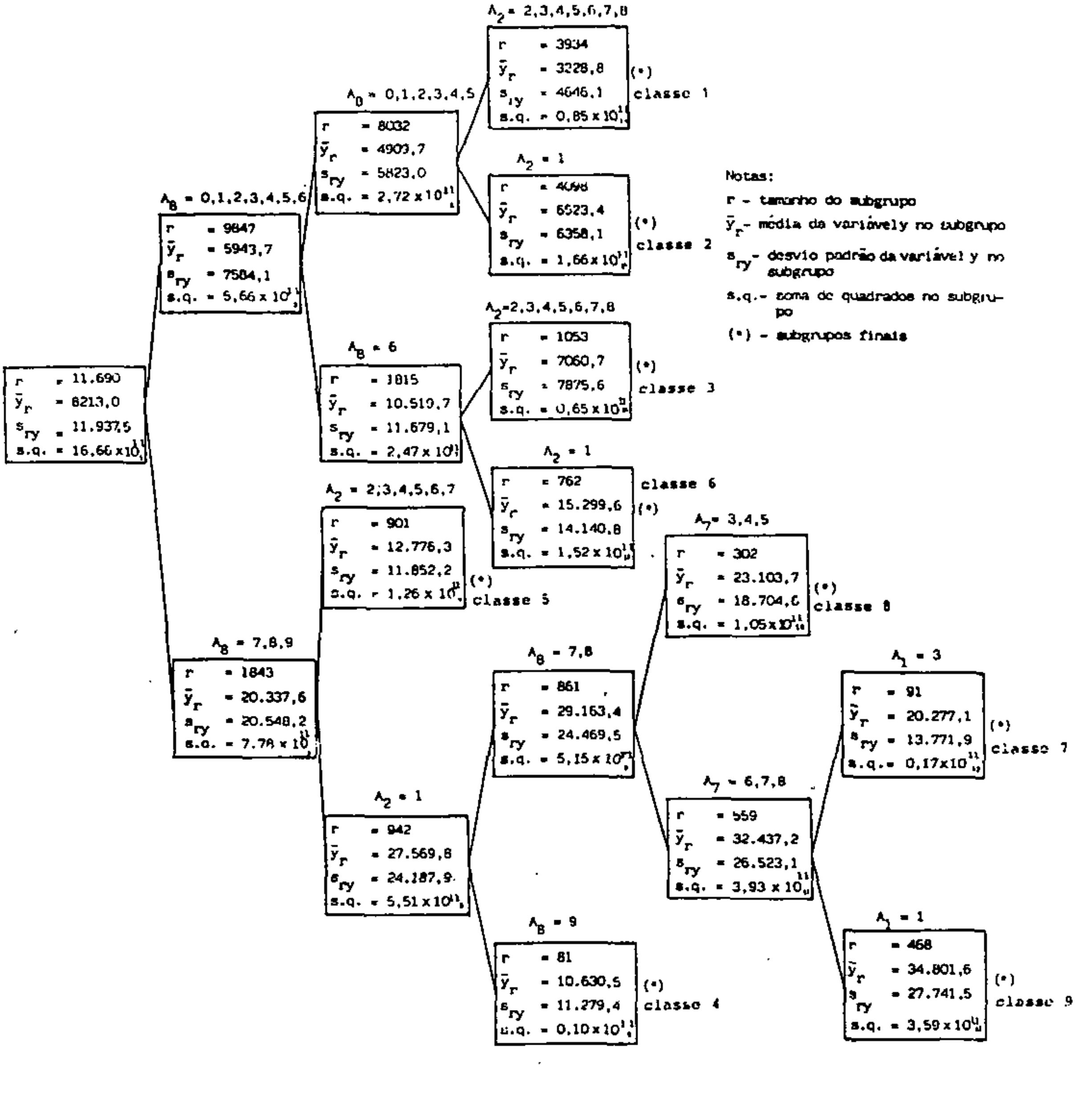
No final do processo, o que se tem é um conjunto de subgrupos cuja definição é clara e o processo pelo qual eles foram isolados é facilmente recuperado.

Em uma primeira aplicação da técnica ao conjunto de dados em análise, foram determinados subgrupos constituídos pelas combinações das categorias das seguintes variáveis: A_1 - Sexo, A_2 - Condição no domicílio e A_8 - Grau de instrução.

Como a técnica despreza os indivíduos para os quais não haja dados para o conjunto completo de variáveis especificadas e a inclusão das 4 variáveis A_3 - Ramo de atividade, A_4 - Posição na ocupação, A_5 - Ganhava por e A_6 - Tinha carteira de trabalho assinada proporcionava uma redução no tamanho da amostra em 2.732 indivíduos, de um total de 11.690 (principalmente por conta da ausência de informações na variável A_6 - Tinha carteira de trabalho assinada, investigada somente para os empregados), foi realizada uma nova aplicação considerando então somente as 3 variáveis destacadas na primeira análise e incluindo a variável A_7 - Faixa de idade.

O resultado dessa aplicação encontra-se no diagrama 5.1 demonstrativo das etapas do processo de isolamento dos subgrupos e nos quadros 5.2, 5.3 e 5.4, que resumem informações sobre os subgrupos, que foram denominados de classes.

Diagrama 5.1 - Demonstrativo das etapas do processo SEARCH para a variável y - rendimento mensal efetivamente recebido em agosto de 1987, na subpopulação de respondentes.



Quadro 5.2 - Composição das classes de imputação segundo as variáveis e suas categorias.

Classe	Variável	Categoria
1	A_2 - Condição no domicílio A_8 - Grau de instrução	(2,3,4,5,6,7,8)- Não chefe (0,1,2,3,4,5) - Sem instrução até 2º grau incompleto
2	A_2 - Condição no domicílio A_8 - Grau de instrução	(1) - Chefe (0,1,2,3,4,5) - Sem instrução até 2º grau incompleto
3	A_2 - Condição no domicílio A_8 - Grau de instrução	(2,3,4,5,6,7,8)- Não chefe (6) - 2º grau completo
4	A_2 - Condição no domicílio A_8 - Grau de instrução	(1) - Chefe (9) - Ignorado
5	A_2 - Condição no domicílio A_8 - Grau de instrução	(2,3,4,5,6,7,8)- Não Chefe (7,8,9) - Superior incompleto ou mais e ignorado

Quadro 5.2 - (continuação)

Classe	Variável	Categoria
6	A_2 - Condição no domicílio	(1) - Chefe
	A_8 - Grau de instrução	(6) - 2º grau completo
7	A_2 - Condição no domicílio	(1) - Chefe
	A_8 - Grau de instrução	(7,8) - Superior incompleto ou mais
	A_7 - Faixa de idade	(6,7,8) - 35 anos ou mais
	A_1 - Sexo	(3) - Feminino
8	A_2 - Condição no domicílio	(1) - Chefe
	A_8 - Grau de instrução	(7,8) - Superior incompleto ou mais
	A_7 - Faixa de idade	(3,4,5) - de 18 a 34 anos
9	A_2 - Condição no domicílio	(1) - Chefe
	A_8 - Grau de instrução	(7,8) - Superior incompleto ou mais
	A_7 - Faixa de idade	(6,7,8) - 35 anos ou mais
	A_1 - Sexo	(1) - Masculino

Quadro 5.3 - Informações sobre as classes na subpopulação de respondentes, na amostra

Classe (h)	r_h	$\bar{y}_{rh}$	s_{rhy}	cv_{rhy} (%)	Soma de Quadrados ($x10^{11}$)
1	3934	3228,8	4646,1	143,9	0,85
2	4098	6523,4	6358,1	97,5	1,66
3	1053	7060,7	7875,6	111,5	0,65
4	81	10630,5	11279,4	106,1	0,10
5	901	12776,3	11852,2	92,8	1,26
6	762	15299,6	14140,8	92,4	1,52
7	91	20277,1	13771,9	67,9	0,17
8	302	23103,7	18704,6	81,0	1,05
9	468	34801,6	27741,5	79,7	3,59
Total	11690	8213,0	11937,5	145,3	16,66

Quadro 5.4 - Análise de variância das classes na subpopulação de respondentes, na amostra

Fonte	Graus de Liberdade	Soma de Quadrados	R^2
Classes	8	$5,80 \times 10^{11}$	0,3478
Erro	11681	$10,86 \times 10^{11}$	-
Total	11689	$16,66 \times 10^{11}$	-

5.3 - A geração de dados ausentes

A fim de tentar compreender o efeito dos métodos de imputação com diferentes taxas de não resposta, foram realizados 3 exercícios de geração de dados ausentes, visando obter taxas globais de não resposta em torno de 5%, 10% e 20%.

Para tanto, o procedimento adotado foi, para cada exercício, multiplicar as taxas de não resposta observadas na amostra original ao longo das classes de imputação consideradas para a aplicação dos métodos MC e RC, por um constante conveniente a fim de atingir a taxa de não resposta global desejada.

Dessa forma, mantém-se o padrão de não resposta observado ao longo das classes. Entretanto, no exercício para obtenção da taxa global de 20%, duas classes de imputação tiveram a taxa de não resposta fixada em 50%, para obedecer à restrição de taxa de não resposta não inferior a 50% para cada classe. Isso, para poder continuar considerando seleção aleatória sem reposição dentre os respondentes para a obtenção de doadores, no método de imputação aleatória dentro de classes.

Os tamanhos das classes nas duas subpopulações nas amostras geradas encontram-se nas tabelas A16.1, A17.1 e A18.1, dos anexos 16, 17 e 18, respectivamente.

De posse das taxas de não resposta desejadas dentro de cada classe, foram selecionados, aleatoriamente dentro de cada classe, respondentes em número correspondente à taxa e, então, o valor do rendimento mensal declarado foi considerado ignorado.

Como já foi visto, o padrão de não resposta recomenda determinado método de imputação. Aqui, o coeficiente de correlação entre a taxa de não resposta e a média da variável y ao longo das classes foi calculado em 0,72, enquanto a correlação entre a taxa de não resposta e o desvio padrão de y foi de 0,63. Por isso, era de se esperar maior adequação dos métodos de imputação por classe. Desse modo, o exercício pretendeu apenas indicar a magnitude dos viéses numa particular situação.

Outro objetivo explorado neste exercício foi a criação de medidas que permitam comparar os resultados dos métodos, e que possam ser úteis em futuros estudos de propriedades através de simulação.

5.4 - Os procedimentos computacionais

5.4.1 - Na seleção das variáveis auxiliares

Para a seleção das variáveis auxiliares utilizadas na formação das classes de imputação, foi utilizado um programa de aplicação do pacote

OSIRIS IV, no qual a técnica SEARCH é uma das opções.

O anexo 13 apresenta o programa utilizado, onde figuram os diversos parâmetros definidos durante a aplicação da técnica.

5.4.2 - Nos procedimentos de imputação e análise

Para a realização da imputação propriamente dita pelos quatro métodos considerados, bem como para a análise dos dados relativos às subpopulações de respondentes e não respondentes, na amostra, antes e depois de efetuada a imputação, foram elaborados programas de aplicação do pacote SAS versão 5.

O anexo 14 apresenta o programa utilizado em sua versão final, qual seja, aquela que considera o exercício de geração de dados ausentes, a aplicação dos métodos de imputação e as estatísticas para a análise.

5.5 Os resultados

Neste item, primeiramente encontram-se descritas as tabelas que resumem os resultados dos métodos de imputação aplicados à amostra original, ou seja, considerando taxa de não resposta $\bar{m} = 1,46\%$. Um conjunto de oito tabelas encontrado no anexo 15. Nos anexos 16, 17 e 18 encontram-se os conjuntos de tabelas análogas referentes aos exercícios de geração de dados ausentes com taxas globais de não resposta, $\bar{m}$, iguais a 4,90%, 9,85% e 18,67%, respectivamente.

É importante frisar que nas tabelas do anexo 15 os resultados obtidos após a imputação efetuada pelos diversos métodos são comparados com os dados dos respondentes da amostra original, e que nos anexos 16, 17 e 18, essas mesmas comparações são feitas com os respondentes das respectivas amostra geradas. Ou seja, cada conjunto de tabelas deve ser analisado isoladamente como se reproduzissem uma situação real de não resposta com taxas de não resposta diferentes.

Em seguida, as tabelas 5.5.1 a 5.5.4 encontradas neste item resumem os resultados da imputação efetuada comparativamente nas diversas amostras e métodos. Ou seja, as comparações têm todas a mesma base: a amostra original de respondentes ($\bar{m} = 1,46\%$), que foi utilizada para a geração dos dados ausentes nos exercícios considerados.

A seguir, a descrição de cada uma das oito tabelas acima referidas.

A tabela A15.1 apresenta as informações sobre o tamanho da amostra original em cada uma das classes de imputação construídas nas duas subpopulações, a de respondentes e a de não respondentes, bem como as taxas de não resposta correspondentes. Nas tabelas A16.1, A17.1 e A18.1 encontram-se as mesmas informações só que relativas às amostras simuladas resultantes após a geração de dados ausentes.

Com o objetivo de analisar a média e o desvio padrão da variável renda, que sofreu imputação, na amostra, as tabelas A15.2, A16.2, A17.2 e A18.2 apresentam estatísticas de tamanho (n), média ($\bar{y}$), desvio padrão ($s_{\bar{y}}$) e coeficiente de variação ($cv_{\bar{y}}$), sobre as duas subpopulações, antes e depois da imputação efetuada pelos quatro métodos.

A fim de verificar ao longo das classes de imputação o comportamento dos métodos que consideram variáveis auxiliares no processo, ou seja, o da média dentro de classes -MC e o aleatório dentro de classes -RC, as tabelas A15.3, A16.3, A17.3 e A18.3 apresentam informações sobre média, desvio padrão e coeficiente de variação da variável renda, para cada classe, antes e depois da aplicação do método da média dentro de classes (MC); e as tabelas A15.4, A16.4, A17.4 e A18.4 são análogas às anteriores, só que para o método aleatório dentro de classes (RC).

Com o objetivo de avaliar o efeito da imputação em tabulações cruzadas entre a variável que sofreu imputação (renda) e outras variáveis investigadas na pesquisa, foi construída a variável número de pessoas ocupadas por faixa de renda (A_9), cujas categorias estão definidas no quadro 5.5; foram selecionadas as seguintes variáveis para fazer as tabulações cruzadas: A_1 - Sexo, A_2 - Condição no domicílio, A_3 - Ramo de atividade, A_7 - Faixa de idade, A_8 - Grau de instrução e Classe, já definidas nos quadros 5.1 e 5.2, e A_{2a}, A_{7a} e A_{8a}, definidas no quadro 5.6, e que foram construídas em função dos resultados da aplicação da técnica SEARCH, ou seja, são as variáveis A_2, A_7 e A_8, respectivamente, só que com algumas de suas categorias agregadas.

A medida utilizada para análise foi o coeficiente de associação C (ou coeficiente de contingência), onde $C = \sqrt{\chi^2/(\chi^2 + N)}$ e χ^2 é o valor da estatística qui-quadrado calculada para uma tabela cruzada com N indivíduos. Essa medida foi escolhida em função de ser limitada, $0 < C < 1$, e ter a vantagem de remover o efeito do tamanho N considerado na tabulação.

Quadro 5.5 - Variável faixa de renda e suas categorias em número de salários mínimos e em cruzados.

Variável	Categorias
A_9 - Faixa de renda	1 - menos de 1 salário mínimo (até Cz$ 1968,00) 2 - 1 a menos de 2 salários mínimos (Cz$ 1969,00 a Cz$ 3937,00) 3 - 2 a menos de 4 salários mínimos (Cz$3938,00 a Cz$ 7875,00) 4 - 4 a menos de 8 salários mínimos (Cz$ 7876,00 a Cz$ 15751,00) 5 - 8 a menos de 20 salários mínimos (Cz$ 15752,00 a Cz$ 39379,00) 6 - 20 salários mínimos ou mais (Cz$ 39380,00 ou mais)

Quadro 5.6 - Variáveis investigadas na pesquisa e suas categorias agregadas após o procedimento SEARCH.

Variável	Categoria
A_{2a} - Condição no domicílio	(1) - Chefe (2 a 8) - não chefe
A_{7a} - Faixa de idade	(1, 2, 3, 4, 5) - até 34 anos (6, 7, 8) - 35 anos ou mais
A_{8a} - Grau de instrução	(0,1,2,3,4,5) - sem instrução até 2º grau incompleto (6) - 2º grau completo (7,8) - superior incompleto ou mais (9) - grau ignorado

Para remover o efeito do número de categorias de cada uma das variáveis consideradas, foi efetuada uma correção na estatística C, que passou a ter a seguinte expressão:

$$C' = (C / \sqrt{(K-1)/K}) \times 100$$ onde

$K = \min (K_1, K_2)$; e

K_1 e K_2 são os números de categorias das variáveis utilizadas no cálculo do coeficiente.

Os valores obtidos para o coeficiente C' antes e depois da imputação efetuada encontram-se nas tabelas A15.5, A16.5 , A17.5 e A18.5 e as relações entre os coeficientes (C' depois/C' antes) encontram-se nas tabelas A15.6, A16.6, A17.6 e A18.6.

Ainda em função da variável número de pessoas ocupadas por faixa de renda (A_9), foi analisada a sua distribuição univariada antes e depois da imputação.

Para essas distribuições, foi construída a seguinte medida de aproximabilidade:

$$\sum_{i=1}^{6} (\text{imputado}_i - \text{real}_i)^2 / \text{real}_i$$

ao longo das 6 faixas de renda da variável A_9, onde imputado_i é o valor do percentual da faixa i após a imputação; e real_i é o valor do percentual da faixa i antes da imputação (só respondentes ou amostra original da PME, conforme o caso).

Os valores dessas distribuições e das medidas de aproximabilidade encontram-se nas tabelas A15.7, A16.7, A17.7 e A18.7.

