4
Informação auxiliar

Segundo D'Orazio et al., 2006, p. 11, o conjunto de modelos identificáveis para $A \cup B$ é pequeno e pode ser inapropriado para o fenômeno em estudo. Para resolver esse problema, conforme citado na seção 1.5, usa-se uma informação auxiliar, além de $A \cup B$. A informação auxiliar, que será descrita a seguir, pode ser paramétrica, ou seja, existe o conhecimento dos valores de alguns dos parâmetros do modelo para (**X**, **Y**, **Z**), ou uma amostra adicional, C que informa sobre a relação de (**Y**, **Z**) ou de (**X**, **Y**, **Z**) .

4.1.
Diferentes tipos de informação auxiliar

Conforme antecipado, na seção 1.5, a suposição de independência condicional entre **Y** e **Z**, dado **X** não pode ser testada nos arquivos A e B . A CIA é uma suposição implícita, e freqüentemente não é um pressuposto correto, para evitar tal pressuposto usa-se a informação auxiliar. Muitos autores descrevem os efeitos da suposição da CIA; dentre esses Sims (1972), Kadane (1978), Cassel (1983), Rodgers (1984), Paass(1986), Barry (1988), Cohen (1991) e Singh et al. (1993).

É fácil entender que estimar a distribuição conjunta de (**X**, **Y**, **Z**) é muito diferente de gerar a distribuição real. Ao discutir esse problema, alguns autores falam do "viés" dos procedimentos estatísticos devido a CIA. Na verdade, os procedimentos de emparelhamento estatístico são não viesados, caso a CIA seja válida. O problema é a especificação errada do modelo que sempre induz a estimativas viesadas.

Quando a CIA não é válida, os parâmetros da relação estatística entre **Y** e **Z** não podem ser estimados a partir dos arquivos A e B . No capítulo 3, uma distribuição normal multivariada foi relatada sob o pressuposto da CIA. Em caso de falha do pressuposto da CIA a solução dada para os parâmetros não estimáveis é:

i. um terceiro arquivo C onde (**X**, **Y**, **Z**) ou (**Y**, **Z**) foram observados conjuntamente;

ii. valores plausíveis dos parâmetros não estimáveis de $(Y,Z \mid X)$ ou (**Y**, **Z**). Por exemplo, no caso de dados normais o coeficiente de correlação parcial ou alguma informação sobre uma categorização de (**X**, **Y**, **Z**), no caso de (**X**, **Y**, **Z**) ser contínua.

Alternativamente, Kadane (1978) propõe o uso de informação auxiliar para evitar o pressuposto da CIA. Usualmente, uma informação auxiliar paramétrica pode ser obtida de uma das duas formas:

(1) uma amostra prévia realizada, arquivos ou coleção de dados;

(2) uma *proxy* de variáveis.

Em ambos os casos, apesar de não ser perfeita, supõe-se que essa informação paramétrica é válida para o modelo que gerou as amostras A e B . A informação paramétrica externa tem um papel importante no contexto de emparelhar estatisticamente, sendo usada para restringir o conjunto de parâmetros possíveis, vide D'orazio et al., 2006, p. 71.

No nosso caso uma *proxy* de variáveis será usada.

4.2.
O caso da normal multivariada

No caso da normal multivariada, em uma abordagem paramétrica usando (ii), a distribuição conjunta de (**X**, **Y**, **Z**) é caracterizada pelos parâmetros:

$$\theta = (\mu, \Sigma) = \left[\begin{pmatrix} \mu_X \\ \mu_Y \\ \mu_Z \end{pmatrix}, \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} & \Sigma_{XZ} \\ \Sigma_{YX} & \Sigma_{YY} & \Sigma_{YZ} \\ \Sigma_{ZX} & \Sigma_{ZY} & \Sigma_{ZZ} \end{pmatrix} \right] \tag{4.2.1}$$

Nesse caso, é suficiente obter informação sobre os parâmetros não estimáveis, isto é, a informação que descreva o relacionamento entre, as variáveis nunca observadas conjuntamente **Y** e **Z**. Podem-se considerar duas possíveis restrições:

I. os coeficientes na matriz de correlação parcial $\rho_{YZ|X}$

II. na matriz de covariância marginal entre **Y** e **Z**, Σ_{YZ}

Coeficientes parciais de correlação conhecidos

De forma geral, esse parâmetro é o mais importante do emparelhamento estatístico. De fato, a distribuição de (4.2.1), pode ser fatorada como:

$$f(x, y, z, \theta) = f_X(x; \theta) f_{YZ|X}(y, z \mid x; \theta_{YZ|X}) \tag{4.2.2}$$

No caso da normal multivariada, $\theta_{YZ|X}$ é a correlação parcial entre **Y** e **Z** dado **X**, o único parâmetro de (4.2.1) que não pode ser estimado para $A \cup B$.

D'Orazio et al., 2006, p.72, provam que o EMV $\hat{\theta}$ de θ é aquele e só aquele compatível com os seguintes EMV:

$\hat{\theta}_X$, calculado em $A \cup B$;

$\hat{\theta}_{Y/X}$, calculado em A;

$\hat{\theta}_{Z/X}$, calculado em B.

