2
Revisão da literatura sobre emparelhamento estatístico

2.1.
Introdução

Na condução de estudos de emparelhamento estatístico, que lidam com pesquisas amostrais, várias abordagens são justificadas de acordo com diferentes paradigmas.

Como antecipado, além de ser uma suposição freqüentemente inválida, a CIA não pode ser testada a partir das amostras A e B disponíveis. Citou-se na seção 1.3 que a metodologia do emparelhamento estatístico depende da combinação entre o objetivo do processo de integração e da informação disponível, onde o mais importante é a representatividade de seu resultado. O objetivo do emparelhamento pode ser micro ou macro e sua abordagem paramétrica alterna entre paramétrica, não-paramétrica ou uma mistura dessas estruturas – método misto.

Os possíveis objetivos micro ou macro depende de se querer obter um arquivo síntese, ou estimar uma característica importante de uma distribuição conjunta predita. Os procedimentos de emparelhamento estatístico e esses podem ser formalizados como paramétrico, onde a família de distribuição conjunta das variáveis, $\Im$, é um conjunto de distribuições paramétricas, ou como não-paramétrico. Uma alternativa adicional, o método misto, com um passo paramétrico, seguido de outro não-paramétrico, também pode ser escolhido. Os três primeiros procedimentos podem ser baseados em inferência usando a verossimilhança, no paradigma bayesiano, na abordagem de modelos assistidos para populações finitas ou no caso não-paramétrico que é o foco desse trabalho.

O emparelhamento estatístico foi proposto originalmente por Okner (1972) e tem sido desenvolvido desde então (veja referências em Rässler, 2002). A maioria dos artigos tem usado o procedimento micro paramétrico para obter um arquivo ampliado: o arquivo síntese. Em outras palavras, são preditos os valores de **Z** faltantes no arquivo A e os valores de **Y** omissos em B.

Os métodos de emparelhamento estatístico podem também ser divididos, grosso modo, em dois grandes grupos, aqueles que se baseiam no modelo específico onde **Y** e **Z** são, probabilisticamente, condicionalmente independentes dado **X** (CIA), e um segundo grupo de métodos que enfrentam o problema usando informação auxiliar (vide seção 1.3.3). Estas são informações externas aos arquivos A e B. No caso paramétrico, referem-se aos parâmetros nas relações estatísticas entre (**Y***,* **Z**) ou na distribuição de (**Y***,* **Z**) ou (**X***,* **Y***,* **Z)** (vide Singh et al., 1993). No caso não-paramétrico referem-se a relação entre as variáveis.

O uso de informação auxiliar é justificável quando conduz a um melhor resultado do que usar o pressuposto da CIA, ou quando a hipótese da CIA não é plausível ou não é natural.

No capítulo 3, considera-se o primeiro grupo de métodos onde $A \cup B$ é uma amostra não observada simultaneamente para os registros *(i.i.d.)* da distribuição f(***x***, ***y***, ***z***). O arquivo A usa a informação predita sobre a variável **Z** e B a informação predita sobre a variável **Y**, ambas predições usadas na estimação de f(***x***, ***y***, ***z***) podendo ser imputada no arquivo *A*, no arquivo *B* ou ambos, na hipótese de CIA, ou seja:

$$f(\mathbf{x}, \mathbf{y}, \mathbf{z}) = f_{Z|X}(\mathbf{z}|\mathbf{x})\, f_{Y|X}(\mathbf{y}|\mathbf{x}))\, f_{X}(\mathbf{x})$$ (ver equação 1.3.3.1).

A seção 2.2 aborda o emparelhamento irrestrito (*unconstrained*) e o restrito (*constrained*), descritos para emparelhar estatisticamente duas amostras aleatórias independentes *A* e *B*, através de um exemplo clássico de Rodgers (1984).

Na seção 2.3 apresentam-se passos de harmonização entre os arquivos *A* e *B*, que algumas vezes, são usados antes do emparelhamento estatístico. Na seção 2.4 e 2.5 discute-se afiliação e dependência positiva, seguida de transformação percentil monotônica.

Tradicionalmente, vários autores vêem o emparelhamento estatístico como uma forma de imputação, dentre eles pode-se citar Barry (1988), Cassell (1983), Cohen (1991a), Paass (1985), Rodgers (1984), Rubin (1976) e Singh et al. (1993).

A imputação e o emparelhamento estatístico têm muito em comum. O termo imputação, usado aqui, tem uma ligação estreita com a técnica para substituir valores faltantes para uma ou mais categorias de respostas. Outros analistas usam o termo em um sentido mais amplo definindo uma técnica de "gerar" todos os valores para uma ou mais variáveis ausentes, que nunca foram perguntadas em uma pesquisa (ou nunca foram coletadas). A imputação deste último tipo — por exemplo, baseada nas estimativas de equações de regressão de outra fonte de dados — pode exibir alguns dos mesmos problemas dos dados do emparelhamento estatístico. Entretanto, o emparelhamento estatístico, apesar de ser um caso particular de imputação, é mais abrangente no seu escopo do que a imputação. A imputação é usada tipicamente para preencher uma porcentagem relativamente pequena dos dados ausentes e o emparelhamento estatístico é usado tipicamente para preencher 100 por cento dos registros, de um bloco ausente.

