NONPARAMETRIC ESTIMATION OF MEAN-SQUARED PREDICTION ERROR IN NESTED-ERROR REGRESSION MODELS

PETER HALL
U MELBOURNE & UC DAVIS

TAPS MAITI
MICHIGAN STATE UNIVERSITY

INTRODUCTION

Unbalanced nested-error regression models arise often in two-stage sample surveys. Besides the noise, a source of variation is added to explain the correlation among observations within clusters, or subjects, and to allow the analysis to borrow strength from other clusters.

Such nested-error regression models are particular cases of general linear mixed models, which often form the basis for inference about small-area means or subject-specific values.

In this talk we propose a new, nonparametric bootstrap technique for estimating the mean-squared error of predictors of mixed effects.

INTRODUCTION (2)

The new method has several attractive properties.

- It does not require specific distributional assumptions about error distributions.
- It produces positive, bias-corrected estimators of mean-squared prediction errors.
- It is easy to apply.

Although our emphasis is on small-area prediction, our methodology is equally useful for other applications, such as estimating subject- or cluster-specific random effects.

Literature Review

As indicated in the Introduction, one of the advantages of our approach is that it produces positive, bias-corrected estimators of mean-squared prediction errors. Bell (2001) and Chen and Lahiri (2002) have discussed the issue of negativity.

Kackar and Harville (1984) and Harville and Jeske (1992) studied various approximations to the mean-squared prediction error of the empirical BLUP, assuming normality in both stages. Prasad and Rao (1990) pointed out that if unknown model parameters are replaced by their estimators, then significant under-estimation of true mean-squared prediction error can still result, and introduced second-order correct mean-squared error estimators under normal models.

Datta and Lahiri (2000) extended the Prasad–Rao approach to cases where model parameters are estimated using maximum likelihood, or restricted maximum likelihood, methods. Das *et al.* (2001) gave rigorous proofs of these results under normality. Bootstrap methods in parametric settings have been suggested, for this problem, by Booth and Hobert (1998) and Lahiri (2003a), for example. See also a review paper by Lahiri (2003b). Jiang *et al.* (2002) proposed an elegant, jackknife-based, parametric approach to bias correction of the mean-squared error estimator.

MODEL

We observe data pairs (X_{ij}, Y_{ij}) generated by the model

$$Y_{ij} = \mu + X_{ij}^{\mathrm{T}}\beta + U_i + s_{ij}\, V_{ij}\,, \quad \text{for} \quad 1 \le i \le n \quad \text{and} \quad 1 \le j \le n_i\,,$$

where each $n_i \ge 2$, Y_{ij} and μ are scalars, X_{ij} is an r-vector, β is an r-vector of unknown parameters, the scalar s_{ij} is known (generally as a function of $X_{i1}, \ldots, X_{in_i}$), the U_is and V_{ij}s are totally independent, the U_is are identically distributed, the V_{ij}s are identically distributed, $E(U_i) = E(V_{ij}) = 0$ for each i, j, $E(U_i^2) = \sigma_U^2$ and $E(V_{ij}^2) = \sigma_V^2$.

All inference will be undertaken conditionally on $\mathcal{X}$, which denotes the set of explanatory data X_{ij} for $1 \le i \le n$ and $1 \le j \le n_i$.

The model (2.1) is a generalisation of the unbalanced nested-error regression model (Stukel and Rao, 1997; Rao, 2003), and is commonly used to model two-level clustered data. See also Battese *et al.* (1988), Datta and Ghosh (1991) and Rao and Choudhry (1995).

Model (2)

The model (2.1) arises through noise, in terms of the V_{ij}s, being added to a value,

$$\Theta_i = \mu + \underline{X}_i'\beta + U_i\,,$$

of the small-area modelling "parameter." Here, $\underline{X}_i = n_i^{-1}\sum_j X_{ij}$.

Our objective is to make inference about estimators of the performance of predictors of the small-area mean Θ_i, or even just the random effect U_i (in the case $\mu = 0$ and $\beta = 0$).

PREDICTORS

Put $\bar{X}_i = a_i^{-1} \sum_j s_{ij}^{-2} X_{ij}$ and $\bar{Y}_i = a_i^{-1} \sum_j s_{ij}^{-2} Y_{ij}$, where $a_i = \sum_j s_{ij}^{-2}$. The best linear unbiased predictor of Θ_i is

$$\Theta_i^{\mathrm{BLUP}} = \mu + \underline{X}_i'\beta + \rho_i \left(\bar{Y}_i - \mu - \bar{X}_i'\beta\right),$$

where $\rho_i = \sigma_U^2/(\sigma_U^2 + a_i^{-1}\,\sigma_V^2)$.

