Supplementary materials for this article are available online. Please click the JASA link at http://pubs.amstat.org.

Model Selection by Testing for the Presence of Small-Area Effects, and Application to Area-Level Data

Gauri S. DATTA, Peter HALL, and Abhyuday MANDAL

The models used in small-area inference often involve unobservable random effects. While this can significantly improve the adaptivity and flexibility of a model, it also increases the variability of both point and interval estimators. If we could test for the existence of the random effects, and if the test were to show that they were unlikely to be present, then we would arguably not need to incorporate them into the model, and thus could significantly improve the precision of the methodology. In this article we suggest an approach of this type. We develop simple bootstrap methods for testing for the presence of random effects, applicable well beyond the conventional context of the natural exponential family. If the null hypothesis that the effects are not present is not rejected then our general methodology immediately gives us access to estimators of unknown model parameters and estimators of small-area means. Such estimators can be substantially more effective, for example, because they enjoy much faster convergence rates than their counterparts when the model includes random effects. If the null hypothesis is rejected then the next step is either to make the model more elaborate (our methodology is available quite generally) or to turn to existing random effects models. This article has supplementary material online.

KEY WORDS: Bootstrap; Confidence intervals; Cumulants; Estimating equation; Mean squared error; p-value.

1. INTRODUCTION

Demand for small-area statistics has seen exponential growth in the thirty years since the publication of the landmark paper by Fay and Herriot (1979). Many government agencies in the United States, Canada, Australia, and elsewhere, and many private sector companies, need reliable estimates of small-area characteristics defined by a state, county, or district at a geographic level, or by age, race, and other variables at demographic level. Since direct estimates at the small-area level are usually not reliable, production of reliable but indirect small-area statistics using suitable administrative records and other surveys, in conjunction with the direct estimates, has necessitated a rapid growth of methodological research in small-area estimation. Models have been proposed connecting the direct estimates (or sometimes the data from surveys producing the direct estimates) with auxiliary variables obtained from other surveys or administrative sources. There are two basic types of small-area models: area-level models and unit-level models (see Rao 2003, chapter 5). While the Fay–Herriot model is a popular area-level model, the nested error regression model, first proposed by Battese, Harter, and Fuller (1988), is a popular unit-level model. Reviews of major developments in small-area estimation include those by Ghosh and Rao (1994), Rao (1999, 2003, 2006), Pfeffermann (2002), Jiang and Lahiri (2006), and Datta (2009).

For reasons of privacy, unit-level data are often not available. For this and other reasons, area-level models enjoy popularity in model-based small-area inference. Area-level models usually model the traditional survey estimators, also known as direct estimators, which are typically computed using survey weights. Most authors using unit-level models in small-area estimation ignore survey weights when developing model-based small-area estimates. See, for example, the work of Battese, Harter, and Fuller (1988), Prasad and Rao (1990), Datta and Ghosh (1991), Singh, Stukel, and Pfeffermann (1998), Datta and Lahiri (2000), Hall and Maiti (2006a, 2006b), and Chatterjee, Lahiri, and Li (2008). Important exceptions include Pfeffermann et al. (1998), Prasad and Rao (1999), and You and Rao (2002, 2003). Bell (2008) and Li and Lahiri (2010) addressed cases where the Fay–Herriot model is entirely model dependent.

Model-based small-area estimates have been developed by shrinking the direct estimates towards synthetic estimates which are formed by fitting suitable regression models. An ingenious approach suggested by Ghosh and Maiti (2004) demonstrates that progress can be made in the case of natural exponential families of distributions (Morris 1982, 1983), which encompass the binomial, negative binomial, Poisson, gamma with known shape parameter, hyperbolic secant, and normal distribution with known variance.

However, other distributions seem out of reach. Moreover, particularly when the number of areas, k say, is only modest in size, this approach to inference can be quite restrictive. The estimators generally converge only at rate $k^{-1/2}$, regardless of the number of sampled units, n_i, per area. However, if the role of small-area random effects V_i could be dispensed with, and the small-area means modeled solely in terms of the design vector $\mathbf{x}_i$, then, as we show in this article, the convergence rate would change to $K'^{-1/2}$, where $K' = \sum_i n_i$. In practical problems many of the n_i's are similar in size to k, and so $K'^{-1/2}$ is

Gauri S. Datta is Professor, Department of Statistics, University of Georgia, Athens, GA 30602 (E-mail: *gauri@stat.uga.edu*). Peter Hall is Professor, Department of Mathematics and Statistics, University of Melbourne, Melbourne, Australia (E-mail: *halpstat@ms.unimelb.edu.au*). Abhyuday Mandal is Assistant Professor, Department of Statistics, University of Georgia, Athens, GA 30602 (E-mail: *amandal@stat.uga.edu*). This research was initiated when Datta visited Professor Peter Hall in fall 2008. He is grateful for the financial support and the warm hospitality of the faculty and staff of Department of Mathematics and Statistics, University of Melbourne. Datta's research was partially supported by National Security Agency and National Science Foundation. Hall's research was partially supported by Australian Research Council. Mandal's research was partially supported by National Science Foundation grant DMS-0905731. The University of Georgia Research Computing Center provided the high-performance computing hardware and network infrastructure required for this research. Authors are grateful to Dr. William Bell of U.S. Bureau of the Census for the data used in this article. Datta acknowledges many insightful discussions with Professor J. N. K. Rao in the course of preparing the manuscript. The authors would also like to thank the Associate Editor and the referees for their constructive comments.

© 2011 American Statistical Association
Journal of the American Statistical Association
March 2011, Vol. 106, No. 493, Theory and Methods
DOI: 10.1198/jasa.2011.tm10036

often much smaller than $k^{-1/2}$; it is never larger than $k^{-1/2}$. For example, if each $n_i = n$ then $K' = kn$. In relatively balanced settings such as this, K' is of the same order as $K = k^2/(\sum_i n_i^{-1})$, and in those cases K is generally a more natural way of representing the convergence rate of estimators. For example, the asymptotic variance of estimators in balanced cases is typically equal to a constant multiple of K^{-1}, where the constant does not depend on k or the n_i's.

Therefore, especially when k is relatively small, inference would be aided significantly if we could test for the presence of the small-area effects V_i, and determine whether they are present. In particular, if those effects could be rendered unnecessary through choice of a suitable model, perhaps a model that did not lie in the natural exponential family, then improvements in performance could be achieved. Moreover, insight into the nature of the data could be gained by showing empirically that it is unnecessary to include the extra layer of complexity required to accommodate the V_i's.

These problems are tackled in the present article. We develop simple bootstrap approaches to hypothesis testing, enabling us to determine whether the small-area effects V_i have a detectable impact on the goodness of fit of the model. Our approach to inference, which is of a new type and is based on simple properties of cumulants, allows quite general models to be considered, well beyond the confines of the natural exponential family. If the bootstrap test fails to reject the null hypothesis of no small-area effects then our methodology for general models immediately gives us estimators of unknown model parameters and estimators of small-area means. Should the null hypothesis be rejected, one could either make the model without small-area effects more elaborate (our methodology is available very generally), so as to extract as much information as possible from the design vectors, or turn to existing methods under additional assumptions on the model, for example, those of Ghosh and Maiti (2004), for inference when small-area effects are present. We experimented with alternative tests based on AIC, BIC, and Wald-type statistics, but found that they had relatively little power.

We are sensitive to the fact that, in the majority of cases where small-area data are analyzed outside national statistical offices, the data are available only in summary form, as weighted averages from more extensive surveys. In fact, the weights themselves are often not accessible; only the sums of the first few powers of the weights may be available. Our cumulant-based approach, and our bootstrap methodology, are not inhibited by this restriction, and so our techniques are applicable quite widely.

Importantly, we believe that testing for the presence of random effects is of interest principally because it can lead to more accurate inference, through permitting the random effects component of a model to be dropped. It is of intellectual interest in its own right, but arguably not of great practical value. In fact, testing for significance of random effects is related to model selection in mixed effects models. In this context, consistent model selection procedures are developed in Jiang and Rao (2003), Jiang et al. (2008), and Jiang, Nguyen, and Rao (2010).

There is a broad range of other contexts, outside small area inference, where testing for random effects could be useful and effective. It includes many applications in econometrics, general problems in longitudinal data analysis, and problems addressing censorship, the analysis of panel data, and probit regression modelling. These typically involve purpose-built models. That makes it difficult to develop, in a single article, an attractive and general methodology for testing for random effects, applicable across a large class of problems. Nevertheless, versions of the methodology suggested in this article, and more generally our viewpoint that testing for random effects is a valuable precursor to modeling, could be investigated in a wide variety of settings.

The organization of our article is as follows. In Section 2 we provide a very general model for unit-level data. The cumulant-based approach suggested there is particularly flexible, and for area-level data by using a suitable discrepancy statistic we develop a test for the absence of small-area effects. We conduct the test using the bootstrap, which introduces further flexibility. Applications and simulations are presented in Section 3; theoretical properties, including convergence rates, are established in Section 4; and formulae useful for practical implementation, as well as technical arguments, are outlined in an Appendix A. Further details are provided in Appendix B of the online supplement.

2. MODEL, ESTIMATORS, AND HYPOTHESIS TESTS

2.1 Data and Model

The random data pairs $(\mathbf{X}_i, Y_{ij})$, for $1 \le i \le k$ and $1 \le j \le n_i$, comprise observations at the unit level. Here, $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}}$ is a p-vector and denotes an area-level covariate for the ith small area, and Y_{ij} is a scalar. Although we conduct inference conditional on the values of $\mathbf{X}_1, \ldots, \mathbf{X}_k$, in theoretical work it is convenient to suppose that those random variables are independent and identically distributed. That condition gives us access to properties of large numbers of design variables. We further assume that each pair $(\mathbf{X}_i, Y_{ij})$ is distributed as $(\mathbf{X}_i, Y_i)$, say, and that, conditional on $\mathbf{X}_i$, the variables Y_{ij}, for $1 \le j \le n_i$, are independent and identically distributed as Y_i given $\mathbf{X}_i$. This hypothesis is typically correct in the absence of small-area random effects, denoted by V_i in the previous paragraph, but generally not if those effects are present.

We may not, for reasons of confidentiality, have access to unit-level data, but it is assumed that we observe area-level data in the form of means:

$$\bar{Y}_i = \sum_{j=1}^{n_i} w_{ij} Y_{ij}, \qquad 1 \le i \le k, \tag{2.1}$$

where the w_{ij}s are nonnegative weights and satisfy $\sum_j w_{ij} = 1$. When applying the bootstrap we need to know the individual weights w_{ij}, but in all other aspects of our work, only the sums

$$W_{ir} = \sum_{j=1}^{n_i} w_{ij}^r, \qquad 1 \le r \le q,$$

are required. Here q, satisfying $1 \le q \le p$, denotes the number of moments or cumulants that we use to construct our estimator of the model parameters.

If the small-area effects V_i were present then the distribution of Y_{ij}, given that $\mathbf{X}_i = \mathbf{x}_i$ and $V_i = v_i$, would depend on $\mathbf{x}_i$, v_i

and a vector of parameters $\boldsymbol{\xi}$, say, the latter not depending on i. By arguing that the V_is can be eliminated, and testing this hypothesis, we are in effect saying that V_i can be replaced by a fixed number v, which we then adjoin to the parameter vector $\boldsymbol{\xi}$ by using $\boldsymbol{\eta}$ to denote the concatenation of v and $\boldsymbol{\xi}$. A simple way in which this can come about is through a regression model, $Y_{ij} = \boldsymbol{\beta}^{\mathrm{T}} Z_i + V_i + \epsilon_{ij}$, where $\boldsymbol{\beta}$ is an unknown vector, not including the intercept which is accommodated via V_i, and Z_i is a design vector. We replace V_i by v, adjoin v to $\boldsymbol{\xi}$ to obtain the p-vector $\boldsymbol{\eta}$, and adjoin the number 1 to the components of each Z_i, thereby obtaining $\mathbf{X}_i$. Of course, there are many other ways in which V_i can appear in a model, and be eliminated through replacing it by a constant; our aim is to construct an empirical test of the appropriateness of this narrowing of the model.

