# On the Use of Recursive Residuals for Testing the Goodness of Fit of Small Area Estimation Models

Y. S. El-Horbaty [1] and L. Zhang[2]

[1] Department of Mathematics, Insurance, and Applied Statistics, Helwan University.
[2] Department of Social Statistics and Demography, University of Southampton.

## Abstract

Correct specification of the model used for small area estimation is important in order to obtain valid predictors of the target quantities and of the prediction of the mean squared error. By using recursive residuals, we construct misspecification tests for the two linear mixed models in common use for small area estimation; the area-level model and the unit-level model. We propose simple formulas for the recursive residuals that do not require repeated estimation of the variance components, and use them to form tests with asymptotic *t* distribution. The proposed tests are most powerful against nonlinear effects of the covariates. Simulation results reveal that under appropriate sorting of the sample observations, the tests possess the correct size under the null hypothesis and good power in detecting misspecification of the linear predictors.

**KeyWords:** Recursive residuals, model testing, area-level model, unit-level model, variance components

## 1. Introduction

Linear mixed models are commonly used for prediction of small area means (Rao, 2003; Jiang and Lahiri, 2006). The best linear unbiased prediction (BLUP) method is used to obtain estimates of the fixed effects and predictions of the unobserved random effects. Clearly, the working model needs to be adequately specified in order to obtain valid estimators of the true population means and of the prediction mean squared error. (PMSE).

The two most popular linear mixed models in small area estimation are the area-level model (Fay and Herriot, 1979) and the unit-level model (Battese et al., 1988). Hereafter, we abbreviate the two models as ALM and ULM, respectively. In order to accommodate for the possibility of misspecified linear predictors under these models, Jiang et al. (2011) propose the use of what they call observed best predictors (OBP). Nevertheless, Jiang et al. (2011) note in numerical examples that the BLUP method can outperform the OBP method, in terms of the PMSE. The authors suggest that testing the working model can be useful in situations where the linear predictor is *severely* misspecified.

Previously, Pan and Lin (2005) proposed the use of the cumulative sums of the estimated raw model residuals, and a simulation based two-sided test statistic. Other techniques are considered in Crainiceanu and Ruppert (2005) and Zhang and Lin (2003), which are based on likelihood ratio and score tests against a broad class of models specified under the alternative hypothesis. The latter class of models are of the semi-parametric type, where the parametric part represents the mean function of the null model, and the other part is a nonparametric function of the available covariates. This nonparametric function

is estimated using the smoothing (penalized) splines, but the power of the tests can be sensitive to the choice of the points that define the splines.

Using recursive residuals, Brown et al. (1975) proposed a test that is easy to compute for detecting shifts in the model parameters over a specific time period. The authors showed that the recursive residuals are independent where each recursive residual represents a standardized one-step-ahead prediction error. Harvey and Collier (1977) used this class of independent residuals to test for polynomial covariate effects in linear regression models with homoscedastic error variance. McGilchrist and Sandland (1979) extended the derivation of the recursive residuals to the general linear model when the variance-covariance matrix is known. See Kianifard and Swallow (1996) for a review of these (and other) uses of recursive residuals. Haslett and Haslett (2007) classified the recursive residuals as conditional residuals, and consider them to be more fundamental and useful in model diagnostics than the least squares residuals that have a marginal nature.

Application of the tests that are based on the recursive residuals to the ALM and ULM presents some potential challenges. On the one hand, the assumption of homoscedastic variance under the multiple linear regression models no longer holds under the ALM. Further, the assumption of independent errors is violated under the ULM. On the other hand, while our models are special cases of the general linear model considered by McGilchrist and Sandland (1979), their approach requires known variance-covariance matrices, which is not the case in practice. Simply replacing the variance components by their estimates can be computationally intensive if the estimation is carried out repeatedly for each recursive residual. It is also unclear whether the test statistic, obtained from plugging in these estimates, is theoretically valid.

We propose simple recursive residuals that are easy to compute. Under the ALM with independent area-level direct estimators, we propose to use the ordinary least squares (OLS) fit of the regression coefficients directly, just like in Brown et al. (1975). The OLS fit allows to establish the asymptotic N(0,1)-distribution of the resulting recursive residuals. Under the ULM, we consider a transformed model as in Fuller and Battese (1973); under which only a function of the unit-level errors remains in the model, thereby avoiding the need for estimating the variance of the random effects. Asymptotic t-tests can be constructed based on recursive residuals computed from the transformed model.

The rest of the paper is organized as follows. In Section 2, the recursive residuals are presented under a class of linear mixed models assuming known variance components. The problems of recursive estimation of the unknown parameters under the ALM and the ULM are discussed in Section 3. The proposed formulas for the recursive residuals are given in Section 4 and the resulting test statistics are presented in Section 5. Simulation results evaluating the empirical size and power of the proposed tests are summarized in Section 6. We provide some concluding remarks and directions for further work in Section 7.

## 2. Recursive residuals under a class of linear mixed models

The derivation in McGilchrist and Sandland (1979) is followed in defining the recursive residuals under a class of linear mixed models, which defines a special form of the linear model with correlated errors within domains. The approach assumes known variance-covariance matrix.