No caso do emparelhamento estatístico, primeiro, tipicamente, imputam-se blocos de registros completos para preencher valores faltantes dentro de outros registros. Em segundo lugar, a função da distância (ou a proximidade) sugerida é, às vezes, definida de modo que a distância entre dois pontos observados seja infinita, isto é, um emparelhamento entre os registros que diferem em certas

variáveis não é permitido. Por exemplo, pode-se desejar que nenhum registro de uma pessoa do sexo masculino seja fundido com um registro de uma pessoa do sexo feminino; e o tratamento do arquivo pode ser assim estruturado para com essas restrições lógicas evitar zeros estruturais. No caso de emparelhamento estatístico os dois arquivos de dados a serem fundidos podem representar amostras de universos ligeiramente diferentes. Por exemplo, uma base pode representar os que pagam imposto de renda, enquanto a outra representa todos os domicílios.

Antes que o emparelhamento estatístico possa ser feito, as amostras devem passar por um passo de harmonização (vide seção 2.3) das mesmas, de modo que os universos representados se tornem homogêneos (vide D'Orazio et al., 2006), já que se supõe que os dois arquivos a serem emparelhados estatisticamente são arquivos de microdados de duas amostras extraídas da mesma população. Esta suposição não é trivial, porque antes que duas bases possam ser combinadas estatisticamente, as mesmas podem requerer algum tratamento, como descrito na seção 2.3.

Outro cuidado necessário antes do emparelhamento estatístico é a meta avaliação dos conceitos das variáveis levantadas nas duas pesquisas. Existem três tipos de variáveis:

(i) com o mesmo conceito nas duas pesquisas;

(ii) com conceitos diferentes, mas harmonizáveis ;

(iii) com conceitos diferentes e não harmonizáveis .

Como exemplo do caso (ii) poder-se-ia citar pesquisas nas quais a definição de pessoa de referência do domicilio fossem diferentes: uma considera um registro administrativo que gerou a amostra e a outra considera a pessoa mais velha, ou com mais escolaridade ou com maior renda. Se na primeira pesquisa

existir também a informação de idade ou de escolaridade ou de renda, os dois conceitos podem ser harmonizados procurando-se no domicilio o individuo com a característica desejada.

O Emparelhamento Exato (*exact matching*) é uma metodologia que equivale ao *merging* ou *record linkage* (veja Fellegi e Sunter (1969), e como já mencionado, é estrategicamente diferente do emparelhamento estatístico, porque visa juntar as mesmas unidades.

Note que a análise de estudos de observações onde se seleciona casos "controle" que são "similares" aos casos "tratamento" tem algumas analogias com o emparelhamento estatístico, sendo denominada emparelhamento amostral (*matched sampling*). Algumas referências para esse procedimento podem ser encontradas em Cochran e Rubin (1973), Rosenbaum e Rubin (1983, 1985) e Rosenbaum (1989).

Para uma descrição ampliada do emparelhamento estatístico, veja Radner et al. (1980) que é uma referência que compara e faz contrastes entre o emparelhamento estatístico e o emparelhamento exato, sendo Goel e Ramalingam (1989) outra referência básica. Também veja Draper et al. (1992) onde o assunto combinação da informação é discutido de forma generalizada.

Outras referências em emparelhamento estatístico são os anais organizados por Scheuren a partir do workshop em modelagem de microsimulaçao realizado em maio de 1988 pelo departamento do Tesouro Americano (*Department of the Treasury*).

Na literatura, também existem procedimentos propostos para emparelhamento estatístico de amostras complexas, vide Renssen (1998), Rodgers (1984) e Rubin (1986). Rubin (1986) trata especificamente de objetivos

micro, ao passo que Renssen (1998) trata de objetivos macro, obtendo estimativas coerentes derivadas das duas amostras.

Radner (1980) descreveu um emparelhamento estatístico em vários estágios que o objetivo macro é obter os elementos de uma tabela cruzada, usando uma função de distância.

Outra estratégia anteriormente usada emprega a informação de **X** para fazer uma tabulação cruzada dos registros usando os dois arquivos, e depois liga os registros segundo uma classificação cruzada com alguma forma de procedimento estocástico ou determinístico, veja Budd (1971), Okner (1972), Alter (1974), Ruggles e Ruggles (1974), Ingram et. al. (2000). Esse procedimento é similar aos métodos de imputação hierárquica, vide Kalton e Kasprzyk (1986).

D'Orazio et al. (2006) menciona os textos mais recentes que usam métodos mistos pode-se citar: Moriarity e Scheuren (2001), Kadane (1978), Rubin (1986) e Rässler (2002).

Nas seções seguintes apresentam-se as estratégias do Emparelhamento Estatístico restrito e irrestrito, os cuidados prévios ao emparelhamento, a afiliação e a dependência positiva, a adição de resíduos aleatórios, e por último a transformação percentil monotônica.

2.2.
Emparelhamento estatístico restrito e irrestrito

Várias estratégias gerais foram usadas para conduzir o emparelhamento estatístico, e novas variações dessas estratégias surgiram onde geralmente um arquivo é tomado como “receptor” ou “base”, e as variáveis do outro considerado “doador” ou “suplementar”, são concatenadas ao arquivo base, gerando um arquivo síntese.

Supõe-se que existe um vetor de variáveis **X** comuns a ambos os arquivos. **X** pode conter variáveis contínuas, categóricas ou uma combinação dos dois tipos. Um arquivo contém (**X***,* **Y**) e outro contém (**X***,* **Z**), e existe um interesse em criar um arquivo contendo (**X***,* **Y***,* **Z**).

Pode-se supor ou não que existe informação auxiliar disponível sobre a distribuição conjunta (**Y***,* **Z**).

