Ensaios sobre Estimação em Pequenos Domínios no INE

Aplicação do Estimador EBLUP e do Estimador sintético da regressão no Inquérito ao Emprego

Pedro Campos[1,2], Luís Correia[1], Paula Marques[1], Jorge M. Mendes[1,3]

[1] *Instituto Nacional de Estatística*

[2] *LIAAD INESC-Porto*

[3] *Centro de Estatística e Gestão de Informação, ISEGI-UNL,*

XVIII Jornadas de Classificação e Análise de Dados
Universidade de Trás-os-Montes e Alto Douro, 7 a 9 de Abril de 2011

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

Índice

- Motivação
- Os estimadores de pequenos domínios
- O Inquérito ao Emprego
- Os estimadores SYNTH_B e o EBLUP
- Resultados
- Conclusões

Motivação

- Os métodos de estimação habituais não permitem, em muitos casos, a produção de indicadores com precisão suficiente
- Os métodos de estimação em pequenos domínios afiguram-se como uma alternativa aos métodos directos ou design-based quando a dimensão da amostra nos domínios não permite a obtenção de resultados fiáveis
- Em certos inquéritos do INE, como no Inquérito ao Emprego (IE), a dimensão dé certas unidades territoriais, NUTS II, para certos indicadores, não permite a produção de informação com precisão suficiente

Os estimadores de pequenos domínios

- Rao (2003) classifica os estimadores de pequenos domínios em **directos, indirectos, sintéticos** e **combinados**
- Neste trabalho utilizam-se estimadores sintéticos e combinados.
  - Os estimadores sintéticos assentam na hipótese de que o domínio de estudo tem características semelhantes a outro domínio de maior dimensão
  - Os estimadores combinados resultam de combinação linear entre um estimador directo e um estimador sintético

Os estimadores de pequenos domínios

- Desenvolvimentos sobre estes estimadores e aplicações, podem ser encontrados em Rao (2003), Fay e Herritot (1979), Longford (1999, 2001) e You (2008).
- No projecto Eurarea (Eurarea, 2004) estudam-se vários estimadores com o objectivo de partilhar informação a nível europeu.

O Inquérito ao Emprego (IE)

- O IE é uma operação estatística que tem como objectivo principal a caracterização do mercado de trabalho.
- Trata-se de um inquérito às famílias, por amostragem de alojamentos, realizado de forma contínua e trimestralmente.
- A amostra total está dividida em seis subamostras e obedece a um esquema de rotação no qual os alojamentos são observados durante seis trimestres consecutivos.
- O Regulamento Comunitário a que está sujeito estabelece níveis de representatividade em nível e em evolução que se traduzem na obrigatoriedade de produzir informação fiável a nível nacional e NUTS II (Nomenclatura das unidades territoriais para fins estatísticos).

Características da amostra do IE 1ºT de 2008 e da população

	Amostra	População
Indivíduos	42 226	≈ 10,6 Milhões
Agregados	15 926	≈ 3,9 Milhões
Domínios		
NUTS III	24*	30
Dim. mínima (ind.)	273 (S Estrela + C Beira)	≈ 41 000 (Pinhal Interior S)
Dim. máxima (ind.)	4749 (Grande Lisboa)	≈ 2 Milhões (G Lisboa)

Estimador sintético de regressão (SYNTH_B)

O estimador sintético ao nível do domínio (SYNTH_B), é definido da seguinte forma

$$\hat{\bar{Y}}_d^{SYNTH\_B} = \overline{\mathbf{X}}_d^T \hat{\beta} = \hat{\bar{Y}}_d$$

Sendo:

$\hat{\bar{Y}}_d$ a estimativa da média populacional no domínio d

$\overline{\mathbf{X}}_d$ a matriz das médias populacionais das variáveis auxiliares ao nível dos domínios

β é um vector de coeficientes de regressão

Estimador sintético de regressão (SYNTH_B)

- O cálculo das estimativas nos domínios depende da estimação dos parâmetros do modelo que são obtidos através do algoritmo “scoring” de Fisher usado para determinar numericamente estimadores de máxima verosimilhança.
- A estimação model-based pressupõe que os valores da variável de estudo resultam da realização de variáveis aleatórias, isto é, de um processo estocástico, cuja distribuição conjunta pode ser explicada através de um modelo

Estimador sintético de regressão (SYNTH_B)

