X Congreso Galego de Estatística e Investigación de Operacións
Pontevedra, 3–4–5 de novembro de 2011

Multinomial-based small area estimation of labour force indicators in Galicia

Esther López Vizcaíno[1], Mª José Lombardía Cortiña[1] and Domingo Morales González[2]

[1]Instituto Galego de Estatística
[2]Universidade da Coruña
[3]Universidad Miguel Hernández de Elche

ABSTRACT

This work deals with small area estimation of labour force characteristics like totals of employed, unemployed and inactive people and unemployment rates. Small area estimators of these quantities are derived from a multinomial logit mixed model with independent random effects on the categories of the response vector. The mean squared errors are estimated by bootstrap methods. Finally, an application to real data from the Spanish Labour Force Survey of Galicia is given.

Palabras e frases crave: Labour Force Survey, small area estimation, area level models, multinomial mixed models, bootstrap, unemployment totals, unemployment rates.

1. INTRODUCTION

Unemployment is nowadays a major problem. Sociological surveys usually place it as one of the main concerns of the citizens and the fight against unemployment is a priority for policy action at all levels of public administration. The availability of a high quality set of labour force indicators, like totals of employed, unemployed and inactive people or unemployment rates, is very important for policy makers. The objective of this work is to estimate these indicators in counties of Galicia by using a model-based approach. In most European countries, the estimation of labour force indicators is done by using the Labour Force Survey (LFS) data. The Spanish LFS (SLFS) uses a stratified design with strata defined by the size of the municipalities. As most municipalities are not represented in the sample, and others have a very small sample size, the estimates at the municipal or county level have a low accuracy. The problem is that domain sample sizes are too small, or even null, to carry out such estimations. This situation may be treated by using small area estimation techniques. Small Area Estimation (SAE) is a part of the statistical science that combines survey sampling and finite population inference with statistical models. See a description of this theory in the monograph of Rao (2003) and more recently Jiang and Lahiri (2006). In this work we propose a multinomial logit mixed model to obtain small area estimates of labour force totals and unemployment rates in Galicia in the fourth quarter of 2008.

2. DATA DESCRIPTION

The SLFS is a quarterly survey that uses a two-stage stratified random sampling to extract samples from each Spanish province. The Primary Sampling Units (PSUs) are Census Sections, which are geographical areas with a maximum of 500 dwellings or approximately 3000 persons. These PSUs are grouped into strata according to the size of the municipality to which they belong. In the fourth quarter of 2008 a stratified sample was extracted without replacement and with probabilities proportional to the number of dwellings from each province of Galicia. The Secondary Sampling Units (SSUs) are dwellings, and in the second sampling stage approximately 4000 dwellings are selected in Galicia. All individuals aged 16 or more in the selected dwelling are interviewed.

In this work the domains of interest are the counties crossed with sex. As there are 48 counties in the SLFS of Galicia, we have $D = 96$ domains P_d partitioned in the subsets P_{d1}, P_{d2} and P_{d3} of employed, unemployed and inactive people. Our target population parameters are the totals of employed and unemployed people and the unemployment rate.

In the fourth quarter of 2008 the distribution of the sample sizes per domains in the SLFS of Galicia has the quantiles $q_{min} = 8$, $q_1 = 24$, $q_2 = 47$, $q_3 = 90$ and $q_{max} = 734$. This means that direct estimates are not reliable. Therefore, small area estimation methods are needed. Small area estimators based on unit level models (models stated for the individual units) are likely to achieve high precision when the model is correctly specified. However, the estimators derived from these models usually need auxiliary data for each unit of the sample along with aggregated data for each domain, something not always available. Area level models need only the domain totals of the auxiliary variables, information that can be usually found in administrative registers. This is why we consider an area-level model using as auxiliary variables the domain proportions of individuals within the categories of the following grouping variables:

- SEX: This variable is coded 1 for men and 2 for women.
- AGE: This variable is categorized in 3 groups with codes 1 for 16-24, 2 for 25-54 and 3 for $\geq$55
- REG: This variable indicates if an individual is registered or not as unemployed in the administrative register of employment claimants.