Duas estratégias relacionadas são denominadas emparelhamento estatístico restrito e irrestrito. Ambas empregam uma função de distância que é definida em **X** por:

$$d\,(\mathbf{x}_A, \mathbf{x}_B) : \mathbf{X}^*\mathbf{X} \rightarrow \Re^+ \qquad (2.2.1)$$

Os dois tipos de emparelhamento estatístico, restrito e irrestrito (*constrained e unconstrained*) têm propriedades diferentes. O emparelhamento irrestrito permite amostragem com reposição do arquivo doador, enquanto o restrito pode ser visto como uma versão modificada de amostragem sem reposição do arquivo doador. Ambas as estratégias tem vantagens e desvantagens que serão discutidas a seguir. Reproduz-se nesta seção o exemplo clássico de Rodgers (1984), que é simples de entender, para ilustrar o emparelhamento irrestrito e restrito.

Nesse exemplo clássico de Rodgers (1984), suponha um arquivo *A*, que possui 8 registros, sendo o arquivo receptor, e um arquivo *B* sendo o arquivo doador com 6 registros, mostrados respectivamente nas tabelas 1 e 3. Suponha-se que existe um interesse em obter algum tipo de análise multivariada envolvendo **Y** do arquivo *A* e **Z** do arquivo *B*. O arquivo *B* é emparelhado

estatisticamente com o arquivo *A*, usando alguma função de distância de **X**, definida como o valor absoluto da diferença entre a variável idade dos dois registros, ou seja:

d ($\mathbf{x}_A$, $\mathbf{x}_B$) = | $\mathbf{x}_A$ - $\mathbf{x}_B$ |, os registros devem ser emparelhados por sexo.

Casos #	Sexo	Idade	Renda (**Y**)	w_j (pesos)
A_1	M	42	9,156	3
A_2	M	35	9,149	3
A_3	F	63	9,287	3
A_4	M	55	9,512	3
A_5	F	28	8,494	3
A_6	F	53	8,891	3
A_7	F	22	8,425	3
A_8	M	25	8,867	3

Tabela 1: Registros do arquivo A

Média da Idade:	40,375
DP (idade):	15,324
Média da Y :	8,973
DP(Y):	0,378

Tabela 2: Estatísticas descritivas de A

Casos #	Sexo	Idade	Despesa (**Z**)	w_i (pesos)
B_1	F	33	**6,932**	4
B_2	M	52	**5,524**	4
B_3	M	28	**4,223**	4
B_4	F	59	**6,147**	4
B_5	M	41	**7,243**	4
B_6	F	25	**3,23**	4

Tabela 3: Registros do arquivo *B*

Média da Idade:	43
DP (idade):	**11,5768**
Média da Z :	5,5496
DP(Z):	1,5669

Tabela 4: Estatísticas descritivas de *B*

Nas Tabelas 2 e 4, $DP(\bullet)$ significa um desvio padrão não ponderado, e dividido por $(n_A - 1)$ e $(n_B - 1)$ respectivamente.

2.2.1. Emparelhamento irrestrito

Um emparelhamento irrestrito (*unconstrained*) permite emparelhamento com o registro "mais próximo", que é definido, por exemplo, por uma função da distância. Por exemplo, como candidatos para o emparelhamento com o registro A_1 (sexo = M e idade = 42) no arquivo A, existem os registros do sexo masculino B_2 (idade = 52), B_3 (idade = 28) e B_5 (idade = 41) . Os valores da função de distância são respectivamente 10, 14 e 1. Neste caso, B_5 é emparelhado com A_1, já que apresenta a menor distância entre as idades. Um dos resultados de emparelhamento irrestrito é:

# Casos Emparelhados	Sexo	Idade	Idade	Renda (**Y**)	Despesa (**Z**)	w_j
A_1 , B_5	M	42	41	9,156	7,243	3
A_2 , B_5	M	35	41	9,149	7,243	3
A_3 , B_4	F	63	59	9,287	6,147	3
A_4 , B_2	M	55	52	9,512	5,524	3
A_5 , B_1	F	28	33	8,484	6,932	3
A_6 , B_4	F	53	59	8,891	6,147	3
A_7 , B_1	F	22	33	8,425	6,932	3
A_8 , B_3	M	25	28	8,867	4,223	3

Tabela 5: Resultado do emparelhamento irrestrito.

Como *A* é o arquivo receptor, sua média e desvio padrão para as variáveis idade e renda (**Y**) permanecem inalterados. Para as variáveis do arquivo *B* (idade e despesa (**Z**)) ocorrem alterações nos valores da média e do desvio padrão:

Média da Idade – arquivo B:	43,25
DP (idade – arquivo B):	12,1037
Média da Z :	6,2989
DP(Z):	1,0379

Tabela 6: Estatísticas descritivas do emparelhamento irrestrito *B* doador

Comparando essas estatísticas com as estatísticas computadas no arquivo *B*, claramente a média e o desvio padrão da variável idade e despesa (**Z**) podem mudar quando um emparelhamento irrestrito for realizado. Note também que o registro B_6 não foi usado no emparelhamento.

2.2.2. Emparelhamento restrito

Um emparelhamento restrito adiciona a restrição de que todos os registros do arquivo doador têm que ser usados. As restrições são:

$$\sum_{j=1}^{n_A} w_{ij} = w_i \quad i = 1,.....,n_B$$

e

$$\sum_{i=1}^{n_B} w_{ij} = w_j \quad j = 1,.....,n_A$$

onde w_i é o peso do $i-$ésimo registro do arquivo B , n_B é o número de registros no arquivo B , w_j é o peso do $j-$ésimo registro do arquivo A , n_A é o número de registros no arquivo A , e w_{ij} representa o peso dado a

combinação do $j-$ésimo registro do arquivo A e o $i-$ésimo registro do arquivo B . Todos os pesos w_{ij} devem ser não negativos.