In the absence of small-area effects the data pairs $(\mathbf{X}_i, Y_{ij})$ are assumed to be generated by a model for which

$$\kappa_r(Y|\mathbf{X} = \mathbf{x}_i) = \lambda_r(\mathbf{x}_i, \boldsymbol{\eta}), \tag{2.2}$$

where $(\mathbf{X}, Y)$ denotes a generic value of $(\mathbf{X}_i, Y_i)$, $\kappa_r(Y|\mathbf{X} = \mathbf{x}_i)$ is the rth cumulant of the distribution of Y, conditional on $X = \mathbf{x}_i$, and $\lambda_r(\mathbf{x}_i, \boldsymbol{\eta})$ is a known function of $\mathbf{x}_i$ and the unknown p-vector $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_p)^{\mathrm{T}}$. Note too that if $\bar{Y}_i$ is given by (2.1) then, in view of the the conditional independence and identical distribution of the Y_{ij}s given $\mathbf{X}_i$, and a standard property of cumulants,

$$\kappa_r(\bar{Y}_i|\mathbf{X}_i = \mathbf{x}_i) = W_{ir}\kappa_r(Y|\mathbf{X}_i = \mathbf{x}_i) = W_{ir}\lambda_r(\mathbf{x}_i, \boldsymbol{\eta}). \tag{2.3}$$

2.2 Properties of Cumulants

For a general random variable U we can write

$$E(U^r) = \sum_{\ell=1}^{r} \sum_{s_1,\ldots,s_\ell} \nu_{s_1,\ldots,s_\ell}(r)\kappa_{s_1}(U)\cdots\kappa_{s_\ell}(U), \tag{2.4}$$

where $\kappa_s(U)$ denotes the sth cumulant of the distribution of U, the coefficients $\nu_{s_1,\ldots,s_\ell}(r)$ are known integers, and the sums are over ℓ and $s_1, \ldots, s_\ell$ are such that each $s_j \geq 1$ and $s_1 + \cdots + s_\ell = r$. In particular, $\nu_r(r) = \nu_{1,1,\ldots,1}(r) = 1$ for all r, and special cases of (2.4) are $E(U) = \kappa_1(U)$, $E(U^2) = \kappa_2(U) + \kappa_1(U)^2$ and $E(U^3) = \kappa_3(U) + 3\kappa_1(U)\kappa_2(U) + \kappa_1(U)^3$. In the absence of small-area effects, on combining (2.2)–(2.4) we see that we can write

$$E(\bar{Y}_i^r|\mathbf{X}_i = \mathbf{x}_i) = \sum_{\ell=1}^{r} \sum_{s_1,\ldots,s_\ell} \nu_{s_1,\ldots,s_\ell}(r) \prod_{t=1}^{\ell} \bigl\{W_{is_t}\lambda_{s_t}(\mathbf{x}_i, \boldsymbol{\eta})\bigr\}. \tag{2.5}$$

Similar properties hold for centred moments. In particular, replacing $\bar{Y}_i^r$ by $\{\bar{Y}_i - \lambda_1(\mathbf{X}_i, \boldsymbol{\eta})\}^r$ on the left-hand side of (2.5) we obtain instead the relation

$$E\bigl[\{\bar{Y}_i - \lambda_1(\mathbf{X}_i, \boldsymbol{\eta})\}^r|\mathbf{X}_i\bigr] = \sum_{\ell=1}^{r} \sum_{s_1,\ldots,s_\ell} \check{\nu}_{s_1,\ldots,s_\ell}(r) \prod_{t=1}^{\ell} \bigl\{W_{is_t}\lambda_{s_t}(\mathbf{X}_i, \boldsymbol{\eta})\bigr\}, \tag{2.6}$$

where $r \geq 2$, the sums are over ℓ and $s_1, \ldots, s_\ell$ such that each $s_j \geq 2$ and $s_1 + \cdots + s_\ell = r$, and the quantities $\check{\nu}_{s_1,\ldots,s_\ell}(r)$ are known integers.

One of the attractions of (2.6) over (2.5) is the greater simplicity of the right-hand side. For example, the versions of (2.6) when $r = 2$, 3, and 4 are

$$E\bigl[\{\bar{Y}_i - \lambda_1(\mathbf{X}_i, \boldsymbol{\eta})\}^2|\mathbf{X}_i\bigr] = W_{i2}\lambda_2(\mathbf{X}_i, \boldsymbol{\eta}), \tag{2.7}$$

$$E\bigl[\{\bar{Y}_i - \lambda_1(\mathbf{X}_i, \boldsymbol{\eta})\}^3|\mathbf{X}_i\bigr] = W_{i3}\lambda_3(\mathbf{X}_i, \boldsymbol{\eta}), \tag{2.8}$$

$$E\bigl[\{\bar{Y}_i - \lambda_1(\mathbf{X}_i, \boldsymbol{\eta})\}^4|\mathbf{X}_i\bigr] = W_{i4}\lambda_4(\mathbf{X}_i, \boldsymbol{\eta}) + 3\{W_{i2}\lambda_2(\mathbf{X}_i, \boldsymbol{\eta})\}^2, \tag{2.9}$$

respectively. The analogous expressions for the uncentred mean $E(\bar{Y}_i^r|\mathbf{X}_i)$ contain two, three, and five terms, respectively, on the right-hand side. On the other hand, the left-hand side of (2.5) does not involve $\boldsymbol{\eta}$, which can simplify inference.

2.3 Methods for Estimating $\boldsymbol{\eta}$

The relation (2.5) motivates us to estimate $\boldsymbol{\eta}$ by, essentially, equating the right-hand side of (2.5) to $\bar{Y}_i^r$. For example, we could solve the following p equations for the p unknowns $\eta_1, \ldots, \eta_p$:

$$\sum_{i=1}^{k} \bar{Y}_i^r = \sum_{i=1}^{k} \sum_{\ell=1}^{r} \sum_{s_1,\ldots,s_\ell} \nu_{s_1,\ldots,s_\ell}(r) \prod_{t=1}^{\ell} \bigl\{W_{is_t}\lambda_{s_t}(\mathbf{X}_i, \boldsymbol{\eta})\bigr\}, \quad 1 \leq r \leq p. \tag{2.10}$$

More generally, defining

$$\begin{aligned} Q_{ir}(\boldsymbol{\eta}) &= \sum_{\ell=1}^{r} \sum_{s_1,\ldots,s_\ell} \nu_{s_1,\ldots,s_\ell}(r) \prod_{t=1}^{\ell} \bigl\{W_{is_t}\lambda_{s_t}(\mathbf{X}_i, \boldsymbol{\eta})\bigr\}, \\ R_{ir} &= \bar{Y}_i^r - Q_{ir}, \end{aligned} \tag{2.11}$$

we can form the q-vector $\mathbf{R}_i = (R_{i1}, \ldots, R_{iq})^{\mathrm{T}}$, multiply $\mathbf{R}_i$ on the left by an appropriate $p \times q$ matrix $\mathbf{A}_i(\mathbf{X}_i, \boldsymbol{\eta})$ which is a function of $\mathbf{X}_i$ and $\boldsymbol{\eta}$ alone, sum over i, and equate to zero, to obtain an estimator $\hat{\boldsymbol{\eta}}$ of $\boldsymbol{\eta}$. That is, $\hat{\boldsymbol{\eta}}$ solves

$$\sum_{i=1}^{k} \mathbf{A}_i(\mathbf{X}_i, \boldsymbol{\eta})\mathbf{R}_i(\boldsymbol{\eta}) = \mathbf{0}. \tag{2.12}$$

Analogously to the definitions of $Q_{ir}(\boldsymbol{\eta})$ and R_{ir} at (2.11), put

$$\begin{aligned} \check{Q}_{ir}(\boldsymbol{\eta}) &= \sum_{\ell=1}^{r} \sum_{s_1,\ldots,s_\ell} \check{\nu}_{s_1,\ldots,s_\ell}(r) \prod_{t=1}^{\ell} \bigl\{W_{is_t}\lambda_{s_t}(\mathbf{X}_i, \boldsymbol{\eta})\bigr\}, \\ \check{R}_{ir} &= \{\bar{Y}_i - \lambda_1(\mathbf{X}_i, \boldsymbol{\eta})\}^r - \check{Q}_{ir}. \end{aligned} \tag{2.13}$$

Define the q-vector $\check{\mathbf{R}}_i = (\check{R}_{i1}, \ldots, \check{R}_{iq})^{\mathrm{T}}$, and, as at (2.12), let $\mathbf{A}_i(\mathbf{X}_i, \boldsymbol{\eta})$ denote a $p \times q$ matrix which is a function of $\mathbf{X}_i$ and $\boldsymbol{\eta}$ alone. We replace (2.12) by

$$\sum_{i=1}^{k} \mathbf{A}_i(\mathbf{X}_i, \boldsymbol{\eta})\check{\mathbf{R}}_i(\boldsymbol{\eta}) = \mathbf{0}, \tag{2.14}$$

which is a system of p equations in the p unknowns $\eta_1, \ldots, \eta_p$. The estimator $\hat{\boldsymbol{\eta}}$ is now taken to be the solution of (2.14) rather than (2.12).

In the contexts of (2.12) and (2.14), specific versions of $\mathbf{A}_i$ are motivated by the optimal estimating function ideas of Godambe and Thompson (1989). In particular, in the case of

(2.14) we can define the p-vector $\check{\mathbf{D}}_{ir}(\boldsymbol{\eta}) = -E\{\partial \check{R}_{ir}(\boldsymbol{\eta})/\partial\boldsymbol{\eta}|\mathcal{X}\}$, the $p \times q$ matrix $\check{\mathbf{D}}_i = (\check{\mathbf{D}}_{i1}, \ldots, \check{\mathbf{D}}_{iq})$ and the $q \times q$ matrix $\boldsymbol{\Sigma}_i = \mathrm{var}(\check{R}_i|\mathcal{X})$, where $\mathcal{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_k\}$. Let

$$\mathbf{A}_i(\mathbf{X}_i, \boldsymbol{\eta}) = \check{\mathbf{D}}_i(\mathbf{X}_i, \boldsymbol{\eta})\boldsymbol{\Sigma}_i(\mathbf{X}_i, \boldsymbol{\eta})^{-1}. \tag{2.15}$$

To assist with calculation we note that

$$(\boldsymbol{\Sigma}_i)_{rs} = \check{Q}_{i,r+s} - \check{Q}_{ir}\check{Q}_{is}. \tag{2.16}$$

Our definitions here have been informed by discussion in Ghosh and Maiti (2004, pp. 98, 99). In a longer version of this article it is shown that, if $\mathbf{A}_i$ is defined by (2.15), then in cases of practical importance the components of $n_i^{-1}\mathbf{A}_i$ are bounded.

The approaches based on (2.12) and (2.14) are equivalent, in the sense that each includes the other, albeit with quite different versions of $\mathbf{A}_i$. The reason for working with both of them is that they motivate different ways of thinking about the problem. In particular, (2.14) is more natural than (2.12) in the setting treated by Godambe and Thompson (1989).

2.4 Bootstrap Methods

To implement the bootstrap we must know the values of the weights w_{ij}. Importantly, we do not need the values of Y_{ij}. First we develop bootstrap tests for the presence of the small-area effects V_i. Indeed, if we have good explanatory information via the $\mathbf{X}_i$s then we may not need a small-area effect in our model. Since the only data available are the weighted survey averages $\bar{Y}_i$ and the design variables $\mathbf{X}_i$, then the test statistic, T below, is founded on the differences of the $\bar{Y}_i$s from their means, under the null hypothesis that small-area effects are not present:

$$T = \sum_{i=1}^{k} \frac{\{\bar{Y}_i - \lambda_1(\mathbf{X}_i, \hat{\boldsymbol{\eta}})\}^2}{W_{i2}\lambda_2(\mathbf{X}_i, \hat{\boldsymbol{\eta}})}.$$

The denominator on the right-hand side takes account of the respective variances of the differences $\bar{Y}_i - \lambda_1(\mathbf{X}_i, \hat{\boldsymbol{\eta}})$. We reject H_0 if T is too large, and we use the parametric bootstrap to calibrate the test, simulating under the null hypothesis when the "true" parameter value is replaced by $\hat{\boldsymbol{\eta}}$. Specifically, we draw resamples $(\mathbf{X}_i, Y_{ij}^*)$, for $1 \le j \le n_i$ and $1 \le i \le k$, from the fitted distribution with parameter $\hat{\boldsymbol{\eta}}$, and we compute the versions $\hat{\boldsymbol{\eta}}^*$ and T^*, of $\hat{\boldsymbol{\eta}}$ and T, respectively, for these resampled data. In particular, $T^* = \sum_i \{\bar{Y}_i^* - \lambda_1(\mathbf{X}_i, \hat{\boldsymbol{\eta}}^*)\}^2/\{W_{i2}\lambda_2(\mathbf{X}_i, \hat{\boldsymbol{\eta}}^*)\}$. Since inference is based on conditioning on the design vectors $\mathbf{X}_i$ then we keep those values fixed in the simulation.

In the context of hypothesis testing we can use either the parametric bootstrap or the "structural" bootstrap, the latter referring to cases where the mean, but not the error, is modeled parametrically. In both approaches we fit a model, for example as discussed in the paragraph containing (2.2) and (2.3). In the fully parametric case we also fit a specified distribution for the error (e.g., normally distributed). When using the structural bootstrap we capture the error distribution by resampling from a set of residuals. These approaches are relatively standard, and so we discuss them further only in our numerical work; see particularly Section 3.3.

Using repeated simulation, and given a nominal significance level α for the test, we compute the nearest solution, $t = \hat{t}_\alpha$ say, of the equation $P(T^* > t|\mathcal{X}) = \alpha$, where $\mathcal{X}$ denotes the set of all data $(\mathbf{X}_i, Y_{ij})$ for $1 \le j \le n_i$ and $1 \le i \le k$. The calculations here are based on repeated numerical simulation, which can be thought of as a computational device for capturing the distribution of T^* conditional on the actual data. The test consists of rejecting H_0 at the α level if $T > \hat{t}_\alpha$. Standard double bootstrap methods can be used to improve the test's level accuracy.