Replacing μ and β by their weighted least-squares estimators, $\tilde{\mu}$ and $\tilde{\beta}$ say, we obtain an empirical version of Θ_i^{BLUP}:

$$\widetilde{\Theta}_i^{\mathrm{BLUP}} = \tilde{\mu} + \underline{X}_i'\tilde{\beta} + \rho_i \left(\bar{Y}_i - \tilde{\mu} - \bar{X}_i'\tilde{\beta}\right).$$

PREDICTORS (2)

Here,

$$\tilde{\mu} = \left(\sum_{i=1}^{n} 1_i^{\mathrm{T}} \mathrm{W}_i^{-1} 1_i \right)^{-1} \sum_{i=1}^{n} 1_i^{\mathrm{T}} \mathrm{W}_i^{-1} \left(Y_i - \mathrm{X}_i^{\mathrm{T}} \tilde{\beta} \right),$$

$$\tilde{\beta} = \left\{ \sum_{i=1}^{n} \left(\mathrm{X}_i - 1_i \bar{X} \right) \mathrm{W}_i^{-1} \left(\mathrm{X}_i - 1_i \bar{X} \right)^{\mathrm{T}} \right\}^{-1} \sum_{i=1}^{n} (\mathrm{X}_i - 1_i \bar{X}) \mathrm{W}_i^{-1} (Y_i - \bar{Y} 1_i),$$

where 1_i is the vector of 1s of length n_i, X_i denotes the $r \times n_i$ matrix with X_{ij} as its jth column, W_i is the $n_i \times n_i$ matrix of which the (j_1, j_2)th component is $\sigma_U^2 + \delta_{j_1 j_2} s_{ij_1}^2 \sigma_V^2$, $\delta_{j_1 j_2}$ is the Kronecker delta, Y_i is the n_i-vector with jth component Y_{ij}, and

$$\bar{X} = \left(\sum_{i=1}^{n} 1_i^{\mathrm{T}} \mathrm{W}_i^{-1} 1_i \right)^{-1} \sum_{i=1}^{n} \mathrm{X}_i \mathrm{W}_i^{-1} 1_i,$$

$$\bar{Y} = \left(\sum_{i=1}^{n} 1_i^{\mathrm{T}} \mathrm{W}_i^{-1} 1_i \right)^{-1} \sum_{i=1}^{n} Y_i^{\mathrm{T}} \mathrm{W}_i^{-1} 1_i$$

denote an r-vector and a scalar, respectively.

PREDICTORS (3)

A practical form of $\widetilde{\Theta}_i^{\mathrm{BLUP}}$ is

$$\widehat{\Theta}_i^{\mathrm{BLUP}} = \hat{\mu} + \underline{X}_i'\hat{\beta} + \hat{\rho}_i \left(\bar{Y}_i - \hat{\mu} - \bar{X}_i'\hat{\beta}\right),$$

where $\hat{\mu}$, $\hat{\beta}$ and $\hat{\rho}_i$ differ from $\tilde{\mu}$, $\tilde{\beta}$ and ρ_i, respectively, in that σ_U^2 and σ_V^2 are replaced by estimators, $\hat{\sigma}_U^2$ and $\hat{\sigma}_V^2$ say (see Stukel and Rao, 1997). We wish to construct a bias-corrected estimator of the mean-squared prediction error,

$$\mathrm{MSE}_i = E\left\{ \left(\widehat{\Theta}_i^{\mathrm{BLUP}} - \Theta_i\right)^2 \,\middle|\, \mathcal{X} \right\}. \tag{1}$$

Formula for Mean Squared Error

Recall the formula for MSE_i:

$$\mathrm{MSE}_i = E\Big\{ \big(\widehat{\Theta}_i^{\mathrm{BLUP}} - \Theta_i\big)^2 \;\Big|\; \mathcal{X} \Big\} . \tag{1}$$

It can be proved that

$$\mathrm{MSE}_i = \psi_0(\xi_0) + n^{-1}\,\psi_1(\xi_1) + O\big(n^{-2}\big)\,,$$

where $\xi_0 = (\sigma_U^2, \sigma_V^2)$, $\xi_1 = (\sigma_U^2, \sigma_V^2, EU^4, EV^4)$,

$$\psi_0(\xi_0) = \frac{\sigma_U^2\, a_i^{-1}\, \sigma_V^2}{\sigma_U^2 + a_i^{-1}\, \sigma_V^2}\,,$$

$a_i = \sum_j \, s_{ij}^{-2}$, and ψ_1 is a known, smooth function.

Crucially, both ξ_0 and ξ_1 depend on the distributions of U and V only through their second and fourth moments.