Paass (1985) ressaltou que existe uma suposição implícita que a soma dos pesos nos dois arquivos são iguais. Goel e Ramalingam (1989) incluíram explicitamente a seguinte restrição:

$$\sum_{i=1}^{n_B} w_i = \sum_{j=1}^{n_A} w_j$$

como uma restrição na sua formulação de emparelhamento restrito. Note que essa suposição é verdade no exemplo que está sendo discutido aqui, tanto no tamanho global como no tamanho por sexo.

Barr e Turner (1978) usaram uma estratégia para lidar com w_{ij}, minimizando a seguinte função objetivo:

$$\sum_{i=1}^{n} \sum_{j=1}^{m} \left(d_{ij} * w_{ij} \right)$$

onde d_{ij} é o valor da função de distância entre do $j-$ésimo registro do arquivo A e o $i-$ésimo registro do arquivo B , e onde os w_{ij} estão sujeitos a restrições estabelecidas previamente.

Os valores de w_{ij} que minimizam a função objetivo de Barr e Turner podem ser encontrados ao resolver-se um problema de programação linear. O tipo de problema de programação linear sendo resolvido é denominado na literatura de “problema de transporte (*transportation problem*)” (consulte Bertsekas 1991 para uma solução do problema); esse termo é um termo histórico sendo devido às restrições dos pesos – a soma das “entradas” devem

ser igual à soma das “saídas”. Rodgers (1984) documentou uma solução, mostrada na tabela 7, que minimiza a função objetivo de Barr e Turner.

# Casos Emparelhados	Sexo	Idade	Idade	Renda (**Y**)	Despesa (**Z**)	w_{ij}
A_1 ,B_2	M	42	52	9,156	5,524	1
A_1 , B_5	M	42	41	9,156	7,243	2
A_2 , B_3	M	35	28	9,149	4,223	1
A_2 , B_5	M	35	41	9,149	7,243	2
A_3 , B_4	F	63	59	9,287	6,147	3
A_4 , B_2	M	55	52	9,512	5,524	3
A_5 , B_1	F	28	33	8,494	6,932	3
A_6 , B_4	F	53	59	8,891	6,147	1
A_6 , B_6	F	53	45	8,891	3,230	2
A_7 , B_1	F	22	33	8,425	6,932	1
A_7 , B_6	F	22	45	8,425	3,230	2
A_8 , B_3	M	25	28	8,867	4,223	3

Tabela 7: Resultado do emparelhamento restrito.

Atenção: nesse exemplo, a solução que minimiza não é única. Outra solução pode ser obtida emparelhando A_5 , B_1 com peso 1, A_5 , B_6 com peso 2, e A_7 , B_1 com peso 3, ao invés de A_5 , B_1 com peso 3, A_7 , B_1 com peso 1 e A_7 , B_6 com peso 2; ambos os emparelhamentos geram o mesmo valor para a

função objetivo, e continua satisfazendo às restrições. A distância entre A_5 e B_1 é 5, entre A_5 e B_6 é 17, entre A_7 e B_1 é 11, e entre A_7 e B_6 é 23.

Para o primeiro emparelhamento, a soma dos três termos $d_{ij} * w_{ij}$ é: (5*1)+ (17*2)+ (11*3)= 72.

Para o segundo emparelhamento, se tem: (5*3)+ (11*1)+ (23*2)= 72.

Em ambos os emparelhamentos, a soma dos pesos para A_5 e A_1 é 3, a soma dos pesos para B_1 é 4, e a soma dos pesos para B_6 é 2.

Tal qual o emparelhamento irrestrito, as médias e os desvios padrões, no arquivo-sintético, para a variável idade e renda (**Y**) permaneceram inalteradas, ou seja, iguais ao do arquivo A . Ao contrário do emparelhamento irrestrito, as médias no arquivo-sintético para a variável idade e despesa (**Z**) são as mesmas do arquivo B . Os desvios padrões para as variáveis no arquivo A e no arquivo B podem ser obtidos a partir do arquivo emparelhado, se os registros emparelhados são reestruturados para refletir o arquivo A ou o arquivo B e uma análise sem ponderação é efetuada, ou se uma análise ponderada é efetuada, seguida de um ajuste para os graus de liberdade.

2.2.3.
Comparação do emparelhamento irrestrito e restrito

O emparelhamento irrestrito trabalha com a associação do registro “mais próximo”, que é medido pela função de distância métrica. O emparelhamento irrestrito pode não “utilizar” todos os registros do arquivo doador. Então não se pode garantir a preservação da distribuição apresentada pelo arquivo doador. Por isso, o emparelhamento irrestrito não é um processo simétrico entre o arquivo doador e receptor; os resultados podem diferir de acordo com quem tenha sido designado para ser o arquivo receptor.

O emparelhamento restrito utiliza todos os registros do arquivo doador e preserva as distribuições marginais apresentas pelo arquivo doador. Então, o emparelhamento restrito é um processo simétrico, o mesmo resultado é obtido, independente de quem tenha sido designado para ser o arquivo receptor. Entretanto, o emparelhamento restrito não permite a associação do registro “mais próximo”.

Rubin (1986) não considerou importante a preservação das distribuições marginais apresentas pelo arquivo doador no procedimento de emparelhamento estatístico, o seu procedimento usado foi o emparelhamento irrestrito.

Ambos os emparelhamentos, irrestrito e restrito, podem ser vistos como uma atribuição de pesos que aperfeiçoa uma função objetivo. O emparelhamento restrito, como o próprio nome indica, introduz restrições que aumentam a carga computacional.

