A comparison of methods to estimate poverty indexes in small samples

Giuseppe Antonaci[1,3], Pedro L. N. Silva[1], Fernando Moura[2]
[1] Brazilian Institute of Geography and Statistics, Brazil
[2] Federal University of Rio de Janeiro, Brazil
[3] Corresponding author: Giuseppe Antonaci, e-mail: gantonaci@gmail.com

Abstract

In recent years, official statistics institutes are facing an increasing demand for more detailed information about the population. In many cases, it would be too costly to provide the required detail using standard direct sample estimators. In this scenario, the development of small area estimation methods has become a necessity and its use more widespread. The World Bank has a poverty mapping project that relies on the small area estimation method developed by Elbers, Lanjouw and Lanjouw (ELL). This approach has been used in several countries, including Brazil. Since the method used has great influence on the estimates and their precision, there is a clear advantage in using the most efficient one. In this paper we compare the ELL method, the Empirical Bayes (EB) method and the Hierarquical Bayes (HB) method, these last two described by Rao. The comparison is done using a super population model where the parameters were based on the 2000 Brazilian Census and the sample design used tries to mimic those of IBGE's surveys. The HB method was the most efficient both in point and interval estimation.

Keywords: Empirical Bayes, Hierarquical Bayes, mixed model, World Bank, poverty mapping

1 Introduction

The main motivation for this work was the publication of the 2008 Brazilian Poverty Map by IBGE (2008) that used the ELL method from Elbers, Lanjouw and Lanjouw (2002). This paper is based on the paper by Rao and Molina (2010), where they compare the ELL and the Empirical Bayes (EB) methods. Our study includes the Hierarchical Bayes (HB) method described by Rao (2003) and we use a super population model that tries to mimics some of IBGE's complex surveys.

2 Small area methods

Before discussing the small area methods we define the FGT poverty index developed by Foster, Greer and Thorbecke (1984). For an area m and considering a poverty line z the FGT index of type α is:

$$F_m(\alpha, z) = \frac{1}{N_m} \sum_{j=1}^{N_m} \left(\frac{z - R_{mj}}{z} \right)^{\alpha} \mathrm{I}(R_{mj} < z) \tag{1}$$

where $\alpha = 0, 1, 2$, $m = 1, \ldots, M$, N_m is the number of units at area m, R_{mj} is a measure of well being of unit j from area m, z is the poverty line, at the same scale as R_{mj}, and $\mathrm{I}(R_{mj} < z)$ is an indicator function that assumes 1 if $R_{mj} < z$ or 0 otherwise.

For this paper we will focus on the FGT index when $\alpha = 0$, which is the proportion of units under the poverty line. The poverty line, z, is defined here

as 60% of the median of the variable used to measure the well being, following what is used in other papers on this subject, eg. Rao and Molina (2010); Osier (2009).

The FGT index depends on a measure of well being, such as the Income ($\boldsymbol{R}$). However, this variable is known only for a small sample of the population, $\boldsymbol{R}_s$, where s denotes the subset of the population in the survey, and precise estimates cannot be obtained. Nevertheless there is a set of variables known for all the population from a census, $\boldsymbol{X}$, that can be used to predict $\boldsymbol{R}$. Also define index r the complement of s.

From the survey data we can fit a model to estimate the relationship between $\boldsymbol{R}_s$ and $\boldsymbol{X}_s$. Applying this model to the census data, we estimate the Income for all units in the population and then poverty indexes for all small areas.

The model we used to relate the income, $\boldsymbol{R}_s$, and known population variables, $\boldsymbol{X}_s$, was a multilevel model with random intercept:

$$Y_{mj} = \mathrm{g}(R_{mj}) = \beta_0 + \beta_1 X_{1,mj} + \cdots + \beta_p X_{p,mj} + u_m + e_{mj} \tag{2}$$

where $X_{p,mj}$ is the value of the p-th variable for unit j in area m and u_m and e_{mj} refers to the random errors of prediction for the area level and for the units respectively. Since we are using Income it is usual that $\mathrm{g}(\cdot) = \log(\cdot)$. We suppose $u_m \sim \mathrm{N}(0, \sigma_u^2)$ and $e_{mj} \sim \mathrm{N}(0, \sigma_e^2)$. Finally, let $\boldsymbol{Y} = \mathrm{g}(\boldsymbol{R}) = \log(\boldsymbol{R})$.

The estimation of the mixed model parameters in the ELL method is done using a procedure developed by Elbers et al. (2002). The EB and the HB methods do not focus on this estimation and recommend the use of a valid and well know estimation method. Their difference is the use of a frequentist method for the EB and a Bayesian method for the HB.