Finalmente, as tabelas A15.8, A16.8, A17.8 e A18.8 que apresentam a porcentagem de valores imputados nas faixas de renda após a imputação efetuada pelos 4 métodos, permitem avaliar comparativamente a distribuição dos valores imputados ao longo das faixas consideradas.

Como já dito, as tabelas deste item 5.5 apresentam os resultados das estatísticas calculadas após a imputação efetuada pelos 4 métodos nos 3 exercícios de simulação, de forma comparativa com a amostra original que serviu de base para a geração de dados ausentes.

Essas tabelas estão aqui descritas e são analisadas no item 5.6.

A tabela 5.5.1 apresenta, para a média e para o desvio padrão da variável renda na amostra, a relação entre o valor obtido depois da imputação efetuada e o respectivo valor observado na amostra original. O objetivo dessas relações é verificar se os métodos aplicados proporcionam a recomposição do valor original da média e do desvio padrão.

A tabela 5.5.2 apresenta, para os quatro métodos e as três taxas de não resposta consideradas, a medida de aproximabilidade entre distribuições definida anteriormente. As distribuições em questão são: a das pessoas ocupadas por faixa de renda antes da simulação, ou seja, dos respondentes da amostra original e a dessas mesmas pessoas depois da imputação efetuada. O objetivo é verificar como os métodos com diversas taxas de não resposta influenciam a distribuição da variável sujeita à imputação.

A tabela 5.5.3 contém as relações entre os coeficientes de contingência C', anteriormente definidos, obtidos após a imputação efetuada e os mesmos coeficientes C' obtidos na amostra original antes da geração de dados ausentes e das imputações. Como dito anteriormente, o objetivo dessas relações é verificar a influência da aplicação dos diversos métodos com diferentes taxas de não resposta em tabulações cruzadas da variável que sofreu imputação com outras variáveis não sujeitas a ausência de resposta.

Para os exercícios de geração de dados ausentes foram calculadas, ainda, três medidas de desvio descritas a seguir:

$$\text{desvio médio} = \sum_{i=1}^{n} (\hat{y}_i - y_i)/n$$

$$\text{desvio absoluto médio} = \sum_{i=1}^{n} |\hat{y}_i - y_i|/n$$

$$\text{desvio quadrático médio} = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2/n$$

onde

n é o número total de respondentes na amostra original;

$\hat{y}_i$ é o valor da renda imputado para o indivíduo i;

y_i é o valor da renda observado na amostra original da PME, para o indivíduo i;

Os resultados obtídos para essa medidas encontram-se na tabela 5.5.4.

5.5.1 - Relação entre os valores obtidos após a imputação efetuada pelos diversos métodos e taxas de não resposta e o valor observado antes das simulações, para as estatísticas média e desvio padrão da variável renda.

Taxa de não resposta %	$\bar{y}$ depois / $\bar{y}$ observado				$s_{\bar{y}}$ depois / $s_{\bar{y}}$ observado			
	MO	RO	MC	RC	MO	RO	MC	RC
4,90	0,971	0,970	1,003	1,005	0,942	0,961	0,964	0,999
9,85	0,934	0,935	1,001	1,006	0,881	0,928	0,928	1,008
18,67	0,877	0,874	0,986	0,984	0,784	0,856	0,862	0,967

5.5.2 - Medida de aproximabilidade entre a distribuição percentual das pessoas ocupadas por faixas de renda antes da geração das amostras (só os respondentes da amostra original) e depois da imputação efetuada pelos diversos métodos, para as diversas taxas de não resposta consideradas.

Métodos de imputação (*)	Taxas de não resposta (%)		
	4,90	9,85	18,67
RC	0,0007	0,0032	0,0717
RO	0,0712	0,2641	1,0026
MC	0,0990	0,4219	1,2819
MO	0,0553	3,3711	11,1908

(*) a ordem de apresentação dos métodos foi alterada em função dos resultados.

Nota: a medida de aproximabilidade foi definida por:

$$\sum_{i=1}^{6} (\text{imputado}_i - \text{real}_i)/\ \text{real}_i \ ,$$

ao longo das 6 faixas de renda da variável A_8, onde imputado_i é o valor do percentual da faixa i após a imputação; e real_i é o valor do percentual da faixa i antes da imputação.

5.5.3 - Relação entre os coeficientes de contingência C' obtidos após a imputação efetuada pelos diversos métodos e taxas de não resposta e os coeficientes C' obtidos na amostra original antes das simulações.

Variáveis auxiliares	Taxas de não resposta											
	$\bar{m}$ = 4,90%				$\bar{m}$ = 9,85%				$\bar{m}$ = 18,67%			
	MO	RO	MC	RC	MO	RO	MC	RC	MO	RO	MC	RC
A_1	0,969	0,961	0,989	1,000	0,981	0,950	0,984	0,984	0,965	0,892	0,988	0,954
A_2	0,988	0,963	1,011	0,994	0,986	0,928	1,035	1,006	0,971	0,884	1,072	0,991
A_{2a}	0,988	0,960	1,016	1,000	0,981	0,923	1,040	1,009	0,975	0,886	1,090	1,006
A_3	0,971	0,962	0,990	0,971	0,904	0,898	0,981	0,937	0,831	0,793	0,957	0,904
A_7	0,985	0,965	0,999	0,979	0,959	0,905	0,999	0,973	0,942	0,837	1,008	0,934
A_{7a}	0,962	0,939	0,989	0,972	0,885	0,802	0,983	0,956	0,863	0,758	1,005	0,945
A_8	0,969	0,939	1,009	0,991	0,910	0,871	1,026	0,982	0,858	0,768	1,055	0,971
A_{8a}	0,972	0,934	1,020	0,999	0,896	0,855	1,040	0,992	0,841	0,737	1,084	0,986
Classe	0,969	0,948	1,020	0,998	0,919	0,881	1,049	0,995	0,885	0,797	1,114	0,998
média	0,975	0,952	1,005	0,989	0,936	0,890	1,015	0,982	0,903	0,817	1,041	0,965
desvio padrão	0,009	0,012	0,012	0,011	0,038	0,042	0,027	0,022	0,056	0,056	0,051	0,032

5.5.4 - Medidas de desvio para os diversos métodos e taxas de não resposta ($\bar{m}$) consideradas.

Métodos de Imputação (*)	Taxa de não resposta								
	$\bar{m}$ = 4,90%			$\bar{m}$ = 9,85%			$\bar{m}$ = 18,67%		
	Desvio médio	Desvio absoluto médio	Desvio quadrático médio	Desvio médio	Desvio absoluto médio	Desvio quadrático médio	Desvio médio	Desvio absoluto médio	Desvio quadrático médio
MC	25,5	364,9	3101,5	7,9	758,9	4307,5	- 116,8	1389,0	5804,4
RC	38,9	491,4	4242,7	46,9	1101,8	6506,4	- 129,1	1901,4	7887,8
MO	- 235,5	452,6	4017,7	- 539,3	952,4	5662,2	- 1010,5	1692,6	7471,9
RO	- 250,2	569,6	4619,3	- 532,2	1189,7	6604,1	- 1036,6	2109,4	8588,9

(*) a ordem de apresentação dos métodos foi alterada em função dos resultados.

5.6 - Os comentários

A seguir são apresentadas algumas observações efetuadas a partir da análise dos resultados apresentados no item 5.5, que contém as comparações entre as amostras geradas e a amostra original de respondentes que as gerou.

Com relação à média da variável y na amostra (tabela 5.5.1) - os métodos que não utilizam classes de imputação, o da média geral (MO) e o aleatório geral (RO), não recompuseram a média observada da variável renda antes da geração. Observou-se uma subestimação sistemática em função da taxa de não resposta considerada. O mesmo não ocorreu com os métodos que utilizam classes de imputação, média dentro de classes (MC) e aleatório dentro de classes (RC). Os valores da média estimada após a imputação ficaram bastante próximos do valor observado, sendo que os dois desvios relativos maiores foram de -1,4% e -1,6% para os métodos MC e RC, respectivamente, com $\bar{m}$=18.6% ; para $\bar{m}$=4,90% e $\bar{m}$=9,85% todos os desvios ficaram abaixo de 1% .

Com relação ao desvio padrão da variável y na amostra (tabela 5.5.1) - como era esperado, há um melhor desempenho dos métodos aleatórios sobre os respectivos determinísticos. Entretanto, o aleatório geral (RO) e o da média dentro de classes (MC) parecem ter o mesmo desempenho. O uso da média geral (MO) neste caso levaria a uma subestimação máxima de 21,6% . Há que se notar a performance positiva do método aleatório dentro de classes (RC), que para as três taxas de não resposta consideradas nesta aplicação apresentou desvios relativos bastante baixos, 0,1%, 0,8% e 3,3%.

Uma explicação para esse resultado está no fato de o vício esperado de s_y^2 ser zero, quando da suposição de modelo MARC - *missing at random within classes*, tal qual mostram os estudos de Santos apresentado no capítulo 4, quadro 4.3, e que a geração de dados ausentes procurou contemplar.

Com relação à distribuição percentual da variável número de pessoas ocupadas ao longo das faixas de renda (tabela 5.5.2) - considerando as medidas de aproximabilidade calculadas, observou-se o seguinte comportamento: a aderência foi maior para a menor

taxa de não resposta e para o método que considera classes de imputação e aleatorização no processo (RC com $\bar{m}$=4,90%).

Aqui também se percebe o grande vício que seria introduzido por usar inadvertidamente o método MO: a medida de aproximabilidade seria 156 vezes maior que para o método RC, considerando-se $\bar{m}$=18,67% .

Por definição, essa medida de aproximabilidade toma valor zero quando os valores imputados coincidirem exatamente com os valores reais observados. Apesar de a inspiração para a sua utilização residir na estatística χ^2, a sua distribuição de probabilidade não é conhecida, para podermos calcular intervalos de confiança e realizar testes de hipóteses. Podemos apenas afirmar que quanto mais distante de zero é a medida, mais distantes são as distribuições percentuais envolvidas no seu cálculo. Podemos também fazer comparações entre os valores obtidos. Nesse sentido, destacaram-se, por estarem distantes de zero, os valores obtidos com $\bar{m}$=18,67% em todos os métodos exceto no aleatório dentro de classes (RC), e o valor com $\bar{m}$=9,85% e o método da média geral (MO). Os demais apresentaram valores bastante próximos de zero e entre si.

Com relação ao coeficiente de contingência corrigido (tabela 5.5.3) - em linhas gerais os resultados dessa medida na aplicacão em estudo apontam o seguinte comportamento: os métodos que não utilizam classes de imputação, o da média geral (MO) e o aleatório geral (RO), subestimaram sempre; no método aleatório dentro de classes (RC), nos poucos casos em que superestimou, o desvio foi menor que 1%; o método da média dentro de classes (MC) foi o que apresentou maior variação de comportamento, ora subestimando ora superestimando. Com exceção de três casos, o desvio relativo, em módulo, aumentou com o aumento da taxa de não resposta para todos os métodos. Em função do coeficiente de contingência, o método aleatório geral (RO) foi o pior para todas as variáveis em todas as taxas de não resposta.

Note-se que o desvio máximo, em módulo, de 26,3% (C'= 0,737), ocorreu no método RO com $\bar{m}$=18,67%, enquanto que o desvio mínimo observado foi 0% no método RC com $\bar{m}$=4,90% (C'= 1,000). E, considerando a média das variáveis, o desvio máximo, em módulo, de 18,3% (C'= 0,817), ocorreu no método RO com $\bar{m}$=18,67% ; o desvio mínimo foi observado de 0,5% no método MC com $\bar{m}$=4,90% (C'= 1,005).

Com relação às medidas de desvio (tabela 5.5.4) - observou-se

que os métodos que não consideram classes de imputação, MD e RD, indicaram desvio negativo, ou seja, subestimação, tal qual observado na análise da média e do desvio padrão; analisando-se as medidas de desvio absoluto ou quadrático médio, observou-se que esses desvios são maiores para taxas de não resposta maiores.

Contrariamente às anteriores, a análise dessas duas últimas medidas de desvio apontou os métodos determinísticos, que não utilizam aleatorização na seleção do doador, como melhores que os estocásticos. Este fato justifica-se pela própria definição das medidas de desvio consideradas. É sabido que o valor de $\hat{y}_i$ que minimiza a expressão do desvio quadrático médio é a média dos valores observados y_i, e que o valor de $\hat{y}_i$ que minimiza a expressão do desvio absoluto médio é a mediana dos valores observados y_i. E, no caso em questão, nos métodos que utilizam aleatorização, os valores de $\hat{y}_i$ são variáveis, enquanto que nos determinísticos não.

5.7 - Conclusões

Das aplicações efetuadas na amostra original e nas três particulares amostras geradas a título de exercício, com três diferentes taxas de não resposta, cujos resultados encontram-se descritos no item 5.5 e analisados no item 5.6, observamos resumidamente que:

- a particular estrutura de não resposta favoreceu os métodos de imputação por classe;

- à medida que a taxa de não resposta aumentou, piorou a performance dos métodos, de maneira mais acentuada para o método da média geral;

- a recomposição da média da variável imputada mostrou que o método da média geral levou a um vício máximo de 13% ;

- a mesma análise sobre o desvio padrão da variável imputada apontou que o vício neste caso é maior do que para a média;

- a análise da distribuição da variável número de pessoas ocupadas ao longo das faixas de renda apontou que o método aleatório dentro de

classes teve ótima capacidade de recompor essa distribuição;

- a análise da relação da variável imputada com outras variáveis sem dados ausentes mostrou também que o método adequado, o aleatório dentro de classes (RC), foi o que menos afetou a relação entre essas variáveis em tabulações cruzadas;

- a análise das medidas de desvio apontou o método da média dentro de classes como o de menor desvio segundo as medidas adotadas;

- e, para todas as medidas analisadas, os métodos que não consideram variáveis auxiliares no processo, ou seja, não utilizam classes de imputação, produziram vícios que podem ser elevados, e levar a interpretações errôneas.

6. Considerações finais

Os nossos estudos teóricos apresentaram resultados que reforçam e clareiam pontos da bibliografia estudada. Por exemplo, eles ratificaram as conclusões de Kalton e Kasprzyk (1982) que no caso de análises univariadas, ou seja, uma única variável com dados ausentes, os métodos determinísticos de imputação servem bem para estimar médias e totais, mas eles distorcem as propriedades distribucionais da variável; os métodos estocásticos são menos eficientes, no sentido de maior erro padrão do estimador, para estimar médias e totais mas eles preservam a variabilidade dos dados dos respondentes.

Vimos também que considerando a relação da variável imputada com outras não sujeitas à imputação, o método que sobressai é estocástico.

Com relação à aplicação em uma pesquisa real e à determinação de qual método melhor se aplica ao problema específico, no estudo propriamente dito com $\bar{m}$ = 1,46% da amostra original, não foi possível concluir por um método melhor que os demais pois a baixa taxa de não resposta ocasionou quase a mesma performance para todos os métodos. Entretanto, o método aleatório dentro de classes-RC apresentou o melhor desempenho, pois incorpora a particular estrutura de não resposta observada nos dados.

Convém lembrar que nas amostras simuladas a geração de dados ausentes foi realizada independentemente dentro das mesmas classes consideradas no processo de imputação, com o objetivo de caracterizar uma situação do tipo MARC - "missing at random within classes", tal qual definido no anexo 12. E essa situação pode ser a explicação para que o método aleatório dentro de classes - RC se destacasse dos demais. Ou seja, o fato de considerar classes de imputação ao invés de imputar na amostra como um todo pode ter sido o fator mais importante para o método ter dado bons resultados nas aplicações realizadas.

Outro objetivo nosso para a realização dos exercícios foi o comportamento dos métodos de imputação frente a taxas de não resposta diferenciadas. Nesse caso, os resultados obtidos não contrariaram os resultados teóricos já esperados, que quanto maior a taxa de não resposta, piores as estimativas obtidas após a imputação efetuada, para qualquer dos métodos considerados.

Com relação à questão sobre as vantagens da utilização de métodos de imputação para tratamento de não resposta a item, parece-nos interessante dispor de métodos de imputação com suas propriedades estudadas e efeitos conhecidos sobre as diversas estatísticas de interesse, tais como média, total, desvio padrão, covariância, etc, a fim de completar um conjunto de dados antes de proceder a análise do mesmo. Aliás, é nesse aspecto que reside o maior interesse nos métodos de imputação: obter o conjunto completo de dados e proceder as análises ignorando o problema da não resposta.

Entretanto, como pudemos verificar, para cada tipo de estimativa desejada na análise dos dados há um método mais adequado e só uma análise individual da variável com dados ausentes e dos objetivos do estudo pode levar a adotar um ou outro método.

Convém lembrar ainda que os nossos estudos só consideraram o caso univariado de não resposta, ou seja, a ausência de resposta em uma única variável de um conjunto de dados multivariados. Não é, porém, o que ocorre na prática das pesquisas por amostra, que freqüentemente são multivariadas e apresentam não resposta a vários itens simultaneamente. E mais, normalmente os resultados dessas pesquisas são apresentados sob forma de diversas tabelas de contingência, na sua maioria simples, mas em muitos casos múltiplas também. Dessa forma, novos estudos precisam ser realizados a fim de contemplar essas outras situações.

Outros métodos ou procedimentos de imputação vêm sendo propostos na literatura, tais como: o método das imputações múltiplas, citado por Kalton e Kasprzyk (1982), que por sua vez referem-se a Rubin (1978). Aliás sobre esse assunto há o livro de Little e Rubin (1987); a utilização de imputação simultânea de mais de um item a partir de um único indivíduo doador, para o caso de não resposta multivariada, também discutida brevemente por Kalton e Kasprzyk (1982); e métodos robustos, tais como o proposto no artigo de Little e Smith (1987) e analisado no trabalho de Silva (1989).

Bibliografia

FORD, B.L. An overview of hot-deck procedures. In: Madow, William G.; Olkim, Ingram; Rubin, Donald B., eds. Incomplete data in sample surveys. New York, Academic Press, 1983, v.2, p.185-207.