- No modelo linear de regressão, as variáveis auxiliares encontram–se agregadas ao nível dos domínios.
- Trata-se de um modelo com uma componente fixa e duas componentes aleatórias:

$$y_d = \overline{\mathbf{X}}_d^T \boldsymbol{\beta} + u_d + \bar{e}_d$$

com $u_d + \bar{e}_d \quad iid \; N(0, \sigma_u^2 + \frac{\sigma_e^2}{n_d})$

sendo *u* e *e* os efeitos aleatórios ao nível do domínio *d (d=1,…,D)* e do indivíduo, respectivamente

- A variância é estimada a partir das variâncias dos m domínios presentes na amostra

$$\hat{\sigma}_e^2 = \frac{1}{n-m} \sum_i \sum_d (y_{id} - \bar{y}_{.d})^2$$

Empirical Best Unbiased Estimator (EBLUP)

- O estimador EBLUP tem como bases de desenvolvimento as do estimador BLUP, (Best Linear Unbiased Prediction), que não impõe pressupostos de normalidade nos efeitos aleatórios u.

$$\widehat{\overline{Y}}_i = \hat{\beta} X_i + \hat{u}_i \qquad \text{assumindo-se} \qquad \widehat{\overline{Y}}_i \sim N(\mu; V)$$

Empirical Best Unbiased Estimator (EBLUP)

- As componentes da variância são estimadas a partir da amostra via ML-Maximum Likelihood ou REML-Residual Maximum Likelihood, dadas pelo erro quadrático médio médio (MSE):

$$\mathrm{MSE}[\widehat{\bar{Y}}_i] \approx G_1(\hat{\varpi}) + G_2(\hat{\varpi}) + 2 \times G_3(\hat{\varpi})$$

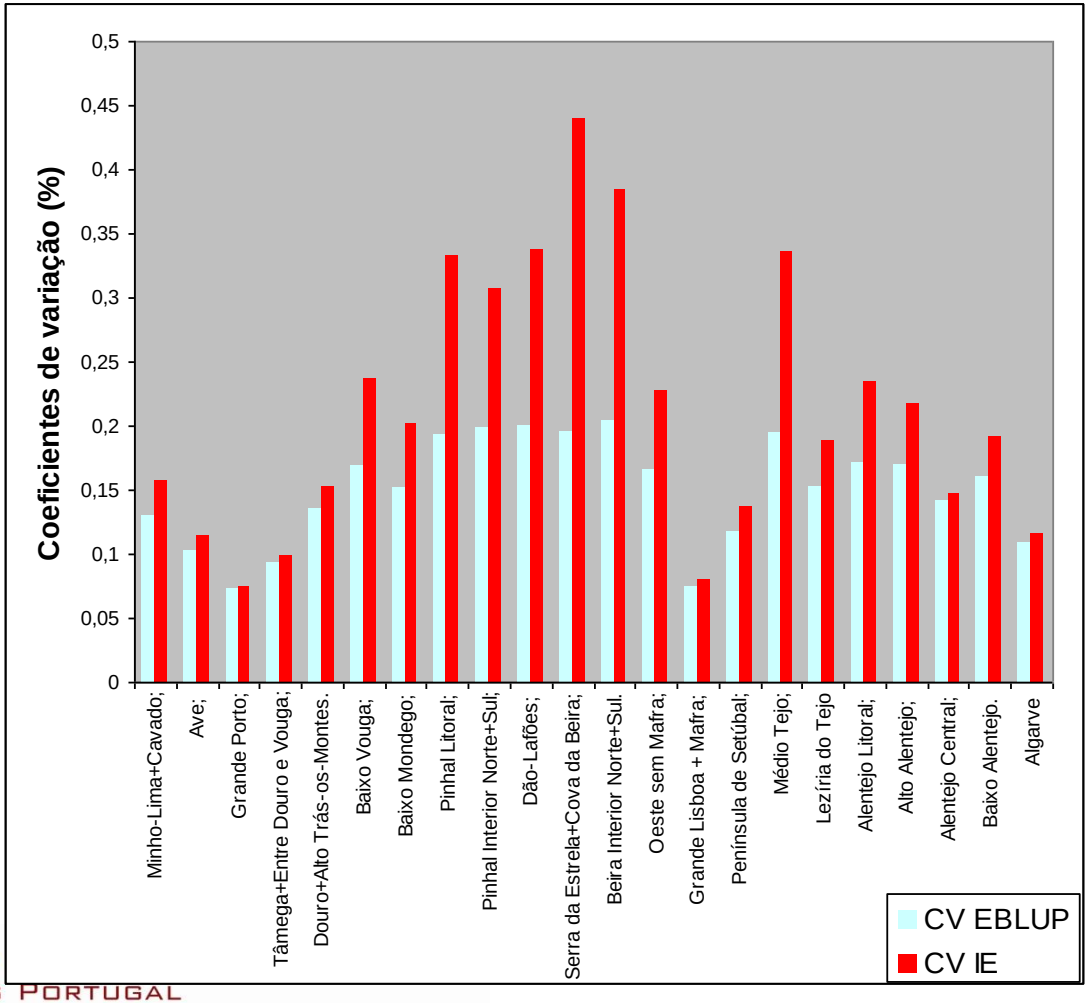
- O estimador EBLUP utilizado é o de nível área (usualmente denominado EBLUP_B) e é dado por

$$\mathrm{EBLUP} = \bar{X}_d^t \beta + \gamma_d (\bar{y}_d - \bar{X}_d^t \hat{\beta})$$

