3 Pressuposto de independência condicional - CIA

A suposição de um modelo específico para (***X***, ***Y***, ***Z***) onde a independência de **Y** e **Z** dado **X** é válida, chamada de suposição de independência condicional - CIA, tem tido um papel extremamente importante no emparelhamento estatístico. Nos primórdios das aplicações de emparelhamento estatístico foi suposta de maneira explícita ou implícita, mesmo que não fosse obrigatoriamente válida. A razão é simples: esse modelo é identificável e estimável diretamente.

De fato, se a CIA for válida, a densidade conjunta f (**x , y, z**) pode ser fatorada como:

$$f(\mathbf{x}, \mathbf{y}, \mathbf{z}) = f_{Z|X}(\mathbf{z}|\mathbf{x})\, f_{Y|X}(\mathbf{y}|\mathbf{x})\, f_X(\mathbf{x}),\ \forall\ \boldsymbol{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y} e \mathbf{z} \in \mathcal{Z}, \tag{3.1}$$

onde $f_{Y|X}$ é a densidade condicional de **Y** dado **X** , e $f_{Z|X}$ é a densidade condicional **Z** dado **X**.

Para estimar (3.1), seria suficiente o processo de concatenar os registros com valores similares de **X**, o vetor de variáveis comuns, das distribuições marginais conjuntas de **X** e **Y** e de **X** e **Z**, usando **X.** Realmente, essa informação está disponível nas amostras distintas A e B.

Conforme antecipado na seção 1.3.3, esse capítulo, na condução da teoria sobre o emparelhamento estatístico, supõe que se tem em mãos uma base de dados onde a CIA é válida, isto é, arquivos, para os quais, esse modelo é verdadeiro. Os efeitos de supor um modelo incorreto foram antecipados na seção 1.3.3.

O citado processo dispõe de uma vastíssima gama de procedimentos disponíveis para resolver o problema de emparelhar estatisticamente a informação $A \cup B$ (Figura 1 na seção 1.2). Em primeiro lugar apresenta uma estimação direta da distribuição conjunta de (3.1) ou de qualquer característica importante da distribuição, por exemplo, o coeficiente de correlação, que seria

uma abordagem macro. Entretanto, desenvolvimentos de emparelhamento estatístico, onde a CIA é válida, tem também dado especial destaque à reconstrução do arquivo sintético, cunhado de abordagem micro, vide seção 1.3.1.

D'Orazio et al., 2006, no seu segundo capítulo, descrevem as abordagens alternativas micro ou macro, aplicáveis tanto para uma estrutura não-paramétrica como para um conjunto, $\Im = \{f\}$, de distribuições paramétricas. Procedimentos mistos e bayesianos também são abordados.

Nessa tese descrevem-se procedimentos micro não-paramétricos.

3.1.
Procedimento micro não-paramétrico sob a CIA

Os primeiros artigos, entre eles o de Okner (1972), foram focados na definição de uma base de dados sintética ampliada através da fusão dos arquivos A e B, sem que qualquer suposição de uma família de distribuição paramétrica particular fosse feita para as variáveis de interesse. Esse objetivo pode ser alcançado de duas formas alternativas:

(i) seleções aleatórias, tendo estimado a distribuição de (***X***, ***Y***, ***Z***) em uma estrutura não-paramétrica;

(ii) Emparelhamento usando média condicional que tem estimado funções de regressão não-paramétricas das variáveis a serem imputadas dado os valores observados.

Todavia, as duas abordagens alternativas podem ser aplicadas usando vários procedimentos de estimação não-paramétricos. A aplicação a um conjunto particular de dados de procedimentos de imputação não-paramétricos é uma prática usual, geralmente denominada procedimento de imputação *hot deck.* Esses procedimentos são caracterizados pelo fato que preenchem os valores faltantes com valores observados do arquivo doador.