HEERINGA, S.G. Nonresponse in the ISDP 1978 research panel: an examination of rates and patterns of nonresponse in the April (ISDP 303) and July (ISDP 403) interviews. Ann Arbor, Survey Research Center, Institute for Social Research, University of Michigan, 1980.

Hendrix, P.E. Investigation of variables relevant to the imputation of selected income and income-related items in the 1978 ISDP research panel. Ann Arbor, Survey Research Center, Institute for Social Research, University of Michigan, 1980.

IBGE. Metodologia da pesquisa mensal de emprego 1980. Rio de Janeiro IBGE, 1983. (Relatórios metodológicos; v.2)

KALSBEEK, W.D. A conceptual review of survey error due to non response. In: Proceedings of Section on Survey Research Methods. Washington, D.C., American Statistical Association. 1980, p. 131-136.

KALTON, G. and KASPRZYK, D. Imputing for missing survey responses. In: Proceedings of the Section on Survey Research Methods. Washington, D.C., American Statistical Association, 1982, p. 22-23.

LIMA, E.L. Curso de análise. Rio de Janeiro, Instituto de Matemática Pura e Aplicada, CNPq, 1976.

LITTLE, R.J.A. and RUBIN, D.B. Missing data in large data sets. In: Wright, Tommy, ed. Statistical methods and the improvement of data quality. Florida, Academic Press, 1983, p. 215-243.

LITTLE, R.J.A. and SMITH, P.J. Editing and imputation for quantitative survey data. Journal of the American Satatistical Association, 82(397): 58-68, 1987.

PERES, A.D.Q. Uma aplicação do algoritmo "EM" no modelo log-linear. Rio de Janeiro, Instituto de Matemática Pura e Aplicada, 1987. Dissertação de mestrado.

PLATEK, R. and GRAY, G.B. On the definitions of response rates. Survey methodology, 12(1): 17-27, 1986.

RAJ, D. Sampling theory. New Tork, McGraw-Hill, 1968.

RIZVI. M.H. Hot-deck procedures: introduction. In: Madow, William G. and Olkim, Ingram, eds. Incomplete data in sample surveys. New York, Academic Press, 1983, v.3, p.351-352.

RUBIN, B.D. Multiple imputations in sample surveys: a phenomenological bayesian approach to nonresponse. In: Proceedings of the Section on Survey Research Methods. Washington, D.C., American Statistical Association, 1978, p. 20-28.

____. Multiple imputations for nonresponse in surveys. New York, John Wiley & Sons. 1987.

SANDE, I.G. Imputation in surveys: coping with reality. The American Statistician, 36(3): 145-152, August, 1982.

____. Hot-deck imputation procedures. In: Madow, William G. and Olkim Ingram, eds. Imcomplete data in sample surveys. New York, Academic Press, 1983, v.3, p. 339-349.

SANDE, G.T. Replacement for a ten-minute gap. In: Madow, Willian G. and Olkim, Ingram, eds. Incomplete data in sample surveys. New York, Academic Press, 1983, v.3, p. 337-338.

SANTOS, R.L. Effects of imputation on complex statistics. Ann Arbor, Survey Research Center, Institute for Social Research, University of Michigan, 1981a.

____. Effects of imputation on regression coefficients. In:Proceedings of the Section on Survey Research Methods. Washington, D.C., American Statistical Association, 1981b,p. 140-145.

SILVA, P.L.N. Crítica e imputação de dados quantitativos utilizando o SAS. Rio de Janeiro:[IMPA],1989. Dissertação de mestrado.

SONQUIST, J.A.; BAKER, E.L.; MORGAN, J.N.Searching for structure. Ann Arbor, Survey Research Center, Institute for SocialResearch, University of Michigan, 1973.

Proposição A1: $\bar{Y} = \bar{R}\bar{Y}_r + \bar{M}\bar{Y}_m$

Demonstração:

$$\bar{Y} = \sum_{i=1}^{N} Y_i/N = \left(\sum_{i=1}^{R} Y_{ri} + \sum_{i=1}^{M} Y_{mi} \right)/N =$$

$$= \frac{R}{N} \sum_{i=1}^{R} \frac{Y_{ri}}{R} + \frac{M}{N} \sum_{i=1}^{M} \frac{Y_{mi}}{M} = \bar{R}\bar{Y}_r + \bar{M}\bar{Y}_m$$

c.q.d.

ANEXO 2

Proposição A2:

$$S_y^2 \doteq \bar{R}\, S_{ry}^2 + \bar{M}\, S_{my}^2 + \bar{R}\,\bar{M}\,(\bar{Y}_r - \bar{Y}_m)^2$$

Demonstração :

$$S_y^2 = \sum_{i=1}^{N} \frac{(Y_i - \bar{Y})^2}{N-1} \qquad \text{e} \quad \bar{Y} = \bar{R}\,\bar{Y}_r + \bar{M}\,\bar{Y}_m$$

$$\frac{N-1}{N} S_y^2 = \sum_{i=1}^{N} \frac{Y_i^2}{N} - \bar{Y}^2 = \frac{\sum_{i=1}^{R} Y_{ri}^2 + \sum_{i=1}^{M} Y_{mi}^2}{N} - \bar{Y}^2 =$$

$$= \frac{R}{N} \sum_{i=1}^{R} \frac{Y_{ri}^2}{R} + \frac{M}{N} \sum_{i=1}^{M} \frac{Y_{mi}^2}{M} - \bar{Y}^2 =$$

$$= \bar{R} \sum_{i=1}^{R} \frac{Y_{ri}^2}{R} - \bar{R}\,\bar{Y}_r^2 + \bar{M} \sum_{i=1}^{M} \frac{Y_{mi}^2}{M} - \bar{M}\,\bar{Y}_m^2 - \bar{Y}^2 + \bar{R}\,\bar{Y}_r^2 + \bar{M}\,\bar{Y}_m^2$$

$$= \bar{R} \sum_{i=1}^{R} \frac{(Y_{ri} - \bar{Y})^2}{R} + \bar{M} \sum_{i=1}^{M} \frac{(Y_{mi} - \bar{Y}_m)^2}{M} - (\bar{R}\,\bar{Y}_r + \bar{M}\,\bar{Y}_m)^2 + \bar{R}\,\bar{Y}^2 + \bar{M}\bar{Y}_m^2$$

$$= \bar{R}\,\frac{(R-1)}{R} S_{ry}^2 + \bar{M}\,\frac{(M-1)}{M} S_{my}^2 - \bar{R}^2\bar{Y}_r^2 - \bar{M}\,\bar{Y}_m^2 + \bar{R}\,\bar{Y}_r^2 + \bar{M}\,\bar{Y}_m^2$$

$$- 2\,\bar{R}\,\bar{M}\,\bar{Y}_r\,\bar{Y}_m =$$

$$= \frac{R-1}{N} S_{ry}^2 + \frac{M-1}{N} S_{my}^2 + \bar{R}\,(1-\bar{R})\,\bar{Y}_r^2 + \bar{M}\,(1-\bar{M})\,\bar{Y}_m^2 -$$

$$- 2\,\bar{R}\,\bar{M}\,\bar{Y}_r\,\bar{Y}_m =$$

$$= \frac{R-1}{N} S_{ry}^2 + \frac{M-1}{N} S_{my}^2 + \bar{R}\,\bar{M}\,\bar{Y}_r^2 + \bar{R}\,\bar{M}\,\bar{Y}_m - 2\,\bar{R}\,\bar{M}\,\bar{Y}_r\,\bar{Y}_m =$$

$$= \frac{R-1}{N} S_{ry}^2 + \frac{M-1}{N} S_{my}^2 + \bar{R}\,\bar{M}(\bar{Y}_r - \bar{Y}_m)^2$$

ou seja,

$$S_y^2 = \frac{R-1}{N-1} S_{ry}^2 + \frac{M-1}{N-1} S_{my}^2 + \frac{N}{N-1} \bar{R}\,\bar{M}\,(\bar{Y}_r - \bar{Y}_m)^2$$

mas, com as aproximações $M - 1 \doteq M$, $N - 1 \doteq N$ e $R - 1 \doteq R$, temos

$$S_y^2 \doteq \bar{R}\, S_{ry}^2 + \bar{M}\, S_{my}^2 + \bar{R}\,\bar{M}\,(\bar{Y}_r - \bar{Y}_m)^2$$

c.q.d.

ANEXO 3

Proposição A3:

$$S_{xy} \doteq \bar{R}\, S_{rxy} + \bar{M}\, S_{mxy} + \bar{R}\,\bar{M}\,(\bar{X}_r - \bar{X}_m)(\bar{Y}_r - \bar{Y}_m)$$

Demonstração:

$$S_{xy} = \sum_{i=1}^{N} \frac{(\bar{Y}_i - \bar{Y})\,(X_i - \bar{X})}{N - 1}$$

$$\bar{Y} = \bar{R}\,\bar{Y}_r + \bar{M}\,\bar{Y}_m$$

$$\bar{X} = \bar{R}\,\bar{X}_r + \bar{M}\,\bar{X}_m$$

$$\frac{N - 1}{N} S_{xy} = \sum_{i=1}^{N} \frac{(Y_i - \bar{Y})(X_i - \bar{X})}{N - 1} = \sum_{i=1}^{N} \frac{X_i\, Y_i}{N} - \bar{X}\,\bar{Y} =$$

$$= \sum_{i=1}^{R} \frac{X_{ri}\, Y_{ri}}{N} + \sum_{i=1}^{M} \frac{X_{mi}\, Y_{mi}}{N} - \bar{X}\,\bar{Y} =$$

$$= \frac{R}{N} \sum_{i=1}^{R} \frac{X_{ri}\, Y_{ri}}{R} + \frac{M}{N} \sum_{i=1}^{M} \frac{X_{mi}\, Y_{mi}}{M} - \bar{X}\,\bar{Y} =$$

$$= \bar{R}\,\left(\sum_{i=1}^{R} \frac{X_{ri}\, Y_{ri}}{R} - \bar{X}_r\,\bar{Y}_r\right) + \bar{M}\,\left(\sum_{i=1}^{M} \frac{X_{mi}\, Y_{mi}}{M} - \bar{X}_m\,\bar{Y}_m\right) -$$

$$- \bar{X}\,\bar{Y} + \bar{R}\,\bar{X}_r\,\bar{Y}_r + \bar{M}\,\bar{X}_m\,\bar{Y}_m =$$

$$= \bar{R}\,\frac{R - 1}{R} S_{rxy} + \bar{M}\,\frac{M - 1}{M} S_{mxy} + \bar{R}\,\bar{X}_r\,\bar{Y}_r + \bar{M}\,\bar{X}_m\,\bar{Y}_m$$

$$- (\bar{R}\,\bar{X}_r + \bar{M}\,\bar{X}_m)\,(\bar{R}\,\bar{Y}_r + \bar{M}\,\bar{Y}_m) =$$

$$= \frac{R - 1}{N} S_{rxy} + \frac{M - 1}{N} S_{mxy} - \bar{R}^2\,\bar{X}_r\,\bar{Y}_r - \bar{R}\,\bar{M}\,\bar{X}_r\,\bar{Y}_m -$$

$$- \bar{M}\,\bar{R}\,\bar{X}_m\,\bar{Y}_r - \bar{M}^2\,\bar{Y}_m\,\bar{X}_m + \bar{R}\,\bar{X}_r\,\bar{Y}_r + \bar{M}\,\bar{X}_m\,\bar{Y}_m =$$

$$= \frac{R - 1}{N} S_{rxy} + \frac{M - 1}{N} S_{mxy} + \bar{R}(1 - \bar{R})\bar{X}_r\bar{Y}_r + \bar{M}(1 - \bar{M})\bar{X}_m\,\bar{Y}_m$$

$$- \bar{R}\ \bar{M}\ \bar{X}_r\ \bar{Y}_m - \bar{M}\ \bar{R}\ \bar{X}_m\ \bar{Y}_r =$$

$$= \frac{R - 1}{N} S_{rxy} + \frac{M - 1}{N} S_{mxy} + \bar{R}\ \bar{M}(\bar{X}_r - \bar{X}_m)\ (\bar{Y}_r - \bar{Y}_m)$$

ou seja,

$$S_{xy} = \frac{R - 1}{N - 1} S_{rxy} + \frac{M - 1}{N - 1} S_{mxy} + \frac{N}{N-1} \bar{R}\ \bar{M}(\bar{X}_r - \bar{X}_m)\ (\bar{Y}_r - \bar{Y}_m)$$

e com as aproximações $R \doteq 1$, $M \doteq M - 1$ e $N \doteq N - 1$, temos

$$S_{xy} \doteq \bar{R}\ S_{rxy} + \bar{M}\ S_{mxy} + \bar{R}\ \bar{M}(\bar{X}_r - \bar{X}_m)\ (\bar{Y}_r - \bar{Y}_m)$$

c.q.d.

ANEXO 4

Definições do termo "hot deck" segundo alguns autores

Little e **Rubin** (1983): apresentam o método"hot deck"como aquele que atribui o valor de y de uma unidade respondente selecionada aleatoriamente dentro de uma subclasse da população a um valor y ausente, ou seja, de um não respondente, pertencente à mesma subclasse da população.

Sande, G.I. (1982 e 1983): define um procedimento de imputação "hot deck"como sendo aquele onde uma resposta incompleta é completada pelo uso de valores de outro ou outros registros da mesma pesquisa, e a escolha desses registros varia com o registro que necessita de imputação. E mais, simplesmente substituir a média de um estrato de dados completos a um campo ausente não é um procedimento "hot deck" pois a escolha dos registros usados na média é independente do registro que necessita de imputação.

Ford, B.L. (1983): define um procedimento"hot deck"como sendo um procedimento que, em geral, é um processo de duplicação - quando um valor é ausente em uma amostra, um valor registrado é duplicado para representá-lo. Também considera que atribuir a média de valores registrados a cada valor ausente não configura um procedimento"hot deck". Entretanto, refere-se ao adjetivo "hot" que associa à utilização de valores da amostra em estudo, em oposição ao adjetivo "cold", de processos"cold-deck",que referem-se à utilização de informações de outras fontes que não a amostra em estudo.

Kalton e **Kasprzyk** (1982): admitem que o termo"hot deck"tem uma variedade de significados, mas utilizam-no para referir-se ao tipo seqüencial, que é algumas vezes conhecido como o procedimento"hot deck"tradicional. O procedimento seqüencial começa com a especificação de classes de imputa

ção e para cada classe a atribuição de um único valor para a variável y a fim de dar início ao processo.

Sande, G.T. (1983): admite que a variedade de técnicas que são chamadas de procedimentos"hot deck"inclui tanto o método seqüencial, como o método que acessa aleatoriamente os dados para selecionar aquele que vai ser utilizado na imputação.

Rizvi, M.H. (1983): apresenta os procedimentos"hot deck"como uma classe de procedimentos que atribuem aos valores ausentes uma função apropriada dos dados da amostra em estudo.

ANEXO 5

Seja H_j $(j = 1,\ldots,n)$ um conjunto de eventos mutuamente exclusivos $[P(H_k \; H_e) = 0]$ dos quais um necessariamente ocorre. Então qualquer evento pode ocorrer em conjunto com algum H_j. E a probabilidade que uma variável aleatória U tome o valor u_i pode ser dada por:

$$P(U = u_i) = \sum_{j=1}^{n} P(U = u_i,\ H = H_j) = \sum_{j=1}^{n} P(H_j)\ P(U = u_i \mid H_j)$$

O valor esperado condicional de U dado H_j pode ser:

$$E(U \mid H_j) = \sum_i u_i\ P(U = u_i \mid H_j)$$

e a notação usada é $E(U \mid H_j) = E_2(U)$.

A variância condicional de U dado H_j pode ser escrita como:

$$V(U \mid H_j) = E_2(U^2) - [E_2(U)]^2$$

e a notação usada é $V(U \mid H_j) = V_2(U)$.

Proposição A5.1: $E(U) = E\ E(U \mid H_j)$

---

e a notação é $E(U) = E_1\ E_2(U)$ onde $E_2(U)$ é o valor esperado condicional de U dado H_1 e E_1 representa a esperança sobre o espaço de H.

Proposição A5.2: $V(U) = E_1V_2(U) + V_1E_2(U)$

---

onde $V_2(U)$ é a variância condicional de U dado H_j e V_1 representa a variância sobre o espaço de H.

As demonstrações de A5.1 e A5.2 encontram-se em Raj (1968).

No caso considerado, os eventos são apenas dois: H_1 - que significa que há seleção de amostra de resíduos sobre a amostra inicial observada, ou seja, o método é estocástico; e H_2 - que significa que não há amostra de resíduos, ou seja, o método é determinístico.

Proposição A6: $E(\bar{y}_r) = \bar{Y}_r$

Demonstração: Temos que (*):

$$E(\bar{y}_r) = E[E(\bar{y}_r \mid r)]$$

mas $E(\bar{y}_r \mid r) = \bar{Y}_r$ devido às suposições efetuadas em 2.1.

Então $E(\bar{y}_r) = E(\bar{Y}_r) = \bar{Y}_r$

c.q.d.

(*) Ressalte-se as suposições efetuadas de r grande e
$P[r > o] = 1$.

