UNIVERSIDADE DA CORUÑA

gmodes DPTO. DE MATEMÁTICAS

SEMINARIO:

"Inferencia en Áreas Pequeñas"

PROGRAMA:

9.30h-10.30h: "*Métodos Estadísticos con Restricciones. Aplicación a la Estimación de Áreas Pequeñas"* (Cristina Rueda Sabater, Universidad de Valladolid)

10.30-11.30h: *"Nonparametric Mixed Effects Models with Applications to Small Area Statistics: Estimation and Parameter Choice"* (María Dolores Martínez Miranda, Universidad de Granada)

11.30-12.00h: Café-Descanso

12.00-13.00h: *"Estimación en Áreas Pequeñas Aplicada a Datos del Ámbito Social y Laboral"* (Roberto Domínguez, Proyecto IGE: Áreas Pequeñas).

Lugar: Aula 2.Grados. Facultade de Informática da UDC
Data: 25 de xaneiro de 2010

TITULO: Métodos Estadísticos con Restricciones. Aplicación a la Estimación de Áreas Pequeñas.

En la primera parte de la charla se introducen las herramientas básicas de los procedimientos de la Inferencia Con Restricciones. Se hace un breve recorrido por los distintos problemas que se resuelven dentro de éste área y se examinan ejemplos ilustrativos. En la segunda parte de la charla se desarrollan los métodos específicos para la resolución de problemas de estimación en áreas pequeñas, poniendo de manifiesto la bondad de los procedimientos con la aplicación de la metodología a conjuntos de datos reales y simulados.

Nonparametric Mixed Effects Models with applications to Small Area Statistics: Estimation and Parameter Choice

María Dolores Martínez Miranda *
Universidad de Granada

Abstract

Nonparametric modelling offers new perspectives for various problems typically faced in small area statistics. A main purpose of small area statistics is the estimation of parameters or prediction of variables for each of the (geographical, climatic, etc.) area the given data are clustered by. In this work we formulate a nonparametric one-way model which allows to analyze data with special correlation structure as small area statistics, among other relevant problems like longitudinal and clustered data. To estimate population parameters as the mean function, the classical approach for small area prediction is based on parametric (linear) mixed models. The flexibility of the nonparametric modelling can play an important role in exploring longitudinal/clustered data, and also it offers new perspectives for various problems typically faced in small area statistics. Among the nonparametric approaches the kernel methods as Local Polynomial Smoothers (Fan and Gijbels 1996) are intuitive and simple exhibiting nice theoretical and practical properties which make them very popular in a wide range of statistics problems. The local polynomial methods have been explored in mixed models for the last three decades. Maybe the paper of Lin and Carroll (2000) was that most popularized these methods in longitudinal/clustered data analysis. The generalized estimating equation (GEE) introduces kernel-weights to take into account only observations in a neighborhood which is controlled by the bandwidth or smoothing parameter, and then a parametric model is assumed only locally. Different ways to introduce the kernel-weights provide different estimators exhibiting different theoretical and practical properties. Several of these estimators are presented and explored in this work and also the problem of choosing the smoothing parameter is addressed. Standard

*Based on the working paper "Bootstrapping Nonparametric Mixed Effects Models", a joint work with Wenceslao González Manteiga, María José Lombardía Cortiña and Stefan Sperlich

nonparametric methods without involving the correlation structure are not suitable because they cannot pick up the extra variability. Previous works in this problem focus mainly in cross-validation strategies (Wu and Zhang 2002, Park and Wu 2006, Gu and Ma 2005, Xu and Zhu 2009). We propose to use resampling methods to solve the bandwidth choice problem in the spirit of the previous works of González-Manteiga et al. (2004) and Martínez-Miranda et al. (2008). Bootstrap approximations of the Mean Squared Error provide simple local bandwidth selectors for the nonparametric estimators.

1 Introduction

Parametric (usually linear) mixed effects models and their extensions, generalized parametric mixed effects models become popular statistical modelling approaches for analyzing data with special correlation structure. To consider the usually called "within-subject correlation" allows to deal with longitudinal and clustered data which naturally arise for example in biomedical studies. Also mixed effects models are particular suitable for small-areas estimation (Jiang and Lahiri 2006). In fact these models incorporate area-specific random effects modelling the additional between area variations which cannot be explained by the fixed effect component. Statistical inference with linear mixed effect models has also been studied a lot in the context of panel data analysis (see for example Verbeke and Molenberghs 2000, Diggle et al. 2002). However parametric formulations may not always be desirable because in many situations the dependence on covariates exhibits more complicated manners. In the last years there has been a notable interest in extending parametric models to more flexible nonparametric formulations (see for example Wu and Zhang 2006 for a recent and complete review).