If k and the value of α that would lead to rejection of H_0 are not too small then we can be reasonably confident that small-area effects are not present in proportions that would require them to be included in the fitted model. Standard parametric bootstrap methods, based on the estimator $\hat{\boldsymbol{\eta}}$, can then be used to construct confidence intervals for $\mu(\mathbf{X}_i) = E(Y_{ij}|\mathbf{X}_i) = \lambda_1(\mathbf{X}_i, \boldsymbol{\eta}^0)$, for any given value of i, and also to compute estimates of the mean squared error of either $\bar{Y}_i$ or $\lambda_1(\mathbf{X}_i, \hat{\boldsymbol{\eta}})$ as an estimator of $\mu(\mathbf{X}_i)$. Here $\boldsymbol{\eta}^0$ is the true unknown parameter. The conventional double bootstrap can be used to improve coverage accuracy of the confidence interval and to reduce bias of the mean squared error estimator.

Depending on the outcome of the test we either use the synthetic estimator or we use the standard small area estimator. In particular, if the null hypothesis of no random effects is not rejected then we use the standard regression synthetic estimator for inference.

3. APPLICATIONS AND SIMULATION RESULTS

3.1 Introduction

In Section 3 we consider two applications of our methodology to estimating small-area means. Subsection 3.2 treats estimation of proportions of individuals in different income segments for various states of the United States, based on data obtained from the Current Population Survey (CPS) and other administrative programs in the United States, and Section 3.3 considers an alternative to the Fay–Herriot model for developing small-area estimates to compare 23 hospitals in terms of care (in this example, successful kidney transplants) that they provide to their patients.

3.2 Small Area Estimation of Poverty Statistics

First we consider estimating the proportion, p_i, of individuals with income below low-income level for each of the 13 western states of the United States. The 13 states are placed together in this category by the Current Population Survey. This is a relatively homogeneous portion of the nation, and comprises more than a quarter of the U.S. population. Data on all 50 states and Washington DC are available, and that larger dataset will be analyzed later in this section in a sampling experiment. If we use all 50 states and DC, a goodness-of-fit test indicates a lack of fit of the regression, suggesting the presence of an unexplained error term. However, when considering only 13 western states the proposed model fits well to the data, and the test statistic does not lead to rejection.

Response variables were drawn from the Annual Social and Economics (ASEC) Supplement of CPS data collected in 2004 based on income year 2003 for the $k = 13$ states: Washington, Oregon, Idaho, Montana, Wyoming, Colorado, Utah, Nevada, California, Arizona, New Mexico, Alaska, and Hawaii. The CPS sample sizes (values of n_i) for these states are all in the thousands. Based on the CPS data the U.S. Census Bureau has considered estimating poverty ratios for different segments of

Table 1. CPS data and estimation of poverty statistics

								Binomial			Trinomial			
								CI			CI			
i	$n_i\bar{y}_{1i}$	$n_i\bar{y}_{2i}$	n_i	x_1	x_2	x_3	x_4	Left	Right	$\widehat{\text{var}}(\hat{p}_i)$ $\times 10^{-5}$	Left	Right	$\widehat{\text{var}}(\hat{p}_{1i})$ $\times 10^{-5}$	VR
1	215	94	1283	0.616	13.43	19.52	9.22	0.1429	0.1763	4.36	0.1422	0.1750	4.35	0.40
2	219	83	1882	−0.088	11.60	13.30	5.16	0.1145	0.1310	1.74	0.1143	0.1321	1.74	0.32
3	140	58	1031	0.293	14.86	11.37	6.52	0.1371	0.1615	2.43	0.1375	0.1610	2.44	0.23
4	178	63	935	0.481	17.13	10.60	8.30	0.1654	0.2073	9.42	0.1646	0.2085	9.41	0.57
5	236	94	1738	0.415	11.21	15.29	5.23	0.1209	0.1425	2.38	0.1189	0.1407	2.38	0.35
6	222	74	1130	−0.194	19.20	14.86	11.37	0.1759	0.2214	13.27	0.1753	0.2220	13.27	0.95
7	136	45	1111	0.634	13.24	9.48	5.05	0.1092	0.1289	2.70	0.1099	0.1299	2.70	0.28
8	147	62	1195	0.459	13.34	7.89	5.10	0.1041	0.1267	3.59	0.1049	0.1297	3.59	0.40
9	162	58	1266	−0.452	10.99	10.78	6.72	0.1190	0.1413	2.98	0.1197	0.1418	2.98	0.34
10	981	387	6422	0.410	13.63	16.32	5.06	0.1320	0.1736	1.72	0.1334	0.1722	1.72	0.85
11	168	56	1332	−1.825	12.60	11.78	7.51	0.1197	0.1358	1.62	0.1189	0.1360	1.62	0.20
12	200	71	1364	0.318	12.86	15.76	11.63	0.1349	0.1683	6.01	0.1362	0.1687	6.01	0.66
13	221	75	1566	0.870	11.11	12.08	7.15	0.1273	0.1495	3.52	0.1267	0.1501	3.52	0.45

the population (such as 0–4 year old children, 5–17 year old children, 18–64 year old adults, etc.) in each state. For this purpose the Census Bureau identified a set of covariates which are available based on past census records (one variable, called the residual, denoted by x_1), and IRS tax returns data (two variables, called the IRS poverty ratio, denoted by x_2, and the percentage of nonfilers, denoted by x_3) and food stamps data (x_4). Explicit values of the x_is are given in Table 1.

To develop our estimates of small-area proportions we consider regression models based on some of these covariates. In estimating the proportions p_i in our case we used area-level data given in terms of state sample proportions $\bar{y}_i$, and we first considered if the random small-area effect terms V_i could be dropped from the logistic mixed model

$$\text{logit}(p_i) = \eta_0 + \eta_1 x_{i1} + \eta_2 x_{i2} + \eta_3 x_{i3} + \eta_4 x_{i4} + V_i, \quad (3.1)$$

where $V_i \sim \text{N}(0, \sigma_v^2)$. The absence of small-area effects corresponds to $\sigma_v^2 = 0$. Based on a total sample size of 22,255 individuals from these 13 states, the test statistic T for testing the absence of small-area effects results in a value of $T = 12.1689$. Based on 1000 bootstrap simulations, the corresponding p-value is 0.205. In particular, the null hypothesis was not rejected at level $\alpha = 0.20$. Values of n_i range from 1.3×10^6 to 6.4×10^6; see Table 1.

On the other hand, after exploring a more complex logistic mixed model given by

$$\text{logit}(p_i) = \eta_0 + \eta_1 x_{i2} + \eta_2 x_{i3} + \eta_3 x_{i4} + \eta_4 x_{i2}^2 + \eta_5 x_{i2} x_{i4} + V_i, \quad (3.2)$$

we obtain an even smaller value of the statistic T for testing the absence of the small-area effects. The corresponding p-value is only 0.617, and so the evidence in favor of no small-area effects is quite convincing for the new model. We used this model to compute estimators of the parameters $\boldsymbol{\eta}$, with no V_i, and from those quantities we computed estimators of $\text{logit}\, p_i$, which we inverted to construct estimators $\hat{p}_i$ of the small-area proportions p_i. The estimated mean squared errors (MSEs) of point estimates and confidence intervals were obtained using formulae in Appendix A.1. The estimated variance associated with the direct estimate $\bar{y}_i$ is given by $\bar{y}_i(1 - \bar{y}_i)/n_i \equiv \psi_i$. Model-based estimates of the p_i's, based on a logistic regression model, have lower estimated MSE than ψ_i, and the ratio, VR, of the former to the latter varies between 0.20 and 0.95 over the small areas. This ratio is given in the last column of Table 1. The table also gives endpoints of confidence intervals.

Small-area estimation of the proportions based on a logistic mixed linear model has been considered earlier by MacGibbon and Tomberlin (1989), Farrell, MacGibbon, and Tomberlin (1997), and Jiang and Lahiri (2001). In a recent article Pfeffermann and Correa (2009) considered estimation of a small-area proportions by the empirical best predictor and the associated MSE of the predictor.

The ASEC Supplement of the CPS classifies an individual into one of several income groups: below low-income level, 100%–125% of low-income level, 125%–150% of low-income level, and above the 150% level. We consider joint estimation of the proportion of individuals below low-income level (p_{1i}), and the proportion in the 125%–150% band (p_{2i}), for each of the 13 western USA states. (In the univariate study presented above we denoted p_{1i} by p_i.) Ghosh and Maiti (2004), who suggested a method based on area-level summary data, modeled only the univariate case, but our approach permits us to undertake multivariate estimation of the proportion vectors $(p_{1i}, p_{2i})^{\text{T}}$ based on the area-level summary data $\bar{y}_{ji}$, denoting sample proportion in the jth category in the ith state for $j = 1, 2$ and $i = 1, \ldots, 13$.

In particular, we consider the multivariate logistic mixed model with small-area effects

$$\text{logit}(p_{ji}) = \eta_{j0} + \eta_{j1} x_{i2} + \eta_{j2} x_{i3} + \eta_{j3} x_{i4} + \eta_{j4} x_{i2}^2 + \eta_{j5} x_{i2} x_{i4} + V_{ji}, \qquad j = 1, 2, \quad (3.3)$$

where $\mathbf{V}_i = (V_{1i}, V_{2i})^{\text{T}} \sim \text{N}_2(\mathbf{0}, \boldsymbol{\Sigma}_v)$ for $i = 1, \ldots, 13$. We test for the absence of small-area effects by testing the null hypothesis that $\boldsymbol{\Sigma}_v$ is a null matrix. The test statistic T corresponds to Pearson's deviance statistic for multivariate logistic regression model. We obtained $T = 12.26$. Based on 1000 bootstrap simulations, and for this value of T the null hypothesis was not rejected even at the level $\alpha = 0.40$. Computation of the

Table 2. Lengths, noncoverage probabilities of the 95% CI's along with eMSE, eRB, and estimated Bias (averaged over 50 states and DC for p's)

	Lengths		Noncoverage probabilities								
			Bootstrap			Normal					
	Bootstrap	Normal	Left	Right	Total	Left	Right	Total	eMSE	eRB	Bias
p	0.1095	0.1040	0.0226	0.0216	0.0442	0.0137	0.0487	0.0624	0.0008	−0.0184	0.0001
η_0	2.1846	2.1877	0.0281	0.0266	0.0547	0.0285	0.0252	0.0537	0.3227	−0.0331	−0.0036
η_1	0.5864	0.5870	0.0258	0.0265	0.0523	0.0246	0.0284	0.0530	0.0229	−0.0194	−0.0049
η_2	0.2114	0.2116	0.0268	0.0265	0.0533	0.0267	0.0257	0.0524	0.0030	−0.0404	0.0000
η_3	0.1499	0.1504	0.0273	0.0272	0.0545	0.0256	0.0246	0.0502	0.0015	−0.0261	−0.0006
η_4	0.1843	0.1847	0.0246	0.0244	0.0490	0.0236	0.0235	0.0471	0.0023	−0.0198	−0.0008

test statistic or the p-value does not require the normality assumption for V_i. As in the univariate case, we use this model to obtain point estimates of small area proportions corresponding to the 13 states. The estimated mean squared errors (MSEs) of the point estimates and the confidence intervals were obtained using results presented in Appendix A.2.

Next we summarize results for data from the model at (3.1), fitted without the small-area effect term V_i to the 87,949 data points from all 50 states and DC. We used the estimated p_i's to generate a new dataset of the same size. For these new data, model (3.1) without the small-area effects is obviously a particularly good fit. We used the new dataset to evaluate properties of our proposed estimators in cases where we know the "true" p_i's. For this purpose we randomly selected $a\%$ of the families from each of the 50 states and DC, to create a "working sample," and applied our methods. We took $a = 1$ and $a = 5$. The results were similar for both of these a's, and hence, for the sake of brevity, we present results only for $a = 1$. We used the same weight for each unit within a small area. For each sample we tested the null hypothesis that there are no small-area effects. As our interest lies in applying the model in cases where it is appropriate, it is reasonable to not study the model in cases where it does not seem to be a good fit. Thus, if the null hypothesis was rejected then we discarded that sample. This process was repeated until we obtained 10,000 working samples.

In Table 2 we present the estimated values of the MSEs, biases, and related quantities of interest. In particular, for each sample we calculated 90% and 95% confidence intervals for the p_i's and η_j's using 1000 bootstrap samples each time, and we report in the table the noncoverage probabilities of the 95% intervals. We also compare our bootstrap confidence intervals with standard confidence intervals obtained from normal approximation. (Both intervals are explained in Appendix A.1.) The results for 90% confidence intervals are very similar.

In the following, θ denotes either p_i or η_j. We define

$$\text{eMSE}(\hat{\theta}) = \frac{1}{s}\sum_{u=1}^{s}(\hat{\theta}_u - \theta)^2, \qquad \overline{\hat{V}}(\hat{\theta}) = \frac{1}{s}\sum_{u=1}^{s}\hat{V}_u(\hat{\theta}),$$

$$\text{eRB}\{\hat{V}(\hat{\theta})\} = \frac{\overline{\hat{V}}(\hat{\theta}) - \text{eMSE}(\hat{\theta})}{\text{eMSE}(\hat{\theta})},$$

where $\hat{V}_u(\hat{\theta})$, the estimated variance of $\hat{\theta}$ for the uth simulation, was obtained using a formula given in Appendix A.1. Here $\text{eMSE}(\hat{\theta})$ and $\overline{\hat{V}}(\hat{\theta})$ are respectively the empirical and model-based estimators of $\text{MSE}(\hat{\theta})$, and $\text{eRB}\{\hat{V}(\hat{\theta})\}$ is a measure of the relative bias of $\hat{V}(\hat{\theta})$ as an estimator of $\text{MSE}(\hat{\theta})$.

