Small area estimation for panel data

Annamaria Bianchi

Abstract Small area estimation based on M-quantile regression has recently been introduced by [2] and it has proved to provide a valid alternative to traditional methods. Thus far, this method has only been applied to cross-sectional data. However, it is well known that the use of panel data may provide significant gains in terms of efficiency of the estimators. This paper explores possible extensions of M-quantile based small area estimators to the panel data context. An application of the methodology to the Kauffman Firm Survey data is envisaged.

1 Introduction

In recent years, there has been an increasing demand by policy makers for estimates of population characteristics at regional level. Unfortunately, limited founding resources for the design of sample surveys often lead to small sample sizes within these domains. As a consequence, direct estimators (which use only data from sample units in the domain) cannot be applied since they yield estimates with unacceptable standard errors. These problems are typically overcome by the use of small area techniques. This is an approach based on models that borrow strength in making an estimate for one small area from sample survey data collected in other small areas and/or at other time periods. The most popular class of models for small area estimation is based on random effects models, which include random area effects to account for between area variation.

In this paper, we focus on models that borrow strength across both small areas and times. These types of estimators are generally based on panel data, that is, sample surveys repeated in time over the same units. Panel data combine the individual dimension with the time dimension, thereby augmenting the information of the data with respect to a cross-section approach. For this reason panel data analysis presents many benefits. It allows to control for individual (and time) unobserved heterogeneity,

Annamaria Bianchi, Dipartimento di Matematica, Statistica, Informatica e Applicazioni, Università degli Studi di Bergamo; email: annamaria.bianchi@unibg.it

Paper supported by the ex 60% 2010 University of Bergamo, Biffignandi grant.

and hence allow to isolate the longitudinal variability of the investigated phenomena from the variability due to the different characteristics of the responding units. Moreover, panel data are more informative since there is more variability and the estimates are therefore more efficient.

In the small area context, it is well known that for such repeated surveys considerable gains in efficiency can be achieved by borrowing strength across both small areas and times. Thus far, the use of longitudinal data for purposes of small area estimation is concentrated mostly on the area level models [3]. The possible reason for this is that in many countries, and especially in the United States, the infrastructure of the official statistics does not support longitudinal data sets at individual level. On the other hand, research on small area estimation from unit level panel data is clearly needed, because aggregating individual level data to adapt for area level models may cause unnecessary loss of information. At the unit level, an appropriate model for panel data must take the covariance of the repeated observations from the same unit into account. One simple model that can be adapted to this purpose is the two-fold nested error regression model proposed by [7]. The small area means then are estimated by the Empirical Best Linear Unbiased Predictor (EPLUP). Refer to [4] for more details.

Recently, an alternative unit-level approach to small area estimation based on M-quantile regression has been proposed by [2]. The advantages of M-quantile based small area estimators with respect to traditional random effects models are that they do not depend on strong distributional assumptions and that they are outlier robust. The initial estimator proposed in [2] has subsequently been extended in various ways [6,8]. However, to the best of out knowledge, up to now this technique has only been applied to cross-sectional data. The gains in efficiency that can be obtained using panel data have not been explored yet. The aim of this research is a theoretical development of M-quantile small area estimators to the panel data context. We also investigate the possibility of applying the methodology to the Kauffman Firm Survey data.

2 M-quantile small area estimation for cross-sectional data

Suppose that a population U of size N is divided into non-overlapping domains of size N_j, $j=1,\dots,d$. Assume that a sample s is available and denote n_j the sample size in area j and s_j (r_j) the sampled (non-sampled) population units in the area. Let y_{ij} denote the value of the variable of interest for unit i belonging to the small area j ($j=1,\dots,d$, $i=1,\dots,n_j$). Assume that the values of y are available for each unit of the sample and that a vector of auxiliary variables x_{ij} is available for each unit of the population. We are interested in predicting small area means for the target variable for each small area: $m_j = N_j^{-1}\sum_{i\in s_j\cup r_j} y_{ij}$ ($j=1,\dots,d$).

A recently introduced approach to small area estimation is based on M-quantile regression. M-quantile regression [1] provides a ‘quantile-like’ generalization of regression based on influence functions. For fixed q ($0<q<1$), the linear M-quantile model of order q (denoted hereafter $Q_q(x;\psi)$) of the conditional distribution of y given x is given by $Q_q\left(x;\psi\right)=x^T\beta_\psi\left(q\right)$, where ψ refers to an appropriately chosen influence function (such as Huber Proposal 2). An estimate $\hat{\beta}_\psi\left(q\right)$ of $\beta_\psi\left(q\right)$ is obtained by solving the following equations (in β)

$$\sum_{j=1}^{d}\sum_{i=1}^{n_j}\psi_q\left(y_{ij}-x_{ij}^T\beta\right)=0,$$

where $\psi_q\left(r\right)=\psi\left(s^{-1}r\right)\left\{qI\left(r>0\right)+\left(1-q\right)I\left(r\le 0\right)\right\}$ and s is a robust estimate of scale such as the mean absolute deviation.