Ambas as estratégias empregam algum tipo de função de distância para definir a similaridade entre os registros. Isso geralmente tem sido feito sem o conhecimento de que o arquivo-sintético resultante da combinação dos registros “similares” seria uma boa estimativa para a distribuição conjunta (**X**, **Y**, **Z**) , vide seção 1. 3.

2.3.
Harmonização das pesquisas antes do emparelhamento

Atualmente, grandes arquivos, gerados por diferentes metodologias, têm sido disponibilizados. Antes do emparelhamento de algumas dessas pesquisas A e B, é fundamental que se verifique a homogeneidade dessas, em relação aos seus conceitos, definições e universo. Emparelhar A e B pode exigir um grande esforço preliminar, em termos de tempo e recursos, para realizar a harmonização de pesquisas de diferentes fontes. Em relação aos arquivos A e B, a escolha das variáveis a serem usadas na concatenação e harmonização dessas bases deve ser executada, sempre que necessário.

Mesmo quando duas pesquisas são conduzidas pela mesma organização, elas podem apresentar incompatibilidades. O Brasil possui uma produção considerável de pesquisas domiciliares e cadastrais. Entretanto cada pesquisa tem seus objetivos bem definidos. Quando é necessário combinar ou comparar duas ou mais fontes, deve-se lidar com as diferentes definições de variáveis. Por exemplo, é fundamental indicar as diferenças nas metodologias da Pesquisa de Orçamentos Familiares (POF) e da Pesquisa Nacional de Amostra de Domicílios (PNAD), que começam nas unidades básicas de informação, que são, respectivamente, a unidade de consumo (UC) e a família (ambos os conceitos aninhados dentro de um mais amplo, o domicilio).

As possibilidades de compatibilizar as bases de dados serão discutidas, após o exemplo a seguir, retirado de D'Orazio et al., 2006.

Renssen (1998) cita a POLS – *Dutch HouseHold Survey on Living Condition* – como uma situação ideal de emparelhamento estatístico, (veja Bakker e Winkels (1998) e Winkels e Everaers (1998) para descrição da

pesquisa). Na verdade, esse é um exemplo de desenho amostral de pesquisas integradas, ou seja, de amostragem matricial (veja seção 1.4). Em outras palavras, essa pesquisa é composta de vários sub-pesquisas ou módulos diferentes, onde cada módulo concentra um aspecto particular das condições de vida do domicílio. Esses módulos têm a importante característica de terem sido integrados, com definições e métodos harmonizados. Eles foram definidos por Winkels e Everaer (1998) e consistem de:

- Um questionário com quesitos demográficos (idade, sexo, lugar de nascimento e etc.) e socioeconômicos (educação, renda do domicílio e etc.)
- Um questionário com poucos quesitos sobre aspectos relevantes de condições de vida.
- Um questionário com muitos quesitos sobre condições de vida.

Os dois primeiros questionários são respondidos por todos os entrevistados. O terceiro é dividido em sub-questionários, tal que cada entrevistado recebe um único desses sub-questionários. Esse último tipo de questionário reduz a carga de resposta, com a finalidade de se obter um painel completo das condições de vida domiciliares; a amostra total é dividida em tantas sub-amostras quanto forem o número de sub-questionários existentes. Cada sub-amostra é associada com um sub-questionário. Então, as primeiras duas partes, dos dois primeiros questionários, representam a variável comum **X** enquanto as variáveis dos sub-questionários da terceira parte, tem o papel das variáveis **Y** e **Z** no emparelhamento estatístico. Esse exemplo é uma aplicação da variante do emparelhamento estatístico descrito em (1.4).

Quando as duas fontes A e B não são planejadas de forma harmônica, diferentes ações devem ser realizadas para a harmonização dessas, tais como as descritas por Van der Laan (2000):

(a) Harmonização das definições das unidades

(b) Harmonização dos períodos de referência

(c) Harmonização das populações de referência e desenho das amostras

(d) Harmonização das variáveis

(e) Harmonização das classificações

(f) Ajustamento para as medidas de erro

(g) Ajustamento para dados faltantes

(h) Derivação das variáveis

Típicas da fase de harmonização, as ações (a)-(e) são do tipo *ad hoc*. As ações (f)-(g) são necessárias quando A e B são afetados por erros não amostrais. O passo (h) é executado para providenciar uma nova variável a partir dos itens dos arquivos A e B.

Na prática, comparam-se alguns conceitos bastante similares entre as duas pesquisas e empiricamente as suas distribuições. A checagem dos questionários não é suficiente. Por exemplo, a fundação IBGE conduz as pesquisas amostrais PME, POF e o Censo Demográfico que podem apresentar caso de inconsistência populacional, quando as amostras são oriundas de diferentes populações de referência ou realizadas em períodos distintos de tempo. A harmonização das variáveis pode ser conseguida através da recodificação das mesmas; por exemplo, quando uma das pesquisas utiliza variáveis contínuas e a outra variáveis categóricas. No caso de diferentes desenhos amostrais para os arquivos A e B , a ação (c) é necessária.

Outras aplicações, além do POLS, citam o mesmo desenho de pesquisa proposto na seção 1.4, visto como uma extensão do tipo Amostragem Matricial Múltipla (*multiple matrix sampling design* - MMS), descrito não só por Shoemaker (1973), mas também por Munger e Lloyd (1988) e Graham et al. 2009, entre outros.

O emparelhamento estatístico quando a pesquisa é planejada usando amostragem matricial é praticamente automático.

2.4.
Afiliação e dependência positiva

Na Estatística vários conceitos têm sido propostos para a noção de dependência positiva.

Para mencionar uma pequena parcela da literatura existente sobre o assunto, De Castro (2006), usa vários conceitos correspondentes à dependência positiva e afiliação.