In this paper we are interested in a one-way model with the following nonparametric formulation:

$$y_{ij} = m(\mathbf{x}_{ij}) + v_i(\mathbf{z}_{ij}) + \varepsilon_{ij}, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, q, \quad \sum_{i=1}^{q} n_i = n, \qquad (1.1)$$

with y_{ij} being the observed responses and $\mathbf{x}_{ij}$ $(k \times 1)$ and $\mathbf{z}_{ij}$ $(r \times 1)$ observable covariates. Here $m(\mathbf{x}_{ij})$ represents the fixed effect or population function, $v_i(\mathbf{z}_{ij})$ are the random-effects functions. The residual errors, ε_{ij}, are supposed to be independent with mean 0 and variances $\sigma^2(\mathbf{x}_{ij})$ (assuming a general heterocedastic situation). Also $v_i(\mathbf{z}_{ij})$ and ε_{ij} are independent, where $v_i(\mathbf{z}_{ij})$ can be considered realizations of a mean 0 smooth process with a covariate function $\gamma(\mathbf{z}_{ij_1}, \mathbf{z}_{ij_2}) = \mathrm{E}[v_i(\mathbf{z}_{ij_1})v_i(\mathbf{z}_{ij_2})]$.

The model (1.1) allows to generalize many interesting problems such as longitudinal data, clustered data, nested-error regression models and random regression

coefficient models, among others. More specifically such models arisen in the following way.

Longitudinal data. Consider a longitudinal study involving q subjects, where y_{ij} is taken from subject i in the time point $\mathbf{x}_{ij} = t_{ij}$ $(j = 1, \ldots, n_i)$. Observations from different subjects are independent, while observations from the same subject are naturally correlated. The intra-subject correlation may be modeled by $v_i(\mathbf{z}_{ij}) = v_i(t_{ij})$, these random effects can be interpreted as the subject effects which are usually called "real effects".

Clustered data. Consider observations from q clusters, such as in multicenter studies, where y_{ij} is taken from cluster c_i with covariate $\mathbf{x}_{ij}$ $(j = 1, \ldots, n_i)$. Observations from different clusters are independent, while observations from the same cluster may be correlated to various degrees. The intra-cluster correlation is usually modeled by $v_i(\mathbf{z}_{ij}) = b_{c_i}$. Here the random effects b_{c_i} are considered "latent effects".

The *Nested-error regression models* proposed by Battese et al. (1988) suppose a population divided into q small areas (geographical areas), being n_i the number of sampled units in the area ith. Let y_{ij} be the character of interest (response variable) observed for the jth sampled unit in the ith sample area and $\mathbf{x}_{ij}$ the corresponding values of a covariate (vector of k auxiliary variables).

A more general model in small areas is the *random regression coefficient model* proposed by Dempster et al. (1981). The model is a (unidimensional) linear regression model including a random slope

$$y_{ij} = \beta x_{ij} + b_i x_{ij} + \varepsilon_{ij}.$$

The flexibility of model (1.1) can play an important role in exploring longitudinal/clustered data. Also nonparametric modelling offers new perspectives for various problems typically faced in small area statistics. Evidently, apart from more flexible modelling of either the variances (González-Manteiga et al. 2009) or the mean function (Opsomer et al. 2008), it can be also used for data mining (Lombardía and Sperlich 2008) and specification testing (see Claeskens and Hart 2009).

Motivated from the above described appealing problems which are globally formulated by (1.1), we aim to first introduce appropriate estimators and predictors. In this sense we focus on kernel estimation and first we discuss about how should be desirable to introduce the kernel weights and the correlation structure in the estimating equations. In this paper we propose two different marginal and joint strategies for the estimation are presented. By exploring the statistical properties of the estimators we second deal with practical issues like feasible inference and necessary parameter choices. Specifically the choice of smoothing (or bandwidth)

parameters becomes a technical but actually crucial task. Differently from data mining, nonparametric estimation of densities or marginal impacts in regression, neither intuition nor eye balling will help here. Also the standard nonparametric methods without involving the correlation structure are not suitable because they cannot pick up the extra variability. When it has been conveniently made it usually be considered cross-validation strategies to provide data-driven bandwidth choices (Wu and Zhang 2002, Park and Wu 2006, Gu and Ma 2005, Xu and Zhu 2009, among others). In this paper we propose to use resampling methods to solve the bandwidth choice problem in the spirit of the previous works of González-Manteiga et al. (2004) and Martínez-Miranda et al. (2008). These are based on bootstrap estimations of the mean squared errors, which could be also used to construct confidence intervals. introduced methods. In small areas statistics little efforts have been spent on studying kernel estimation and bandwidth selection under nonparametric models like (1.1).

Along the paper we deal with the estimation of several functions and parameters. We aim to consider relevant problems in the above appealing situations (longitudinal/clustered data, nested models, small areas etc.). In this sense we specifically deal with the estimation of the population function, $m(\mathbf{x})$, the mixed effects or individual functions, $\eta_i(\mathbf{x}, \mathbf{z}) = m(\mathbf{x}) + v_i(\mathbf{z})$, and population parameters such as $\Theta_i = m(\dot{\mathbf{x}}_i) + v_i(\dot{\mathbf{z}}_i)$ with $m(\dot{\mathbf{x}}_i) = \sum_{j=1}^{n_i} m(\mathbf{x}_{ij})/n_i$ and $v_i(\dot{\mathbf{z}}_i) = \sum_{j=1}^{n_i} v_i(\mathbf{z}_{ij})/n_i$, which arise mainly in small area statistics.