It is clear from the first panel of Table 2 that the lengths of the confidence intervals obtained from the normal approximation and from the bootstrap approximation are similar. However, from the middle panel of Table 2, and Figure 1, it is clear that the normal CIs perform much worse than the bootstrap CIs in terms of noncoverage probabilities. Table 2 and Figure 1 show

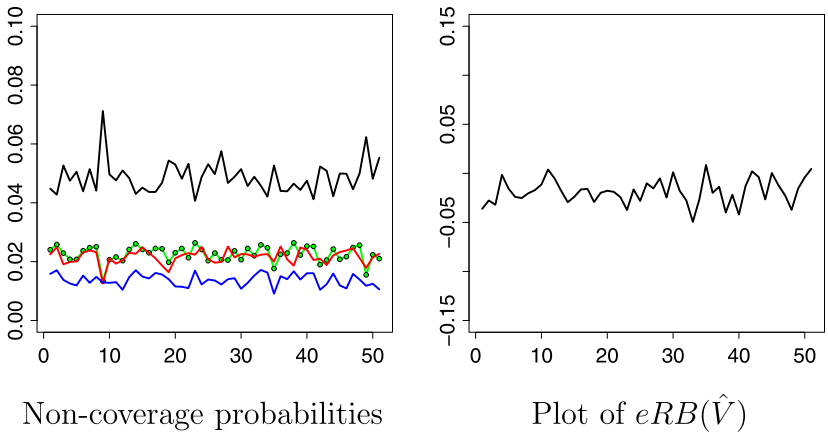

Figure 1. Noncoverage probabilities and $\text{eRB}(\hat{V})$ for 50 states and DC. The online version of this figure is in color.

empirical relative biases of the estimates of the MSE of $\hat{p}_i$ small (not greater than 5% in absolute terms). In the left panel of Figure 1, the top line represents the right noncoverage probabilities for the normal approximation and the bottom one represents the left noncoverage probabilities for the same. The two middle lines represents the left and right noncoverage probabilities for the proposed bootstrap method, where the beaded and solid lines correspond to the left and right noncoverage probabilities, respectively.

Our estimates and estimated MSEs are valid under the null hypothesis. To check performance of our estimators when there is a random effect, we simulated the data with $A = \bar{D} = 0.06$ and performed testing using χ^2 cutoffs at level $\alpha = 0.2$. If the null hypothesis $H_0 : A = 0$ was not rejected, we calculated the MSE of our estimates and compared it with the "true" MSE, calculated by simulating in the case where the random effects are known to be present. In this case the "true" MSE was the average of the squared difference of the estimator and the parameter $\theta_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + v_i$. Our estimated MSE was smaller than the "true" MSE, as expected. This result reflects the high power of our test procedure; if indeed there is a need for a random effect, our test indicates that fact with high probability.

3.3 An Alternative to the Fay–Herriot Model

Jiang and Tang (2011) analyzed data reported by Morris and Christiansen (1995) from a medical survey comparing surgery success rates of several hospitals. Data involve 23 hospitals (out of a total of 219 hospitals) where each hospital performed at least 50 kidney transplants during a 27 month period (see Table 3). They reported failure rates, that is, values of $\hat{p}_i$, for kidney transplant operations, where $\hat{p}_i =$ (number of graft failures)$/n_i$ and n_i was the number of kidney transplants at hospital i during the period. The variance of $\hat{p}_i$, denoted by D_i^*, is approximated by $(0.2)(0.8)/n_i$, where 0.2 is the observed failure rate combining all hospitals. A covariate in the form of severity index x_i is also reported for each hospital; it equals the average fraction of females, African Americans, children and extremely ill kidney recipients at hospital i. Here the hospitals represent the small areas. The goal is to compare hospitals based on the true graft failure rates p_i, or equivalently, in terms of logit(p_i).

Since the data are proportions then a logit transformation might offer an appropriate way of implementing the Fay and Herriot (1979) model in this case. Note that, since individual n_i's are relatively large, variances of the logit transformed data can be approximated by the delta method. In particular, if $E(\hat{p}_i) = p_i$ and $\mathrm{var}(\hat{p}_i) = D_i^*$, then $\mathrm{var}\{\mathrm{logit}(\hat{p}_i)\} \approx D_i^*/\{p_i(1-p_i)\}^2$. We denote logit$(\hat{p}_i)$ by y_i and $D_i^*/\{p_i(1-p_i)\}^2$ by D_i. For this example, Jiang, Nguyen, and Rao (2010) proposed the following Fay–Herriot model:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + V_i + \epsilon_i,$$

with $i = 1, \ldots, k = 23$, where β_j, for $j = 0, 1, 2, 3$, are are unknown coefficients and everything else is as in a Fay–Herriot model. The V_i's are usually assumed to be independent, identically normally distributed random variables. However, our goal is to test the null hypothesis of whether the random effects V_i are necessary. For that purpose the normality assumption is not required. Under this hypothesis our model is

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + e_i, \qquad i = 1, \ldots, k, \quad (3.4)$$

where $e_i \sim \mathrm{N}(0, D_i)$.

Table 3. Hospital data of Morris and Christiansen (1995)

Area	y_i	x_i	D_i	MSE_i^S	$\mathrm{MSE}_i^{\mathrm{PR}}$
1	−0.838	0.112	0.118	0.0084	0.0315
2	−1.815	0.206	0.110	0.0070	0.0299
3	−1.368	0.104	0.106	0.0073	0.0305
4	−0.695	0.168	0.106	0.0047	0.0279
5	−0.632	0.337	0.086	0.0857	0.0947
6	−1.289	0.169	0.083	0.0046	0.0280
7	−1.688	0.211	0.083	0.0085	0.0313
8	−1.791	0.195	0.083	0.0047	0.0280
9	−1.266	0.221	0.076	0.0122	0.0344
10	−1.355	0.077	0.076	0.0086	0.0324
11	−1.331	0.195	0.069	0.0047	0.0279
12	−1.015	0.185	0.066	0.0039	0.0272
13	−1.153	0.202	0.066	0.0060	0.0289
14	−1.036	0.108	0.051	0.0079	0.0299
15	−1.782	0.204	0.051	0.0065	0.0287
16	−2.031	0.072	0.048	0.0117	0.0334
17	−1.380	0.142	0.043	0.0083	0.0294
18	−1.313	0.136	0.040	0.0088	0.0293
19	−1.457	0.172	0.038	0.0043	0.0261
20	−1.313	0.202	0.033	0.0060	0.0262
21	−1.614	0.087	0.033	0.0061	0.0266
22	−1.565	0.177	0.028	0.0040	0.0240
23	−1.621	0.072	0.024	0.0117	0.0268

Let $\mathbf{D} = \text{diag}\{D_1, \ldots, D_k\}$, $\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{D}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{D}^{-1}\mathbf{y}$, and

$$T = \sum_i \frac{(y_i - \mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}})^2}{D_i} = \mathbf{y}^{\mathrm{T}}[\mathbf{D}^{-1} - \mathbf{D}^{-1}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{D}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{D}^{-1}]\mathbf{y}. \quad (3.5)$$

Here $\mathbf{x}_i^{\mathrm{T}}$, the ith row of $\mathbf{X}$, is given by $(1, x_i, x_i^2, x_i^3)$. Under the model in (3.4), T will be distributed as χ^2 with $23 - 4 = 19$ degrees of freedom. For these data, $T_{\text{obs}} = 23.66$ and the corresponding p-value is 21.0%. The p-value is large enough to justify excluding the random effects terms V_i.

We dispensed with the normality assumption on the sampling errors and used the bootstrap to calculate the p-value of the test statistic. In particular we defined $\hat{\hat{e}}_i = D_i^{-1/2}(y_i - \mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}})/\sqrt{T/k}$ and chose k e_i^*'s by simple random sampling, with replacement, from the set of all $\hat{\hat{e}}_i$'s, to get the y_i^*'s, where $y_i^* = \mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}} + D_i^{1/2}\mathbf{e}_i^*$. Then we calculated the T^*'s by replacing $\mathbf{y}$ with $\mathbf{y}^*$ in Equation (3.5). We calculated 1000 T^*'s, and found that 13.1% of them were greater than the value of T calculated from the real dataset. As this p-value is rather large, there is enough evidence not to reject the null hypothesis that the random effects V_i are absent from the model.

Based on the simpler model which does not include any random effects for small areas, we estimated the ith small area mean, namely $\mu_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}$, by $\mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}$ $(= \hat{\mu}_i^S$, say). The associated estimated mean squared error was $\text{MSE}_i^S = \mathbf{x}_i^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{D}^{-1}\mathbf{X})^{-1}\mathbf{x}_i$. In comparison, the estimated mean squared error of the EBLUP of $\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + V_i$, based on the Fay–Herriot model, was

$$\text{MSE}_i^{\text{PR}} = g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2) + 2g_{3i}(\hat{\sigma}_v^2),$$

with $g_{1i}(\sigma_v^2) = \sigma_v^2 D_i/(\sigma_v^2 + D_i)$, $g_{2i}(\sigma_v^2) = \{D_i^2/(\sigma_v^2 + D_i)^2\}\mathbf{x}_i^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{x}_i$ and $g_{3i}(\sigma_v^2) = \{D_i^2/(\sigma_v^2 + D_i)^3\}\,\text{var}(\hat{\sigma}_v^2)$ where $\hat{\sigma}_v^2 = \{\mathbf{y}^{\mathrm{T}}(\mathbf{I} - \mathbf{H})\mathbf{y} - \sum D_i(1 - h_{ii})\}/(k - 4)$ is an unbiased estimator of σ_v^2, the variance of the small-area effects. Here $\mathbf{H} = (h_{ij}) = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$, $\Omega = D + \sigma_v^2 I_k$, and $\text{var}(\hat{\sigma}_v^2) \approx 2\sum(\sigma_v^2 + D_i)^2/k^2$. For details the reader is referred to Prasad and Rao (1990).

For the given data, $\hat{\sigma}_v^2 = 0.0177$, and including the random effects increases the estimated MSE; the percentage increase is given by $100(\text{MSE}_i^{\text{PR}} - \text{MSE}_i^S)/\text{MSE}_i^S$. In our example, the percentage increase has a median value of 325%, and so by not including the small-area effects, which are most likely not contributing to the fit of the Fay–Herriot model, we can obtain significantly more accurate estimates of the small-area means of the hospitals. It should be noted that if indeed a random small area effect is present, then it is expected that our estimated MSE will be smaller than the true MSE, which is verified by simulations. However, we feel that this underestimation is not of much concern since we decide not to use the random effects only after testing for its nonsignificance with a large probability of Type I error. If indeed there is a need for a random effect, our test indicates that fact with high probability, and in that case we recommend using suitable small area methods available in the literature.

To explore the robustness of our methodology we replaced Y_i by $\mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}} + \text{sgn}(r_i)M \times D_i^{1/2}$, where r_i was the residual $Y_i - \mathbf{X}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}$; and we did this for $i = 1, \ldots, 23$, changing only one Y_i at a time and recomputing the test statistic in each instance. When $M = 1.25$ the standardized residuals were between -2 and 2, and even for this moderate perturbation, 13 of the 23 values of the test statistic were significant at the 20% level. When $M = 2$ the figure was 22 out of 23. These results indicate that the test is sensitive to outlying values of Y_i. That is not unexpected, since our simulation studies show that our test is particularly powerful, and as a result will interpret outliers as evidence of departure from the null hypothesis. This type of sensitivity is a general issue with powerful tests. In a different context, Bell and Huang (2006) discussed the impact of outliers in the context of small area testing.

4. THEORETICAL PROPERTIES OF MODEL PARAMETERS ESTIMATORS

4.1 Properties When $\hat{\boldsymbol{\eta}}$ Is Based on Uncentred Moments

Throughout Section 4 we consider the area sample sizes n_i to be functions of k, which we take to diverge to infinity. In particular, for each k we choose a new sequence $n_1(k), \ldots, n_k(k)$, and so we are in effect addressing a triangular array. Section 4.1 treats the case where $\hat{\boldsymbol{\eta}}$ is defined by (2.12), and gives theoretical properties, which we elucidate in Section 4.2. There we discuss the convergence rates summarized in Section 1. Section 4.3 addresses the case of centered moments, where $\hat{\boldsymbol{\eta}}$ is defined by (2.14).