Afiliação é amplamente usada em Estatística, teoria da confiabilidade (*reliability theory*) e muitas outras áreas de Ciências Sociais e Economia; possivelmente sob outros nomes. Quando existe uma função de densidade, em Estatística, a propriedade de afiliação é conhecida como razão de verossimilhança de dependência positiva (*positive likehood ratio dependence* – PLRD) esse nome foi dado por Lehmann (1966), quando introduziu o conceito de PLRD.

PLRD é largamente conhecido pelos estatísticos como uma propriedade forte e muitos artigos usam algumas de suas condições mais fracas.

Também usada em Estatística, a Afiliação é uma generalização de dependência positiva, introduzida por Milgrom e Weber (1982a, p. 1096), sendo mais abrangente que a correlação positiva.

Muitos processos em teoria Econômica são explicados por funções não-lineares, devido à natureza das variáveis usadas como rendimentos, despesas, aluguéis, salários, juros, dividendos, transferências para a seguridade social, rendimentos de capital, rendimentos de mudanças patrimoniais e outras. A relação entre duas variáveis ainda que possa ser aproximada por uma equação linear, freqüentemente, requer transformações nas variáveis em passos prévios,

para garantir esta aproximação linear. Em geral, quando o pressuposto de linearidade é inválido, a CIA também não é válida. Um processo de emparelhamento estatístico onde a CIA é válida apresenta variáveis linearmente associadas. O nosso caso é uma aplicação ao nível de domicilio, usando as variáveis renda e aluguel, onde não se pode pressupor a independência condicional - CIA, por serem essas variáveis relacionadas de forma não-linear. Assumir a CIA conduziria a resultados viesados no relacionamento conjunto dessas variáveis, no arquivo sintético.

Uma das mais importantes considerações quando se estuda as variáveis da renda e do aluguel é que, de uma forma geral, essas variáveis apresentam uma dependência positiva. Mas isso não significa que sempre um maior valor da renda implique em um maior gasto com aluguel, mas intuitivamente esperam-se valores maiores do aluguel, quando os valores da renda aumentarem. Então, ao invés da dependência positiva, prefere-se usar o conceito de afiliação para as variáveis observadas Y e Z, caso exista uma função monotônica crescente, não-linear, entre essas variáveis. A suposição de ordenamento das mesmas, ao nível de domicílio, por exemplo, é uma indicação confiável com que se pode contar. Uma vantagem do uso do conceito de afiliação que se opõe à covariância é que a relação é invariante a transformações monotônicas. Por exemplo, a transformação via o logaritmo das variáveis renda e aluguel não altera esse comportamento.

Formalize-se a definição 2.4.1:

> Se Y e Z são afiliadas estocasticamente e se $g(.)$ e $f(.)$ são funções monotônicas então $g(Y)$ e $f(Z)$ são também afiliadas estocasticamente.

As variáveis renda e aluguel são variáveis aleatórias com uma dada distribuição conjunta. Essas variáveis apresentam um modelo comportamental, que no processo de decisão sobre a escolha de uma alternativa de aluguel, é influenciado por fatores racionais e subjetivos. Os fatores racionais são aqueles explicados a partir de características sócio-econômicas dos indivíduos residentes nos domicílios. Os fatores subjetivos são aqueles que não são expressos diretamente a partir de conceitos econômicos, advindos de fatores aleatórios, da decisão subjetiva associada a cada domicílio i ou j.

$$Y_j \qquad j = 1,.....,n_A$$

$$Z_i \qquad i = 1,.....,n_B$$

Formalizando o conceito de afiliação estocástica, para dois domicílios quaisquer i , j, formalize-se a definição 2.4.2::

$$\forall \varepsilon > 0 \qquad \exists \quad \delta > 0 \quad \ni$$

$$\text{Se} \quad Y_i > Y_j + \delta \Rightarrow P(Z_i > Z_j) < \varepsilon$$

Diz-se então que Y e Z são estocasticamente afiliados.

Uma definição formal de afiliação é dada a seguir onde comparam-se o conceito de afiliação e outras definições de dependência:

Suponha-se um caso bivariado e associe-se que as variáveis aleatórias X e Y tem uma distribuição conjunta F e uma função de densidade f estritamente positiva[12]. Os seguintes conceitos são formalizações da noção de dependência positiva:

[12] A hipótese de densidade estritamente positiva é feita somente para simplificar.

Propriedade **I** : X e Y são positivamente correlatadas (PC) se $\text{cov}(X, Y) \geq 0$.

Propriedade **II** : X e Y são ditas positivamente dependentes no quadrante (PQD) se $\text{cov}(g(X), h(Y)) \geq 0$ para toda g e h não-decrescente.

Propriedade **III** : os valores reais variáveis aleatórias X e Y são ditas associadas (As) se $\text{cov}(g(X,Y), h(X,Y)) \geq 0$ para toda g e h não-decrescente.

Propriedade **IV** : Y é dito decrescente a esquerda em X (denotado por LTD(Y|X)) se $\Pr[Y \leq y \mid X \leq x] \geq 0$ é não-decrescente em x para todo y . X e Y satisfazem a propriedade IV se LTD(Y|X) e de LTD(X|Y) foram válidos.

Propriedade **V** : Y é dito positivamente dependente na regressão em X (denotado por PRD(Y|X)) se $\Pr[Y \leq y \mid X \leq x] = F_{Y|X}(y \mid x)$ é não-decrescente em x para todo y . X e Y satisfazem a propriedade V se PRD(Y|X) e de PRD(X|Y) foram válidos.