Proposição A7:

$$B(\bar{y}_{MC}) = \bar{M}(\bar{Y}_r - \bar{Y}_m) - \sum_{h=1}^{H} \frac{W_h}{\bar{R}} (\bar{Y}_{rh} - \bar{Y}_r)(\bar{R}_h - \bar{R})$$

$$= B(\bar{y}_{MO}) - \sum_{h=1}^{H} \frac{W_h}{\bar{R}} (\bar{Y}_{rh} - \bar{Y}_r)(\bar{R}_h - \bar{R})$$

Demonstração : Podemos escrever:

$$\sum_{h=1}^{H} \frac{W_h}{\bar{R}} (\bar{Y}_{rh} - \bar{Y}_r)(\bar{R}_h - \bar{R}) = \sum_{h=1}^{H} \left(\frac{R_h}{R} - \frac{N_h}{N}\right) \bar{Y}_{rh}$$

pois $\sum_{h=1}^{H} W_h \left(\frac{\bar{R}_h}{\bar{R}} - 1\right)\bar{Y}_r = \bar{Y}_r \sum_{h=1}^{H} \left(\frac{R_h}{R} - \frac{N_h}{N}\right) = 0$

Então,

$$\bar{M}\ (\bar{Y}_r - \bar{Y}_m) - \sum_{h=1}^{H} \frac{W_h}{\bar{R}} (\bar{Y}_{rh} - \bar{Y}_r)(\bar{R}_h - \bar{R}) =$$

$$= \bar{M} \sum_{h=1}^{H} \frac{R_h}{R} \bar{Y}_{rh} - \bar{M} \sum_{h=1}^{H} \frac{M_h}{M} \bar{Y}_{mh} - \sum_{h=1}^{H} \left(\frac{R_h}{R} - \frac{N_h}{N}\right) \bar{Y}_{rh} =$$

$$= - \bar{M} \sum_{h=1}^{H} \frac{M_h}{M} \bar{Y}_{mh} + \sum_{h=1}^{H} \left(\bar{M}\ \frac{R_h}{R} - \frac{R_h}{R} + \frac{N_h}{N}\right) \bar{Y}_{rh} =$$

$$= - \sum_{h=1}^{H} \frac{M_h}{N} \bar{Y}_{mh} + \sum_{h=1}^{H} \left[\frac{N_h}{N} - \frac{R_h}{R} (1 - \bar{M})\right] \bar{Y}_{rh} =$$

$$= - \sum_{h=1}^{H} \frac{M_h}{N} \bar{Y}_{mh} + \sum_{h=1}^{H} \left(\frac{N_h}{N} - \frac{R_h}{N}\right) \bar{Y}_{rh} = \sum_{h=1}^{H} \frac{M_h}{N} (\bar{Y}_{rh} - \bar{Y}_{mh})$$

c.q.d.

Proposição A8: $E(\frac{1}{r}) \doteq \frac{1}{E(r)}$

Demonstração: Seja $\Delta r = \frac{r - E(r)}{E(r)}$

então $r = E(r)\ (1 + \Delta r)$

Sabemos por Lima (1976) que a série $\sum_{n=o}^{\infty} (-\Delta r)^n$ é convergente para $\{ |\Delta r| < 1 \}$ *, ou seja,

$$\sum_{n=o}^{\infty} (-\Delta r)^n = 1 - \Delta r + (\Delta r)^2 - (\Delta r)^3 + \ldots = \frac{1}{1+\Delta r}$$

Então: $E\ (\frac{1}{r}) = E\ \frac{1}{E(r)\ (1+\Delta r)} = \frac{1}{E(r)}\ E(\frac{1}{1+\Delta r}) =$

$$= \frac{1}{E(r)}\ E\ [\ 1 - \Delta r + (\Delta r)^2 - (\Delta r)^3 + \ldots] \doteq \frac{1}{E(r)}$$

quando são desprezados os fatores de o (n^{-1}) ou menores.

c.q.d.

* A condição $|\Delta r| < 1 \iff |\bar{r} - \bar{R}| < \bar{R}$, que ocorre quando $\bar{r} \leqslant \bar{R}$ ou $\bar{R} < \bar{r} < 2\ \bar{R}$.

Proposição A9:

$$\mathrm{Var}\,(\bar{y}_{MO}) - \mathrm{Var}\,(\bar{y}_{MC}) \doteq \sum_{h=1}^{H} \frac{W_h\,(\bar{Y}_{rh} - \bar{Y}_r)^2}{n}\left(\frac{1}{\bar{R}} - 1\right) +$$

$$+ \sum_{h=1}^{H} \frac{W_h\,S^2_{rhy}}{n}\left(\frac{1}{\bar{R}} - \frac{1}{\bar{R}_h}\right)$$

Demonstração: podemos decompor S^2_{ry} em função da variabilidade entre e dentro das classes, da seguinte forma:

$$S^2_{ry} \doteq \sum_{h=1}^{H} W_h\,(\bar{Y}_{rh} - \bar{Y}_r)^2 + \sum_{h=1}^{H} W_h\,S^2_{rhy}$$

Então, $\mathrm{Var}(\bar{y}_{MO}) - \mathrm{Var}(\bar{y}_{MC}) \doteq \dfrac{S^2_{ry}}{n\bar{R}} - \sum_{h=1}^{H} \dfrac{W_h(\bar{Y}_{rh} - \bar{Y}_r)^2}{n} -$

$$- \sum_{h=1}^{H} \frac{W_h\,S^2_{rhy}}{n\bar{R}_h} = \sum_{h=1}^{H} \frac{W_h(Y_{rh} - Y_r)^2}{n\bar{R}} + \sum_{h=1}^{H} \frac{W_h\,S^2_{rhy}}{n\bar{R}}$$

$$- \sum_{h=1}^{H} \frac{W_h\,S^2_{rhy}}{n\bar{R}_h} - \sum_{h=1}^{H} \frac{W_h\,(\bar{Y}_{rh} - \bar{Y}_r)^2}{n}$$

$$= \sum_{h=1}^{H} \frac{W_h\,(\bar{Y}_{rh} - \bar{Y}_r)}{n}\left(\frac{1}{\bar{R}} - 1\right) + \sum_{h=1}^{H} \frac{W_h\,S^2_{rhy}}{n}\left(\frac{1}{\bar{R}} - \frac{1}{\bar{R}_h}\right)$$

c.q.d.

Proposição A10: Se k = 1

$$Var(\bar{y}_{MR}) \pm \frac{S^2_{ry.z}}{n\bar{R}} + \left[\frac{\bar{M}\ (1 - \bar{M})}{n} + \bar{M}^2\right]_r \frac{(\bar{z}_m - \bar{z}_r)^2}{\sum_{i=1}^{r}(z_{ri} - \bar{z}_r)^2} S^2_{ry.z} +$$

$$+ \beta^2_{r1}\ (\bar{z}_m - \bar{z}_r)^2 \frac{\bar{M}\ (1 - \bar{M})}{n}$$

Demonstração: do item 3.1.5 temos

$$\bar{y}_{MR} = \bar{r}\ \bar{y}_r + \bar{m}\ \bar{y}_{mMR}$$

onde $\bar{y}_{mMR} = b_{ro} + b_{r1}\ \bar{z}_m$

e. $b_{ro} = \bar{y}_r - b_{r1}\ \bar{z}_r$

então, $\bar{y}_{MR} = \bar{y}_r + \bar{m}\ b_{r1}\ (\bar{z}_m - \bar{z}_r)$

Considerando Z = z, fixo, temos:

$$Var\ (\bar{y}_{MR}) = E\left[\ Var(\bar{y}_{MR}|r)\right] + Var\left[E\ (\bar{y}_{MR}|r)\right]$$

Cálculo de $E[Var\ (\bar{y}_{MR}|r)]$:

$$Var\ (\bar{y}_{MR}|r) = Var\left[(\bar{y}_r + \bar{m}\ b_{r1}\ (\bar{z}_m - \bar{z}_r))|r\right] =$$

$$= Var(\bar{y}_r|r) + \bar{m}^2\ (\bar{z}_m - \bar{z}_r)^2\ Var(b_{r1}|r)$$

pois $Cov(\bar{y}_r,\ b_{r1}|r) = 0$

Mas $Var(\bar{y}_r|r) = \frac{S^2_{ry.z}}{r}$

e $Var(b_{r1}|r) = \frac{S^2_{ry.z}}{\sum_{i=1}^{r}(z_{ri} - \bar{z}_r)^2}$

pois pela suposição do modelo de regressão

$$Y_{mi} = \beta_{ro} + \beta_{r1} z_{mi} + \epsilon_{mi} \qquad \text{com } \epsilon_{mi} \sim f(0, S^2_{ry.z})$$

$\forall$ f não especificada.

Então, $\quad Var(\bar{y}_{MR}|r) = \dfrac{S^2_{ry.z}}{r} + \bar{m}^2 (\bar{z}_m - \bar{z}_r)^2 \dfrac{S^2_{ry.z}}{\sum_{i=1}^{r} (z_{ri} - \bar{z}_r)^2}$

e, $\quad E[Var(\bar{y}_{MR}|r)] \doteq \dfrac{S^2_{ry.z}}{n\bar{R}} + \dfrac{(\bar{z}_m - \bar{z}_r)^2}{\sum_{i=1}^{r} (z_{ri} - \bar{z}_r)^2} S^2_{ry.z} \left[\dfrac{\bar{M}(1 - \bar{M})}{n} + \bar{M}^2\right]$

pois $\quad E\left(\dfrac{1}{r}\right) \doteq \dfrac{1}{E(r)} = \dfrac{1}{n\bar{R}}$ como visto anteriormente

e $\qquad E(\bar{m}^2) = [Var(\bar{m}) + E^2(\bar{m})] \doteq$

$$\doteq \frac{\bar{M}(1 - \bar{M})}{n} + \bar{M}^2$$

onde $E(\bar{m})$ e $Var(\bar{m})$ justificam-se pelo fato de m seguir uma distribuição hipergeométrica com parâmetros N, M, n.

Cálculo de $Var[E(\bar{y}_{RM}|r)]$:

$$E(\bar{y}_{MR}|r) = E[(\bar{y}_r + \bar{m}\, b_{r1}(\bar{z}_m - \bar{z}_r)|r] = \bar{Y}_r + \bar{m}\, \beta_{r1}(\bar{z}_m - \bar{z}_r)$$

então, $\quad Var[E(\bar{y}_{MR}|r)] = \beta^2_{r1}(\bar{z}_m - \bar{z}_r)^2 \, Var(\bar{m}) \doteq$

$$\doteq \beta^2_{r1}(\bar{z}_m - \bar{z}_r) \frac{\bar{M}(1 - \bar{M})}{n}$$

Voltando à expressão inicial, temos:

$$Var(\bar{y}_{MR}) \doteq E[Var(\bar{y}_{MR}|r)] + Var[E(\bar{y}_{MR}|r)] =$$

$$= \frac{S^2_{ry.z}}{n\bar{R}} + \frac{(\bar{z}_m - \bar{z}_r)^2}{\sum_{i=1}^{r} (z_{ri} - \bar{z}_r)^2} S^2_{ry.z} \left[\frac{\bar{M}(1 - \bar{M})}{n} + \bar{M}^2\right] + \beta^2_{r1} (\bar{z}_m - \bar{z}_r)^2 \frac{\bar{M}(1 - \bar{M})}{n}$$

c.q.d.

ANEXO 11

Proposição A11: Se k = 1

$$Var(\bar{y}_{RR}) \doteq Var(\bar{y}_{MR}) + \frac{\bar{M}\ S^2_{ry.z}}{n}$$

Demonstração : do item 3.1.6 temos

$$\bar{y}_{RR} = \bar{r}\ \bar{y}_r + \bar{m}\ \bar{y}_{mRR}$$

onde $\bar{y}_{mRR} = b_{ro} + b_{r1}\ \bar{z}_m + \bar{e}_{mRR}$

e $b_{ro} = \bar{y}_r - b_{r1}\ \bar{z}_r$

então $\bar{y}_{RR} = \bar{y}_r + \bar{m}\ b_{r1}\ (\bar{z}_m - \bar{z}_r) + \bar{m}\ \bar{e}_{mRR}$

sabemos que:

$$Var(\bar{y}_{RR}) = E\ [Var(\bar{y}_{RR}|r)] + Var\ [E(\bar{y}_{RR}|r)]$$

Utilizando os operadores E_1 , E_2 , V_1 e V_2 anteriormente definidos, temos:

$$E[\bar{y}_{RR}|r] = E_1[E_2\ (\bar{y}_{RR}|r)]$$

$$Var(\bar{y}_{RR}|r) = V_1[E_2(\bar{y}_{RR}|r)] + E_1[\ V_2(\bar{y}_{RR}|r)]$$

Cálculo de $V_2(\bar{y}_{RR}|r)$:

$$V_2(\bar{y}_{RR}|r) = V_2[(\bar{y}_r + \bar{m}\ b_{r1}(\bar{z}_m - \bar{z}_r) + \bar{m}\ \bar{e}_{mRR})|r] =$$

$$= V_2(\bar{m}\ \bar{e}_{mRR}|r) = \bar{m}^2\ \frac{(r-1)s^2_{ry.z}}{rm}$$

pois $e_{miRR} \sim f(0,\ \frac{r-1}{r}\ s^2_{ry.z})$

Cálculo de $E_1[V_2(\bar{y}_{RR}|r)]$:

$$E_1[V_2\ (\bar{y}_{RR}|r)] = E_1[\bar{m}^2\ \frac{(r-1)}{rm}\ s^2_{ry.z}|r] =$$

$$= \frac{\bar{m}^2\ (r-1)}{rm}\ S^2_{ry.z}$$

Cálculo de $E_2(\bar{y}_{RR}|r)$:

$$E_2(\bar{y}_{RR}|r) = E_2[(\bar{y}_r + \bar{m}\, b_{r1}(\bar{z}_m - \bar{z}_r) + \bar{m}\, \bar{e}_{mRR})|r] =$$

$$= \bar{y}_r + \bar{m}\, b_{r1}(\bar{z}_m - \bar{z}_r)$$

pois $E_2(\bar{e}_{mRR}|r) = 0$

Cálculo de $V_1[E_2(\bar{y}_{RR}|r)]$:

$$V_1[E_2(\bar{y}_{RR}|r)] = V_1[(\bar{y}_r + \bar{m}\, b_{r1}(\bar{z}_m - \bar{z}_r)|r] =$$

$$= V_1(\bar{y}_r|r) + \bar{m}^2(\bar{z}_m - \bar{z}_r)^2 V_1(b_{r1}|r) =$$

$$= \frac{S^2_{ry.z}}{r} + \bar{m}^2(\bar{z}_m - \bar{z}_r)^2 \frac{S^2_{ry.z}}{\sum_{i=1}^{r}(z_{ri} - \bar{z}_r)^2}$$

pois $Cov_1(\bar{y}_r, b_{r1}|r) = 0$

Cálculo de $E(\bar{y}_{RR}|r)$:

$$E(\bar{y}_{RR}|r) = E_1[E_2(\bar{y}_{RR}|r)] = E_1[(\bar{y}_r + \bar{m}\, b_{r1}(\bar{z}_m - \bar{z}_r))|r]$$

$$= \bar{Y}_r + \bar{m}\, \beta_{r1}(\bar{z}_m - \bar{z}_r)$$

pois $E_1(\bar{y}_r|r) = \bar{Y}_r$

e $E_1(b_{r1}|r) = \beta_{r1}$

Cálculo de $Var(\bar{y}_{RR}|r)$:

$$Var(\bar{y}_{RR}|r) = V_1[E_2(\bar{y}_{RR}|r)] + E_1[V_2(\bar{y}_{RR}|r)] =$$

$$= \frac{S^2_{ry.z}}{r} + \bar{m}^2(\bar{z}_m - \bar{z}_r)^2 \frac{S^2_{ry.z}}{\sum_{i=1}^{r}(z_{ri} - \bar{z}_r)^2} + \frac{\bar{m}^2(r-1)}{rm} S^2_{ry.z}$$

Voltando à expressão inicial, temos:

$$Var(\bar{y}_{RR}) = E[Var(\bar{y}_{RR}|r)] + Var[E(\bar{y}_{RR}|r)] \pm$$

$$\doteq \frac{S^2_{ry.z}}{n\bar{R}} + \frac{S^2_{ry.z}}{\sum_{i=1}^{r}(z_{ri}-\bar{z}_r)^2}(\bar{z}_m - \bar{z}_r)^2\left[\frac{\bar{M}(1-\bar{M})}{n} + \bar{M}^2\right] + S^2_{ry.z}\,\frac{\bar{M}}{n}$$

$$+ \beta^2_{r1}\,(\bar{z}_m - \bar{z}_r)^2\,\frac{\bar{M}(1-\bar{M})}{n}$$

pois $\quad E\left(\frac{S^2_{ry.z}}{r}\right) = S^2_{ry.z}\,E\left(\frac{1}{r}\right) \doteq \frac{S^2_{ry.z}}{n\bar{R}}$

$$E(\bar{m}^2) \doteq \frac{\bar{M}(1-\bar{M})}{M} + \bar{M}^2$$

$$E\left(\frac{\bar{m}^2(r-1)}{rm} - S^2_{ry.z}\right) \doteq S^2_{ry.z}\,E\left(\frac{m}{n^2}\right) = S^2_{ry.z}\,\frac{\bar{M}}{n}$$

$$Var(\bar{m}) \doteq \frac{\bar{M}(1-\bar{M})}{n}$$

todos esses resultados já tendo sido utilizados anteriormente.

Ou seja:

$$Var(\bar{y}_{RR}) \doteq Var(\bar{y}_{MR}) + S^2_{ry.z}\,\frac{\bar{M}}{n}$$

c.q.d.