Os procedimentos de imputação da família *hot deck* são atraentes porque prescindem de qualquer especificação de família de distribuição, ou seja, são não-paramétricos, e não necessitam de qualquer estimativa da função de distribuição ou de qualquer outra de suas características. Não obstante, será visto que os métodos *hot deck* assumem, implicitamente uma estimativa particular, seja uma distribuição ou uma função de média condicional. Os procedimentos serão descritos em certo nível de detalhe.

A maioria dos artigos de emparelhamento estatístico que fazem uso dos métodos *hot deck*, dentre eles Singh et al. (1993), tem adotado implicitamente a CIA. Nesse caso, são atribuídos diferentes papéis às duas amostras A e B. Uma amostra é suposta como arquivo receptor: os itens omissos em cada registro do arquivo receptor são imputados, usando itens adequadamente escolhidos da outra amostra que é o arquivo doador.

A escolha de qual dos arquivos será o receptor ou doador depende de muitos fatores. O mais importante são os fenômenos em estudo e a qualidade da informação contida nos dois arquivos.

Observe que a prática padronizada, para decidir quem será o doador ou receptor, para os arquivos com dados de qualidade confiável, depende do seu tamanho amostral. Por exemplo, se o menor dos arquivos for o doador, alguns dos seus registros serão imputados mais de uma vez no arquivo receptor, que artificialmente poderá modificar a variabilidade da distribuição da variável imputada no arquivo síntese final.

Três métodos *hot deck* têm sido usados no emparelhamento estatístico (Singh et al.,1993):

(i) *Random hot deck* (*hot deck* aleatório);

(ii) *Rank hot deck* (*hot deck* por posto);

(iii) *Distance hot deck* (*hot deck* por distância).

Esses métodos são considerados não-paramétricos, podendo ser vistos como uma contra parte aos procedimentos paramétricos com abordagem micro.

3.2.
Técnicas *hot deck*

3.2.1.
Random hot deck

O *random hot deck* consiste na escolha aleatória do registro a ser doado, no arquivo doador, para cada registro no arquivo receptor. Em particular, algumas vezes, uma escolha aleatória é realizada em subconjuntos adequados de unidades nos arquivos doadores. Geralmente, as unidades de ambos os arquivos são grupados em subconjuntos homogêneos, de acordo com algumas características, por exemplo: unidades da mesma região geográfica, indivíduos com as mesmas características demográficas, etc. Esses subconjuntos são denominados classes de doadores.

A predição via *random hot deck*, nas classes de doadores definidas através de **X**, que é suposta variável categórica, é equivalente à estimação da distribuição condicional de **Z** dado **X** em B (arquivo doador), onde uma observação do mesmo é selecionada. Quando **Z** é uma variável contínua, a distribuição de **Z** dado **X**, $F_{Z/X}$ é estimada pela função de distribuição cumulativa empírica $\hat{F}_{Z/X}$. O mesmo é válido quando **Z** é categórica.

Teoricamente, existem $n_B^{n_A}$ possíveis conjuntos de registros que podem ser escolhidos do B (arquivo doador) como doadores.

Quando o *random hot deck* aleatório é executado, sem considerar as classes de doadores, está se assumindo que **Z** e **X** são independentes. Então, a distribuição marginal empírica de **Z** em B é usada para gerar os valores a serem imputados.

3.2.2. *Rank hot deck*

Caso exista uma variável de emparelhamento ordenável **X***,* essa pode ser usada para selecionar os doadores de B a serem atribuídos aos registros de A , mesmo que não sejam conceitualmente idênticas. Nesta situação, é possível explorar a relação de ordenamento entre os valores de **X**: uma aplicação desse método está em Singh et al., 1990 *apud* D'orazio et al., 2006, p. 39, supondo $n_B = n_A$.

As unidades são ordenadas, separadamente, em ambos os arquivos, de acordo com os valores de X_a^A e X_b^B. Os arquivos são concatenados usando a associação dos registros, que possuam o mesmo posto.