Este resultado e a restrição (I) podem conduzir a passos para obter estimativas dos parâmetros de regressão. De acordo com Seber (1977) e Cox e Wermuth (1996) é possível obter as estimativas dos parâmetros de regressão de **Z** dado **X** e **Y** usando estimativas EMV. Observe que a informação auxiliar paramétrica dos coeficientes de correlação parcial de **Y** e **Z** dado **X** tem sido usado por Rubin (1986) e Rässler (2002) com pequenas diferenças entre si. Ver detalhes em D'orazio et al., 2006, p.74.

Covariâncias marginais conhecidas

Suponha que Σ_{YZ} é conhecido e fixo em um valor Σ^*_{YZ} . Dado que as propriedades dos EMV irrestritos dos coeficientes de 4.2.1 são coerentes com as estimativas sob CIA, existem duas possibilidades:

(i) As estimativas dos parâmetros estimáveis (4.2.1) obtidas usando os passos MV citados anteriormente são coerentes com as restrições impostas, em outras palavras a matriz de covariância:

$$\begin{pmatrix} \hat{\Sigma}_{XX} & \hat{\Sigma}_{XY} & \hat{\Sigma}_{XZ} \\ \hat{\Sigma}_{YX} & \hat{\Sigma}_{YY} & \hat{\Sigma}_{YZ} \\ \hat{\Sigma}_{ZX} & \hat{\Sigma}_{ZY} & \hat{\Sigma}_{ZZ} \end{pmatrix} \qquad (4.2.3)$$

É semi-definida positiva. Note que essa solução é para as estimativas MV irrestritas.

(ii) A matriz (4.2.3) é definida negativa. Isso significa que o Σ^*_{YZ} imposto não é compatível com a CIA. Nesse caso, as soluções MV restritas não são válidas mas transformações adequadas do algoritmo EM para levar em conta essa restrição pode ser feita para que se obtenha uma solução.

Ressalte-se que essa restrição tem sido aplicada na estrutura geral dos trabalhos de Moriarity e Scheuren (2001, 2003, 2004), seguindo as idéias de Kadane (1978), com objetivo micro.

Note que Kadane (1978) sugere o uso de estimadores consistentes para obter os parâmetros estimáveis. Moriarity e Scheuren (2001) adotam estimadores consistentes, mas diferentes dos estimadores MV. A diferença foi sugerir que ao computar a matriz de covariância residual $\hat{\Sigma}_{ZZ|XY}$ para a regressão de **Z** em **Y** e **X,** essa fosse usada para imputar, somando, o ruído residual nos valores da regressão, antes do emparelhamento, conforme formalizado na seção 2.5.

4.3.Procedimento micro não-paramétrico usando informação auxiliar

A exploração de informação auxiliar de um terceiro arquivo C que possui registros a nível individual em (**X**, **Y**, **Z**) ou (**Y**, **Z**) com os métodos *hot deck* pode ser representada como um procedimento de dois passos[14].

Seja A o arquivo receptor e B o arquivo doador

(i) para cada $a = 1,.....,n_A$, um valor observado Z_c^* em C, é imputado no $a-ésimo$ registro de A usando um dos procedimentos *hot deck* descritos na seção 3.2.

Note que, quando o *distance hot deck* é usado, os seguintes procedimentos são aplicados:

(a) quando C contém informações nas variáveis (**X**, **Y**, **Z**) a distância é computada em relação a (**X**, **Y)**;

(b) quando C contém a informação nas variáveis (**Y**, **Z**), a distancia é computada usando somente **Y**.

(ii) para cada $a = 1,.....,n_A$, é imputado o valor final observado Z_b^*, correspondente ao vizinho mais próximo b^* em B , com respeito à distância que também considera os valores intermediários $d((x_a, z_c^*),(x_b, z_b))$ previamente determinados.

[14] Note que este procedimento difere do *hot deck* sob a hipótese de CIA.

Essa técnica foi proposta por Singh et al. (1993) *apud* D'orazio et al., 2006, p. 84, e uma técnica mais geral foi introduzida por Paass (1986) envolvendo técnicas não-paramétricas baseadas nos métodos *kNN*. Essa forma de usar a informação auxiliar explora o relacionamento entre X, Y e Z observado nesse terceiro arquivo C.

4.4.
Métodos mistos

Quando existe informação auxiliar disponível, a maioria das técnicas é baseada, essencialmente, nos métodos mistos que possuem dois passos principais, conforme citado em 1.3.1.

Passo 1 : estimação de parâmetros de um modelo;

Passo 2 : uso de técnicas *hot deck* condicionadas ao primeiro passo.

As diferenças entre as técnicas de *hot deck* dependem da natureza da informação auxiliar. A informação auxiliar pode ser:

- a nível micro, isto é, um arquivo externo C;
- na forma paramétrica para os parâmetros chave do modelo sob estudo;
- na forma paramétrica mas considerando a informação sobre outros parâmetros que não os chave.

Nesse trabalho, na seção 5.1 introduz-se um procedimento que utiliza informação auxiliar na forma não-paramétrica. A informação é baseada em alguma teoria estabelecida sobre a distribuição de (X,Y,Z) mais precisamente sobre a relação dos percentis de Y e de Z, claramente não é um dos parâmetros chaves da distribuição de (X,Y,Z). No nosso caso esta informação auxiliar é oriunda da teoria econômica.