Propriedade **VI** : Y é dito ter função risco inversamente dependente decrescente em X (denotado por IHRD(Y|X)) se $\dfrac{F_{Y|X}(y \mid x)}{f_{Y|X}(y \mid x)}$ é não-decrescente em x para todo y , onde $f_{Y|X}(y \mid x)$ é a função de densidade de probabilidade de Y condicionado a X . X e Y satisfazem a propriedade VI se IHRD (Y|X) e de IHRD (X|Y) foram válidos.

Teorema 1: Seja Afiliação a Propriedade **VII.** Então as propriedades acima são sucessivamente mais fortes e ilustra-se como afiliação é uma propriedade mais abrangente:

$$(VII) \Rightarrow (VI) \Rightarrow (V) \Rightarrow (IV) \Rightarrow (III) \Rightarrow (II) \Rightarrow (I)$$

2.5.
Adição de resíduos aleatórios

O procedimento de soma de resíduos usado nessa tese é uma inovação de Moriarity e Scheuren, 2001, que revisaram e aprimoraram as metodologias desenvolvidas por Kadane (1978) e Rubin (1986). Foram formalizados detalhes importantes e indicados acertos para algumas falhas encontradas nessas metodologias. As fórmulas usadas pelos autores foram simplificadas por Moriarity e Scheuren.

O mais importante nesse trabalho de Moriarity e Scheuren mostra que os processos descritos por Kadane e Rubin não são confiáveis para preservar a matriz de covariância (correlação) entre Y e Z, conforme originalmente dito. A inovação essencial foi somar os resíduos às estimativas da regressão, antes de realizar o emparelhamento estatístico para tornar possível a preservação da matriz de covariância (correlação), especificada no estudo de simulação.

Os três métodos usam uma abordagem mista (seção 1.3) e permitem que várias suposições sejam feitas sobre a distribuição de (Y, Z). Executam um emparelhamento estatístico correspondente a cada uma das suposições, para então avaliar a variação das estimativas realizadas pelos grupos de arquivos criados por esse procedimento. Esse processo exibe a quantidade de incerteza das estimativas associada ao emparelhamento estatístico realizado.

2.5.1.
Método de Kadane

Kadane (1978) apresenta uma metodologia de emparelhamento estatístico onde o vetor (X, Y, Z) é suposto ter uma distribuição normal trivariada com matriz de covariância (correlação):

$$\Sigma = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} & \Sigma_{XZ} \\ \Sigma_{YX} & \Sigma_{YY} & \Sigma_{YZ} \\ \Sigma_{ZX} & \Sigma_{ZY} & \Sigma_{ZZ} \end{pmatrix}. \qquad (1.3.3.2)$$

Note que todos os elementos de Σ podem ser estimados a partir dos arquivos A (Σ_{XY}) ou B (Σ_{XZ}) exceto Σ_{YZ} e a sua transposta Σ_{ZY}. Conforme a seção (1.5), em geral, não é possível construir apuradamente a distribuição original de (X, Y, Z) usando a distribuição de (X, Y) do arquivo A e a distribuição (X, Z) do arquivo B, sendo ausente a informação sobre a distribuição de (Y, Z). A priori, pouca ou nenhuma informação sobre a distribuição de (Y, Z) está disponível.

No método de Kadane, no caso univariado, um valor admissível de Σ_{YZ} é escolhido. Valor admissível é um valor que faça Σ ser definida positiva. Σ_{YZ} pode ser generalizado para o caso multivariado.

Esse dado valor de Σ_{YZ} é usado nas regressões dos arquivos A e B produzindo arquivos aumentados $(X, Y, \hat{Z})$ (arquivo A) e $(X, \hat{Y}, Z)$ (arquivo B). Os arquivos aumentados são emparelhados usando a distância de Mahalanobis e os valores de Y e Z são alterados nos registros emparelhados para obter-se os registros aumentados $(X_j, Y_j, \hat{Z}_i)$ (arquivo A) e $(X_i, \hat{Y}_j, Z_i)$ (arquivo B), onde o *j-ésimo* registro do arquivo A foi emparelhado com o *i-ésimo* registro do arquivo B. O emparelhamento descrito por Kadane é restrito, ou seja, todos os registros dos dois arquivos têm que ser usados no emparelhamento. O resultado final é um arquivo síntese formado pelos registros $(X_j, Y_j, \hat{Z}_i)$ (arquivo A) e $(X_i, \hat{Y}_j, Z_i)$ (arquivo B).

Kadane recomenda que esse procedimento seja repetido para vários valores de Σ_{YZ} que gera os respectivos arquivos síntese para cada valor admissível de Σ_{YZ}.

A especificação de Σ_{YZ}, no caso de uma distribuição normal trivariada (X,Y,Z) não singular, para que a incerteza possa ser medida é dada pelo intervalo de (1.6.1); com a exigência de que a matriz de covariância de (X,Y,Z), Σ, da equação (1.3.3.2) deve ser positiva definida, e para isso a $Cor(Y,Z)$ deve estar contida no intervalo:

$$(Cor(X,Y)*Cor(X,Z)) \pm \sqrt{(1-(Cor(X,Y))^2)*(1-(Cor(X,Z))^2)} \qquad (1.6.1)$$

Se $Cor(Y,Z)$ é igual a $(Cor(X,Y)*Cor(X,Z))$ temos a independência condicional de (Y,Z) dado X.

Nos dois passos do método misto, de Kadane, primeiro o passo de regressão e depois de emparelhamento serão especificados.