3. THE MULTINOMIAL MIXED MODEL

This section introduces the multinomial logit mixed model that will be used on the estimation of domain totals of employed, unemployed and inactive peoples and of unemployment rates. We first introduce some notation and assumptions. Let us use indexes $k = 1, \dots, q-1$ and $d = 1, \dots, D$ for the $q-1$ multinomial categories of the target variable and for the D domains respectively. We have $q = 3$ categories in the real data presented in Section 2, i.e. employed, unemployed and inactive people. However, there are only $q - 1 = 2$ categories in the multinomial model, i.e. employed and unemployed people. Let u_{dk} be the random effect associated to category k and domain d and define $\boldsymbol{u}_d = \underset{1\le k\le q-1}{\text{col}}(u_{dk})$. We assume that $\boldsymbol{u}_d \sim N(\boldsymbol{0}, \boldsymbol{V}_{u_d})$, are independent with covariance matrices $\boldsymbol{V}_{u_d} = \underset{1\le k\le q-1}{\text{diag}}(\varphi_k)$. We further assume that the response vectors $\boldsymbol{y}_d = \underset{1\le k\le q-1}{\text{col}}(y_{dk})$, conditioned to $\boldsymbol{u}_d$, are independent with multinomial distributions

$$\boldsymbol{y}_d|_{\boldsymbol{u}_d} \sim \text{M}(\nu_d, p_{d1}, \dots, p_{dq-1}),\ \ d = 1, \dots, D, \tag{0.1}$$

where the ν_d's are known integer numbers. The covariance matrix of $\boldsymbol{y}_d$ conditioned to $\boldsymbol{u}_d$ is $\boldsymbol{W}_d = \text{var}(\boldsymbol{y}_d|\boldsymbol{u}_d) = \nu_d[\text{diag}(\boldsymbol{p}_d) - \boldsymbol{p}_d\boldsymbol{p}_d']$, where $\boldsymbol{p}_d = \underset{1\le k\le q-1}{\text{col}}(p_{dk})$ and $\text{diag}(\boldsymbol{p}_d) = \underset{1\le k\le q-1}{\text{diag}}(p_{dk})$. For the natural parameters $\eta_{dk} = \log(\frac{p_{dk}}{p_{dq}})$, we assume the model

$$\eta_{dk} = \boldsymbol{x}_{dk}\boldsymbol{\beta}_k + u_{dk},\ \ d = 1, \dots, D,\ k = 1, \dots, q-1, \tag{0.2}$$

where $\boldsymbol{x}_{dk} = \underset{1\le r\le t_k}{\text{col}'}(x_{dkr})$, $\boldsymbol{\beta}_k = \underset{1\le r\le t_k}{\text{col}}(\beta_{kr})$ and $t = \sum_{k=1}^{q-1} t_k$. The mean and the variance of y_{dk}, conditioned to $\boldsymbol{u}_d$, are $\mu_{dk} = \nu_d p_{dk}$ and $\omega_{dk} = \nu_d p_{dk}(1 - p_{dk})$, respectively. The probability of the multinomial category k in the domain d is

$$p_{dk} = \frac{\exp\{\eta_{dk}\}}{1 + \sum_{\ell=1}^{q-1}\exp\{\eta_{d\ell}\}},\ d = 1, \dots, D,\ k = 1, \dots, q-1.$$

To fit the model we combine the PQL method for estimating and predicting the $\boldsymbol{\beta}_k$'s and the $\boldsymbol{u}_d$'s, with the REML method for estimating the variance components φ_k's. The presented method is based on a normal approximation to the joint probability distribution of the vector $(\boldsymbol{y}, \boldsymbol{u})$. The combined algorithm was first introduced by Schall (1991) and later used by Molina et al. (2007) in applications of generalized linear mixed models to small area estimation problems. In this work, we adapt the combined algorithm to the multinomial logit mixed models. Concerning the estimation of the mean squared error of $\hat{p}_{dk}$, we can also use the approach of González-Manteiga et al. (2008) by introducing the parametric bootstrap method.

4. APLICATION TO REAL DATA

For each category, Table 1 presents the estimates of the regression parameters and their standard deviations. It also presents the p-values for testing the hypothesis $H_0 : \beta_{kr} = 0$. The estimates of the model variances and their standard deviations (in parentheses) are $\hat{\varphi}_1 = 0.17$ (0.03) and $\hat{\varphi}_2 = 0.50$ (0.13). Confidence intervals for $\hat{\varphi}$ do not contain the 0 which confirms the need to include two random effects in the model proposed.