ANEXO 12

Os modelos MAR e MARC

10.1 Os dados são ditos MAR (missing at random), para uma variável específica Y, se a subpopulação de respondentes é uma amostra aleatória simples de tamanho R retirada da população total de tamanho N. Assim, sob esse modelo, as distribuições de Y nas subpopulações de respondentes e de não respondentes coincidirão em média ao longo de repetidas amostras de tamanho R. Ainda mais, se os tamanhos dos subpopulações de respondentes e de não respondentes, R e M, forem grandes relativamente aos tamanhos amostrais, r e m, a variabilidade amostral devida ao processo MAR pode ser considerada desprezível. Então, os parâmetros das duas subpopulações serão aproximadamente iguais, ou seja:

$$\bar{Y}_r \doteq \bar{Y}_m, \quad S^2_{ry} \doteq S^2_{my} \quad , \quad S_{rxy} \doteq S_{mxy}$$

10.2 Os dados são ditos MARC (missing at random within cells), para uma variável especifica Y, se a subpopulação de respondentes dentro de uma classe h (h=1,...,H) é uma amostra aleatória simples de tamanho R_h ($\sum_{h=1}^{H} R_h = R$) da população total de tamanho N_h naquela classe. ($\Sigma\ N_h = N$). Ou seja, para valores grandes de R_h e M_h relativamente aos tamanhos de amostra r_h e m_h na classe h, temos as aproximações, $\forall$ h, (h=1,...,H):

$$\bar{Y}_{rh} \doteq \bar{Y}_{mh} \quad , \quad S^2_{ryh} \doteq S^2_{myh} \quad , \quad S_{rxyh} \doteq S_{mxyh}$$

ANEXO 13

Programa de aplicação da técnica SEARCH, do pacote OSIRIS IV.

```
εDICT DICTOUT = - DICT
CRIACAO DO DICIONARIO
PRINT = OUTD
V =  1   NAME = CONTROLE  COL = 1   WI = 6
V =  2   NAME = SERIE     WI = 2
V =  3   NAME = VAR109    WI = 2
V =  4   NAME = VAR110    WI = 2
V =  5   NAME = V201
V =  6   NAME = V202  WI = 1
V =  7   NAME = V203
V =  8   NAME = V208
V =  9   NAME = V210
V = 10   NAME = V211
V = 11   NAME = V301
V = 12   NAME = V302
V = 13   NAME = V303  WI = 3
V = 14   NAME = V304
V = 15   NAME = V305  WI = 1   MD1 = 0   MD2 = 9
V = 16   NAME = V306  MD1 = 0   MD2 = 9
V = 17   NAME = V307  MD1 = 0   MD2 = 9
V = 18   NAME = V308  MD1 = 0   MD2 = 9
V = 19   NAME = V309  WI = 9    MD1 = NONE   MD2 = NONE
V = 20   NAME = V310  WI = 2
V = 21   NAME = V256  WI = 3
V = 22   NAME = FAIXAID  WI = 2   MD1 = 31
V = 23   NAME = FAIXAN   WI = 1   MD1 =  9
V = 24   NAME = SUBPOP   WI = 1   COL = 50   MD1 = NONE
V = 25   NAME = GRAU MD1 = 9
εSEARCH DICTIN = - DICT   DATAIN = - ARQ
INCLUDE V24 = 1
ANALISANDO O RENDIMENTO MENSAL
DEPV = V19   ANALYSIS = MEANS   IDVAR = V1   PRINT = (FULLTR,TABLE)
VARS = V6, V7   F
VARS = V23   CODES = (1 - 9)  M
VARS = V25   CODES = (0 - 9)  M  END
```

ANEXO 14

```
//SUL20 JOB (SAX01SUL,20,9,0),'DPI.SONIA.1026',TIME=(2,00),
//    NOTIFY=SUL,MSGCLASS=A
//STEP      EXEC SAS,REGION=3072K
//WORK       DD UNIT=WORK,SPACE=(CYL,(10,10),,,ROUND)
//FT11F001   DD SYSOUT=*,DCB=(BLKSIZE=141,LRECL=137,RECFM=VBA)
//FT12F001   DD SYSOUT=*,DCB=(BLKSIZE=141,LRECL=137,RECFM=VBA),
//           COPIES=1
//PME87      DD DSN=SUL.FIX01.DATA,DISP=SHR
//SYSIN      DD *
  DATA GERAL ;
  INFILE PME87 ; /* ARQUIVO COM OS DADOS DA PME COM AS CARACTERISTICAS*/
                 /* DAS PESSOAS INVESTIGADAS */
       INPUT   (CONTROLE N_SERIE  VAR_109 VAR_110 V201-V203 V209-V211
                V301-V310 V256 FAIXAID FAIXAN SUBPOP GRAU)
               ( 6. 2. 2. 2. 2. 1. 1. 1. 1. 1. 1. 1. 3. 3. 1.
                 1. 1. 1. 9. 2. 3. 2. 2. 1. 1.);
DATA ARQ1; SET GERAL; /* ARQ1 - ARQUIVO COM CLASSES CONSTRUIDAS */
  IF V203=1  THEN DO;
     IF GRAU=9 THEN CLASSE=4;
     ELSE IF GRAU=6 THEN CLASSE=6;
     ELSE IF GRAU=0 | GRAU=1 | GRAU=2 |
             GRAU=3 | GRAU=4 | GRAU=5 THEN CLASSE=2;
     ELSE IF GRAU=7 | GRAU=8 THEN DO;
            IF FAIXAN =20 | FAIXAN=30 | FAIXAN=40 | FAIXAN=50
               THEN CLASSE=8;
            ELSE IF FAIXAN=60 | FAIXAN=70 | FAIXAN=80 THEN DO;
                 IF V202=3 THEN CLASSE=7;
                 ELSE IF V202=1 THEN CLASSE=9;
                                                      END;
                                 END;
                    END;
   ELSE  IF V203=2 | V203=3 | V203=4 | V203=5 |
              V203=6 | V203=7 | V203=8 THEN DO;
         IF GRAU=0 | GRAU=1 | GRAU=2 |
            GRAU=3 | GRAU=4 | GRAU=5 THEN CLASSE=1;
         ELSE IF GRAU=6 THEN CLASSE=3;
         ELSE IF GRAU=7 |  GRAU=8 | GRAU=9 THEN CLASSE=5;
                                              END;
/* GERAÇÃO DE DADOS AUSENTES */
DATA ARQ1; SET ARQ1;IF CLASSE=. THEN DELETE;
                    ELSE IF SUBPOP=3 THEN DELETE;
                    ELSE REAL=V309;
PROC UNIVARIATE DATA=ARQ1;
 VAR REAL ;
DATA CL1; SET ARQ1;IF CLASSE=1 ;
     ALEAT1=RANUNI(0);
  PROC SORT DATA=CL1; BY ALEAT1;
  DATA CL1; SET CL1;
    DROP ALEAT1;
    X= 20*.51/1.46;
    IF (_N_<=X*39.34) & (_N_<=50*39.34) THEN DO;
      V309=.;
      SUBPOP=3;
                          END ;
DATA CL2; SET ARQ1;IF CLASSE=2 ;
     ALEAT2=RANUNI(0);
  PROC SORT DATA=CL2; BY ALEAT2;
  DATA CL2; SET CL2;
    DROP ALEAT2;
    X= 20*1.13/1.46;
```

```
    IF (_N_<=X*40.98) & (_N_<=50*40.98) THEN DO;
      V309=.;
      SUBPOP=3;
                          END ;
DATA CL3; SET ARQ1;IF CLASSE=3 ;
     ALEAT3=RANUNI(0);
  PROC SORT DATA=CL3; BY ALEAT3;
  DATA CL3; SET CL3;
    DROP ALEAT3;
    X= 20*1.59/1.46;
    IF (_N_<=X*10.53) & (_N_<=50*10.53) THEN DO;
      V309=.;
      SUBPOP=3;
                          END ;
DATA CL4; SET ARQ1;IF CLASSE=4 ;
     ALEAT4=RANUNI(0);
  PROC SORT DATA=CL4; BY ALEAT4;
  DATA CL4; SET CL4;
    DROP ALEAT4;
    X= 20*2.41/1.46;
    IF (_N_<=X*0.91) & (_N_<=50*0.91) THEN DO;
      V309=.;
      SUBPOP=3;
                          END ;
DATA CL5; SET ARQ1;IF CLASSE=5 ;
     ALEAT5=RANUNI(0);
  PROC SORT DATA=CL5; BY ALEAT5;
  DATA CL5; SET CL5;
    DROP ALEAT5;
    X= 20*3.53/1.46;
    IF (_N_<=X*9.01) & (_N_<=50*9.01) THEN DO;
      V309=.;
      SUBPOP=3;
                          END ;
DATA CL6; SET ARQ1;IF CLASSE=6 ;
     ALEAT6=RANUNI(0);
  PROC SORT DATA=CL6; BY ALEAT6;
  DATA CL6; SET CL6;
    DROP ALEAT6;
    X= 20*2.18/1.46;
    IF (_N_<=X*7.62) & (_N_<=50*7.62) THEN DO;
      V309=.;
      SUBPOP=3;
                          END ;
DATA CL7; SET ARQ1;IF CLASSE=7 ;
     ALEAT7=RANUNI(0);
  PROC SORT DATA=CL7; BY ALEAT7;
  DATA CL7; SET CL7;
    DROP ALEAT7;
    X= 20*6.19/1.46;
    IF (_N_<=X*0.91) & (_N_<=50*0.91) THEN DO;
      V309=.;
      SUBPOP=3;
                          END ;
DATA CL8; SET ARQ1;IF CLASSE=8 ;
     ALEAT8=RANUNI(0);
  PROC SORT DATA=CL8; BY ALEAT8;
  DATA CL8; SET CL8;
    DROP ALEAT8;
    X= 20*1.95/1.46;
```

```
    IF (_N_<=X*3.02) & (_N_<=50*3.02) THEN DO;
      V309=.;
      SUBPOP=3;
                          END ;
DATA CL9; SET ARQ1;IF CLASSE=9 ;
     ALEAT9=RANUNI(0);
  PROC SORT DATA=CL9; BY ALEAT9;
  DATA CL9; SET CL9;
    DROP ALEAT9;
    X= 20*5.07/1.46;
    IF (_N_<=X*4.68) & (_N_<=50*4.68) THEN DO;
      V309=.;
      SUBPOP=3;
                          END ;
DATA ARQ1; SET CL1 CL2 CL3 CL4 CL5 CL6 CL7 CL8 CL9;
/* ARQ1 - ARQUIVO COM DADOS AUSENTES GERADOS */
DATA ARQ1; SET ARQ1;/* CONSTRUçãO DE VARIAVEIS AUXILIARES*/
KEEP V309 V202 V203 FAIXAN GRAU V305 CLASSE SUBPOP
              V203A FAIXANA GRAUA REAL;
IF FAIXAN=10 | FAIXAN=20 | FAIXAN=30 | FAIXAN=40 | FAIXAN=50
             THEN FAIXANA=1;
ELSE IF FAIXAN=60 | FAIXAN=70 | FAIXAN=80 THEN FAIXANA=2;
IF V203=1 THEN V203A=1;
ELSE V203A=2;
IF GRAU=9 THEN GRAUA=19;
ELSE IF GRAU=6 THEN GRAUA=16;
ELSE IF GRAU=7 | GRAU=8 THEN GRAUA=17;
ELSE GRAUA=15;
DATA ARQ1;SET ARQ1;/*CONSTRUçãO DE FAIXAS DE RENDA - SALARIO MINIMO*/
IF SUBPOP=1  THEN DO;
IF V309 <=1968 THEN SM=1;
ELSE IF 1969 <= V309 <= 3937 THEN SM=2;
ELSE IF 3938 <= V309 <= 7875 THEN SM=3;
ELSE IF 7876 <= V309 <= 15751 THEN SM=4;
ELSE IF 15752 <= V309 <= 39379 THEN SM=5;
ELSE IF  V309 >= 39380 THEN SM=6;
                    END;
PROC FREQ DATA= ARQ1;
TABLES  V202 V203 FAIXAN GRAU V305 CLASSE
            V203A FAIXANA GRAUA SUBPOP SM;
TABLES V202*SUBPOP;
TABLES V203*SUBPOP;
TABLES V203A*SUBPOP;
TABLES FAIXAN*SUBPOP;
TABLES FAIXANA*SUBPOP;
TABLES GRAU*SUBPOP;
TABLES GRAUA*SUBPOP;
TABLES V305*SUBPOP;
TABLES CLASSE*SUBPOP;
TABLES SM*SUBPOP;
DATA ARQ2; SET ARQ1;
 IF SUBPOP=1;
PROC FREQ DATA= ARQ2;
TABLES V202*SM/ALL;
TABLES V203*SM/ALL;
TABLES V203A*SM/ALL;
TABLES FAIXAN*SM/ALL;
TABLES FAIXANA*SM/ALL;
TABLES GRAU*SM/ALL;
TABLES GRAUA*SM/ALL;
```

```
 TABLES V305*SM/ALL;
 TABLES CLASSE*SM/ALL;
/*                                                                  */
/*         ATENCAO   METODO DA MEDIA  DENTRO DE CLASSES             */
/*                                                                  */
   DATA ARQC; SET ARQ1; /* ARQUIVO TOTAL CLASSIFICADO */
   PROC SORT DATA=ARQC; BY  CLASSE;
   DATA ARQRC; SET ARQC; /* ARQUIVO DE RESPONDENTES CLASSIFICADO*/
     IF SUBPOP=1;
   PROC MEANS DATA=ARQRC NOPRINT;
    VAR V309 ;
    BY  CLASSE;
    OUTPUT OUT=SAIDARC N=R_CLAS
                       MEAN=MED_CLAS
                       STD=DES_CLAS
                       SUM=TOT_CLAS
                       VAR=VAR_CLAS
                       USS=SQ_CLAS;
/* SAIDARC=ESTATISTICAS DO ARQUIVO DE RESPONDENTES CLASSIFICADO*/
  PROC PRINT DATA=SAIDARC;
/* JUNTAC=ARQUIVO DE ESTATISTICAS DE RESPONDENTES CLASSIFICADO*/
/*          JUNTO COM ARQUIVO TOTAL CLASSIFICADO*/
  DATA JUNTAC ;
   MERGE ARQC SAIDARC;
   BY  CLASSE;
  DATA IMPUTAC; SET JUNTAC; /* ARQUIVO PARA A IMPUTAÇÃO*/
    IF SUBPOP=3 THEN V309= MED_CLAS;
  DATA IMPNRC;SET IMPUTAC; IF SUBPOP=3;
  PROC MEANS DATA=IMPNRC NOPRINT;
   VAR V309 ;
   BY  CLASSE;
   OUTPUT OUT=SAI_IMP N=R_CLAS
                      MEAN=MED_CLAS
                      STD=DES_CLAS
                      SUM=TOT_CLAS
                      VAR=VAR_CLAS
                      USS=SQ_CLAS;
PROC PRINT DATA=SAI_IMP;
DATA FINAL;SET IMPUTAC; /* ARQUIVO COM A IMPUTAÇÃO */
PROC MEANS DATA=FINAL NOPRINT;
 VAR V309 ;
 BY  CLASSE;
 OUTPUT OUT=SAIDA  N=R_CLAS
                   MEAN=MED_CLAS
                   STD=DES_CLAS
                   SUM=TOT_CLAS
                   VAR=VAR_CLAS
                   USS=SQ_CLAS;
PROC PRINT DATA=SAIDA;
PROC SORT DATA=FINAL;
   BY SUBPOP;
PROC MEANS DATA=FINAL NOPRINT;
 VAR V309 ;
 BY  SUBPOP;
 OUTPUT OUT=SAIDA1  N=R_CLAS
                    MEAN=MED_CLAS
                    STD=DES_CLAS
                    SUM=TOT_CLAS
                    VAR=VAR_CLAS
                    USS=SQ_CLAS;
```

```
   PROC PRINT DATA=SAIDA1;
   PROC MEANS DATA=FINAL ;
    VAR V309 ;
DATA FINAL; SET FINAL;
IF SUBPOP=3 THEN DO;
 IF V309 <=1968 THEN SM=1;
 ELSE IF 1969 <= V309 <= 3937 THEN SM=2;
 ELSE IF 3938 <= V309 <= 7875 THEN SM=3;
 ELSE IF 7876 <= V309 <= 15751 THEN SM=4;
 ELSE IF 15752 <= V309 <= 39379 THEN SM=5;
 ELSE IF  V309 >= 39380 THEN SM=6;
                      END;
 PROC FREQ DATA= FINAL;
 TABLES V202*SM/ALL;
 TABLES V203*SM/ALL;
TABLES V203A*SM/ALL;
TABLES FAIXAN*SM/ALL;
TABLES FAIXANA*SM/ALL;
TABLES GRAU*SM/ALL;
TABLES GRAUA*SM/ALL;
TABLES V308*SM/ALL;
TABLES CLASSE*SM/ALL;
TABLES SM*SUBPOP;
/*                                                              */
/*      ATENCAO  METODO ALEATORIO DENTRO DE CLASSES              */
/*                                                              */
DATA ARQR; SET ARQ1;/* ARQR - ARQUIVO COM RESPONDENTES*/
IF SUBPOP=1;
PROC UNIVARIATE DATA= ARQR PLOT;
VAR V309;
PROC SORT DATA=ARQR;
 BY CLASSE;
 PROC MEANS DATA=ARQR;
  VAR   V309 ;
   BY CLASSE;
DATA ARQNR; SET ARQ1; IF SUBPOP=3;/* ARQNR - ARQUIVO COM NÃO RESPON
                                     DENTES */
DATA IMCLA1; SET ARQR; IF CLASSE=1;
DATA GER1; SET IMCLA1;
    ALEAT1= RANUNI(0);
PROC SORT DATA= GER1; BY ALEAT1;
DATA A1(RENAME=(V309=RENDA)); SET GER1;
DATA ARQNR1; SET ARQNR; IF CLASSE=1;
DATA JUNT1; MERGE A1 ARQNR1;
DATA IMP1; SET JUNT1;
   IF SUBPOP=3 THEN V309=RENDA;
   ELSE DELETE;
DATA IMP1; SET IMP1; DROP ALEAT1 RENDA;
PROC APPEND BASE=IMCLA1 DATA=IMP1;
DATA IMCLA2; SET ARQR; IF CLASSE=2;
DATA GER2; SET IMCLA2;
    ALEAT2= RANUNI(0);
PROC SORT DATA= GER2; BY ALEAT2;
DATA A2(RENAME=(V309=RENDA)); SET GER2;
DATA ARQNR2; SET ARQNR; IF CLASSE=2;
DATA JUNT2; MERGE A2 ARQNR2;
DATA IMP2; SET JUNT2;
   IF SUBPOP=3 THEN V309=RENDA;
   ELSE DELETE;
DATA IMP2; SET IMP2; DROP ALEAT2 RENDA;
```

```
PROC APPEND BASE=IMCLA2 DATA=IMP2;
DATA IMCLA3; SET ARQR; IF CLASSE=3;
DATA GER3; SET IMCLA3;
    ALEAT3= RANUNI(0);
PROC SORT DATA= GER3; BY ALEAT3;
DATA A3(RENAME=(V309=RENDA)); SET GER3;
DATA ARQNR3; SET ARQNR; IF CLASSE=3;
DATA JUNT3; MERGE A3 ARQNR3;
DATA IMP3; SET JUNT3;
   IF SUBPOP=3 THEN V309=RENDA;
   ELSE DELETE;
DATA IMP3; SET IMP3; DROP ALEAT3 RENDA;
PROC APPEND BASE=IMCLA3 DATA=IMP3;
DATA IMCLA4; SET ARQR; IF CLASSE=4;
DATA GER4; SET IMCLA4;
    ALEAT4= RANUNI(0);
PROC SORT DATA= GER4; BY ALEAT4;
DATA A4(RENAME=(V309=RENDA)); SET GER4;
DATA ARQNR4; SET ARQNR; IF CLASSE=4;
DATA JUNT4; MERGE A4 ARQNR4;
DATA IMP4; SET JUNT4;
   IF SUBPOP=3 THEN V309=RENDA;
   ELSE DELETE;
DATA IMP4; SET IMP4; DROP ALEAT4 RENDA;
PROC APPEND BASE=IMCLA4 DATA=IMP4;
DATA IMCLA5; SET ARQR; IF CLASSE=5;
DATA GER5; SET IMCLA5;
    ALEAT5= RANUNI(0);
PROC SORT DATA= GER5; BY ALEAT5;
DATA A5(RENAME=(V309=RENDA)); SET GER5;
DATA ARQNR5; SET ARQNR; IF CLASSE=5;
DATA JUNT5; MERGE A5 ARQNR5;
DATA IMP5; SET JUNT5;
   IF SUBPOP=3 THEN V309=RENDA;
   ELSE DELETE;
DATA IMP5; SET IMP5; DROP ALEAT5 RENDA;
PROC APPEND BASE=IMCLA5 DATA=IMP5;
DATA IMCLA6; SET ARQR; IF CLASSE=6;
DATA GER6; SET IMCLA6;
    ALEAT6= RANUNI(0);
PROC SORT DATA= GER6; BY ALEAT6;
DATA A6(RENAME=(V309=RENDA)); SET GER6;
DATA ARQNR6; SET ARQNR; IF CLASSE=6;
DATA JUNT6; MERGE A6 ARQNR6;
DATA IMP6; SET JUNT6;
   IF SUBPOP=3 THEN V309=RENDA;
   ELSE DELETE;
DATA IMP6; SET IMP6; DROP ALEAT6 RENDA;
PROC APPEND BASE=IMCLA6 DATA=IMP6;
DATA IMCLA7; SET ARQR; IF CLASSE=7;
DATA GER7; SET IMCLA7;
    ALEAT7= RANUNI(0);
PROC SORT DATA= GER7; BY ALEAT7;
DATA A7(RENAME=(V309=RENDA)); SET GER7;
DATA ARQNR7; SET ARQNR; IF CLASSE=7;
DATA JUNT7; MERGE A7 ARQNR7;
DATA IMP7; SET JUNT7;
   IF SUBPOP=3 THEN V309=RENDA;
   ELSE DELETE;
DATA IMP7; SET IMP7; DROP ALEAT7 RENDA;
```

```
PROC APPEND BASE=IMCLA7 DATA=IMP7;
DATA IMCLA8; SET ARQR; IF CLASSE=8;
DATA GER8; SET IMCLA8;
    ALEAT8= RANUNI(0);
PROC SORT DATA= GER8; BY ALEAT8;
DATA A8(RENAME=(V309=RENDA)); SET GER8;
DATA ARQNR8; SET ARQNR; IF CLASSE=8;
DATA JUNT8; MERGE A8 ARQNR8;
DATA IMP8; SET JUNT8;
   IF SUBPOP=3 THEN V309=RENDA;
   ELSE DELETE;
DATA IMP8; SET IMP8; DROP ALEAT8 RENDA;
PROC APPEND BASE=IMCLA8 DATA=IMP8;
DATA IMCLA9; SET ARQR; IF CLASSE=9;
DATA GER9; SET IMCLA9;
    ALEAT9= RANUNI(0);
PROC SORT DATA= GER9; BY ALEAT9;
DATA A9(RENAME=(V309=RENDA)); SET GER9;
DATA ARQNR9; SET ARQNR; IF CLASSE=9;
DATA JUNT9; MERGE A9 ARQNR9;
DATA IMP9; SET JUNT9;
   IF SUBPOP=3 THEN V309=RENDA;
   ELSE DELETE;
DATA IMP9; SET IMP9; DROP ALEAT9 RENDA;
PROC APPEND BASE=IMCLA9 DATA=IMP9;
DATA FIMRO; SET IMCLA1 IMCLA2 IMCLA3 IMCLA4 IMCLA5
                IMCLA6 IMCLA7 IMCLA8 IMCLA9;
     /* FIMRO - ARQUIVO COM A IMPUTAÇãO */
PROC SORT  DATA=FIMRO;
       BY  CLASSE;
PROC MEANS DATA=FIMRO;
 VAR   V309 ;
  BY CLASSE;
PROC SORT  DATA=FIMRO;
       BY  SUBPOP;
PROC MEANS DATA=FIMRO;
 VAR   V309 ;
  BY SUBPOP;
PROC MEANS DATA=FIMRO;
 VAR   V309 ;
DATA FIMRO;SET FIMRO;
IF SUBPOP=3 THEN DO;
 IF V309 <=1968 THEN SM=1;
 ELSE IF 1969 <= V309 <= 3937 THEN SM=2;
 ELSE IF 3938 <= V309 <= 7875 THEN SM=3;
 ELSE IF 7876 <= V309 <= 15751 THEN SM=4;
 ELSE IF 15752 <= V309 <= 39379 THEN SM=5;
 ELSE IF  V309 >= 39380 THEN SM=6;
                    END;
 PROC FREQ DATA= FIMRO;
TABLES V202*SM/ALL;
TABLES V203*SM/ALL;
TABLES V203A*SM/ALL;
TABLES FAIXAN*SM/ALL;
TABLES FAIXANA*SM/ALL;
TABLES GRAU*SM/ALL;
TABLES GRAUA*SM/ALL;
TABLES V305*SM/ALL;
TABLES CLASSE*SM/ALL;
TABLES SM*SUBPOP;
```

```
/*                                                                  */
/*            ATENCAO   METODO DA  MEDIA GERAL                      */
/*                                                                  */
DATA ARQY; SET ARQ1;/* ARQY - ARQUIVO COM RESPONDENTES*/
IF SUBPOP=1;
PROC UNIVARIATE DATA=ARQY ;
 VAR V309;
PROC MEANS DATA=ARQY NOPRINT;
 VAR V309;
 OUTPUT OUT=SAIDA1 N=R_GER
                   MEAN=MED_GER
                   STD=DES_GER
                   SUM=TOT_GER
                   VAR=VAR_GER
                   USS=SQ_GER;
  /* SAIDA1 - ARQUIVO COM ESTATISTICAS DE TODOS OS RESPONDENTES*/
  DATA JUNTA1; /* JUNTA A VARIAVEL RENDA MEDIA DOS RESPONDENTES AO
                  ARQUIVO COMPLETO*/
   IF _N_=1 THEN SET SAIDA1;
                 SET ARQ1;
  DATA IMPUTA1;SET JUNTA1;
   IF SUBPOP=3 THEN V309= MED_GER;
   /* IMPUTA1 - ARQUIVO COM A IMPUTAÇÃO */
  PROC UNIVARIATE DATA=IMPUTA1;
   VAR V309;
  PROC SORT DATA=IMPUTA1;
   BY SUBPOP;
  PROC UNIVARIATE DATA=IMPUTA1;
   BY SUBPOP;
   VAR V309;
DATA IMPUTA1; SET IMPUTA1;
IF SUBPOP=3 THEN DO;
IF V309 <=1968 THEN SM=1;
ELSE IF 1969 <= V309 <= 3937 THEN SM=2;
ELSE IF 3938 <= V309 <= 7875 THEN SM=3;
ELSE IF 7876 <= V309 <= 15751 THEN SM=4;
ELSE IF 15752 <= V309 <= 39379 THEN SM=5;
ELSE IF  V309 >= 39380 THEN SM=6;
                     END;
PROC FREQ DATA= IMPUTA1;
TABLES V202*SM/ALL;
TABLES V203*SM/ALL;
TABLES V203A*SM/ALL;
TABLES FAIXAN*SM/ALL;
TABLES FAIXANA*SM/ALL;
TABLES GRAU*SM/ALL;
TABLES GRAUA*SM/ALL;
TABLES V305*SM/ALL;
TABLES CLASSE*SM/ALL;
TABLES SM*SUBPOP;
/*                                                                  */
/*              ATENCAO  METODO  ALEATORIO GERAL                    */
/*                                                                  */
DATA ARQW; SET ARQ1;/* ARQW - ARQUIVO COM RESPONDENTES*/
IF SUBPOP=1;
PROC MEANS DATA=ARQW;
VAR V309;
DATA GERAUNI; SET ARQW;/* GERACAO DE ALEATORIOS UNIFORMES
                          PARA CADA RESPONDENTE */
KEEP  V309   ALEA;
```

```
      ALEA= RANUNI(0);
PROC SORT DATA=GERAUNI; /* ORDENACAO DO ARQUIVO PELOS ALEATORIOS*/
  BY ALEA;
/* AUX-ARQUIVO AUXILIAR QUE REBATIZA V309 E
                              SELECIONA VALORES A IMPUTAR */
DATA AUX(RENAME=(V309=RENDA));SET GERAUNI;
DATA ARQNRX; SET ARQ1; /* ARQUIVO COM NAO RESPONDENTES */
    IF SUBPOP=3;
DATA JUNTAX;/* JUNTA A RENDA DOS RESP. AO ARQ. DE  NAO RESPONDENTES*/
     MERGE   AUX ARQNRX ;
DATA IMPUTAX; SET JUNTAX;/* ARQUIVO COM A IMPUTAÇÃO (SO NAO RESP.) */
     IF SUBPOP=3 THEN V309=RENDA;
       ELSE DELETE;
DATA IMPUTAX;SET IMPUTAX; DROP ALEA RENDA;
PROC APPEND BASE=ARQW DATA=IMPUTAX; /* CONSTROI ARQUIVO COMPLETO*/
/* ARQW - ARQUIVO COM AS IMPUTAÇÕES          */
PROC MEANS DATA=ARQW;
VAR V309;
BY SUBPOP;
PROC MEANS DATA=ARQW;
VAR V309;
DATA ARQW; SET ARQW;
IF SUBPOP=3 THEN DO;
 IF V309 <=1968 THEN SM=1;
 ELSE IF 1969 <= V309 <= 3937 THEN SM=2;
 ELSE IF 3938 <= V309 <= 7875 THEN SM=3;
 ELSE IF 7876 <= V309 <= 15751 THEN SM=4;
 ELSE IF 15752 <= V309 <= 39379 THEN SM=5;
 ELSE IF  V309 >= 39380 THEN SM=6;
                       END;
 PROC FREQ DATA= ARQW;
TABLES V202*SM/ALL;
TABLES V203*SM/ALL;
TABLES V203A*SM/ALL;
TABLES FAIXAN*SM/ALL;
TABLES FAIXANA*SM/ALL;
TABLES GRAU*SM/ALL;
TABLES GRAUA*SM/ALL;
TABLES V305*SM/ALL;
TABLES CLASSE*SM/ALL;
TABLES SUBPOP*SM;
/* PARA ANALISE DE VICIO REAL APOS A IMPUTAÇÃO EFETUADA*/
DATA FINAL ; SET FINAL;
     DIFMC= (V309-REAL);
     ADIFMC= ABS (DIFMC);
     SDIFMC= (DIFMC)**2;
DATA FIMRO;SET FIMRO;
     DIFRC= (V309-REAL);
     ADIFRC= ABS(DIFRC);
     SDIFRC= (DIFRC)**2;
DATA IMPUTA1; SET IMPUTA1;
     DIFMO= (V309-REAL);
     ADIFMO= ABS  (DIFMO);
     SDIFMO= (DIFMO)**2;
DATA ARQW; SET ARQW;
     DIFRO= (V309-REAL);
     ADIFRO= ABS (DIFRO);
     SDIFRO= (DIFRO)**2;
PROC UNIVARIATE DATA=FINAL;
   VAR DIFMC ADIFMC SDIFMC;
```

```
PROC UNIVARIATE DATA=FIMRO;
   VAR DIFRC ADIFRC SDIFRC;
PROC UNIVARIATE DATA=IMPUTA1;
   VAR DIFMO ADIFMO SDIFMO;
PROC UNIVARIATE DATA=ARQW;
   VAR DIFRO ADIFRO SDIFRO;
DATA FINAL3; SET FINAL; IF SUBPOP=3;
DATA FIMRO3; SET FIMRO; IF SUBPOP=3;
DATA IMPUTA13; SET IMPUTA1; IF SUBPOP=3;
DATA ARQW3; SET ARQW; IF SUBPOP=3;
PROC UNIVARIATE DATA=FINAL3 FREQ;
   VAR DIFMC ADIFMC SDIFMC ;
PROC UNIVARIATE DATA=FIMRO3 FREQ;
   VAR DIFRC ADIFRC SDIFRC;
   PROC UNIVARIATE DATA=IMPUTA13 FREQ;
      VAR DIFMO ADIFMO SDIFMO;
   PROC UNIVARIATE DATA=ARQW3 FREQ;
      VAR DIFRO ADIFRO SDIFRO;
   PROC SORT DATA=FINAL; BY CLASSE;
   PROC SORT DATA=FIMRO; BY CLASSE;
   PROC MEANS DATA=FINAL; BY CLASSE;
     VAR DIFMC ADIFMC SDIFMC;
   PROC MEANS DATA=FIMRO; BY CLASSE;
     VAR DIFRC ADIFRC SDIFRC;
   PROC SORT DATA=FINAL3; BY CLASSE;
   PROC SORT DATA=FIMRO3; BY CLASSE;
   PROC MEANS DATA=FINAL3;BY CLASSE;
     VAR DIFMC ADIFMC SDIFMC;
   PROC MEANS DATA=FIMRO3;BY CLASSE;
     VAR DIFRC ADIFRC SDIFRC;
/*
//
```