Resultados

- A variável Y a estimar representa a taxa de desemprego ao nível das NUTS III (ou agregações de NUTS III).
- O modelo proposto recorre aos dados (administrativos) do Instituto do Emprego e Formação Profissional, nomeadamente o número de desempregados inscritos nos Centros de Emprego.

Comparação gráfica entre os coeficientes de variação obtidos pelo estimador EBLUP e pelo estimador directo actualmente utilizado no IE.



Estimativas e coeficientes de variação da taxa de desemprego (estimadores directo, sintético e EBLUP ao nível das NUTS III)

NUTS III	Dimensão amostra	Estimativa (%)			Coeficiente de variação (%)		
		DIRECT	SYNTH_B	EBLUP_B	DIRECT	SYNTH_B	EBLUP_B
1 Minho-Lima+Cávado	2143	5.0	5.9	5.5	15.7	17.8	13.0
2 Ave	1625	9.8	9.1	9.1	11.4	12.7	10.3
3 Grande Porto	4134	11.2	11.5	10.9	7.4	10.0	7.2
4 Tâmega+Entre D e V	2924	7.7	7.5	7.5	9.9	15.4	9.2
5 Douro + Alto T-os-M	1320	7.3	5.6	6.8	15.2	19.1	13.5
6 Baixo Vouga	909	5.6	5.7	6.1	23.7	19.3	16.8
7 Baixo Mondego	916	4.1	5.9	4.9	20.1	18.3	15.2
8 Pinhal Litoral	653	4.1	4.5	5.3	33.3	23.8	19.2
9 Pinhal Interior N + S	503	4.0	4.1	5.0	30.7	27.4	19.9
10 Dão-Lafões	776	3.6	4.1	4.9	33.7	27.2	20.0
11 S Estrela + C Beira	273	8.4	8.6	6.7	44.0	13.7	19.5
12 Beira Interior N + S	508	5.8	6.2	6.0	38.5	18.0	20.5
13 Oeste sem Mafra	907	6.2	6.2	6.3	22.7	15.8	16.6
14 G Lisboa + Mafra	4749	8.8	8.5	9.1	8.0	12.6	7.4
15 P de Setúbal	2105	8.0	8.4	7.9	13.8	12.9	11.7
16 Médio Tejo	678	5.7	5.3	6.1	33.6	20.0	19.5
17 Lezíria do Tejo	1170	7.3	6.5	6.8	18.9	16.2	15.3
18 Alentejo Litoral	698	6.2	5.8	6.2	23.5	20.8	17.2
19 Alto Alentejo	982	7.6	8.6	6.6	21.7	12.1	17.0
20 Alentejo Central	1284	11.4	9.7	8.0	14.7	11.5	14.1
21 Baixo Alentejo	1057	8.2	9.5	7.0	19.2	11.3	16.0
22 Algarve	4261	8.0	6.3	7.4	11.5	16.9	10.8
23 Açores	3961	5.6	5.4	-	15.9	21.9	-
24 Madeira	3690	6.2	6.9	-	16.1	16.8	-

Conclusões

- Foram aplicados dois estimadores ao IE.
- A variável Y a estimar representa a taxa de desemprego ao nível das NUTS III (ou agregações de NUTS III).
- O modelo proposto recorre aos dados (administrativos) do Instituto do Emprego e Formação Profissional, nomeadamente o número de desempregados inscritos nos Centros de Emprego.
- Os resultados apresentados mostram que é possível obter ganhos de precisão nas estimativas produzidas que correspondem às regiões de menor dimensão.

Conclusões

- Assim sendo, a utilização destes estimadores faz melhorar a precisão das estimativas. Pode-se constatar que a diferença entre as estimativas obtidas pelo estimador sintético e as obtidas pelo estimador directo, são, em quase todas as regiões, inferiores à unidade em valor absoluto.
- Relativamente à precisão das mesmas nota-se um ganho generalizado, em especial no estimador EBLUP. Este ganho é mais acentuado nas regiões cuja dimensão amostral é menor.