	Employed people			Unemployed people		
Variable	Estimate	Std. Dev.	p-value	Estimate	Std. Dev.	p-value
CONSTANT	-1.415	0.255	0.000	-5.236	0.567	0.000
REG	-1.680	1.161	0.148	6.089	2.347	0.009
AGE=1	2.378	1.172	0.042	2.891	2.408	0.230
AGE=2	3.274	0.552	0.000	4.490	1.132	0.000
SEX=1	-0.520	0.115	0.000	-0.352	0.231	0.127

Table 1. Parameter estimates $\widehat{\beta}_{kr}$, $k=1,2$, $r=1,\dots,5$.

Figure 0.1 plots the domain residuals

$$r_{dk} = \frac{y_{dk} - \hat{y}_{dk}}{y_{dk}}, \quad k=1,2, \quad d=1,\dots,96,$$

versus the predicted sample totals of employed and unemployed people. We observe that in general the model-based estimates are smother than the direct ones. Figure

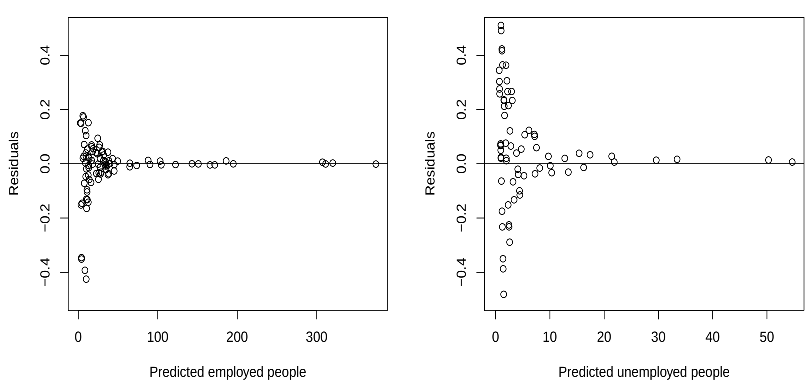

Figure 0.1: Domain residuals versus predicted values.

0.2 plots the direct ($\hat{Y}_{dk}^{dir}$) versus the model-based estimates ($\hat{N}_d^{dir}\hat{p}_{dk}$) of the totals of employed and unemployed people. We observe that both type of estimates behave quite similarly for employed people. This is because there are enough observations for these estimates. However, their behaviour presents slightly greater differences for unemployed people, which are due to the lower number of observations Figures 0.3 and 0.4 plot the estimated employment totals and unemployment rates respectively for men (left) and women (right), with counties sorted by sample size. We observe that both estimates (direct and model-based) tend to be closer as soon as the sample size increases. Figure 0.5 plot the parametric bootstrap estimates of the relative mean squared errors (RMSE) of the model-based estimates of the unemployment rates. The RMSEs of the corresponding direct estimates are much greater than their counterparts of the model-based ones. This is why they are not plotted.

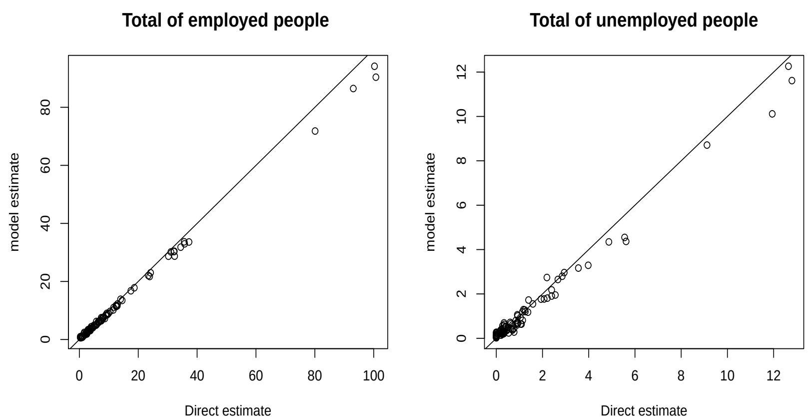

Figure 0.2: Direct versus model-based estimates.