Para simplificar, se A é o arquivo receptor e $n_B = kn_A$, com k inteiro, arquivos são emparelhados pela associação dos registros com os mesmos postos (*rank*). Quando os arquivos contêm um número de registros diferentes, efetua-se o emparelhamento a partir da função de distribuição cumulativa empírica de **X** no arquivo receptor:

$$\hat{F}_X^A(x) = \frac{1}{n_A}\sum_{a=1}^{n_A} \mathrm{I}\left(x_a^A \le x\right), \qquad \boldsymbol{x} \in \mathcal{X} \tag{3.2.2.1}$$

e no arquivo doador:

$$\hat{F}_X^B(x) = \frac{1}{n_B}\sum_{b=1}^{n_B} \mathrm{I}\left(x_b^B \le x\right), \qquad \boldsymbol{x} \in \mathcal{X}. \tag{3.2.2.2}$$

Então, cada $a = 1,.....,n_A$ é associado com o registro b^* em B tal que:

$$\left|\hat{F}_X^A\left(x_a^A\right) - \hat{F}_X^B\left(x_{b^*}^B\right)\right| = \min_{1\le b\le n_B}\left|\hat{F}_X^A\left(x_a^A\right) - \hat{F}_X^B\left(x_b^B\right)\right|. \tag{3.2.2.3}$$

3.2.3.
Distance hot deck

Esse tipo de técnica de emparelhamento foi usada amplamente nas aplicações iniciais de emparelhamento estatístico, citam-se Okner (1972), Ruggles e Ruggles (1974) e Rodgers (1984), dentre outros. Cada registro no arquivo receptor é concatenado com o registro mais próximo, existente no arquivo doador, de acordo com uma medida de distância, computada usando as variáveis de concatenação. Por exemplo, no caso mais simples de variáveis contínuas X, significa que o doador para o $a-$*ésimo* registro no arquivo receptor A é escolhido tal que:

$$d_{ab^*} = \left| x_a^A - x_{b^*}^B \right| = \min_{1 \le b \le n_B} \left| x_a^A - x_b^B \right| \qquad \forall a \in A \tag{3.2.3.1}$$

Em geral, quando dois ou mais doadores estão igualmente distantes do registro receptor, a escolha entre eles ocorre de maneira aleatória.

O processo caracterizado pela equação 3.2.3.1 é usualmente chamada de *unconstrained distance hot deck* ou *hot deck* irrestrito por distância. É denominada irrestrito porque cada registro do arquivo doador B pode ser doado mais de uma vez.

Outro tipo de *distance hot deck* é o *constrained distance hot deck* ou *hot deck* restrito por distância. Essa abordagem restringe que cada registro do arquivo doador B , escolhido para a doação, seja doado apenas uma vez. O *constrained hot deck* necessita que o número de doadores seja maior ou igual ao número de receptores, $n_A \le n_B$. No caso mais simples, o número de unidades em ambos os arquivos é igual, $n_A = n_B$, então a figura de mérito a ser minimizada seria:

$$\sum_{a=1}^{n_A} \sum_{b=1}^{n_B} (d_{ab} \, \mathrm{I}_{ab}) \tag{3.2.3.2}$$

sujeito às restrições:

$$\sum_{b=1}^{n_B} \mathrm{I}_{ab} = 1 \qquad a = 1,.....,n_A \qquad (3.2.3.3)$$

$$\sum_{a=1}^{n_A} \mathrm{I}_{ab} = 1 \qquad b = 1,.....,n_B \qquad (3.2.3.4)$$

$$\mathrm{I}_{ab} \in \{0;1\}$$

onde I_{ab} é uma função indicadora da doação é igual a 1, se o par (a,b) é emparelhado e $\mathrm{I}_{ab} = 0$ quando não emparelha (Kadane, 1978).