Although in principle ψ_1 is known, it is generally a very complex function of the X_{ij}s and s_{ij}s, and so estimating MSE_i by $\widetilde{\mathrm{MSE}}_i = \psi_0(\hat{\xi}_0) + n^{-1}\,\psi_1(\hat{\xi}_1)$, for estimators $\hat{\xi}_0$ and $\hat{\xi}_1$, is not attractive. We suggest instead a bootstrap approach where ξ_1 is estimated implicitly.

USING THE BOOSTRAP TO ESTIMATE MEAN SQUARED ERROR

In a bootstrap approach to this problem it is sufficient to resample from empirical "approximations" to the distributions of U and V that have first, second and fourth moments which are root-n consistent for the corresponding moments of U and V.

In particular, we do not need the distributions from which we resample to actually be consistent for the distributions of U and V.

This is a variant of the moment-matching, or "wild," bootstrap method, which almost invariably addresses first, second and third, rather than first, second and fourth, moments. Examples of applications of the moment-matching bootstrap can be found in work of Fan and Li (2002), Flachaire (2002), Domínguez and Lobato (2003), Prášková (2003), Kauermann and Opsomer (2003), Li *et al.* (2003) and González Manteiga *et al.* (2004).

Moment Matching Bootstrap Algorithm

Given $z_2, z_4 > 0$ with $z_2^2 \leq z_4$, let $D(z_2, z_4)$ denote the distribution of a random variable Z, say, for which $E(Z) = 0$ and $E(Z^j) = z_j$ for $j = 2, 4$. Let $\mathcal{D}$ denote a class of such distributions, with exactly one member $D(z_2, z_4)$ for each pair (z_2, z_4).

Given estimators $\hat{\sigma}_U^2$ and $\hat{\sigma}_V^2$ of σ_U^2 and σ_V^2, for example those given by Stukel and Rao (1997), as well as estimators $\hat{\gamma}_U$ and $\hat{\gamma}_V$ of $\gamma_U = E(U^4)$ and $\gamma_V = E(V^4)$, satisfying the standard moment conditions $\hat{\sigma}_U^4 \leq \hat{\gamma}_U$ and $\hat{\sigma}_V^4 \leq \hat{\gamma}_V$, draw resamples $\mathcal{U}^* = \{U_1^*, \ldots, U_n^*\}$ and $\mathcal{V}^* = \{V_{ij}^* : 1 \leq i \leq n, 1 \leq j \leq n_i\}$ by sampling independently from the distributions $D(\hat{\sigma}_U^2, \hat{\gamma}_U)$ and $D(\hat{\sigma}_V^2, \hat{\gamma}_V)$, respectively, the distributions being the uniquely determined members of $\mathcal{D}$.

Mimicking the model given earlier, define

$$Y_{ij}^* = \hat{\mu} + X_{ij}^{\mathrm{T}} \hat{\beta} + U_i^* + s_{ij}\, V_{ij}^*\,, \quad \text{for} \quad 1 \leq i \leq n \quad \text{and} \quad 1 \leq j \leq n_i\,.$$

MOMENT MATCHING BOOTSTRAP ALGORITHM (2)

Let $\mathcal{Z}$ and $\mathcal{Z}^*$ denote the set of all pairs (X_{ij}, Y_{ij}), and the set of all pairs (X_{ij}, Y_{ij}^*), respectively. Using the data in $\mathcal{Z}^*$, compute the bootstrap versions $\hat{\mu}^*$, $\hat{\beta}^*$, $\hat{\sigma}_U^*$, $\hat{\sigma}_V^*$, $\hat{\gamma}_U^*$, $\hat{\gamma}_V^*$ and $\widehat{\Theta}_i^{*\mathrm{BLUP}}$ of $\hat{\mu}$, $\hat{\beta}$, $\hat{\sigma}_U$, $\hat{\sigma}_V$, $\hat{\gamma}_U$, $\hat{\gamma}_V$ and $\widehat{\Theta}_i^{\mathrm{BLUP}}$, respectively, and put

$$\widehat{\mathrm{MSE}}_i = E\left\{ \left(\widehat{\Theta}_i^{*\mathrm{BLUP}} - \Theta_i^*\right)^2 \,\middle|\, \mathcal{Z} \right\}; \tag{2}$$

compare (1). In (2), $\Theta_i^* = \hat{\mu} + \underline{X}_i'\hat{\beta} + U_i^*$.

The quantity $\widehat{\mathrm{MSE}}_i$ in (2) is our basic estimator of MSE_i. It can be shown to have bias of order n^{-1}.

BIAS-CORRECTING $\widehat{\mathrm{MSE}}_i$

To bias-correct $\widehat{\mathrm{MSE}}_i$ we use the double bootstrap, as follows.