ANEXO 15

Resultados da aplicação na amostra original, com $\bar{m}$ = 1,46%.

Tabela A 15.1 - Tamanho das classes das duas subpopulações na amostra original ($\bar{m}$ = 1,46%)

Classe	Total (n_h)	Respondentes (r_h)	Não Respondentes (m_h)	(m_h/n_h) (%)
1	3954	3934	20	0,51
2	4145	4098	47	1,13
3	1070	1053	17	1,59
4	83	81	2	2,41
5	934	901	33	3,53
6	779	762	17	2,18
7	97	91	6	6,19
8	308	302	6	1,95
9	493	468	25	5,07
Total	11863	11690	173	1,46

Tabela A 15.2 - Informações sobre as subpopulações antes e depois da imputação efetuada pelos diversos métodos ($\bar{m}$ = 1,46%)

Estatística	Antes da imputação	Após a imputação							
	Respondentes	Não respondentes				Total			
		MO	RO	MC	RC	MO	RO	MC	RC
n	11690	173	173	173	173	11863	11863	11863	11863
$\bar{y}$	8213,0	8213,0 (1,000)	9436,0 (1,149)	13436,4 (1,636)	14511,6 (1,767)	8213,0 (1,000)	8230,9 (1,002)	8289,2 (1,009)	8304,9 (1,011)
$s_{\bar{y}}$	11937,0	- (-)	14255,3 (1,194)	10047,4 (0,842)	19607,5 (1,643)	11850,1 (0,993)	11974,7 (1,003)	11928,1 (0,999)	12106,6 (1,014)
$cv_{\bar{y}}$ (%)	145,3	- (-)	151,1 (1,040)	74,8 (0,515)	135,1 (0,930)	144,3 (0,993)	145,5 (1,001)	143,9 (0,990)	145,8 (1,003)

Nota: Os números entre parênteses representam a quociente entre as informações após a imputação e antes da imputação.