For $i = 1,...,m$, and $j = 1,...,n_i$, $n = \sum_{i=1}^{m} n_i$, let $y_{ij}$ be the $j^{th}$ response value from the $i^{th}$ domain, and let $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ be the corresponding $p \times 1$ and $q \times 1$ vectors of covariates associated with the fixed effects and random effects, respectively. The linear mixed model takes the form

[1] $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$,

where $\mathbf{Y} = (\mathbf{Y}_1^T,...,\mathbf{Y}_m^T)^T$, $\mathbf{Y}_i = (y_{i1},...,y_{in_i})^T$, $\mathbf{X} = (\mathbf{X}_1^T,...,\mathbf{X}_m^T)^T$, $\mathbf{X}_i = (\mathbf{x}_{i1},...,\mathbf{x}_{in_i})^T$, $\mathbf{Z} = diag(\mathbf{Z}_1,...,\mathbf{Z}_m)$, $\mathbf{Z}_i = (\mathbf{z}_{i1},...,\mathbf{z}_{in_i})^T$, $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects, $\mathbf{u} = (\mathbf{u}_1^T,...,\mathbf{u}_m^T)^T$, $\mathbf{u}_i$ is a $q \times 1$ vector of unobservable random effects in the $i^{th}$ domain, $\mathbf{e} = (\mathbf{e}_1^T,...,\mathbf{e}_m^T)^T$, $\mathbf{e}_i = (e_{i1},...,e_{in_i})^T$, $e_{ij}$ is the $j^{th}$ residual error in the $i^{th}$ domain. Both $\mathbf{u}$ and $\mathbf{e}$ are independently distributed where $\mathbf{u} \sim N(\mathbf{0},\mathbf{G})$, $\mathbf{G} = diag(\mathbf{G}_i)$, $\mathbf{G}_i = \text{var}(\mathbf{u}_i)$, and $\mathbf{e} \sim N(\mathbf{0},\mathbf{R})$ where $\mathbf{R}$ is not necessarily diagonal. It follows that the covariance matrix of $\mathbf{Y}$ in [1] is $\text{var}(\mathbf{Y}) = \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R}$.

Representation of the recursive estimation of the model fixed effects requires the following additional notation. Let $b_{(h)}$ denote the $h^{th}$ element of a vector $\mathbf{b}$, $\mathbf{b}_{(h)}$ denotes the first $h$ elements of the vector $\mathbf{b}$, let $\mathbf{B}_{(h)}$ denotes the first $h$ rows of a matrix $\mathbf{B}$, and $\mathbf{B}_{(h,h)}$ denotes the first $h$ rows and columns of the matrix $\mathbf{B}$. Further, let $\hat{\mathbf{b}}_h$ denote an estimated vector obtained as a function of the first $h$ observed data points. Then, model [1] can be represented in the following form

[2] $\mathbf{Y}_{(n)} = \mathbf{X}_{(n)}\boldsymbol{\beta} + \mathbf{Z}_{(n)}\mathbf{u} + \mathbf{e}_{(n)}$,

and the covariance matrix is given by $\text{var}(\mathbf{Y}_{(n)}) = \mathbf{V}_{(n,n)} = \mathbf{Z}_{(n)}\mathbf{G}\mathbf{Z}_{(n)}^T + \mathbf{R}_{(n,n)}$, where

[3] $\mathbf{V}_{(h+1,h+1)} = \begin{bmatrix} \mathbf{V}_{(h,h)} & \mathbf{v}_{(h)} \\ \mathbf{v}_{(h)}^T & v_{(h+1,h+1)} \end{bmatrix}$,

$\mathbf{v}_{(h)} = (v_{h+1,1},...,v_{h+1,h})^T$ represents the vector of covariance between the observations in the $(h+1)^{th}$ row of $\mathbf{V}_{(h+1,h+1)}$ such that $v_{t,s} = \text{cov}(y_{(t)}, y_{(s)})$; $t,s = 1,...,n$, and $v_{(h+1,h+1)}$ is the $(h+1)^{th}$ diagonal element in $\mathbf{V}_{(h+1,h+1)}$.

Under [2], the recursive residuals are defined as

[4] $W_h = \dfrac{R_{1h} - R_{2h}}{\sqrt{c_{1h} + c_{2h}}}$,

where $h > p$, $R_{1h} = y_{(h+1)} - \mathbf{Y}_{(h)}^T\mathbf{g}_{1h}$, $\mathbf{g}_{1h} = \mathbf{V}_{(h,h)}^{-1}\mathbf{v}_{(h)}$, $R_{2h} = \mathbf{Y}_{(h)}^T\mathbf{V}_{(h,h)}^{-1}\mathbf{X}_{(h)}\mathbf{g}_{2h}$, $\mathbf{g}_{2h} = (\mathbf{X}_{(h)}^T\mathbf{V}_{(h,h)}^{-1}\mathbf{X}_{(h)})^{-1}\boldsymbol{\kappa}_h$, $\boldsymbol{\kappa}_h = \mathbf{x}_{(h+1)} - \mathbf{X}_{(h)}^T\mathbf{g}_{1h}$, $c_{1h} = v_{(h+1,h+1)} - \mathbf{v}_{(h)}^T\mathbf{g}_{1h}$, and $c_{2h} = \mathbf{g}_{2h}^T\boldsymbol{\kappa}_h$. Thus, $W_h$ represents a standardized difference between the observed