The estimation of $\boldsymbol{Y}$ and the FGT index has similarities between the various methods, but again with some differences. All of the methods estimate the distribution of $\boldsymbol{Y}$ given $\boldsymbol{X}$. Since the transformation done by the FGT index is too complex the distribution of the index, or even its mean, cannot be directly obtained.

To solve this problem all three methods take replicas of $\boldsymbol{Y}$ based on its distribution, calculate the FGT index on each replica and take the mean across all replicas as an estimate.

The ELL method estimates the parameters of the mixed model and then it produces the replicas directly from the distribution of $\boldsymbol{Y}|\hat{\boldsymbol{\beta}}, \hat{\sigma}_u, \hat{\sigma}_e$.

The EB method depends on the errors being normally distributed, so that the distribution of $\boldsymbol{Y}$ given $\boldsymbol{X}$ is also Normal. It then uses properties of the multivariate normal distribution and the knowledge of $\boldsymbol{Y}_s$, so the replicas are taken from $\boldsymbol{Y}_r|\boldsymbol{Y}_s, \hat{\boldsymbol{\beta}}, \hat{\sigma}_u, \hat{\sigma}_e$.

For the HB method we already obtain the posterior distribution of the parameters so we can take the replicas for $\boldsymbol{Y}_r$ from

$$f(\boldsymbol{Y}_r|\boldsymbol{Y}_s) = \int_{(\boldsymbol{\beta},\sigma_u,\sigma_e)} f(\boldsymbol{Y}_r|\boldsymbol{\beta}, \sigma_u, \sigma_e, \boldsymbol{Y}_s) \cdot f(\boldsymbol{\beta}, \sigma_u, \sigma_e|\boldsymbol{Y}_s) d(\boldsymbol{\beta}, \sigma_u, \sigma_e)$$

An important part of each method is how they estimate the precision of their estimators. The ELL and the HB method both use the FGT index estimates from the replicas to estimate the mean square error (MSE) of the estimator. The EB method uses a parametric bootstrap to estimate the MSE.

3 Comparison of the methods

To compare the methods we used a superpopulation model from which 3000 simulated populations were created. From each of these populations a sample

similar to those used in IBGE's household surveys was drawn.

Since in this case we have information about all population and its sample it was possible to verify the real estimation error for each method. Moreover we tested if the MSE estimators for each method were efficient in estimating the true MSEs.

3.1 Simulated Population

We used a super population model with two levels and three independent variables. The model was:

$$Y_{mj} = \beta_0 + \beta_1 X_{1,mj} + \beta_2 X_{2,mj} + \beta_3 X_{3,mj} + u_m + e_{mj} \tag{3}$$

Variables $X_{1,mj} \sim Ber((m/160)^2)$ and $X_{2,mj} \sim Ber\left(\frac{(1-m/160)}{2}\right)$ change across the areas so the mean income changes for different areas.

$\boldsymbol{X}_3$ is a dummy variable that assumes 1 if the unit is in an urban area and 0 if it is in an rural area. Therefore β_3 accounts for the difference in the logarithm of the income for living in a urban area.

The values of the fixed parameters were $\beta_0 = 3$, $\beta_1 = 1$, $\beta_2 = -1.2$ and $\beta_3 = 0.4$ and for the random parameters we used $u_m \sim N(0, \sigma_u^2)$ and $e_{mj} \sim N(0, \sigma_e^2)$ where $\sigma_u^2 = 0.1$ and $\sigma_e^2 = 0.65$.

We simulated data for 160 areas, each area with N_m units where

$$N_m \sim \begin{cases} Gamma\left(\frac{250^2}{5800}, \frac{250}{5800}\right), & \text{if } m \text{ is an urban area} \\ Gamma\left(\frac{115^2}{4130}, \frac{115}{4130}\right), & \text{if } m \text{ is a rural area} \end{cases}$$

The values of the parameters for the simulation were chosen to mimic a real life population. The starting point was the model used by Rao and Molina (2010) and we adapted it for our objectives and reality. The data used was the 2000 Demographic Census for the state of Minas Gerais. This state was chosen because it is considered a good proxy of the entire country, having borders with the richer municipalities at the south and with some of the poorest ones at the north.

The greatest modification was introducing a variable to classify urban and rural areas. This was done because of the great influence this characteristic has on the Income and because it is used as a stratification variable in our surveys. Of the 160 simulated areas 47 were randomly flagged as rural and kept fixed across all simulations.

The values of β_3, the weight of the random parameters in the total variance, the proportion of rural areas as well as their distribution of number of households were also based on data from the 2000 Demographic Census for Minas Gerais.