Conditional on $\mathcal{U}^*$ and $\mathcal{V}^*$, draw resamples $\{U_1^{**}, \ldots, U_n^{**}\}$ and $\{V_{ij}^{**} : 1 \le i \le n, 1 \le j \le n_i\}$ by sampling independently from the distributions $D\{(\hat{\sigma}_U^*)^2, \hat{\gamma}_U^*\}$ and $D\{(\hat{\sigma}_V^*)^2, \hat{\gamma}_V^*\}$, respectively.

Let

$$Y_{ij}^{**} = \hat{\mu}^* + X_{ij}^{\mathrm{T}}\hat{\beta}^* + U_i^{**} + s_{ij}\, V_{ij}^{**}\,, \quad \text{for} \quad 1 \le i \le n \quad \text{and} \quad 1 \le j \le n_i\,,$$

and from the data pairs (X_{ij}, Y_{ij}^{**}), compute the double-bootstrap version $\widehat{\Theta}_i^{**\mathrm{BLUP}}$ of $\widehat{\Theta}_i^{\mathrm{BLUP}}$.

Define

$$\widehat{\mathrm{MSE}}_i^{\,*} = E\Big\{ \big(\widehat{\Theta}_i^{**\mathrm{BLUP}} - \Theta_i^{**}\big)^2 \,\Big|\, \mathcal{X}, \mathcal{Z}^* \Big\}\,,$$

where $\Theta_i^{**} = \hat{\mu}^* + \underline{X}_i'\hat{\beta}^* + U_i^{**}$. Then, $\widehat{\mathrm{MSE}}_i^{\,*}$ is the direct bootstrap analogue of $\widehat{\mathrm{MSE}}_i$.

BIAS-CORRECTING $\widehat{\mathrm{MSE}}_i$ (2)

The bias of $\widehat{\mathrm{MSE}}_i$ is estimated by

$$\widehat{\mathrm{bias}}_i = E\big(\widehat{\mathrm{MSE}}_i^* \;\big|\; \mathcal{Z}\big) - \widehat{\mathrm{MSE}}_i \,.$$

A simple bias-corrected estimator is

$$\widehat{\mathrm{MSE}}_i^{\mathrm{bc}} = \widehat{\mathrm{MSE}}_i - \widehat{\mathrm{bias}}_i = 2\,\widehat{\mathrm{MSE}}_i - E\big(\widehat{\mathrm{MSE}}_i^* \;\big|\; \mathcal{Z}\big) \,.$$

Other approaches can also be used.

Distributions $D(z_2, z_4)$

The simplest example of an appropriate distribution $D(1, p^{-1})$ of a random variable Z is perhaps the three-point distribution,

$$P(Z = 0) = 1 - p\,, \quad P\big(Z = \pm p^{-1/2}\big) = \tfrac{1}{2}\,p\,,$$

where $0 < p < 1$. Here, $E(Z) = 0$, $E(Z^2) = 1$ and $E(Z^4) = p^{-1}$. Therefore we may take $D(z_2, z_4)$ to be the distribution of $z_2^{1/2} Z$ when $p = z_2^2/z_4$.

The Pearson family of distributions also has potential for fitting the first four moments. If (a) the first and third moments are zero, (b) the second is $z_2 = 1$, and (c) the fourth is $z_4 > 3$, implying that tails are heavier than those of the normal distribution, then the Pearson family distribution is rescaled Student's t. The number of degrees of freedom, r, is not necessarily an integer, and is given by $z_4 = 3(r-2)/(r-4)$.

Either approach is effective in practice. While Student's t can be employed only when kurtosis is positive, this is the case in many practical situations.

Theoretical and Numerical Properties

Details are given in:

HALL, P. AND MAITI, T. Nonparametric estimation of mean-squared prediction error in nested-error regression models. *Ann. Statist.* **34**, 1733–1750.

The estimators $\widehat{\mathrm{MSE}}_i^{\mathrm{bc}}$ and $\widehat{\mathrm{bias}}_i$ converge at optimal rates, and in particular, $\widehat{\mathrm{MSE}}_i^{\mathrm{bc}} - \mathrm{MSE}_i$ and $\widehat{\mathrm{bias}}_i - \mathrm{bias}_i$ are both of order n^{-2}, as $n \to \infty$.

Numerical properties reflect this performance. In particular, in the models we treated, while the naive estimator of mean squared error suffers from substantial under-estimation, in the range 8% to 20%, the average relative bias of the bootstrap approach varies from less than 5% to less than 10% in the same setting. (The naive estimator of mean squared error is $\psi_0(\hat{\xi}_0)$, where $\hat{\xi}_0$ is obtained by replacing σ_U^2 and σ_V^2 by $\hat{\sigma}_U^2$ and $\hat{\sigma}_V^2$, respectively, in a formula given earlier for ξ_0.)