BANCO DE PORTUGAL
EUROSISTEMA

Aspetos práticos associados à utilização da base de dados do ISFF

1. Ficheiros da base de dados

A base de dados do Inquérito à Situação Financeira das Famílias (ISFF) disponibilizada pelo INE está organizada por tipos de variáveis e inclui nove ficheiros com o seguinte conteúdo:

1) **Variáveis S**: variáveis de identificação da amostra inicial (país, edição, identificador do alojamento, etc.), do registo dos contactos efetuados pelos entrevistadores e das caraterísticas do alojamento; estes dados referem-se a todos os alojamentos da amostra inicial (amostra bruta).
2) **Variáveis H:** variáveis da entrevista ao nível da família e ponderador final da amostra; estes dados referem-se às famílias respondentes (amostra líquida).
3) **Variáveis R:** variáveis da entrevista ao nível do indivíduo, relativas a questões demográficas; estes dados referem-se a todos os membros das famílias respondentes.
4) **Variáveis P:** variáveis da entrevista ao nível do indivíduo, excluindo questões demográficas; estes dados referem-se a todos os membros das famílias respondentes com idade igual ou superior a 16 anos.

5, 6, 7 e 8) **Variáveis FS, FH, FR e FP:** variáveis sombra (*flag*s) das variáveis S, H, R e P, respetivamente.

9) **Variáveis W:** 1000 réplicas do ponderador final; estes dados referem-se às famílias respondentes.

A listagem de todas as variáveis que fazem parte dos ficheiros S, H, P e R, assim como os códigos das variáveis sombra estão disponíveis na página do ISFF no *site* do Banco de Portugal, na entrada "Base de dados".

Na base de dados do ISFF não existem valores em falta na maior parte das variáveis, uma vez que, em caso de não resposta, as variáveis foram imputadas. Foi utilizado um método estocástico de imputação múltipla que origina cinco estimativas alternativas (cinco *implicates*) para cada valor em falta. Os ficheiros relativos aos dados da entrevista às famílias (H, R, P, FH, FR e FR) têm assim, para cada variável, um número de observações que corresponde a cinco vezes o número de respondentes (famílias ou indivíduos). Para uma mesma família, as cinco observações das variáveis H, R e P são iguais entre si, no caso de a resposta ter sido recolhida durante a entrevista, e diferentes, no caso de a mesma ter sido imputada.

2. Variáveis de identificação e ligação dos ficheiros

A base de dados do ISFF inclui as seguintes variáveis de identificação:

- SA0010: identificador da família
- RA0010: identificador do indivíduo dentro da família
- IM0100: identificador do *implicate*

A variável SA0010 consta dos nove ficheiros. A variável RA0010 consta dos ficheiros com dados ao nível do indivíduo (R, FR, P e FP). A variável IM0100 consta dos ficheiros que se referem a variáveis recolhidas junto das famílias (H, FH, R, FR, P e FP). Assim, por exemplo, a chave de ligação dos ficheiros H e P é constituída pelas variáveis SA0010 e IM0100 e a dos ficheiros R e P é constituída pelas variáveis SA0010, RA0010 e IM0100.

3. Convenções relativas aos nomes das variáveis H, P e R

O **primeiro caractere** do nome das variáveis da entrevista às famílias identifica a unidade de referência dessa variável: H-Família; R-Todos os indivíduos; P-Indivíduos com idade igual ou superior a 16 anos.

No ISFF, as variáveis H, P e R podem ser nucleares (*core*, na designação em inglês), não nucleares (*non-core*, na designação em inglês) ou nacionais. As variáveis nucleares e não nucleares têm definições harmonizadas entre os variáveis países que fazem parte do HFCS. As nucleares estão disponíveis para todos os países e as não nucleares apenas para os países que as decidiram adotar. As variáveis nacionais são específicas do ISFF e não fazem parte da base de dados do HFCS. No caso das variáveis não nucleares e nacionais, o **segundo caractere** da sua designação é N e O, respetivamente.

O **segundo (terceiro) caractere** do nome das variáveis nucleares (variáveis não nucleares e nacionais) identifica a secção do questionário à qual a variável pertence, de acordo com o seguinte lista:

- A - Aspetos sociodemográficos
- B - Ativos reais e empréstimos garantidos por imóveis
- C - Outras dívidas e acesso ao crédito
- D - Negócios e ativos financeiros
- E - Situação no mercado de trabalho
- F - Direitos sobre pensões
- G - Rendimentos
- H - Heranças e doações
- I - Consumo e poupança

As **perguntas de escolha múltipla** originam diferentes variáveis cujos nomes apenas diferem no **último caractere**, o qual é uma letra.