Tabela A 15.3 - Informações sobre as classes na amostra antes e após a imputação efetuada pelo método da média dentro de classes - MC ($\bar{m}$ = 1,46%)

Classe	$\bar{y}$			$s_{\bar{y}}$			$cv_{\bar{y}}$		
	antes	depois	depois / antes	antes	depois	depois / antes	antes	depois	depois / antes
1	3228,8	3228,8	1,000	4646,1	4634,3	0,997	143,9	143,5	0,997
2	6523,4	6523,4	1,000	6358,1	6321,9	0,994	97,5	96,9	0,994
3	7060,7	7060,7	1,000	7875,6	7812,7	0,992	111,5	110,7	0,992
4	10630,5	10630,5	1,000	11279,4	11141,0	0,988	106,1	104,8	0,988
5	12776,3	12776,3	1,000	11852,2	11640,7	0,982	92,8	91,1	0,982
6	15299,6	15299,6	1,000	14140,8	13985,5	0,989	92,4	91,4	0,989
7	20277,1	20277,1	1,000	13771,9	13334,5	0,969	67,9	65,8	0,969
8	23103,7	23103,7	1,000	18704,6	18521,0	0,990	81,0	80,2	0,990
9	34801,6	34801,6	1,000	27741,5	27027,5	0,974	79,7	77,7	0,974

Tabela A 15.4 - Informações sobre as classes na amostra antes e depois da imputação efetuada pelo método aleatório dentro de classes - RC ($\bar{m}$ = 1,46%)

Classe	$\bar{y}$			$s_{\bar{y}}$			$cv_{\bar{y}}$		
	antes	depois	depois / antes	antes	depois	depois / antes	antes	depois	depois / antes
1	3228,8	3228,6	1,000	4646,1	4643,1	0,999	143,9	143,8	0,999
2	6523,4	6527,6	1,001	6358,1	6340,7	0,997	97,5	97,1	0,996
3	7060,7	7072,0	1,002	7875,6	7854,0	1,003	111,5	111,1	0,996
4	10630,5	10561,1	0,993	11279,4	11149,9	0,989	106,1	105,6	0,995
5	12776,3	12754,5	0,998	11852,2	11867,9	1,001	92,8	93,1	1,003
6	15299,6	15384,9	1,006	14140,8	14270,4	1,009	92,4	92,7	1,003
7	20277,1	20700,8	1,021	13771,9	13912,4	1,010	67,9	67,2	0,990
8	23103,7	23040,6	0,997	18704,6	18578,1	0,993	81,0	80,6	0,995
9	34801,6	34994,6	1,006	27741,5	28108,2	1,013	79,7	80,3	1,008

Tabela A 15.5 - Coeficiente de contingência corrigido (C') entre a variável número de pessoas ocupadas por faixa de renda (A_9) e algumas variáveis auxiliares na amostra na subpopulação de respondentes (A_r) e na amostra após a imputação efetuada pelos diversos métodos ($\bar{m}$ = 1,46%)

Variáveis auxiliares	A_r	MO	RO	MC	RC
A_1	36,63	36,48	36,35	36,63	36,49
A_2	37,90	37,79	37,57	38,01	37,90
A_{2a}	45,68	45,40	45,11	45,68	45,69
A_3	23,14	22,92	22,81	23,14	22,92
A_7	38,60	38,60	38,34	38,60	38,60
A_{7a}	25,74	25,88	25,60	25,88	26,02
A_8	59,48	59,04	58,50	59,81	59,37
A_{8a}	57,97	57,62	56,93	58,43	58,08
Classe	66,49	65,95	65,62	66,93	66,60

Tabela A 15.6 - Relação entre os coeficientes de contingência C' obtidos após a imputação efetuada pelos diversos métodos e os coeficientes C' obtidos na amostra na subpopulação de respondentes($\bar{m}$=1,46%)

Variáveis auxiliares	$\frac{C'_{MO}}{C'_{Ar}}$	$\frac{C'_{RO}}{C'_{Ar}}$	$\frac{C'_{MC}}{C'_{Ar}}$	$\frac{C'_{RC}}{C'_{Ar}}$
A_1	0,996	0,992	1,000	0,996
A_2	0,997	0,991	1,003	1,000
A_{2a}	0,994	0,988	1,000	1,000
A_3	0,990	0,986	1,000	0,990
A_7	1,000	0,993	1,000	1,000
A_{7a}	1,005	0,995	1,005	1,011
A_8	0,993	0,984	1,006	0,998
A_{8a}	0,994	0,982	1,008	1,002
Classe	0,992	0,987	1,007	1,002
Média das relações	0,996	0,989	1,003	1,000

Tabela A 15.7 - Distribuição da variável número de pessoas ocupadas por faixa de renda (A_9) na amostra na subpopulação de respondentes (A_r) e na amostra total após a imputação efetuada pelos diversos métodos ($\bar{m}$ = 1,46%)

Faixas de renda (A_9)	A_r		MO		RO		MC		RC	
	Valores Absolutos	Valores Relativos (%)	Valores Absolutos	Valores Relativos (%)	Valores Absolutos	Valores Relativos (%)	Valores Absolutos	Valores Relativos (%)	Valores Absolutos	Valores Relativos (%)
1	1263	10,8	1263	10,7	1284	10,8	1263	10,7	1275	10,7
2	3904	33,4	3904	32,9	3952	33,3	3924	33,1	3933	33,2
3	2951	25,2	2951	24,9	2997	25,3	3015	25,4	2986	25,2
4	2160	18,5	2333	19,6	2192	18,5	2212	18,6	2214	18,7
5	1056	9,0	1056	8,9	1075	9,1	1093	9,2	1086	9,1
6	456	3,1	356	3,0	363	3,0	356	3,0	369	3,1
Total	11690	100,0	11863	100,0	11863	100,0	11863	100,0	11863	100,0
Medida de aproximabilidade			0,0817		0,0050		0,0134		0,0054	

Tabela A 15.8 - Porcentagem de valores imputados nas faixas de de renda após a imputação efetuada pelos diver<u>s</u>os métodos ($\bar{m}$ = 1,46%)

Faixa de renda	Métodos de imputação			
	MO	RO	MC	RC
1	0	1,64	0	0,94
2	0	1,21	0,51	0,74
3	0	1,53	2,12	1,17
4	7,42	1,46	2,35	2,44
5	0	1,73	3,39	2,76
6	0	1,93	0	3,52
Total	1,46	1,46	1,46	1,46

ANEXO 16

Resultados da aplicação na amostra simulada, com $\bar{m}$ = 4,90%.

Tabela A 16.1 - Tamanho das classes das duas subpopulações na amostra, após a simulação com taxa de não resposta $\bar{m}$ = 4,90%

Classe	Total (n_h)	Respondentes (r_h)	Não Respondentes (m_h)	(m_h/n_h) (%)
1	3934	3866	68	1,73
2	4098	3940	158	3,86
3	1053	996	57	5,41
4	81	75	6	7,41
5	901	793	108	11,99
6	762	706	56	7,35
7	91	72	19	20,88
8	302	282	20	6,62
9	468	387	81	17,31
Total	11690	11117	573	4,90

Tabela A 16.2 - Informações sobre as subpopulações antes e depois da imputação efetuada pelos diversos métodos. ($\bar{m}$ = 4,90%)

Estatística	Antes da imputação	Após a imputação							
	Respondentes	Não respondentes				Total			
		MO	RO	MC	RC	MO	RO	MC	RC
n	11117	573	573	573	573	11690	11690	11690	11690
$\bar{y}$	7977,5	7977,5 (1,000)	7677,2 (0,962)	13301,7 (1,667)	13575,0 (1,702)	7977,5 (1,000)	7962,8 (0,998)	8238,5 (1,033)	8251,9 (1,034)
$s_{\bar{y}}$	11529,6	- (-)	10368,6 (0,899)	9872,3 (0,856)	17273,8 (1,498)	11243,4 (0,975)	11475,2 (0,995)	11511,1 (0,998)	11936,4 (1,035)
$cv_{\bar{y}}$ (%)	144,5	- (-)	135,1 (0,935)	74,2 (0,513)	127,2 (0,880)	140,9 (0,975)	144,1 (0,997)	139,7 (0,967)	144,7 (1,001)

Nota: Os números entre parênteses representam o quociente entre as informações após a imputação e antes da imputação.

Tabela A 16.3 - Informações sobre as classes na amostra antes e após a imputação efetuada pelo método da média detro de classes - MC ($\bar{m}$ = 4,90%)

Classe	$\tilde{y}$			$s_{\bar{y}}$			$cv_{\bar{y}}$		
	antes	depois	depois / antes	antes	depois	depois / antes	antes	depois	depois / antes
1	3234,1	3234,1	1,000	4678,2	4637,5	0,991	144,7	143,4	0,991
2	6551,0	6551,0	1,000	6388,2	6263,8	0,981	97,5	95,6	0,981
3	7108,0	7108,0	1,000	8042,4	7821,5	0,973	113,1	110,0	0,973
4	10979,3	10979,3	1,000	11584,9	11142,0	0,962	105,5	101,5	0,962
5	12887,5	12887,5	1,000	12135,1	11383,8	0,938	94,2	88,3	0,938
6	15351,6	15351,6	1,000	14465,5	13923,1	0,963	94,2	90,7	0,963
7	20262,2	20262,2	1,000	13564,2	12047,7	0,888	66,9	59,5	0,888
8	23493,3	23493,3	1,000	18902,9	18264,1	0,966	80,5	77,7	0,966
9	34437,8	34437,8	1,000	26777,1	24344,3	0,909	77,8	70,7	0,909

Tabela A 16.4 - Informações sobre as classes na amostra antes e depois da imputação efetuada pelo método aleatório dentro de classes - RC ($\bar{m}$ = 4,90%)

Classe	$\bar{y}$			$s_{\bar{y}}$			$cv_{\bar{y}}$		
	antes	depois	depois / antes	antes	depois	depois / antes	antes	depois	depois / antes
1	3241,1	3245,9	1,001	4678,2	4679,4	1,000	144,7	144,2	0,997
2	6551,0	6522,4	0,996	6388,2	6322,2	0,990	97,5	96,9	0,994
3	7108,0	7247,6	1,020	8042,4	8362,3	1,040	113,1	115,4	1,020
4	10979,3	11085,8	1,010	11584,9	11568,4	0,999	105,5	104,4	0,990
5	12887,5	12854,6	0,997	12135,1	12479,4	1,028	94,2	97,1	1,031
6	15351,6	15345,8	0,999	14465,5	14545,8	1,006	94,2	94,8	1,006
7	20262,2	20445,9	1,009	13564,2	13917,1	1,026	66,9	68,1	1,018
8	23493,3	23428,4	0,997	18902,9	18783,3	0,994	80,5	80,2	0,996
9	34437,8	34669,8	1,007	26777,1	26451,9	0,988	77,8	76,3	0,981

Tabela A 16.5 - Coeficiente de contingência corrigido (C') entre a variável número de pessoas ocupadas por faixa de renda (A_9) e algumas variáveis auxiliares na amostra na subpopulação de respondentes (A_r) e na amostra após a imputação efetuada pelos diversos métodos ($\bar{m}$ = 4,90%)

Variáveis auxiliares	A_r	MO	RO	MC	RC
A_1	37,05	35,50	35,21	36,21	36,63
A_2	38,12	37,46	36,46	38,34	37,68
A_{2a}	45,96	45,11	43,84	46,39	45,68
A_3	23,03	22,47	22,25	22,92	22,47
A_7	38,34	38,01	37,25	38,56	37,79
A_{7a}	24,89	24,75	24,18	25,46	25,03
A_8	58,94	57,62	55,87	60,03	58,94
A_{8a}	57,39	56,35	54,16	59,12	57,85
Classe	65,95	64,41	62,88	67,81	66,38

Tabela A 16.6 - Relação entre os coeficientes de contingência C' obtidos após a imputação efetuada pelos diversos métodos e os coeficientes C' obtidos na amostra na subpopulação de respondentes($\bar{m}$=4,90%)

Variáveis auxiliares	$\frac{C'_{MO}}{C'_{Ar}}$	$\frac{C'_{RO}}{C'_{Ar}}$	$\frac{C'_{MC}}{C'_{Ar}}$	$\frac{C'_{RC}}{C'_{Ar}}$
A_1	0,958	0,950	0,977	0,989
A_2	0,983	0,957	1,006	0,989
A_{2a}	0.982	0.954	1,009	0,994
A_3	0,976	0,966	0,995	0,976
A_7	0,991	0,971	1,006	0,986
A_{7a}	0,994	0,972	1,023	1,006
A_8	0,978	0,948	1,019	1,000
A_{8a}	0,982	0,944	1,030	1,008
Classe	0,977	0,953	1,028	1,007
Média das relações	0,980	0,957	1,010	0,995

Tabela A 16.7 - Distribuição da variável número de pessoas ocupadas por faixa de renda (A_9) na amostra na subpopulação de respondentes (A_r) e na amostra total após a imputação efetuada pelos diversos métodos ($\bar{m}$ = 4,90%)

Faixas de renda (A_9)	A_r		MO		RO		MC		RC	
	Valores Absolutos	Valores Relativos (%)	Valores Absolutos	Valores Relativos (%)	Valores Absolutos	Valores Relativos (%)	Valores Absolutos	Valores Relativos (%)	Valores Absolutos	Valores Relativos (%)
1	1224	11,0	1224	10,5	1291	11,1	1224	10,5	1261	10,8
2	3779	34,0	3779	32,3	3979	34,0	3847	32,9	3891	33,3
3	2816	25,3	2816	24,1	2956	25,3	3031	25,9	2953	25,3
4	2020	18,2	2593	22,2	2124	18,2	2190	18,7	2169	18,5
5	960	8,6	960	8,2	1008	8,6	1080	9,3	1050	9,0
6	318	2,9	318	2,7	332	2,8	318	2,7	366	3,1
Total	11117	100,0	11690	100,0	11690	100,0	11690	100,00	11690	100,00
Medida de aproximabilidade			1,0762		0,0044		0,1571		0,0554	

Tabela A 16.8 - Porcentagem de valores imputados nas faixas de renda após a imputação efetuada pelos diversos métodos ($\bar{m}$ = 4,90%)

Faixa de renda	Métodos de imputação			
	MO	RC	MC	RC
1	0	5,19	0	2.93
2	0	5,03	1,77	2,88
3	0	4,74	7,09	4,64
4	22,10	4,90	7,76	5,37
5	0	4,76	11,11	8,57
6	0	4,22	0	13,11
Total	4,90	4,90	4,90	4,90

ANEXO 17

Resultados da aplicação na amostra simulada, com $\bar{m}$ = 9,85%.

Tabela A 17.1 - Tamanho das classes das duas subpopulações na amostra, após a simulação com taxa de não resposta $\bar{m}$ = 9,85%

Classe	Total (n_h)	Respondentes (r_h)	Não Respondentes (m_h)	(m_h/n_h) (%)
1	3934	3797	137	3,48
2	4098	3781	317	7,74
3	1053	939	114	10,83
4	81	68	13	16,05
5	901	684	217	24,08
6	762	649	113	14,83
7	91	53	38	41,76
8	302	262	40	13,25
9	468	306	162	34,62
Total	11690	10539	1151	9,85

Tabela A 17.2 - Informações sobre as subpopulações antes e depois da imputação efetuada pelos diversos métodos ($\bar{m}$ = 9,85%)

Estatística	Antes da imputação	Após a imputação							
	Respondentes	Não respondentes				Total			
		MO	RO	MC	RC	MO	RO	MC	RC
n	10539	1151	1151	1151	1151	11690	11690	11690	11690
$\bar{y}$	7673,8	7673,8 (1,000)	7745,7 (1,009)	13231,5 (1,724)	13627,6 (1,776)	7673,8 (1,000)	7680,8 (1,001)	8221,0 (1,071)	8260,0 (1,076)
$s_{\bar{y}}$	11082,7	- (-)	11056,6 (0,998)	9757,5 (0,880)	17728,1 (1,600)	10522,9 (0,949)	11079,6 (0,999)	11083,3 (1,000)	12033,2 (1,086)
$cv_{\bar{y}}(\%)$	144,4	- (-)	142,7 (0,988)	73,7 (0,510)	130,1 (0,901)	137,1 (0,949)	144,3 (0,999)	134,8 (0,934)	145,7 (1,009)

Nota: Os números entre parenteses representam o quociente entre as informações após a imputação e antes da imputação.

Tabela A 17.3 - *Informações sobre as classes na amostra antes e após a imputação efetuada pelo método da média dentro de classes - MC* ($\bar{m}$ = 9,85%)

Classe	$\bar{y}$			$s_{\bar{y}}$			$cv_{\bar{y}}$		
	antes	depois	$\frac{\text{depois}}{\text{antes}}$	antes	depois	$\frac{\text{depois}}{\text{antes}}$	antes	depois	$\frac{\text{depois}}{\text{antes}}$
1	3218,1	3218,1	1,000	4681,8	4599,5	0,982	145,5	142,9	0,982
2	6557,2	6557,2	1,000	6386,2	6134,2	0,960	97,4	93,5	0,960
3	7164,7	7164,7	1,000	7889,7	7450,0	0,945	110,1	104,0	0,945
4	11209,4	11209,4	1,000	11898,0	10888,5	0,915	106,1	97,1	0,915
5	12845,2	12845,2	1,000	12187,0	10616,6	0,871	94,9	82,7	0,871
6	15366,6	15366,6	1,000	14602,2	13474,5	0,923	95,0	87,7	0,923
7	19850,3	19850,3	1,000	14246,5	10829,0	0,760	71,8	54,6	0,760
8	23282,5	23282,5	1,000	18670,8	17386,0	0,931	80,2	74,7	0,931
9	34185,5	34185,5	1,000	27057,3	21866,3	0,809	79,1	64,0	0,809

Tabela A 17.4 - Informações sobre as classes na amostra antes e depois da imputação efetuada pelo método aleatório dentro de classes - RC ($\bar{m}$ = 9,85%)

Classe	$\bar{y}$			$s_{\bar{y}}$			$cv_{\bar{y}}$		
	antes	depois	depois / antes	antes	depois	depois / antes	antes	depois	depois / antes
1	3218,1	3228,0	1,003	4681,8	4645,8	0,992	145,5	143,9	0,989
2	6557,2	6585,3	1,004	6386,2	6379,1	0,999	97,4	96,9	0,995
3	7164,7	7114,2	0,993	7889,7	7808,0	0,990	110,1	109,8	0,997
4	11209,4	11769,7	1,050	11898,0	12779,3	1,074	106,1	108,6	1,024
5	12845,2	12914,9	1,005	12187,0	12392,2	1,017	94,9	96,0	1,012
6	15366,6	15449,7	1,005	14602,2	14823,0	1,015	95,0	95,9	1,009
7	19850,3	20128,4	1,014	14246,5	14938,5	1,049	71,8	74,2	1,033
8	23282,5	23185,0	0,996	18670,8	18569,1	0,995	80,2	80,1	0,999
9	34185,5	34586,2	1,012	27057,3	27672,2	1,023	79,1	80,8	1,011

Tabela A 17.5 - Coeficiente de contingência corrigido (C'), entre a variável número de pessoas ocupadas por faixa de renda (A_9) e algumas variáveis auxiliares na amostra na subpopulação de respondentes (A_r) e na amostra após a imputação efetuada pelos diversos métodos ($\bar{m}$ = 9,85%)