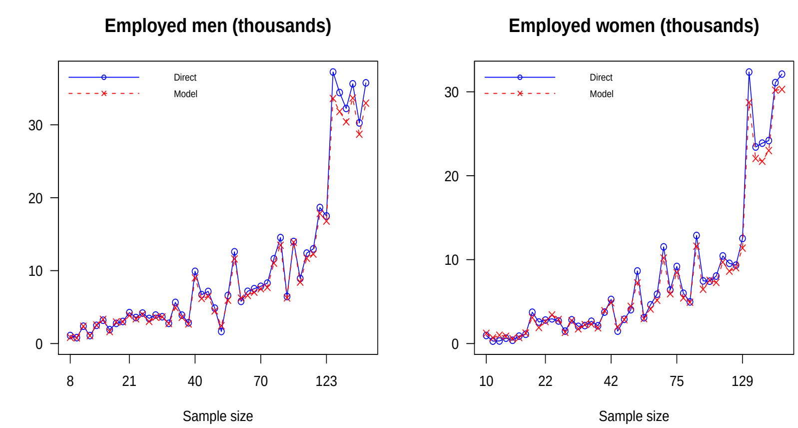

Figure 0.3: Men and women employment totals.

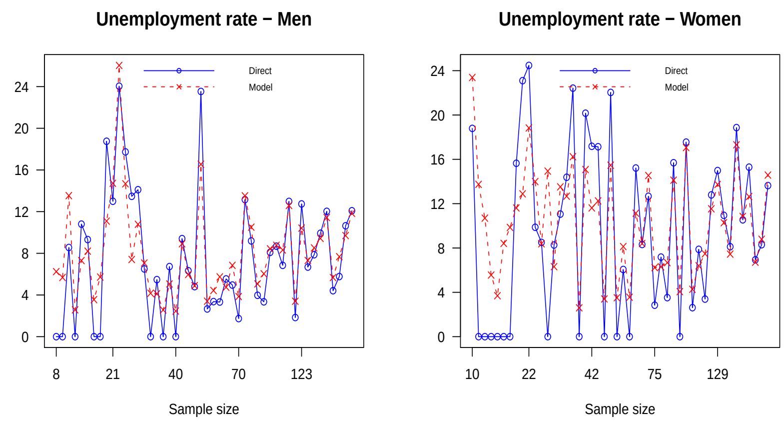

Figure 0.4: Men and women unemployment rates.

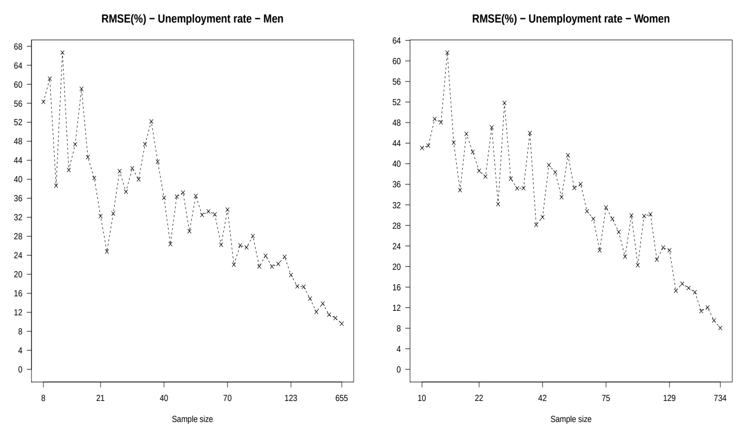

Figure 0.5: RMSEs of model-based estimates of unemployment rates.

REFERENCES

González-Manteiga, W., Lombardía, M.J., Molina, I., Morales, D. and Santamaría, L. (2007). Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model. *Computational Statistics and Data Analysis*, **51**, 2720–2733.

Jiang, J. and Lahiri, P. (2006). Mixed model prediction and small area estimation. *Test*, **15**, 1-96.

Molina, I., Saei, A. and Lombardía, M.J. (2007). Small area estimates of labour force participation under multinomial logit mixed model. *The Journal of the Royal*

Statistical Society, series A, **170**, 975–1000.

Rao, J.N.K. (2003). *Small area estimation.* John Wiley and Sons, New-York.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*,