We introduce two matrices, $\mathbf{M}$ and $\mathbf{L}$, as follows. Let $\boldsymbol{\eta}^0$ denote the true value of $\boldsymbol{\eta}$, and write $\boldsymbol{\lambda}_s'$ for the p-vector of first derivatives of λ_s. Define the p-vector

$$\mathbf{u}_{ir} = \sum_{\ell=1}^{r} \sum_{s_1,\ldots,s_\ell} \nu_{s_1,\ldots,s_\ell}(r) \left(\prod_{t=1}^{\ell} W_{is_t}\right) \times \sum_{t=1}^{\ell} \left\{ \prod_{u:u\neq t} \lambda_{s_u}(\mathbf{X}_i, \boldsymbol{\eta}^0) \right\} \boldsymbol{\lambda}_{s_t}'(\mathbf{X}_i, \boldsymbol{\eta}^0) = (u_{ir1}, \ldots, u_{irp})^{\mathrm{T}} \quad (4.1)$$

for $1 \le i \le k$ and $1 \le r \le q$, and the $p \times p$ matrix $M = (m_{st})$, where

$$m_{st} = E(b_{st}) \quad (4.2)$$

and

$$b_{st} = \frac{1}{k}\sum_{i=1}^{k}\sum_{r=1}^{q} a_{isr}(\mathbf{X}_i, \boldsymbol{\eta}^0)u_{irt}. \quad (4.3)$$

Here a_{isr} denotes the (s, r)th component of the $p \times q$ matrix A_i. Define too the $p \times p$ matrix $\mathbf{L} = (\ell_{st})$, where

$$\ell_{st} = b_{st} - E(b_{st}) - \frac{1}{k}\sum_{i=1}^{k}\sum_{r=1}^{q}\{\bar{Y}_i^r - Q_{ir}(\boldsymbol{\eta}^0)\} \times \{\partial a_{isr}(\mathbf{X}_i, \boldsymbol{\eta}^0)/\partial\boldsymbol{\eta}^0\}_t \quad (4.4)$$

and $\{\partial a_{isr}(\mathbf{X}_i, \boldsymbol{\eta})/\partial\boldsymbol{\eta}\}_t$ denotes the tth component of the p-vector obtained by differentiating the (s, r)th component of $\mathbf{A}_i(\mathbf{X}_i, \boldsymbol{\eta})$ with respect to $\boldsymbol{\eta}$.

We assume that

$$\begin{aligned}&\lambda_r(\mathbf{x},\boldsymbol{\eta}) \text{ for } 1 \le r \le q, \text{ and each component of } A_i(\mathbf{x},\boldsymbol{\eta}), \text{ have two derivatives in } \boldsymbol{\eta}, \text{ and those}\\ &\text{derivatives are uniformly bounded for } \mathbf{x} \text{ in any given compact set, for } \boldsymbol{\eta} \text{ in some neighbourhood}\\ &\text{of } \boldsymbol{\eta}^0, \text{ and in } i \ge 1; \text{ the distribution of } \mathbf{X} \text{ is com-}\\ &\text{pactly supported; for each } c > 0, E(|Y|^c|\mathbf{X}=\mathbf{x}) \text{ is}\\ &\text{bounded uniformly in } \mathbf{x} \text{ in the support of the dis-}\\ &\text{tribution of } \mathbf{X}; \text{ and the eigenvalues of } \mathbf{M}^{\mathrm{T}}\mathbf{M} \text{ are}\\ &\text{bounded away from zero.}\end{aligned} \tag{4.5}$$

The assumption in (4.5) that the distribution of $\mathbf{X}$ is compactly supported can be relaxed in exchange for stronger conditions on the functions λ_r and $\mathbf{A}_i$ and their derivatives, at extreme values of the arguments of those functions. If $\mathbf{A}_i(\mathbf{x},\boldsymbol{\eta})$ does not actually depend on $\boldsymbol{\eta}$ then the assumption in (4.5) that $E(|Y|^c|\mathbf{X}=\mathbf{x})$ be bounded in $\mathbf{x}$ for each $c > 0$ can be replaced by the condition that $E(|Y|^p|\mathbf{X}=\mathbf{x})$ be bounded in $\mathbf{x}$.

Next we elucidate the condition in (4.5) on the eigenvalues of $\mathbf{M}^{\mathrm{T}}\mathbf{M}$. If $\bar{Y}_i$ were a simple average, meaning that $w_{ij} = n_i^{-1}$ for each j, then W_{ir} would equal simply n_i^{1-r}. If in addition the area sample sizes n_i were all equal to n, then W_{ir} would equal n^{1-r} for each i and each r. To simplify some of our discussion we shall suppose that the weights and sample sizes do not depart greatly from this norm, in the sense that there exists a function $n = n(k) \ge 1$ such that

$$\begin{aligned}&\text{for constants } C_1, C_2 \text{ satisfying } 0 < C_1 < C_2 <\\ &\infty \text{ we have, for each } 1 \le r \le p \text{ and all } k, C_1 \le\\ &n^{r-1}\min_{1\le i\le k} W_{ir} \le n^{r-1}\max_{1\le i\le k} W_{ir} \le C_2.\end{aligned} \tag{4.6}$$

The simplest case is that where the matrix $\mathbf{A}_i$ does not depend on i; say, $\mathbf{A}_i(\mathbf{x},\boldsymbol{\eta})$ is the matrix with (s,r)th component equal to $a_{sr}(\mathbf{x},\boldsymbol{\eta})$. Then, if (4.6) holds, $m_{st} = E(b_{st}) = m_{st}^0 + O(n^{-1})$, where

$$m_{st}^0 = \sum_{r=1}^{q} E\left\{a_{sr}(\mathbf{X},\boldsymbol{\eta}^0)\frac{\partial}{\partial\eta_t}\lambda_1(X,\boldsymbol{\eta})^r\Big|_{\boldsymbol{\eta}=\boldsymbol{\eta}^0}\right\}.$$

For example, when $p = q$ and each $\mathbf{A}_i = \mathbf{I}_p$, in which case $\hat{\boldsymbol{\eta}}$ is the estimator obtained by solving equations (2.10), we have simply $m_{st}^0 = E\{(\partial/\partial\eta_t)\lambda_1(\mathbf{X},\boldsymbol{\eta})^s\}|_{\boldsymbol{\eta}=\boldsymbol{\eta}^0}$. More generally, if we define $\mathbf{M}^0 = (m_{rj}^0)$ then a sufficient condition for the assumption in (4.5) that the eigenvalues of $\mathbf{M}^{\mathrm{T}}\mathbf{M}$ are bounded away from zero is that (4.6) holds with $n = n(k) \to \infty$ as $k \to \infty$, and the eigenvalues of $\mathbf{M}^{0\mathrm{T}}\mathbf{M}^0$ are bounded away from zero.

Under condition (4.5) the matrices $\mathbf{M}$ and $\mathbf{M}^{-1}$ are genuinely of order 1, in that none of their components diverges to infinity. In particular, it is not necessary to impose a boundedness condition on $\mathbf{M}$. However, the matrices may not actually converge; that requires, for example, a degree of regularity in the values of $n_i(k)$ which is not implied by (4.5).

Define the random p-vector

$$\mathbf{S} = \frac{1}{k}\sum_{i=1}^{k}\mathbf{A}_i(\mathbf{X}_i,\boldsymbol{\eta}^0)\mathbf{R}_i(\boldsymbol{\eta}^0) = (S_1,\ldots,S_p)^{\mathrm{T}}. \tag{4.7}$$

Theorem 1. Assume that (4.5) holds. Then for each $D_1 > 0$ there exist constants $D_2, D_3, D_4 > 0$ such that, with probability at least $1 - D_2k^{-D_1}$, equations (2.12) have a solution $\hat{\boldsymbol{\eta}}$ satisfying $\|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^0\| \le D_3$, and moreover, any such solution satisfies $\|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^0 - (\mathbf{M}+\mathbf{L})^{-1}\mathbf{S}\| \le D_4\|\mathbf{S}\|^2$ and $\|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^0 - \mathbf{M}^{-1}\mathbf{S}\| \le D_4\|\mathbf{S}\|(\|\mathbf{S}\| + \|\mathbf{L}\|)$.

4.2 Elucidation of Theorem 1

First we describe the variance of $\mathbf{S}$, and hence that of $\hat{\boldsymbol{\eta}}$, in the "balanced" case where the area sample sizes n_i, and hence also the matrices $\mathbf{A}_i$, are similar. By its definition, S has zero mean. It is also clear from the definition of S_r at (4.7) that, if the $\mathbf{A}_i$s are all of approximately the same order (e.g., if the $\mathbf{A}_i$s are all identical), then $\mathrm{var}(S_r) = O(k^{-1})$. However, the variance can be of strictly smaller order than k^{-1} if the n_is diverge. Indeed, taking $p = q$ and $\mathbf{A}_i \equiv \mathbf{I}$ for simplicity, it can be shown that, under the conditions of Theorem 1 and if $\min_{1\le i\le k} n_i(k) \to \infty$ as $k \to \infty$,

$$\begin{aligned}\mathrm{cov}(S_r, S_s) = rsK^{-1}E\{\lambda_1(\mathbf{X},\boldsymbol{\eta})^{r+s-2}\,\mathrm{var}(Y|\mathbf{X})\}\\ + o(K^{-1}),\end{aligned} \tag{4.8}$$

where

$$K^{-1} = k^{-2}\sum_{i=1}^{k} n_i^{-1}. \tag{4.9}$$

In particular, if each $n_i = n$ then $K = kn$. Analogous results hold if the matrices $\mathbf{A}_i$ are different but of similar orders for different values of i. In this context, Theorem 1 implies that $\hat{\boldsymbol{\eta}} = \boldsymbol{\eta}^0 + (\mathbf{M}+\mathbf{L})^{-1}\mathbf{S} + O_p(K^{-1})$. The matrix L converges to zero only at rate $k^{-1/2}$, which is potentially much slower than $K^{-1/2}$. (The rate $k^{-1/2}$ for L is concise, not just an upper bound.) As a result, $\hat{\boldsymbol{\eta}} = \boldsymbol{\eta}^0 + \mathbf{M}^{-1}\mathbf{S} + O_p\{(kK)^{-1/2}\}$, rather than having a remainder of K^{-1}. Assumptions (4.5), and the condition that the A_is are of similar orders and $\min_{1\le i\le k} n_i(k) \to \infty$, are sufficient to imply that $K^{1/2}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^0)$ is asymptotically normally distributed with zero mean and covariance matrix $\mathbf{M}^{-1}\boldsymbol{\Sigma}(\mathbf{M}^{-1})^{\mathrm{T}}$, where $\boldsymbol{\Sigma}$ is the asymptotic covariance matrix of $K^{1/2}\mathbf{S}$ and has (r,s)th component given by K times the first term in the right-hand side of (4.8).

Next we address the "unbalanced" setting, where the n_is are quite different and different variance properties can arise. Cases of this type include those where A_i is of size n_i and the n_is take widely differing values. There, $\mathrm{var}(S_r) = O(t_1)$ where $t_1 = k^{-2}\sum_i n_i$, the components of the matrix $\mathbf{M}$ are of size $t_2 = k^{-1}\sum_i n_i$, and $\mathbf{L} = O_p(t_2t_3)$ where $t_3 = (\sum_i n_i^2)^{1/2}/(\sum_i n_i)$. In consequence, provided that $t_3 = o(t_2)$, which ensures negligibility of $\mathbf{L}$ relative to $\mathbf{M}$, we have $\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^0 = O_p(t_1^{1/2}/t_2) = O_p(K'^{-1/2})$, where $K' = \sum_i n_i$. If the integers n_i are of similar sizes then $K \asymp K'$, that is, the ratio K/K' is bounded away from zero and infinity as k diverges (e.g., if each $n_i = n$ then $K = K' = kn$). However, if the n_is are of widely differing sizes then K' can be an order of magnitude larger than K, and then the corresponding estimator $\hat{\boldsymbol{\eta}}$, constructed using $\mathbf{A}_i$s of respective sizes n_i, is relatively accurate. To take an extreme case, if $\sup_{k;i\ge 2} n_i(k) < \infty$, but $n_1 = n_1(k)$ where $n_1(k)/k \to \infty$ as $k \to \infty$, then $K \asymp k$ but $K' \asymp n_1(k)$, which diverges more rapidly than k, implying that $\hat{\boldsymbol{\eta}} \to \boldsymbol{\eta}^0$ more rapidly under the regime where the A_is are of respective sizes n_i. [Note too that in this case, $t_3 = o(t_2)$ and so $\mathbf{L}$ is negligible relative to $\mathbf{M}$.]

This context, where the matrices $\mathbf{A}_i$ are of order n_i, arises naturally when $\hat{\boldsymbol{\eta}}$ is defined in terms of estimating functions; for discussion, see the paragraph containing (2.15) and (2.16). The theory we have given in Theorem 1 applies without change to this context, provided the n_is are of similar sizes, in particular where

$$\sup_{k\ge 1}\Big\{\max_{1\le i\le k} n_i(k)\Big/\min_{1\le i\le k} n_i(k)\Big\} < \infty. \tag{4.10}$$

Nevertheless, to connect the discussion in the previous paragraph to the theorem we need to rescale the matrix $\mathbf{M}$. For example, in the discussion it was convenient to interpret the components of M as being of size t_2, rather than 1, so we need to divide M by t_2 to get a matrix for which (4.5) holds. Similar normalisations have to be applied to $\mathbf{L}$ and $\mathbf{S}$. Only if the n_is are of very widely different sizes, as in the example where $n_1(k)/k$ diverges, is a different approach to the theorem necessary.

The following corollary to Theorem 1 can also be proved. It is stated under assumption (4.5), which continues to apply (after the rescalings discussed in the previous paragraph) if the matrices A_i are of sizes n_i and (4.10) holds. Alternative proofs give analogues of the corollary in cases where the $\mathbf{A}_i$s are of very different sizes.

Corollary 1. If (4.5) holds, if $\max_{1\le i\le k} n_i = O(k^C)$ for some $C > 0$, and if the function f of p variables is bounded and has two bounded derivatives in a neighbourhood of $\boldsymbol{\eta}^0$, then $E\{f(\hat{\boldsymbol{\eta}})\} = f(\boldsymbol{\eta}^0) + O(K^{-1})$.