3.2 Complex survey design

The areas were stratified considering the urban and rural classification and the area mean of the household income. We created 3 strata for urban areas and 2 for rural areas. We selected a sample of 40 areas where Neyman allocation of the sample to the strata was used. In each sampled area we selected 25 units. In total, for each simulated population, we had a sample of 1,000 units from a population of approximately 33,000 units.

Since we used a complex survey design the units have different selection probabilities and this may have an influence on the point and interval estimates (Cochran, 1977). However, there is no consensus if the weights must be used to estimate the parameters in a mixed model or what is the best method to use them.

We tested a few methods to estimate the model parameters and compared their results with the true values. The methods used were Restricted Maximum Likelihood (REML), Pseudo Maximum Likelihood (PML), Probability Weighted Iteratively Generalized Least Square suggested by Pffefermann et al. (1998) and Gibbs sampling (MCMC).

All methods did really well on estimating all parameters except for σ_u. The PML was discarded after our first test since its estimator of σ_u was both biased and imprecise. The REML and PWIGLS were close so further tests were performed with them. After testing 30 different combinations of population and sample size the REML performed a little better than the PWIGLS overall and was therefore chosen.

3.3 Point estimation of poverty indexes

For each of the 3,000 populations and samples simulated we estimated the proportion in poverty ($\hat{F}_m(0,z)$) for each of the 160 areas using the methods discussed. For comparison we also calculated the true value of the proportion in poverty in each populations area. We then have

$$\hat{F}^{\text{ELL}}_{i,m}(0,z) \qquad \hat{F}^{\text{EB}}_{i,m}(0,z) \qquad \hat{F}^{\text{HB}}_{i,m}(0,z) \qquad F_{i,m}(0,z)$$

where $i = 1, \ldots, 3000$ and $m = 1, \ldots, 160$.

We calculated the mean of estimates for each area across all simulations and the results are presented in figure 1. We can see that all methods, across several simulations, managed to efficiently estimate the true value of the poverty indexes.

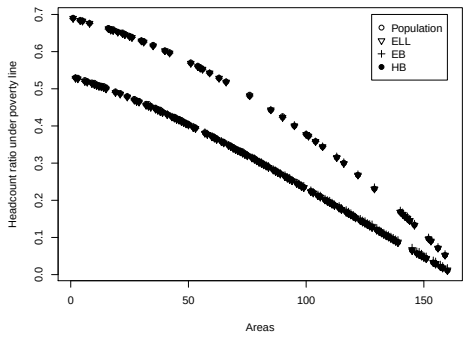

Figure 1: Proportion in poverty by area

It is easy to notice in figure 1 how the effect of being a rural area increases its proportion of the population in poverty.

Getting near the true value across several simulations is not enough, we also need the estimators to be as near as possible of the true value at each iteration of the superpopulation model. We calculated the square error between the estimates and the true value for each simulated population and each area. Then the 3,000 square errors for each area were divided in two groups, when the area was sampled and when the area wasn't and we took the mean in each group. So we have:

$$\text{MSE}^{\text{ELL}}_{s,m} = \frac{1}{\sum_{i=1}^{3000} \text{I}(m \in s_i)} \sum_{i=1}^{3000} \left(\hat{F}^{\text{ELL}}_{i,m}(0,z) - F_{i,m}(0,z) \right)^2 \text{I}(m \in s_i)$$

where s_i is the subset of area indexes sampled in population i and $\text{I}(m \in s_i)$ is an indicator function that assumes 1 if m is in s_i and 0 otherwise. The formulas for the MSE for other methods and for when the areas were not sampled are similar.

In figure 2 we present the true MSE for the areas.

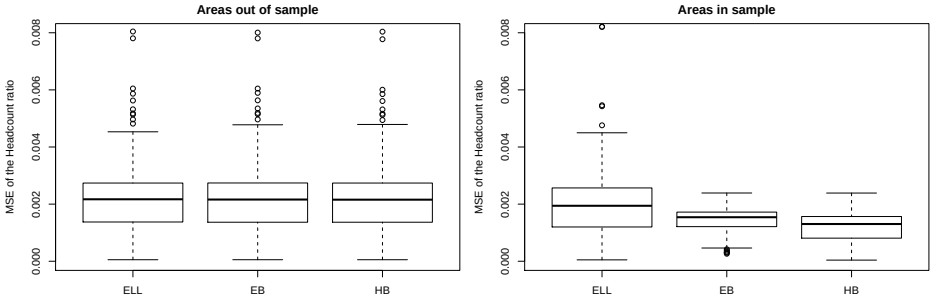

Figure 2: True MSE of the Proportion in poverty

For the areas not sampled the estimator for the 3 methods is basically the same, but the EB and HB methods use the information available for units sampled in the sampled areas. Both have a much smaller MSE than the ELL method in this case.