No ISFF existem, para alguns tópicos, perguntas que se repetem num **ciclo com três iterações** (por exemplo, no caso de empréstimos com garantia da residência principal, são efetuadas perguntas sobre as caraterísticas de cada um dos três empréstimos de maior valor). As respostas à mesma pergunta em cada uma das três iterações são registadas em variáveis cujo nome apenas difere no **último caractere** (1, 2 ou 3, identificando a iteração). Por exemplo, HB1701, HB1702 e HB1703 correspondem aos valores em dívida de cada um dos três maiores empréstimos com garantia da residência principal. Nas perguntas de escolha múltipla que fazem parte de ciclos, o número da iteração corresponde ao **penúltimo caractere** do nome da variável. Por exemplo, as variáveis HB1201A,…, HB1201I correspondem aos motivos pelos quais o empréstimo de maior valor foi contratado.

4. Utilização da imputação múltipla, dos ponderadores e das réplicas

Na utilização dos dados do ISFF devem-se levar em consideração as seguintes caraterísticas técnicas do inquérito:

i) A amostra do ISFF é uma amostra probabilística e, portanto, para se obterem resultados extrapolados para o total das famílias residentes em Portugal, cada observação tem que ser ponderada pelo ponderador da família em causa;

ii) Os cinco *implicates* devem ser utilizados em conjunto por forma a se levar em conta a incerteza associada ao processo de imputação, tanto na obtenção de estimativas pontuais como no cálculo de variâncias.

iii) As réplicas dos ponderadores devem ser utilizadas no cálculo de variâncias das estimativas pontuais, por forma a se levar em conta que essa estimativa depende da amostra que está a ser utilizada.

Para o cálculo de uma estimativa de um determinado parâmetro θ (por exemplo: a média, a mediana, ou o parâmetro de uma regressão) obtêm-se, numa primeira etapa,

os valores das estimativas desse parâmetro para cada um dos cinco *implicates* ($\hat{\theta}_m$) e, numa segunda etapa, calcula-se a média aritmética das cinco estimativas, ou seja:

$$\bar{\theta} = \frac{1}{5}\sum_{m=1}^{5} \hat{\theta}_m \qquad (1)$$

Por sua vez, de modo a levar em conta a incerteza associada ao processo de imputação, a variância deste estimador (T) corresponde a uma combinação da variância de cada *implicate* (variância *within* - *W*) e da variância entre *implicates* (variância *between* - *B*).

$$T = W + \left(1 + \frac{1}{5}\right) B \qquad (2)$$

$$W = \frac{1}{5}\sum_{m=1}^{5} \hat{V}_m(\hat{\theta}_m) \qquad (3)$$

$$B = \frac{1}{4}\sum_{m=1}^{5} (\hat{\theta}_m - \bar{\theta})^2 \qquad (4)$$

Quando se pretende estimar parâmetros para a população, no cálculo das estimativas pontuais deverá ser utilizado o ponderador final da amostra e, no cálculo da variância dessas estimativas, deverão ser utilizadas as 1000 réplicas desse ponderador. Assim, por exemplo, a estimativa do rendimento médio das famílias obtida com o *implicate* m é dada por:

$$\hat{\theta}_m = \frac{1}{N}\sum_{i=1}^{n} w_i y_{im} \qquad (5)$$

em que w_i é o ponderador final da amostra para a família *i*, y_{im} é o rendimento da família *i* de acordo com o implicate *m*, *n* é o número de famílias da amostra líquida, *N* o número de famílias residentes em Portugal.

A variância da estimativa para o rendimento médio é dada por:

$$V_m(\hat{\theta}_m) = \frac{1}{1000-1}\sum_{b=1}^{1000} (\theta^*_{mb} - \bar{\theta}^*_m)^2 \qquad (6)$$

θ^*_{mb} corresponde ao rendimento médio estimado com o *implicate m* e utilizando como ponderador a réplica *b:*

$$\theta^*_{mb} = \frac{1}{N}\sum_{i=1}^{n} w_{ib} y_{im} \qquad (7)$$

$\bar{\theta}^*_m$ é a média dos rendimentos estimados com cada uma das 1000 réplicas dos ponderadores e usando os dados relativos ao rendimento do *implicate m:*

$$\bar{\theta}^*_{mb} = \frac{1}{1000}\sum_{b=1}^{1000} \theta^*_{mb} \qquad (8)$$

Com o objetivo de facilitar a utilização deste tipo de dados, alguns programas estatísticos incluem comandos específicos para dados de inquérito e para dados com imputação múltipla. Estes comandos levam em conta o facto dos cinco *implicates* não poderem ser tratados como independentes entre si, uma vez que tal origina uma sobrestimação da significância dos coeficientes estimados[1].

[1] Alguns exemplos úteis de códigos de Stata podem ser consultados em Albacete, N., S. Thandi Dippenaar, P. Lindner, K. Wagner, "*Methodological notes for Austria: Eurosystem Household Finance and Consumption Survey 2017*", January 2019.