An immediate implication is that if we substitute $\hat{\boldsymbol{\eta}}$ for $\boldsymbol{\eta} = \boldsymbol{\eta}^0$ in a function, $f(\boldsymbol{\eta})$, representing the coverage error of a confidence or prediction interval, or a mean squared error; if that function is bounded and has two bounded derivatives in a neighborhood of $\boldsymbol{\eta}^0$; and, when the target is a prediction interval, if the random variable whose value we are predicting is statistically independent of $\hat{\boldsymbol{\eta}}$; then the coverage error, or the expected value of the mean squared error, is in error by only $O(K^{-1})$.

4.3 Properties When $\hat{\boldsymbol{\eta}}$ Is Based on Centred Moments

Here we treat the estimator defined by (2.14). Let $\check{R}_{ir}$ be as at (2.13). In place of the definition of $\mathbf{u}_{ir}$ at (4.1), put

$$\check{\mathbf{u}}_{ir} = (\check{u}_{ir1}, \dots, \check{u}_{irp})^{\mathrm{T}} = -E\left\{\frac{\partial}{\partial\boldsymbol{\eta}}\check{R}_{ir}(\boldsymbol{\eta})\,\middle|\,\mathbf{X}_i\right\}\Big|_{\boldsymbol{\eta}=\boldsymbol{\eta}^0}.$$

Define $\check{b}_{st}$ as at (4.3) but with u_{irt} there replaced by $\check{u}_{irt}$, and let $\check{m}_{st} = E(\check{b}_{st})$ and

$$\check{\ell}_{st} = \check{b}_{st} - E(\check{b}_{st}) - \frac{1}{k}\sum_{i=1}^{k}\sum_{r=1}^{q}\check{R}_{ir}(\boldsymbol{\eta}^0)\{\partial a_{isr}(\mathbf{X}_i,\boldsymbol{\eta}^0)/\partial\boldsymbol{\eta}^0\}_t,$$

as indicated by the first identity in (4.2) and by (4.4). Analogously to (4.7), put $\check{\mathbf{S}} = (\check{S}_1, \dots, \check{S}_p)^{\mathrm{T}} = k^{-1}\sum_i \mathbf{A}_i(\mathbf{X}_i,\boldsymbol{\eta}^0) \times \check{\mathbf{R}}_i(\boldsymbol{\eta}^0)$, and write $\check{\mathbf{M}}$ for the $p\times p$ matrix with (s,t)th component m_{st}. Redefine $\hat{\boldsymbol{\eta}}$ to be a solution of (2.14) rather than (2.12). The analogue of Theorem 1 in this setting is the following result.

Theorem 2. Assume that (4.5) holds, but with the condition that eigenvalues of $\mathbf{M}^{\mathrm{T}}\mathbf{M}$ be bounded away from zero replaced by the same assumption on the eigenvalues of $\check{\mathbf{M}}^{\mathrm{T}}\check{\mathbf{M}}$. Then for each $D_1 > 0$ there exist constants $D_2, D_3, D_4 > 0$ such that, with probability at least $1 - D_2k^{-D_1}$, equations (2.14) have a solution $\hat{\boldsymbol{\eta}}$ satisfying $\|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^0\| \le D_3$, and moreover, any such solution also satisfies $\|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^0 - (\check{\mathbf{M}} + \check{\mathbf{L}})^{-1}\check{\mathbf{S}}\| \le D_4\|\check{\mathbf{S}}\|^2$ and $\|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^0 - \check{\mathbf{M}}^{-1}\check{\mathbf{S}}\| \le D_4\|\check{\mathbf{S}}\|(\|\check{\mathbf{S}}\| + \|\check{\mathbf{L}}\|)$.

The discussion in the paragraph containing (4.10), and in the paragraph previous to that one, also applies here. Additionally, the analogue of Corollary 1 holds, with an almost identical proof. If $\min_{1\le i\le k} n_i(k)$ diverges to infinity as $k\to\infty$, and if (for simplicity) we take $\mathbf{A}_i$ to be the identity and assume (4.10), then the asymptotic covariance of $\mathbf{S}$ is given by $\mathrm{cov}(\check{S}_r, \check{S}_s) = o(K^{-1})$ if at least one of r and s does not equal 1, and

$$\mathrm{var}(\check{S}_1) = \mathrm{var}(S_1) = K^{-1}E\{\mathrm{var}(Y|\mathbf{X})\} + o(K^{-1}), \tag{4.11}$$

where K is given by (4.9). Compare (4.11) with (4.8). At another extreme, if $n_1, n_2, \dots$ are uniformly bounded as $k\to\infty$, and if (again for simplicity) we assume that (4.10) holds, $p = q$ and $\mathbf{A}_i = \mathbf{I}$, then $\mathrm{cov}(\check{S}_r, \check{S}_s) = c_{rs}(k)k^{-1} + o(k^{-1})$ where $c_{rs}(k)$ has a moderately complex formula but is bounded uniformly in r and s as $k\to\infty$, and, generally, $\inf_{1\le r\le q, k\ge 1} c_{rr}(k) > 0$.

Next we elucidate the condition, imposed in Theorem 2, that the eigenvalues of $\mathbf{M}^{\mathrm{T}}\mathbf{M}$ are bounded away from zero. Assume that (4.6) holds, and recall from the definitions above Theorem 2 that

$$\begin{aligned}\check{b}_{st} &= \frac{1}{k}\sum_{i=1}^{k}\sum_{r=1}^{q} a_{isr}(\mathbf{X}_i,\boldsymbol{\eta}^0)E\left\{\frac{\partial}{\partial\eta_t}\check{R}_{ir}(\boldsymbol{\eta})\,\middle|\,\mathbf{X}_i\right\}\Big|_{\boldsymbol{\eta}=\boldsymbol{\eta}^0}\\ &= -\frac{1}{k}\sum_{i=1}^{k}\sum_{r=1}^{q} a_{isr}(\mathbf{X}_i,\boldsymbol{\eta}^0)\\ &\qquad\times E\big[r\{\bar{Y}_i - \lambda_1(\mathbf{X}_i,\boldsymbol{\eta}^0)\}^{r-1}\lambda'_{1t}(\mathbf{X}_i,\boldsymbol{\eta}^0) + \check{Q}'_{irt}(\boldsymbol{\eta}^0)|\mathbf{X}_i\big],\end{aligned}$$

where $\check{Q}'_{irt}(\boldsymbol{\eta})$ denotes the derivative of $\check{Q}_{ir}(\boldsymbol{\eta})$ with respect to η_t. If $r = 1$ then $\check{Q}_{ir}$, and hence also $\check{Q}'_{irt}$, vanishes. If $r\ge 2$ then it follows from the definition of $\check{Q}_{ir}$, at (2.13), and from (4.6), that $E\{\check{Q}'_{irt}(\boldsymbol{\eta}^0)\} = O(n^{-1})$. If $r\ge 2$ then $E[\{\bar{Y}_i - \lambda_1(\mathbf{X}_i,\boldsymbol{\eta}^0)\}^{r-1}|\mathbf{X}_i] = O(n^{-1})$, and when $r = 1$ the conditional expected value equals 1. Combining these results we deduce that when (4.6) holds, $\check{m}_{st} = E(\check{b}_{st}) = \check{m}^0_{st} + O(n^{-1})$, where $\check{m}^0_{st} = -k^{-1}\sum_i E\{a_{is1}(\mathbf{X},\boldsymbol{\eta}^0)\lambda'_{1t}(\mathbf{X},\boldsymbol{\eta}^0)\}$. Therefore the condition that the eigenvalues of $\check{\mathbf{M}}^{\mathrm{T}}\check{\mathbf{M}}$ be bounded away from zero is, for all sufficiently large n, equivalent to the same assumption on eigenvalues of $\check{\mathbf{M}}^{0\mathrm{T}}\check{\mathbf{M}}^0$, where $\check{\mathbf{M}}^0 = (\check{m}^0_{st})$. If the matrix-valued function $\mathbf{A}_i(\mathbf{x},\boldsymbol{\eta})$ does not depend on i, and has (r,s)th component equal to $a_{rs}(\mathbf{x},\boldsymbol{\eta})$, then $\check{m}^0_{st} = -E\{a_{s1}(\mathbf{X},\boldsymbol{\eta}^0)\lambda'_{1t}(\mathbf{X},\boldsymbol{\eta}^0)\}$.

Abbreviated proofs of Theorem 1 and Corollary 1 are given in Appendix A. More detailed arguments, together with proofs of (4.8) and Theorem 2, appear in Appendix B of the online supplement.

APPENDIX A: THEOREMS AND PROOFS

A.1 Estimation of MSE and Construction of Confidence Interval (Binomial Case)

First we use the delta method to obtain an approximate estimator of the MSE of $\hat{p}_i$. Estimator $\hat{\boldsymbol{\eta}}$ satisfies the estimating equation $\mathbf{S}(\boldsymbol{\eta}) = \mathbf{0}$,

where from (4.7) $\mathbf{S}(\boldsymbol{\eta}) = \sum_{i=1}^{k} W_{i2}^{-1}(\bar{y}_i - p_i)\mathbf{x}_i$. Here, $\bar{y}_i$ is the proportion of families below low-income level in state i, $p_i = p_i(\boldsymbol{\eta}) = \frac{\exp\{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\eta}\}}{1+\exp\{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\eta}\}}$ and $\mathbf{x}_i$ is the covariate vector including a one for the intercept. Note that

$$\frac{\partial \mathbf{S}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^{\mathrm{T}}} = -\sum_{i=1}^{k} W_{i2}^{-1} p_i(1-p_i)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}.$$

We know that $\operatorname{var}(\bar{y}_i) = W_{i2}p_i(1-p_i)$. Then, $\operatorname{var}(\mathbf{S}(\boldsymbol{\eta})) = -\frac{\partial \mathbf{S}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^{\mathrm{T}}}$. Let $\hat{p}_i = p_i(\hat{\boldsymbol{\eta}})$. Then, by delta method,

$$\operatorname{var}(\hat{p}_i) = \{p_i(1-p_i)\}^2\{\mathbf{x}_i^{\mathrm{T}}\operatorname{var}(\hat{\boldsymbol{\eta}})\mathbf{x}_i\} + o(K^{-1}).$$

Estimated variance–covariance matrix of $\hat{\boldsymbol{\eta}}$ and of $\hat{p}_i$, respectively, are

$$\begin{aligned}\widehat{\operatorname{var}}(\hat{\boldsymbol{\eta}}) &\approx \left\{\left(\frac{\partial \mathbf{S}}{\partial \hat{\boldsymbol{\eta}}^{\mathrm{T}}}\right)^{-1}\right\}\left[\operatorname{var}(\mathbf{S}(\boldsymbol{\eta}))|_{\boldsymbol{\eta}=\hat{\boldsymbol{\eta}}}\right]\left\{\left(\frac{\partial \mathbf{S}}{\partial \hat{\boldsymbol{\eta}}}\right)^{-1}\right\}^{\mathrm{T}} \\ &= \left(\sum_{i=1}^{k} W_{i2}^{-1}\hat{p}_i(1-\hat{p}_i)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)^{-1}, \\ \widehat{\operatorname{var}}(\hat{p}_i) &= \{\hat{p}_i(1-\hat{p}_i)\}^2\mathbf{x}_i^{\mathrm{T}}\widehat{\operatorname{var}}(\hat{\boldsymbol{\eta}})\mathbf{x}_i.\end{aligned}$$

Using the estimated variance of $\hat{p}_i$ and normal approximation, an approximate $100(1-\alpha)\%$ CI for p_i is given by $\hat{p}_i \pm z_{\alpha/2}\sqrt{\widehat{\operatorname{var}}(\hat{p}_i)}$. However, we can also create a bootstrap confidence interval. To that goal, first, using $\hat{\boldsymbol{\eta}}$, estimated $\boldsymbol{\eta}$ from data, we generate B bootstrap samples $\{Y_{ij}^*, j = 1,\ldots,n_i, i = 1,\ldots,k\}$ and calculate $\hat{\boldsymbol{\eta}}^*$ for each bootstrap sample. Then we calculate $\hat{p}_i^* = p_i(\hat{\boldsymbol{\eta}}^*)$ and $\widehat{\operatorname{var}}^*(\hat{p}_i^*) = \{\hat{p}_i^*(1-\hat{p}_i^*)\}^2\{\mathbf{x}_i^{\mathrm{T}}\widehat{\operatorname{var}}^*(\hat{\boldsymbol{\eta}}^*)\mathbf{x}_i\}$ where $\widehat{\operatorname{var}}^*(\hat{\boldsymbol{\eta}}^*)$ is obtained from the formula of $\widehat{\operatorname{var}}(\hat{\boldsymbol{\eta}})$ by replacing $\hat{\boldsymbol{\eta}}$ by $\hat{\boldsymbol{\eta}}^*$. Then we calculate the $B\alpha/2$th and $B(1-\alpha/2)$th order statistics of the standardized values $(\hat{p}_i^* - \hat{p}_i)/\sqrt{\widehat{\operatorname{var}}^*(\hat{p}_i^*)}$. We call them q_{li} and q_{ui} respectively. We define a $100(1-\alpha)\%$ CI for p_i as $\{\hat{p}_i - q_{ui}\sqrt{\widehat{\operatorname{var}}(\hat{p}_i)}, \hat{p}_i - q_{li}\sqrt{\widehat{\operatorname{var}}(\hat{p}_i)}\}$.