2.5.1.1.Passo de regressão

No passo de regressão, para um valor admissível de Σ_{YZ} especificado, o procedimento inicia a estimação dos valores faltantes nos dois arquivos usando a expectância condicional, isto é a regressão. Por exemplo, se Z for a variável ausente e todas as quantidades necessárias forem conhecidas, então segundo Anderson 1984, página 36, podemos formalizar:

$$\hat{Z}_j = \mu_Z + \begin{pmatrix}\Sigma_{ZX} & \Sigma_{ZY}\end{pmatrix}\begin{pmatrix}\Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY}\end{pmatrix}^{-1}\begin{pmatrix}X_j - \mu_X \\ Y_j - \mu_Y\end{pmatrix} \qquad (2.5.1.1)$$

Nesta aplicação, todas as quantidades que não sejam referentes a Σ_{ZY} podem ser estimadas usando um ou ambos arquivos. Para um dado valor de Σ_{ZY}, esse procedimento é usado para o arquivo A, e uma rotina similar é realizada em caso de Y faltante no arquivo B.

Como estabelecido por Kadane, pode ser provado que a distribuição conjunta de $(X_j, Y_j, \hat{Z}_j)$ é normal com média (μ_X, μ_Y, μ_Z) e matriz de covariância singular:

$$S_1 = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} & \phi_1' \\ \Sigma_{YX} & \Sigma_{YY} & \phi_2' \\ \phi_1 & \phi_2 & \phi_3 \end{pmatrix} \qquad (2.5.1.2)$$

De forma análoga Kadane prova que a distribuição conjunta de $(X_i, \hat{Y}_i, Z_i)$ tem matriz de covariância singular:

$$S_2 = \begin{pmatrix} \Sigma_{XX} & \phi_4' & \Sigma_{XZ} \\ \phi_4 & \phi_6 & \phi_5' \\ \Sigma_{ZX} & \phi_5 & \Sigma_{ZZ} \end{pmatrix} \qquad (2.5.1.3)$$

Para simplificar as fórmulas de Kadane, Moriarity 2001 usa as fórmulas:

$$(\phi_1 \ \phi_2) = (\Sigma_{ZX} \ \Sigma_{ZY}) \qquad (2.5.1.4)$$

$$\phi_3 = (\Sigma_{ZX}\Sigma_{ZY}) \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_{XZ} \\ \Sigma_{YZ} \end{pmatrix}$$

$$= (\Sigma_{ZX.Y}(\Sigma_{XX.Y})^{-1}\Sigma_{XZ} + \Sigma_{ZY.X}(\Sigma_{YY.X})^{-1}\Sigma_{YZ} \qquad (2.5.1.5)$$

ode as fórmulas de Kadane foram simplificadas para $\phi_i, i = 1,2...,6$, veja demonstração em Moriarity 2001, sendo úteis na metodologia de Moriarity.

2.5.1.2. Emparelhamento

Suponha $W_j = (X_j, Y_j, \hat{Z}_j)$ do arquivo A e $V_i = (X_i, \hat{Y}_j, Z_i)$ do arquivo B, sendo $W_j - V_i$ um vetor de zeros. Kadane demonstra que a matriz de covariância de $W_j - V_i$ é a soma da duas matrizes singulares de covariância S_1 e S_2, e é não singular, vide Moriarity 2001 para esse resultado.

Dada a não singularidade da matriz $(S_1 + S_2)$ e usando a distância de Mahalanobis em (X, Y, Z), sugerida por Kadane, para realizar um emparelhamento restrito do arquivo A com o arquivo B. A distância de Mahalanobis a ser minimizada é notada matricialmente como:

$$(W_j - V_i)' (S_1 + S_2)^{-1} (W_j - V_i) \qquad (2.5.1.6)$$

e nesse caso, o uso de um emparelhamento restrito equivale a um "problema de transporte" que é um tipo de problema de programação linear, vide Barr e Tunner (1978) e Bertesekas (1991).

Moriarity, 2001 apresenta uma simulação para verificar se um dado valor especificado de Σ_{ZY} é preservado após a aplicação desses dois passos do método de Kadane. As alternativas adicionais a esse método é a adição de resíduos às estimativas de regressão no primeiro passo, antes de executar o segundo passo que passa a emparelhar em (Y, Z), que é a segunda alternativa.

Formalizações com as referidas demonstrações devidas a essas alternativas encontram-se em Moriarity e Scheuren, 2001.

2.6.
Transformação Percentil Monotônica

Para uma variável aleatória X com função de densidade de probabilidade F_X, a inversa da função de densidade de probabilidade ou transformação percentil é definida por:

$$F_X^{-1}(p) = \inf_x \{x : F(x) > p\} \qquad \forall \ p \in [0,1].$$

Se F for monotônica não crescente com densidade f, para

$\forall\ p$, $0 < p < 1$, o $p-ésimo$ percentil é definido como $F^{-1}(p)$.

A estimativa de transformação[13] da função *F* é $\hat{F}_n^{-1}(p)$.

Em muito experimentos ou situações de pesquisa, se requer um estimador para a função $\hat{F}_n^{-1}(p)$.

Algum cuidado deve ser tomado, pois $\hat{F}_n$ não é obrigatoriamente invertível. Para evitar ambigüidades defini-se:

$$\hat{F}_n^{-1}(p) = \inf_x \{x : \hat{F}_n(x) > p\} \qquad \forall \ p \in [0,1].$$

o $100 * p - ésimo$ percentil amostral é definido por $\hat{F}_n^{-1}(p)$.

[13] Vide Wasserman, 2003, p. 102. Casella, 2002, p. 54, o teorema 2.1.10 sobre transformações de variáveis, que também pode ser usado para gerar números aleatórios para uma certa distribuição.