The idea underlying M-quantile based small area estimation is the following. The conditional variability across the population can be characterized by the so-called M-quantile coefficients of the population units. For unit i in small area j with values $\left(x_{ij},y_{ij}\right)$, this coefficient is defined as the value q_{ij} such that $Q_{q_{ij}}\left(x_{ij};\psi\right)=y_{ij}$ -- that is, q_{ij} is the order of the M-quantile passing through the point $\left(x_{ij},y_{ij}\right)$. If a hierarchical structure does explain part of the variability in the population, then it is expected that units belonging to the same area have similar coefficients. It is therefore natural to characterize each small area j by means of an indicator θ_j defined here as the mean of the population M-quantile coefficients belonging to that area $\theta_j=N_j^{-1}\sum_{i\in s_j\cup r_j}q_{ij}$. This coefficient identifies an M-quantile regression plane $Q_{\theta_j}\left(x;\psi\right)$ characteristic of that area, which allows the prediction of unobserved data in the area. Such predicted values are then used to construct estimates of m_j . When $\beta_\psi\left(q\right)$ is a sufficiently smooth function of q , the following bias-adjusted estimator has been proposed [8]:

$$\hat{m}_j=N_j^{-1}\left[\sum_{i\in s_j}y_{ij}+\sum_{i\in r_j}x_{ij}^T\hat{\beta}_\psi\left(\hat{\theta}_j\right)+\frac{N_j-n_j}{n_j}\sum_{i\in s_j}\left\{y_{ij}-x_{ij}^T\hat{\beta}_\psi\left(\hat{\theta}_j\right)\right\}\right].$$

Refer to [8,6] for other possible estimators, together with the corresponding MSE estimators.

3 M-quantile small area estimation for panel data

Let now y_{ijt} denote the value of the variable of interest for unit i belonging to the small area j at time t ($j=1,\ldots,d$, $t=1,\ldots,T$, $i=1,\ldots,n_{jt}$). Denote x_{ijt} the corresponding covariates known at the population level. We are now interested in predicting small area means for the target variable at the final time T : $m_{jT}=N_j^{-1}\sum_{i\in s_i\cup r_i}y_{ijT}$ ($j=1,\ldots,d$).

In order to extend the M-quantile based small area technique to panel data, the first step is to extend M-quantile regression to panel data. For a given q , the simplest M-quantile panel data model is $Q_{qt}\left(x;\psi\right)=x^T\beta_\psi\left(q;t\right)$ ($t=1,\dots,T$). The natural estimator for $\beta_\psi\left(q;t\right)$ is then the pooled M-quantile regression estimator $\hat{\beta}_\psi\left(q;t\right)$, which solves

$$\sum_{j=1}^{d}\sum_{t=1}^{T}\sum_{i=1}^{n_{jt}}\psi_q\left(y_{ijt}-x_{ijt}^T\beta\right)=0.$$

This kind of regression allows to look at the dynamic relationship and is expected to increase the efficiency of estimators. However, it does not allow us to control for unobserved cross-section heterogeneity. An improvement of the β estimates is expected by allowing explicitly for unobserved effects. For a given q, a natural specification that incorporates strict exogeneity is $Q_{qt}\left(x,c;\psi\right)=x^T\beta_\psi\left(q;t\right)+c$, where c denotes an unobserved, time-constant variable called unobserved effect. Unfortunately, unlike in the case of estimating effects on the conditional mean, it is not possible to proceed without further assumptions. The same problem is faced by quantile regression for panel data and it has been treated recently. Refer to [9] and references therein.

In the present research, we intend to give conditions allowing the estimation of the M-quantile model. Once estimators are well defined we extend the M-quantile based small area estimators to panel data. Of course, also the problem of MSE estimation has to be addressed, taking into account the covariance of the repeated observations from the same unit. Next we assess possibilities of application of the methodology to the Kauffman Firm Survey data sponsored by the Ewing Marion Kauffman Foundation. The KFS is the largest longitudinal survey of new businesses in the world. It consists of a cohort of 4,928 firms that were founded in 2004 in the United States and tracked over the years [5]. The analysis will allow us to perform comparative economic analysis of different areas in the U.S. related to businesses in their early years of operation.

References

1. Breckling, J., Chambers, R.: M-quantiles, Biom. **75**, 761—771 (1988)
2. Chambers, R., Tzavidis, N.: M-quantile models for small area estimations. Biom. **93**, 255—268 (2006)
3. Rao, J.N.K., Yu, M.: Small area estimation by combining time series and cross-sectional data. Can. J. Stat. **22**, 511—528 (1994)
4. Rao, J.N.K.: Small Area Estimation. Wiley, New York (2003)
5. Robb, A., Reedy, E.J., Ballou, J., Des Roches, D., Potter, F., Zhao, Z.: An Overview of the Kauffman Firm Survey. Results from the 2004-2008 Data. Available at http://www.kauffman.org/uploadedFiles/kfs_2010_report.pdf
6. Salvati, N., Tzavidis, N., Pratesi, M., Chambers, R.: Small area estimation via M-quantile geographically weighted regression. Test (2010) doi: 10.1007/s11749-010-0231-1
7. Stukel, D.M., Rao, J.N.K.: On small-area estimation under two-fold nested error regression models. J. Stat. Plan. Inference **78**, 131—147 (1999)
8. Tzavidis, N., Marchetti, S., Chambers, R.: Robust estimation of small area means and quantiles. Aust. N. Z. J. Stat. **52**, 167—186 (2010)
9. Wooldridge, J.: Econometric analysis of cross section and panel data, MIT Press, Cambridge (2010)