A.2 Estimation of MSE and Construction of Confidence Interval (Multi/Trinomial Case)

Here $p_{i1} = p_{i1}(\boldsymbol{\eta}) = \frac{\exp\{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\eta}_1\}}{1+\exp\{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\eta}_1\}+\exp\{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\eta}_2\}}$ and $p_{i2} = p_{i2}(\boldsymbol{\eta}) = \frac{\exp\{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\eta}_2\}}{1+\exp\{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\eta}_1\}+\exp\{\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\eta}_2\}}$. We write $\boldsymbol{\eta} = (\boldsymbol{\eta}_1^{\mathrm{T}}, \boldsymbol{\eta}_2^{\mathrm{T}})^{\mathrm{T}}$. Note that

$$\mathbf{S}(\boldsymbol{\eta}) = \begin{pmatrix}\mathbf{S}^{(1)}(\boldsymbol{\eta}) \\ \mathbf{S}^{(2)}(\boldsymbol{\eta})\end{pmatrix} = \begin{pmatrix}\sum_{i=1}^{k} W_{i2}^{-1}(\bar{y}_{i1} - p_{i1})\mathbf{x}_i \\ \sum_{i=1}^{k} W_{i2}^{-1}(\bar{y}_{i2} - p_{i2})\mathbf{x}_i\end{pmatrix},$$

with

$$\begin{aligned}\frac{\partial \mathbf{S}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}^{\mathrm{T}}} &= -\sum_{i=1}^{k} W_{i2}^{-1}\begin{bmatrix}p_{i1}(1-p_{i1}) & -p_{i1}p_{i2} \\ -p_{i1}p_{i2} & p_{i2}(1-p_{i2})\end{bmatrix} \otimes \mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} \\ &= -\operatorname{cov}(\mathbf{S}(\boldsymbol{\eta})).\end{aligned}$$

By delta method,

$$\begin{aligned}\widehat{\operatorname{cov}}(\hat{\boldsymbol{\eta}}) &\approx \left\{\left(\frac{\partial \mathbf{S}}{\partial \hat{\boldsymbol{\eta}}^{\mathrm{T}}}\right)^{-1}\right\}\left[\operatorname{cov}(\mathbf{S}(\boldsymbol{\eta}))|_{\boldsymbol{\eta}=\hat{\boldsymbol{\eta}}}\right]\left\{\left(\frac{\partial \mathbf{S}}{\partial \hat{\boldsymbol{\eta}}}\right)^{-1}\right\}^{\mathrm{T}} \\ &= \left(-\frac{\partial \mathbf{S}(\boldsymbol{\eta})}{\partial \hat{\boldsymbol{\eta}}^{\mathrm{T}}}\right)^{-1}, \\ \widehat{\operatorname{cov}}\begin{pmatrix}\hat{p}_{i1} \\ \hat{p}_{i2}\end{pmatrix} &\approx \begin{bmatrix}\hat{p}_{i1}(1-\hat{p}_{i1}) & -\hat{p}_{i1}\hat{p}_{i2} \\ -\hat{p}_{i1}\hat{p}_{i2} & \hat{p}_{i2}(1-\hat{p}_{i2})\end{bmatrix} \\ &\quad \times \mathbf{V}\begin{bmatrix}\hat{p}_{i1}(1-\hat{p}_{i1}) & -\hat{p}_{i1}\hat{p}_{i2} \\ -\hat{p}_{i1}\hat{p}_{i2} & \hat{p}_{i2}(1-\hat{p}_{i2})\end{bmatrix}^{\mathrm{T}},\end{aligned}$$

where

$$\mathbf{V} = \begin{bmatrix}\mathbf{x}_i^{\mathrm{T}} & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0}^{\mathrm{T}} & \mathbf{x}_i^{\mathrm{T}}\end{bmatrix}\widehat{\operatorname{cov}}(\hat{\boldsymbol{\eta}})\begin{bmatrix}\mathbf{x}_i^{\mathrm{T}} & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0}^{\mathrm{T}} & \mathbf{x}_i^{\mathrm{T}}\end{bmatrix}^{\mathrm{T}}.$$

A.3 Proof of Theorem 1

Put $\boldsymbol{\Delta} = \boldsymbol{\eta} - \boldsymbol{\eta}^0$ and $\mathbf{A}_i = (a_{isr})$, and observe that, by (4.5), for $\boldsymbol{\eta}$ within any given but fixed radius of $\boldsymbol{\eta}^0$,

$$\begin{aligned}\lambda_s(\mathbf{X}_i, \boldsymbol{\eta}) &= \lambda_s(\mathbf{X}_i, \boldsymbol{\eta}^0) + \boldsymbol{\lambda}_s'(\mathbf{X}_i, \boldsymbol{\eta}^0)^{\mathrm{T}}\boldsymbol{\Delta} \\ &\quad + \Theta_{is}(\boldsymbol{\eta})\|\boldsymbol{\eta} - \boldsymbol{\eta}^0\|^2, \\ a_{isr}(\mathbf{X}_i, \boldsymbol{\eta}) &= a_{isr}(\mathbf{X}_i, \boldsymbol{\eta}^0) + \mathbf{a}_{isr}'(\mathbf{X}_i, \boldsymbol{\eta}^0)^{\mathrm{T}}\boldsymbol{\Delta} \\ &\quad + \Theta_{isr}(\boldsymbol{\eta})\|\boldsymbol{\eta} - \boldsymbol{\eta}^0\|^2,\end{aligned} \tag{A.1}$$

where $\mathbf{a}_{isr}'(\mathbf{X}_i, \boldsymbol{\eta}) = \partial a_{isr}(\mathbf{X}_i, \boldsymbol{\eta})/\partial\boldsymbol{\eta}$ and, here and below, $\Theta_{i\ldots}(\boldsymbol{\eta})$ denotes a generic random variable satisfying, with probability 1, $|\Theta_{i\ldots}(\boldsymbol{\eta})| \le C_1$ for $\|\boldsymbol{\eta} - \boldsymbol{\eta}^0\| \le C_2$ and $\|\mathbf{X}_i\| \le C_2$, for any $C_2 > 0$, where C_1 depends on C_2 but not on i. A Taylor expansion argument can be used to show that $Q_{ir}(\boldsymbol{\eta}) = Q_{ir}(\mathbf{X}_i, \boldsymbol{\eta}^0) + \mathbf{u}_{ir}^{\mathrm{T}}\boldsymbol{\Delta} + \Theta_i(\boldsymbol{\eta})\|\boldsymbol{\eta} - \boldsymbol{\eta}^0\|^2$. This result and (A.1) imply that

$$\begin{aligned}&\{\mathbf{A}_i(\mathbf{X}_i, \boldsymbol{\eta})\mathbf{R}_i(\boldsymbol{\eta})\}_s \\ &= \{\mathbf{A}_i(\mathbf{X}_i, \boldsymbol{\eta}^0)\mathbf{R}_i(\boldsymbol{\eta}^0)\}_s \\ &\quad + \sum_{r=1}^{q}\left[\{\bar{Y}_i^r - Q_{ir}(\boldsymbol{\eta}^0)\}\mathbf{a}_{isr}'(\mathbf{X}_i, \boldsymbol{\eta}^0) - a_{isr}(\mathbf{X}_i, \boldsymbol{\eta}^0)\mathbf{u}_{ir}\right]^{\mathrm{T}}\boldsymbol{\Delta} \\ &\quad + \Theta_{is}(\boldsymbol{\eta})(|\bar{Y}_i^q| + 1)\|\boldsymbol{\eta} - \boldsymbol{\eta}^0\|^2.\end{aligned}$$

Summing over i, and equating to zero as entailed by (2.12), it can be shown that $\boldsymbol{\eta} = \hat{\boldsymbol{\eta}}$ satisfies

$$(\mathbf{M} + \mathbf{L})(\boldsymbol{\eta} - \boldsymbol{\eta}^0) + \boldsymbol{\Theta}(\boldsymbol{\eta})\|\boldsymbol{\eta} - \boldsymbol{\eta}^0\|^2 = \mathbf{S}. \tag{A.2}$$

In (A.2), $\boldsymbol{\Theta}$ denotes a p-vector with, as its sth component, a random variable Θ_s satisfying, with probability 1, $|\Theta_s(\boldsymbol{\eta})| \le p^{-1/2}C_1$ for $\|\boldsymbol{\eta} - \boldsymbol{\eta}^0\| \le C_2$ and C_2 sufficiently small, where C_1 depends on C_2.

It can be shown from (4.5) and Rosenthal's inequality that for all $C_3, C_4 > 1$,

$$P(\|\mathbf{L}\| > C_3) \le C_5(C_4)\left(C_3k^{1/2}\right)^{-C_4}, \tag{A.3}$$

where $C_5(C_4) > 0$ depends on C_4 but not on C_3 or k. [In (A.3), and below, we write the norm $\|\mathbf{Q}\|$ of a $p \times p$ matrix to denote the supremum of $\|\mathbf{Q}\mathbf{v}\|$ over all p-vectors $\mathbf{v}$ for which $\|\mathbf{v}\| = 1$.]

Let $c \in (0,1)$ denote a lower bound to the least eigenvalue of $\mathbf{M}^{\mathrm{T}}\mathbf{M}$, and put $C_3 = \frac{1}{3}c$. It can be shown that equations (2.12) have a solution whenever $\|\mathbf{S}\| \le C_3\min(C_1^{-1}C_3, C_2)$. Hence, the probability of a solution is not less than $\pi_k \equiv 1 - C_6k^{-C_4/2}$, for some $C_6 > 0$. Moreover, it can be deduced from (A.2) that under the same conditions, any solution $\hat{\boldsymbol{\eta}}$ of (A.1) satisfies

$$\|(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^0) - (M+L)^{-1}S\| \le (3/2c)C_1\|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^0\|^2, \tag{A.4}$$

and therefore, if $\|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^0\| \le C_3$, we can see from (A.2) that

$$\begin{aligned}\frac{1}{2}\|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^0\| &\le \|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^0\|\{1 - (3/2c)\|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^0\|\} \\ &\le \|(\mathbf{M}+\mathbf{L})^{-1}\mathbf{S}\| \le (3/2c)\|\mathbf{S}\|.\end{aligned} \tag{A.5}$$

Together, (A.4) and (A.5) imply that if $\|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^0\| \le \min(C_1^{-1}C_3, C_2, C_3)$ then with probability not less than π_k, $\|(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^0) - (\mathbf{M}+\mathbf{L})^{-1}\mathbf{S}\| \le (3/2c)C_1C_3^{-2}\|\mathbf{S}\|^2$, from which follows the second-last inequality in Theorem 1. The last inequality in the theorem follows via a Taylor expansion, enabled by (A.3).

A.4 Proof of Corollary 1

Define the event $\mathcal{E} = \{\|\mathbf{S}\| \le \epsilon, \|L\| \le \epsilon\}$, and let $\widetilde{\mathcal{E}}$ denote the complement of $\mathcal{E}$. Let $\epsilon > 0$, let $C > 0$ be as in the condition $\max_{1\le i\le k} n_i = O(k^C)$ in the statement of the corollary, and note that

$$(\sup|f|)^{-1}E\{|f(\hat{\boldsymbol{\eta}})|I(\widetilde{\mathcal{E}})\} \le P(\widetilde{\mathcal{E}}) \le P(\|\mathbf{S}\| > \epsilon) + P(\|\mathbf{L}\| > \epsilon) = O(K^{-1}). \quad (A.6)$$

Assume that f has two bounded derivatives within radius C_7 of $\boldsymbol{\eta}^0$, where $C_7 > 0$, and let D_4 be as in Theorem 1. Let $c \in (0, 1)$ be a lower bound to the least eigenvalue of $\mathbf{M}^{\mathrm{T}}\mathbf{M}$, and write f' for the p-vector of first derivatives of f. By choosing $\epsilon = \epsilon(C_7)$ sufficiently small we can show, as in the proof of Theorem 1, that when $\mathcal{E}$ obtains we have $\|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^0\| \le \frac{1}{2}C_7$, $\|\{(\mathbf{M}+\mathbf{L})^{-1}\mathbf{S}\}^{\mathrm{T}}f'(\boldsymbol{\eta}^0)\| \le \frac{1}{2}C_7$ and $D_4\|\mathbf{S}\|^2 \le \frac{1}{2}C_7$, and that the least eigenvalues of $(\mathbf{M}+\mathbf{L})^{\mathrm{T}}(\mathbf{M}+\mathbf{L})$ and $(\mathbf{I}+\mathbf{L}\mathbf{M}^{-1})^{\mathrm{T}}(\mathbf{I}+\mathbf{L}\mathbf{M}^{-1})$ exceed $\frac{1}{2}c$. Hence, by Taylor expansion of $f(\hat{\boldsymbol{\eta}})$ about $\boldsymbol{\eta}^0$, we deduce from Theorem 1 that there exists a constant $C_8 > 0$ such that, provided $\mathcal{E}$ obtains, $|f(\hat{\boldsymbol{\eta}}) - f(\boldsymbol{\eta}^0) - \{(\mathbf{M}+\mathbf{L})^{-1}\mathbf{S}\}^{\mathrm{T}}f'(\boldsymbol{\eta}^0)| \le C_8\|\mathbf{S}\|^2$. Therefore,

$$\left|E\{f(\hat{\boldsymbol{\eta}})I(\mathcal{E})\} - f(\boldsymbol{\eta}^0)P(\mathcal{E}) - E\left[\{(\mathbf{M}+\mathbf{L})^{-1}\mathbf{S}\}^{\mathrm{T}}I(\mathcal{E})\right]f'(\boldsymbol{\eta}^0)\right| = O(K^{-1}). \quad (A.7)$$