O problema de programação linear muda, ligeiramente, quando existem mais doadores que receptores, $n_B > n_A$. Neste caso o conjunto de restrições se transforma para:

$$\sum_{b=1}^{n_B} \mathrm{I}_{ab} = 1 \qquad a = 1,.....,n_A \qquad (3.2.3.5)$$

$$\sum_{a=1}^{n_A} \mathrm{I}_{ab} \leq 1 \qquad b = 1,.....,n_B$$

$$\mathrm{I}_{ab} \in \{0;1\}. \qquad (3.2.3.6)$$

Esse sistema de restrições implica que:

$$\sum_{a=1}^{n_A} \sum_{b=1}^{n_B} \mathrm{I}_{ab} = n_A \ .$$

Esta situação é um caso particular do descrito na seção 2.2.2, onde cada observação representa só a si mesmo $w_i = w_j = 1$ e os pesos w_{ij} ficam também reduzidos a 0 ou 1.

A busca do melhor emparelhamento restrito entre dois arquivos corresponde a minimizar a soma das distâncias entre os seus registros. Do ponto de vista matemático, esse problema de otimização é um clássico problema de programação linear, veja Burkard e Derigs (1980). Esse problema exige um considerável esforço computacional para problemas de maiores dimensões.

Numa abordagem micro não-paramétrica, um estimador de regressão não-paramétrico freqüentemente usado é o *kNN* - *k-Nearest Neighbor*, que é obtido por uma média ponderada dos $\text{k vizinhos mais próximos}$ com pesos definidos pelo posto da distância de cada vizinho ao ponto em questão (vide Chen, J. e Shao J. (2000, 2001)).

Para $k = 1$ este procedimento é idêntico a um *distance hot deck.* O estimador *kNN* corresponde à estimação da média condicional de **Z** dado **X** =x.

A questão sobre a qualidade do arquivo síntese e sua habilidade em preservar a distribuição da variável imputada, isto é, **Z** dado **X** deve ser sempre discutida,sendo aqui apresentada na seção 6.1.1.2.

3.2.4.
Matching noise

O emparelhamento estatístico gera o arquivo síntese. Tal arquivo é apropriado para análises estatísticas ulteriores, quando a distribuição de probabilidade conjunta das variáveis de interesse neste arquivo coincide, ou fica próxima, da distribuição de probabilidade das mesmas variáveis na população. A discrepância entre essas distribuições é o *matching noise*, vide D'Orazio et al. (2006, p. 33 e 83).

Scanu et al., (2006) apresentam um artigo sobre o assunto para as distribuições normal e uniforme, pois quando a variável de emparelhamento X é contínua, pode surgir um problema de distância para as diferenças entre $(x_a^A) = (x_{a1}^A \ldots\ldots\ldots\ldots x_{aP}^A)$ e $(x_b^B) = (x_{b1}^B \ldots\ldots\ldots\ldots x_{bP}^B)$.

Segundo D'Orazio et al., 2006, p. 45, quando **X** é categórica e um método *random hot deck* condicionado a **X** é aplicado tem-se um procedimento livre de ruído de emparelhamento (*free* of m*atching noise),* ou seja, o vetor **X** da distribuição original "coincide" com o vetor **X** da distribuição do arquivo síntese.

No nosso caso, para **X**, **Y** e **Z** univariados, A receptor e B doador, o arquivo resultante é $(x_a, y_a, \tilde{z}_a)$, $a = 1,\ldots\ldots\ldots\ldots,n_A$, onde $\tilde{z}_a = z_b$ para algum b em B, escolhido de acordo com um método da família *hot deck*.

Dado que (x_a, y_a) é uma amostra gerada por $f_{XY}(x;y)$ e supondo a CIA, pode ser provado que $(x_a, \tilde{z}_a)$ é gerado a partir da $f_{XZ}(x;z)$ ou, em outras palavras, que a distribuição $(X, \tilde{Z})$ é igual à distribuição (**X**, **Z**) (vide demonstração em D'Orazio et al., 2006, p. 46).