Variáveis auxiliares	A_r	MO	RO	MC	RC
A_1	37,90	35,92	34,79	36,06	36,06
A_2	38,56	37,36	35,16	39,22	38,12
A_{2a}	46,25	44,83	42,15	47,52	46,10
A_3	22,92	20,91	20,79	22,70	21,69
A_7	38,01	37,03	34,95	38,56	37,57
A_{7a}	23,62	22,77	20,65	25,31	24,61
A_8	58,17	54,12	51,82	61,02	58,39
A_{8a}	56,23	51,96	49,54	60,28	57,50
Classe	65,18	61,13	58,61	69,78	66,17

Tabela A 17.6 - Relação entre os coeficientes de contingência C' obtidos após a imputação efetuada pelos diversos métodos e os coeficientes C' obtidos na amostra na subpopulação de respondentes($\bar{m}$=9,85%)

Variáveis auxiliares	$\frac{C'_{MO}}{C'_{Ar}}$	$\frac{C'_{RO}}{C'_{Ar}}$	$\frac{C'_{MC}}{C'_{Ar}}$	$\frac{C'_{RC}}{C'_{Ar}}$
A_1	0,948	0,918	0,951	0,951
A_2	0,969	0,912	1,017	0,988
A_{2a}	0,969	0,911	1,027	0,996
A_3	0,912	0,907	0,990	0,946
A_7	0,974	0,919	1,014	0,988
A_{7a}	0,964	0,874	1,071	1,041
A_8	0,930	0,891	1,048	1,003
A_{8a}	0,924	0,881	1,071	1,022
Classe	0,938	0,899	1,070	1,015
Média das relações	0,948	0,901	1,029	0,994

Tabela A 17.7 - Distribuição da variável número de pessoas ocupadas por faixa de renda (A_9) na amostra na subpopulação de respondentes (A_r) e na amostra total após a imputação efetuada pelos diversos métodos ($\bar{m}$ = 9,85%)

Faixas de renda (A_9)	A_r		MO		RO		MC		RC	
	Valores Absolutos	Valores Relativos (%)	Valores Absolutos	Valores Relativos (%)	Valores Absolutos	Valores Relativos (%)	Valores Absolutos	Valores Relativos (%)	Valores Absolutos	Valores Relativos (%)
1	1183	11,2	1183	10,1	1297	11,1	1183	10,1	1241	10,6
2	3644	34,6	3644	31,2	4030	34,5	3781	32,3	3873	33,1
3	2689	25,5	3840	32,9	3013	25,8	3120	26,7	2977	25,5
4	1908	18,1	1908	16,3	2108	18,0	2251	19,3	2184	18,7
5	842	8,0	842	7,2	938	8,0	1082	9,3	1046	8,9
6	273	2,6	273	2,3	304	2,6	273	2,3	369	3,2
Total	10539	100,0	11690	100,0	11690	100,0	11690	100,0	11690	100,00
Medida de aproximabilidade			2,8832		0,0053		0,6428		0,3568	

Tabela A 17.8 - Porcentagem de valores imputados nas faixas de renda após a imputação efetuada pelos diversos métodos ($\bar{m}$ = 9,85%)

Faixa de renda	Métodos de imputação			
	MO	RO	MC	RC
1	0	8,79	0	4,67
2	0	9,58	3,62	5,91
3	29,97	10,75	13,81	9,67
4	0	9,49	15,24	12,64
5	0	10,23	22,18	19,50
6	0	10,20	0	26,02
Total	9,85	9,85	9,85	9,85

ANEXO 18

Resultados da aplicação na amostra simulada, com $\bar{m}$ = 18,67%.

Tabela A 18.1 - Tamanho das classes das duas subpopulações na amostra, após a simulação com taxa de não resposta $\bar{m}$ = 18,67%

Classe	Total (n_h)	Respondentes (r_h)	Não Respondentes (m_h)	(m_h/n_h) (%)
1	3934	3660	274	6,96
2	4098	3464	634	15,47
3	1053	824	229	21,75
4	81	55	26	32,10
5	901	466	435	48,28
6	762	535	227	29,79
7	91	46	45	49,45
8	302	222	80	26,49
9	468	235	233	49,79
Total	11690	9507	2183	18,67

Tabela A 18.2 - Infomrações sobre as subpopulações antes e depois da imputação efetuada pelos diversos métodos. ($\bar{m}$ = 18,67%)

Estatística	Antes da imputação	Após a imputação							
	Respondentes	Não respondentes				Total			
		MO	RO	MC	RC	MO	RO	MC	RC
n	9507	2183	2183	2183	2183	11690	11690	11690	11690
$\bar{y}$	7202,7	7202,5 (1,000)	7062,8 (0,981)	11988,5 (1,665)	11922,6 (1,655)	7202,5 (1,000)	7176,4 (0,996)	8090,2 (1,124)	8083,9 (1,122)
$s_{\bar{y}}$	10384,0	- (-)	9486,9 (0,914)	8863,7 (0,854)	15051,5 (1,449)	9364,3 (0,902)	10222,2 (0,984)	10287,6 (0,991)	11548,3 (1,112)
$cv_{\bar{y}}$	144,2	- (-)	134,3 (0,931)	73,9 (0,512)	126,2 (0,875)	130,0 (0,901)	142,4 (0,988)	127,1 (0,881)	142,9 (0,991)

Nota: Os números entre parênteses representam o qüociente entre as informações após a imputação e antes da imputação.

Tabela A 18.3 - Informações sobre as classes na amostra antes e após a imputação efetuada pelo método da média dentro de classes - MC. ($\bar{m}$ = 18,67%)

Classe	$\bar{y}$			$s_{\bar{y}}$					$\bar{y}$
	antes	depois	$\frac{\text{depois}}{\text{antes}}$	antes	depois	$\frac{\text{depois}}{\text{antes}}$	antes	depois	$\frac{\text{depois}}{\text{antes}}$
1	3248,2	3248,2	1,000	4785,8	4616,1	0,965	147,3	142,1	0,965
2	6548,8	6548,8	1,000	6438,7	5919,6	0,919	98,3	90,4	0,919
3	6998,4	6998,4	1,000	7868,3	6959,4	0,884	112,4	99,4	0,884
4	9799,6	9799,6	1,000	9654,7	7932,2	0,822	98,5	80,9	0,822
5	11939,1	11939,1	1,000	10315,7	7414,9	0,719	86,4	62,1	0,719
6	15110,6	15110,6	1,000	14796,2	12394,4	0,838	97,9	82,0	0,838
7	21581,7	21581,7	1,000	14064,7	9945,2	0,707	65,2	46,1	0,707
8	22606,3	22606,3	1,000	18522,1	15871,0	0,857	81,9	70,2	0,857
9	33769,5	33769,5	1,000	26105,2	18478,9	0,708	77,3	54,7	0,708

Tabela A 18.4 - Informações sobre as classes na amostra antes e após a imputaçã efetuada pelo método aleatório dentro de classes - RC ($\bar{m}$ = 18,67%)

Classe	$\bar{y}$			$s_{\bar{y}}$			$\bar{y}$		
	antes	depois	depois / antes	antes	depois	depois / antes	antes	depois	depois / antes
1	3248,2	3237,9	0,999	4785,8	4666,2	0,975	147,3	144,1	0,978
2	6548,8	6570,3	1,003	6438,7	6519,6	1,013	98,3	99,2	1,009
3	6998,4	6856,9	0,980	7868,3	7340,8	0,933	112,4	107,1	0,953
4	9799,6	10474,2	1,069	9654,7	9894,9	1,025	98,5	94,5	0,959
5	11939,1	11956,6	1,001	10315,7	10323,7	1,001	86,4	86,3	0,999
6	15110,6	14959,7	0,990	14796,2	14464,9	0,978	97,9	96,7	0,988
7	21581,7	21217,1	0,983	14064,7	13618,0	0,968	65,2	64,2	0,985
8	22606,3	22698,6	1,004	18522,1	18347,8	0,991	81,9	80,8	0,987
9	33769,5	33785,6	1,001	26105,2	26123,7	1,001	77,3	77,3	1,000

Tabela A 18.5 - Coeficiente de contingência corrigido (C') entre a variável número de pessoas ocupadas por faixa de renda (A_9) e algumas variáveis auxiliares na amostra na subpopulação de responden[illegible] efetuada pelos diversos métodos ($\bar{m}$ = 18,67%)

Variáveis auxiliares	A_r	MO	RO	MC	RC
A_1	39,17	35,36	32,67	36,20	34,93
A_2	39,55	36,81	33,52	40,64	37,57
A_{2a}	47,94	44,55	40,45	49,78	45,96
A_3	22,70	19,23	18,34	22,14	20,91
A_7	37,90	36,37	32,32	38,89	36,04
A_{7a}	24,04	22,21	19,52	25,88	24,32
A_8	57,40	51,05	45,68	62,77	57,73
A_{8a}	55,54	48,73	42,73	62,82	57,16
Classe	64,63	58,83	53,02	74,05	66,38

Tabela A 18.6 - Relação entre os coeficientes de contingência C' obtidos após a imputação efetuada pelos diversos métodos e os coeficientes C'obtidos na amostra na subpopulação de respondentes($\bar{m}$=8,67%)

Variáveis auxiliares	$\frac{C'_{MO}}{C'_{Ar}}$	$\frac{C'_{RO}}{C'_{Ar}}$	$\frac{C'_{MC}}{C'_{Ar}}$	$\frac{C'_{RC}}{C'_{Ar}}$
A_1	0,903	0,834	0,924	0,892
A_2	0,931	0,850	1,028	0,950
A_{2a}	0,929	0,844	1,038	0,959
A_3	0,847	0,808	0,975	0,921
A_7	0,960	0,853	1,026	0,951
A_{7a}	0,924	0,812	1,076	1,012
A_8	0,889	0,796	1,094	1,006
A_{8a}	0,877	0,769	1,131	1,029
Classe	0,910	0,820	1,146	1,027
Média das relações	0,908	0,821	1,049	0,972

Tabela A 18.7 - Distribuição da variável número de pessoas ocupadas por faixa de renda (A_9) na amostra na subpopulação de respondentes (A_r) e na amostra total após a imputação efetuada pelos diversos métodos. ($\bar{m}$ = 18,67%)

Faixas de renda (A_9)	A_r		MO		RO		MC		RC	
	Valores Absolutos	Valores Relativos (%)	Valores Absolutos	Valores Relativos (%)	Valores Absolutos	Valores Relativos (%)	Valores Absolutos	Valores Relativos (%)	Valores Absolutos	Valores Relativos (%)
1	1138	12,0	1138	9,7	1399	12,0	1138	9,7	1300	11,1
2	3409	35,9	3409	29,2	4196	35,9	3683	31,5	3877	33,2
3	2416	25,4	4599	39,3	2980	25,5	3279	28,1	2950	25,2
4	1631	17,2	1631	14,0	1995	17,1	2319	19,8	2153	18,4
5	716	7,5	716	6,1	878	7,5	1074	9,2	1090	9,3
6	197	2,1	197	1,7	242	2,1	197	1,7	320	2,7
Total	9507	100,0	11690	100,0	11690	100,0	11690	100,0	11690	100,0
Medida de aproximabilidade			10,2301		0,0010		2,1217		0,9593	

Tabela A 18.8 - Porcentagem de valores imputados nas faixas de renda após a imputação efetuada pelos diversos métodos ($\bar{m}$ = 18,67%)

Faixa de renda	Métodos de imputação			
	MO	RO	MC	RC
1	0	18,66	0	12,46
2	0	18,76	7,44	12,07
3	0	18,93	26,32	18,10
4	47,47	18,25	29,67	24,25
5	0	18,45	33,33	34,31
6	0	18,60	0	38,44
Total	18,67	18,67	18,67	18,67

INFORMES' DE MATEMÁTICA

SÉRIE D

001/82 - Sobre a Função Custo de um Sistema de Componentes Independentes de Tempo Máximo de Não Funcionamento.

Rubens Wolf Francisco

002/82 - Estudo Comparativo de Estimadores do ED50.

Henrique Westenberger

003/83 - Um Método de Aproximação Estocástica usando Regressão Isotônica.

Gregorio Saravia Atuncar

004/84 - Equilíbrio Geral e o Core em Economias com Infinitos Bens.

Paulo Klinger Monteiro

005/84 - Alguns Aspectos do Cálculo de Confiabilidade em Redes de Comunicação.

Margarida Pinheiro Mello

006/84 - Sistemas Markovianos com Infinitas Componentes - O Problema de Existência do Processo Associado.

Sara Carmona

007/84 - Considerações sobre Modelos de Oferta de Produto Agrícola: Soja no Estado de São Paulo 1973-1980.

Leandro Vicente Fernandes Maniero

008/84 - Intervalo de Confiança de Comprimento Pré-Fixado para o Processo de Robbins-Monro Adaptado.

Deborah Pereira Medeiros

009/84 - As Estatísticas de Kolmogorov-Smirnov Generalizada e Máximo do χ^2: Distribuições e Aplicações.

Cibele Comini Cesar

010/84 - Regressão com Dados Censorados.

José Eduardo Corrente

011/84 - Alguns Modelos para Interação em Análise de Variância com 2 Fatores e uma Observação por Cela

Francisco Antônio Morais de Souza

012/85 - Discriminação e Classificação em Curvas de Crescimento

Pedro Alberto Barbetta

013/85 - Comparação entre Análise Univariada e Multivariada para alguns Modelos de Medidas Repetidas

Alba Maria Brazão Teixeira

014/85 - A Eficiência Assintótica do Estimador de Kaplan-Meier

Cristina Costa Neto

015/85 - Seleção de Amostra com Probabilidade Desiguais - O Método dos Grupos Aleatórios

José Matias de Lima

016/85 - Um Estudo Monte Carlo sobre o Comportamento de Diversos Algoritmos para o Cálculo de Médias e Variâncias Amostrais

Nuno Duarte da Costa Bittencourt

017/85 - Aplicações de Controle Ótimo ao Mercado de Petróleo

Roberto Apelfeld

018/85 - Alguns Aspectos da Teoria dos Jogos e uma Aplicação ao Problema da Dívida Externa

Solange de Oliveira Roza

019/85 - Um Algoritmo para a Análise de Experimentos Planejados

José Carlos da Rocha Castelar Pinheiro

020/86 - Estimação em Pequeno Domínio

Rosely Moraes Garcia

021/86 - Teste de Significância Repetido Modificado no Caso de Duas Médias Binomiais.

Tania Guillén de Torres

022/86 - A Indústria de Papel e Celulose: Um Modelo de Alocação de Fábricas de Celulose Fibra Curta Autosuficientes em Energia e Matéria-Prima

Maria de Fátima Salles Abreu Passos

023/86 - Teoremas Sobre o Tempo de Ocupação no Modelo do Votante

Claudio Landim

024/87 - Uma Aplicação do Algoritmo "EM" no Modelo Log-Linear

Antonieta D'Alcântara Q. Peres

025/87 - Resíduos Recursivos

Renato Martins Assunção

026/89 - Cálculo de Medidas Transitórias de Confiabilidade e Desempenho para Sistemas de Computação

Moacyr Pereira dos Santos

027/89 - A Reposição de Dados Omissos em Experimentos Planejados

Antonio Carlos M. Ponce de Leon

028/89 - Crítica e Imputação de Dados Quantitativos Utilizando o SAS

Pedro Luis do Nascimento Silva

029/89 - Análise Comparativa de Testes para Detecção de Periodicidades em Séries Temporais

Antonio Marcos Duarte Júnior

030/89 - Um Modelo Para Confiabilidade de Componentes
Sylvia Susana Campodónico Viacava

031/89 - Um Estimador para o Vício da Taxa Aparente de Erro na Regressão Logistica

Beatriz Mendes Luz

032/89 - Tópicos de Robustez Quantitativa na Família das Distribuições Exponenciais a um Parâmetro

Rodrigo Salgado Labouriau

033/90 - Implementações de tarjan para Métodos de Caminhamento em Grafos.

Enio Emanuel Ramos Russo

034/90 - Estatística na Análise e na Síntese de Imagens

Alejandro César Frery

035/90 - Um Modelo de Otimização para o Planejamento da Expansão da Capacidade de Geração de um Sistema Hidrotérmico

Nora Marcela Campodónico Viacava

036/90 - Variação do Nível Socio Econômico dos Classificados no Vestibular Cesgranrio entre 1975 e 1981.

Rosana de Freitas Castro

037/90 - Comparação entre o MQ, MML, M e GM-Estimadores para o Modelo de Regressão

Glaura C. Franco

038/90 - Modelagem de Persistência a Longo Prazo em Séries Temporais

Antonio Fernando Pêgo e Silva

039/90 - Modelos Estocasticos em Finanças

Armando Rodrigues Gomes Junior

040/90 - Métodos Alternativos para Identificação da Ordem de Modelos Arma (P,Q)

Renata Patricia Lima Jeronymo

041/91 - Um Estudo Sobre Alguns Problemas de Escalonamento (Scheduling)

José Carlos Amaral de Almeida

042/91 - Modelos de Associação e Modelos de Correlação para Análise de Tabelas de Contingência com duas Entradas

Celso Rômulo Barbosa Cabral

043/91 - Um Pacote em Basic para Estimação Robusta nos Modelos de Posição e de Regressão

Paulo Roberto Medeiros de Azevedo

044/91 - Algoritmos para o Problema de Caminhos Mínimos na Presença de ciclos negativos

Jorge José Correia Salles

045/91 - Alguns Algoritmos de Pontos Interiores para o Problema de Programação Linear

José Luiz Zava Coelho

046/91 - "Bidask Spread" e o Problema do "Market Maker"
Marcelo de Oliveira e Costa Nazareth

047/92 - Estimação na Família Tweedie e uma Aplicação à Seguros de Automóveis

Gabriel Yáñez Canal

048/92 - A Ausência de Resposta em Pesquisas: uma Aplicação de Métodos de Imputação

Sonia Albieri

049/92 - Sobre a Unicidade Global do Equilíbrio Walrasiano

Alejandro Felipe Lugos Ceruti

Impresso na Gráfica do
impa

pelo Sistema Xerox/1065