Let $j_0 \ge 1$ denote an integer. The properties discussed in the previous paragraph can be applied to prove that

$$\begin{aligned} &E\{(\mathbf{M}+\mathbf{L})^{-1}\mathbf{S}I(\mathcal{E})\} \\ &= \mathbf{M}^{-1}\sum_{j=0}^{j_0}\left[E\{(-\mathbf{L}\mathbf{M}^{-1})^j\mathbf{S}\} - E\{(-\mathbf{L}\mathbf{M}^{-1})^j\mathbf{S}I(\|\mathbf{S}\| > \epsilon)\}\right] \\ &\quad + O\left\{\sum_{j=0}^{j_0}\left[E\left\{\|\mathbf{L}\|^{2j}I(\|\mathbf{L}\| > \epsilon)\right\}\right]^{1/2}(E\|\mathbf{S}\|^2)^{1/2}\right. \\ &\quad \left. + E\{\|\mathbf{L}\|^{j_0+1}\|\mathbf{S}\|I(\mathcal{E})\}\right\}, \end{aligned} \quad (A.8)$$

where, here and below, order-of-magnitude expressions for vectors are interpreted component by component. We can further expand terms in $\mathbf{L}$, writing $\mathbf{L} = \mathbf{L}_1 + \mathbf{L}_2$ where $\mathbf{L}_1$ is the $p \times p$ matrix with (s, t)th component equal to $b_{st} - E(b_{st})$, and $\mathbf{L}_2 = \mathbf{L} - \mathbf{L}_1$ [see (4.4) for a definition of $\mathbf{L}$]. Condition (4.5) implies that $E\|\mathbf{L}_1\|^j = O(k^{-j/2})$ and $E\|\mathbf{L}_2\|^j + E\|\mathbf{S}\|^j = O(K^{-j/2})$ for all integers j, and so we can deduce from (A.8) that the same expansion holds if we replace $\mathbf{L}$ by $\mathbf{L}_1$, replace ϵ by $\frac{1}{2}\epsilon$ in one place, and add $O(K^{-1} + k^{-(j_0+1)}K^{-1/2})$ to the right-hand side:

$$\begin{aligned} &E\{(\mathbf{M}+\mathbf{L}_1)^{-1}\mathbf{S}I(\mathcal{E})\} \\ &= \mathbf{M}^{-1}\sum_{j=0}^{j_0}\left[E\{(-\mathbf{L}_1\mathbf{M}^{-1})^j\mathbf{S}\} - E\left\{(-\mathbf{L}_1\mathbf{M}^{-1})^j\mathbf{S}I(\|\mathbf{S}\| > \epsilon)\right\}\right] \\ &\quad + O\left\{\sum_{j=0}^{j_0}\left[E\left\{\|\mathbf{L}_1\|^{2j}I\left(\|\mathbf{L}_1\| > \frac{1}{2}\epsilon\right)\right\}\right]^{1/2}(E\|\mathbf{S}\|^2)^{1/2}\right. \\ &\quad \left. + E\{\|\mathbf{L}_1\|^{j_0+1}\|\mathbf{S}\|I(\mathcal{E})\}\right\} + O\left(K^{-1} + k^{-(j_0+1)}K^{-1/2}\right). \end{aligned} \quad (A.9)$$

The result $E\{\|\mathbf{L}_1\|^{j_0+1}\|\mathbf{S}\|I(\mathcal{E})\} = O\{(E\|\mathbf{L}_1\|^{2(j_0+1)})^{1/2}K^{-1/2}\}$ follows from the property $E(\|\mathbf{S}\|^2) = O(K^{-1})$. Using the argument leading to (A.6) we can show that if j_0 is sufficiently large, $E(\|L_1\|^{2(j_0+1)}) = O(k^{-(j_0+1)}) = O(K^{-1})$, and therefore, $E\{\|\mathbf{L}_1\|^{j_0+1}\|\mathbf{S}\|I(\mathcal{E})\} = O(K^{-1})$. Moreover, $E(\mathbf{S}|\mathcal{F}_X) = \mathbf{0}$, where $\mathcal{F}_X$ denotes the sigma field generated by $\mathbf{X}_1, \mathbf{X}_2, \ldots$, and so we have $E\{(-\mathbf{L}_1\mathbf{M}^{-1})^j\mathbf{S}\} = E\{(-\mathbf{L}_1\mathbf{M}^{-1})^j E(\mathbf{S}|\mathcal{F}_X)\} = \mathbf{0}$. Also, for $j \ge 1$,

$$\begin{aligned} E\{(-\mathbf{L}_1\mathbf{M}^{-1})^j\mathbf{S}I(\|\mathbf{S}\| > \epsilon)\} &= O\left[E\left\{\|\mathbf{L}_1\|^jE(\|\mathbf{S}\|^2|\mathcal{F}_X)\right\}\right] \\ &= O(K^{-1}). \end{aligned}$$

Using (A.3) we deduce that $E\{\|\mathbf{L}_1\|^{2j}I(\|\mathbf{L}_1\| > \frac{1}{2}\epsilon)\} = O(k^{-2j} \times k^{-C_4/2})$, provided that $C_4 \ge 2(C+1)$; see (A.6). Additionally, by (A.6), $P(\mathcal{E}) = 1 - P(\widetilde{\mathcal{E}}) = 1 - O(K^{-1})$. Combining the results in this paragraph, (A.7) and (A.9), we deduce that $|E\{f(\hat{\boldsymbol{\eta}})I(\mathcal{E})\} - f(\boldsymbol{\eta}^0)| = O(K^{-1})$. The corollary follows from the latter identity and (A.6).

SUPPLEMENTARY MATERIALS

Appendix B: It contains the additional proofs. (supplement.pdf)

[Received January 2010. Revised August 2010.]

REFERENCES

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988), "An Error Components Model for Prediction of County Crop Area Using Survey and Satellite Data," *Journal of the American Statistical Association*, 83, 28–36. [362]

Bell, W. R. (2008), "Examining Sensitivity of Small Area Inferences to Uncertainty About Sampling Error Variances," in *Proceedings of Section on Survey Research Methods*, Joint Statistical Meetings 2008, Denver, Colorado; Alexandria, VA: American Statistical Association, pp. 327–333. [362]

Bell, W. R., and Huang, E. T. (2006), "Using the t-Distribution to Deal With Outliers in Small Area Estimation," in *Proceedings of Statistics Canada Symposium 2006: Methodological Issues in Measuring Population Health*, Gatineau, Quebec, Canada; Ottawa, Ontario: Statistics Canada. [369]

Chatterjee, S., Lahiri, P., and Li, H. (2008), "Parametric Bootstrap Approximation to the Distribution of EBLUP, and Related Prediction Intervals in Linear Mixed Models," *The Annals of Statistics*, 36, 1221–1245. [362]

Datta, G. S. (2009), "Model-Based Approach to Small Area Estimation," in *Handbook of Statistics: Sample Surveys: Inference and Analysis*, Vol. 29B, eds. D. Pfeffermann and C. R. Rao, Amsterdam, The Netherlands: North-Holland, pp. 251–288. [362]

Datta, G. S., and Ghosh, M. (1991), "Bayesian Prediction in Linear Models: Applications to Small Area Estimation," *The Annals of Statistics*, 19, 1748–1770. [362]

Datta, G. S., and Lahiri, P. (2000), "A Unified Measure of Uncertainty of Estimated Best Linear Unbiased Predictors in Small Area Estimation Problems," *Statistica Sinica*, 10, 613–627. [362]

Farrell, P. J., MacGibbon, B., and Tomberlin, T. J. (1997), "Empirical Bayes Estimators of Small Area Proportions in Multistage Designs," *Statistica Sinica*, 7, 1065–1083. [366]

Fay, R. E., and Herriot, R. A. (1979), "Estimates of Income for Small Places: An Application of James–Stein Procedure to Census Data," *Journal of the American Statistical Association*, 74, 269–277. [362,368]

Ghosh, M., and Maiti, T. (2004), "Small Area Estimation Based on Natural Exponential Family Quadratic Variance Function Models and Survey Weights," *Biometrika*, 91, 95–112. [362,363,365,366]

Ghosh, M., and Rao, J. N. K. (1994), "Small Area Estimation: An Appraisal" (with discussion), *Statistical Science*, 9, 55–93. [362]

Godambe, V. P., and Thompson, M. E. (1989), "An Extension of Quasi-Likelihood Estimation," *Journal of Statistical Planning and Inference*, 22, 137–152. [364,365]

Hall, P., and Maiti, T. (2006a), "Nonparametric Estimation of Mean-Squared Prediction Error in Nested-Error Regression Models," *The Annals of Statistics*, 34, 1733–1750. [362]

——— (2006b), "On Parametric Bootstrap Methods for Small-Area Prediction," *Journal of the Royal Statistical Society, Ser. B*, 68, 221–238. [362]

Jiang, J., and Lahiri, P. S. (2001), "Empirical Best Prediction for Small Area Inference With Binary Data," *Annals of the Institute of Statistical Mathematics*, 53, 217–243. [366]

——— (2006), "Mixed Model Prediction and Small Area Estimation," *Test*, 15, 1–96. [362]

Jiang, J., and Rao, J. S. (2003), "Consistent Procedures for Mixed Linear Model Selection," *Sankhyā*, 65, 23–42. [363]

Jiang, J., and Tang, E. (2011), "The Best EBLUP in the Fay–Herriot Model," *Annals of the Institute of Statistical Mathematics*, to appear. [368]

Jiang, J., Nguyen, T., and Rao, J. S. (2010), "Fence Method for Nonparametric Small Area Estimation," *Survey Methodology*, 36, 3–11. [363,368]

Jiang, J., Rao, J. S., Gu, Z., and Nguyen, T. (2008), "Fence Methods for Mixed Model Selection," *The Annals of Statistics*, 36, 1669–1692. [363]

Li, H., and Lahiri, P. (2010), "An Adjusted Maximum Likelihood Method for Solving Small Area Estimation Problems," *Journal of Multivariate Analysis*, 101, 882–892. [362]

MacGibbon, B., and Tomberlin, T. J. (1989), "Small Area Estimates of Proportions via Empirical Bayes Techniques," *Survey Methodology*, 15, 237–252. [366]

Morris, C. N. (1982), "Natural Exponential Families With Quadratic Variance Functions," *The Annals of Statistics*, 10, 65–80. [362]

——— (1983), "Natural Exponential Families With Quadratic Variance Functions: Statistical Theory," *The Annals of Statistics*, 11, 515–529. [362]

Morris, C. N., and Christiansen, C. L. (1995), "Hierarchical Models for Ranking and for Identifying Extremes With Applications," in *Bayesian Statistics*, Vol. 5, eds. J. M. Bernardo, J. O. Bereger, A. P. Dawid, and A. F. M. Smith, New York: Oxford Univ. Press. [368]

Pfeffermann, D. (2002), "Small Area Estimation: New Developments and Directions," *International Statistical Review*, 70, 125–143. [362]

Pfeffermann, D., and Correa, S. (2009), "New Bootstrap Bias Corrections With Application to Estimation of Prediction MSE in Small Area Estimation," preprint, Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, United Kingdom. [366]

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., and Rasbash, J. (1998), "Weighting for Unequal Selection Probabilities in Multilevel Models," *Journal of the Royal Statistical Society, Ser. B*, 60, 23–40. [362]

Prasad, N. G. N., and Rao, J. N. K. (1990), "The Estimation of Mean Squared Error of Small Area Estimators," *Journal of the American Statistical Association*, 85, 163–171. [362,369]

——— (1999), "On Robust Small Area Estimation Using a Simple Random Effects Model," *Survey Methodology*, 25, 67–72. [362]

Rao, J. N. K. (1999), "Some Recent Advances in Model-Based Small Area Estimation," *Survey Methodology*, 25, 175–186. [362]

——— (2003), *Small Area Estimation*, New York: Wiley. [362]

——— (2006), "Inferential Issues in Small Area Estimation: Some New Developments," *Statistics in Transition*, 7, 513–526. [362]

Singh, A. C., Stukel, D. M., and Pfeffermann, D. (1998), "Bayesian versus Frequentist Measures of Error in Small Area Estimation," *Journal of the Royal Statistical Society, Ser. B*, 60, 377–396. [362]

You, Y., and Rao, J. N. K. (2002), "A Pseudo-Empirical Best Linear Unbiased Prediction Approach to Small Area Estimation Using Survey Weights," *Canadian Journal of Statistics*, 30, 431–439. [362]

——— (2003), "Pseudo-Hierarchical Bayes Small Area Estimation Combining Unit Level Models and Survey Weights," *Journal of Statistical Planning and Inference*, 111, 197–208. [362]