3.4 Estimating the MSE

Almost as important as having a low estimation error is knowing it. When working with real data we don't have access to the true value of the estimators, so we can't calculate the estimation errors like they were just presented and we need a way to estimate them. In this subsection we will estimate the MSE as proposed by each method and compare them to the true errors.

Just like we did with the true errors, we estimated the errors for each population and area, when areas were sampled and when they were not. We have:

$$\hat{\text{mse}}^{\text{ELL}}_{s,m} = \frac{1}{\sum_{i=1}^{3000} \text{I}(m \in s_i)} \sum_{i=1}^{3000} \hat{\text{mse}}^{\text{ELL}}_{i,m} \cdot \text{I}(m \in s_i)$$

and similar formulas for others methods and when the area wasn't sampled.

To analyze the precision of the estimation of the errors we calculated the ratio of the estimated value of the error for each area and the true value obtained in subsection 3.3.

$$\text{ratio}^{\text{ELL}}_{s,m} = \frac{\hat{\text{mse}}^{\text{ELL}}_{s,m}}{\text{MSE}^{\text{ELL}}_{s,m}}$$

with similar formulas for other methods and areas and for when the areas were not sampled.

Ideally the ratios would be near one. Smaller values mean that the error is underestimated and greater values that they are overestimated.

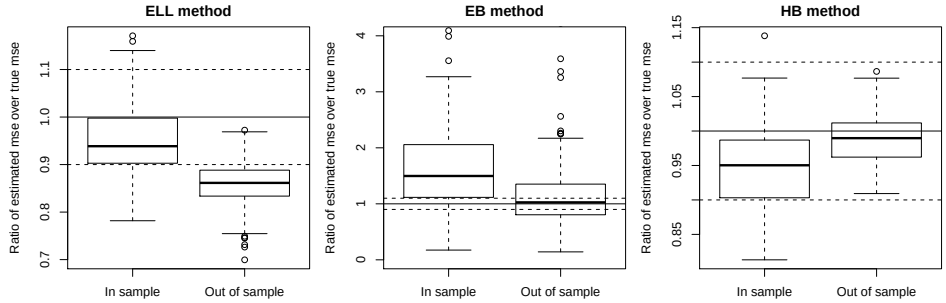

Figure 3: Boxplot of the ratios of MSE for sampled and unsampled areas

We can see how efficiently the ELL method estimated the estimation error. The errors for the sampled areas were a little under estimated, but most still fall

inside a 10% interval. However, for most of the areas not sampled the errors were under estimated. This is dangerous since we tend to believe that our estimates are more precise than they really are.

The results found for the EB method are inconclusive. The bootstrap method to estimate the erros has been shown by Rao and Molina (2010) to correct estimate the errors for a simpler population model. Apparently the complexity introduced in the population model and the survey design greatly increased the time required to estimate the errors using the bootstrap approach.

For the sampled areas the HB method underestimated a little the MSE but around 75% of the areas had their errors estimated into the 10% range. For the unsampled areas, all errors were well estimated. Not one area had their mse estimated farther than 10% of the real value.

4 Conclusion

In this comparison the HB method was clearly superior. It was the one with lower mean square error and the one to better estimate this error.

The bootstrap procedure to estimate the MSEs of the EB method seems to have problems to converge with the population and sample we used. This can probably be solved with more computational power, but the EB method already took ten times longer than the HB method to get to the results presented here.

Even if the ELL method did estimate well its error this advantage is lost since the method was not as precise as the EB or HB method when estimating the poverty index.

References

Cochran, W. G. (1977) *Sampling Techniques*. Probability and Mathematical Statistics-Applied. John Wiley & Sons, Inc., 3 edn.

Elbers, C., Lanjouw, J. O. and Lanjouw, P. (2002) Micro-level estimation of welfare. *Policy Research Working Paper 1*, The World Bank.

Foster, J., Greer, J. and Thorbecke, E. (1984) A class of decomposable poverty measures. *Econometrica*, **52**, 761–766.

IBGE (2008) Mapa de pobreza e desigualdade: municípios brasileiros. *Technical report*, Instituto Brasileiro de Geografia e Estatística.

Osier, G. (2009) Variance estimation for complex indicators of poverty and inequality using linearization techniques. *Survey Research Methods*, **3**, 167–195.

Pffefermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H. and Rasbash, J. (1998) Weighting for unequal selection probabilities in multilevel model. *Journal of the Royal Statistical Society series B*, **60**, 23–40.

Rao, J. N. K. (2003) *Small Area Estimation*. Wiley Series in Survey Methodology. John Wiley & Sons, Inc.

Rao, J. N. K. and Molina, I. (2010) Small area estimation of poverty indicators. *The Canadian Journal of Statistics*, **38**, 369–385.