References

[1] Battese, G.E., Harter, R.M. and Fuller, W.A. (1988) "An error-component model for prediction of county crop areas using survey and satellite data",*Journal of the American Statistical Association*, 83, 28–36.

[2] Claeskens and Hart (2009) "Goodness-of-fit tests in mixed models", *Test*, 18, 213–239.

[3] Dempster, A.P., Rubin, D.B. and Tsutakawa, R.K. (1981) "Estimation in covariance component models", *Journal of the American Statistical Association*, 76, 341–353.

[4] Diggle, P.J., Heagerty, P., Liang, K.-Y. and Zeger, S.L. (2002) *Analysis of Longitudinal Data*, Second Edition, Oxford: Oxford University Press.

[5] Fan, J. and Gijbels, I. (1996) *Local Polynomial Modelling and its Applications*, London: Chapman and Hall.

[6] González-Manteiga, W., Martínez-Miranda, M.D. and Pérez-González, A. (2004), "The choice of smoothing parameter in nonparametric regression

through wild bootstrap", *Computational Statistics and Data Analysis*, 47, 487–515.

[7] González-Manteiga, W., Lombardía M.J., Molina I., Morales D. and Santamaría L. "Small Area Estimation under Fay–Herriot Models with Nonparametric Estimation of Heteroscedasticity", *Statistical Modelling*, to appear.

[8] Gu, C. and Ma, P. (2005) "Optimal smoothing in nonparametric mixed-effect models", *The Annals of Statistics,* 33, 1357—1379.

[9] Jiang, J. and Lahiri, P. (2006) "Mixed Model Prediction and Small Area Estimation", *Test*, 15, 1–96.

[10] Lin, X. and Carroll, R.J. (2000) "Nonparametric Function Estimation for Clustered Data When Predictor is Measured Without/With Error", *Journal of the American Statistical Association*, 95, 520–534.

[11] Lombardía, M.J. and Sperlich, S. (2008) "Semiparametric inference in generalized mixed effects models", *Journal of the Royal Statistical Society, B*, 70, 913–930.

[12] Martínez-Miranda, M.D., Raya-Miranda, R., González-Manteiga, W. and González-Carmona, A. (2008) A bootstrap local bandwidth selector for additive models, *Journal of Computational and Graphical Statistics*, 17, 38-55.

[13] Opsomer, J., Claeskens, G., Ranalli, M.G., Kauermann, G., and Breidt, F.J. (2005), "Nonparametric Small Area Estimation Using Penalized Spline Regression", *Journal of the Royal Statistical Society, B,* 70, 265—286.

[14] Park, J.G. and Wu, H. (2006) "Backfitting and local likelihood methods for nonparametric mixed-effects models with longitudinal data", *Journal of Statistical Planning and Inference*, 136, 3760–3782.

[15] Verbeke, G. and Molenberghs, G. (2000) *Linear mixed models for longitudinal data*, Springer-Verlag, New York, Inc.

[16] Wu, H. and Zhang, J.T. (2006) *Nonparametric Regression Methods for Longitudinal Data Analysis*, Wiley Series in Probability and Statistics, USA.

[17] Wu, H. and Zhang, J.T. (2002) "Local Polynomial Mixed-Effects Models for Longitudinal Data", *Journal of the American Statistical Association*, 97, 883–897.

[18] Xu, W. and Zhu, L. (2009) "Kernel-based Generalized Cross-validation in Non-parametric Mixed-effect Models", *Scandinavian Journal of Statitics*, in press.

Estimación en áreas pequeñas aplicada a datos del ámbito social y laboral.

Roberto Domínguez Gómez[1], María José Lombardía Cortiña[2], Esther López Vizcaíno[3], Wenceslao González Manteiga[1], and José Manuel Prada Sánchez[1]

[1]Universidade de Santiago de Compostela
[2]Universidade da Coruña
[3]Instituto Galego de Estatística

A Coruña, 25 de Enero de 2010

Resumen: El ingreso medio mensual por hogar y la tasa de desempleo son indicadores de la situación socioeconómica y por lo tanto es de principal interés para la sociedad en general, y en particular para la administración local y regional que necesitan la información para diferentes programas económicos y sociales. La efectividad de estos programas depende del conocimiento de la situación socioeconómica a través de información estadística fiable. En consecuencia, hoy en día los estudios y las investigaciones a nivel regional y local son de gran interés. Un estudio de simulación compara los resultados de la estimación basada en el diseño con la estimación basada en el modelo. Finalmente se aplican los estimadores con datos reales.

Palabras clave: Encuesta de Condiciones de Vida, Enuesta de Población Activa, modelo lineal mixto, bootstrap.