$(h+1)^{th}$ value of $Y$ and its predicted value using the previous $h$ observations where the linear, unbiased, and minimum mean squared error predictor of $y_{(h+1)}$ based on $\mathbf{Y}_{(h)}$ is given by $\mathbf{Y}_{(h)}^T\mathbf{g}_{1h} + R_{2h}$. Assuming known variance components, which represent the parameters included in $\mathbf{G}$ and $\mathbf{R}_{(h,h)}$, it follows that $W_h \overset{ind}{\sim} N(0,1)$, $h > p$. Note that in theory, the parameters of $\mathbf{V}_{(h,h)}$ has to be known. However, in practice $\mathbf{V}_{(h,h)}$ needs to be estimated accurately in order that $W_h$ converges to the normal distribution. The problems of estimating $\mathbf{V}_{(h,h)}$ under the ALM and the ULM are discussed next.

## 3. Recursive residuals under small area models

In the previous section, the definition of $W_h$ in [4] requires known covariance $\mathbf{V}_{(h,h)}$. A plug-in expression using the estimated values would require recursive estimation based on the first $h$ observations according to an appropriate sorting. Discussion of the choice of data sorting is postponed to Section 5. In this section, the ALM and the ULM are represented as special cases of [1], and the problems of computing the recursive residuals are discussed.

Consider first the ALM and denote by $\tilde{y}_i$ the direct survey estimate of the $i^{th}$ domain mean, by $\mathbf{x}_i$ the associated $p\times 1$ vector of covariates measured at the domain level, and by $e_i$ the associated sampling error, where $i = 1,...,m$ and $m > p$. The ALM is given by

[5] $\tilde{\mathbf{Y}}_{(m)} = \mathbf{X}_{(m)}\boldsymbol{\beta} + \mathbf{u}_{(m)} + \mathbf{e}_{(m)}$,

where $\tilde{\mathbf{Y}}_{(m)} = (\tilde{y}_1,...,\tilde{y}_m)^T$, $\mathbf{X}_{(m)} = (\mathbf{x}_1,...,\mathbf{x}_m)^T$, $\mathbf{u}_{(m)} = (u_1,...,u_m)^T$, $\mathbf{e}_{(m)} = (e_1,...,e_m)^T$. We assume that $\mathbf{u}_{(m)}$ and $\mathbf{e}_{(m)}$ are independent such that $\mathbf{u}_{(m)} \sim N(\mathbf{0}, A\mathbf{I}_{(m,m)})$ and $\mathbf{e}_{(m)} \sim N(\mathbf{0}, \mathbf{D}_{(m,m)})$ where $A = \text{var}(u_i)$, $\mathbf{D}_{(m,m)} = diag(D_1,...,D_m)$ and the $D_i's$ are known sampling variances. Note that in model [1] if $n_i = 1$, for $i = 1,...,m$, and $q = 1$, then $\mathbf{G}_i$ reduces to $A$, $\mathbf{G} = A\mathbf{I}_{(m,m)}$, and $\mathbf{R}_{(n,n)}$ reduces to $\mathbf{D}_{(m,m)}$. Thus, model [5] represents a special case of model [1] with $\mathbf{Z}$ reduced to the identity matrix of order $m$. Let $\text{var}(\tilde{\mathbf{Y}}_{(m)}) = A\mathbf{I}_{(m,m)} + \mathbf{D}_{(m,m)} = \mathbf{Q}_{(m,m)}$. Under model [5], the recursive residuals extend the definition in Brown et al. (1975) and are given by

[6] $$W_h^{ALM} = \frac{\tilde{y}_{(h+1)} - \mathbf{x}_{(h+1)}^T\hat{\boldsymbol{\beta}}_h}{\sqrt{\text{var}(\tilde{y}_{(h+1)} - \mathbf{x}_{(h+1)}^T\hat{\boldsymbol{\beta}}_h)}}, \quad h = p+1,...,m$$

where

[7] $\hat{\boldsymbol{\beta}}_h = (\mathbf{X}_{(h)}^T\mathbf{Q}_{(h,h)}^{-1}\mathbf{X}_{(h)})^{-1}\mathbf{X}_{(h)}^T\mathbf{Q}_{(h,h)}^{-1}\tilde{\mathbf{Y}}_{(h)}$,

and

[8] $\text{var}(\tilde{y}_{(h+1)} - \mathbf{x}_{(h+1)}^T\hat{\boldsymbol{\beta}}_h) = (A + D_{h+1}) + \mathbf{x}_{(h+1)}^T(\mathbf{X}_{(h)}^T\mathbf{Q}_{(h,h)}^{-1}\mathbf{X}_{(h)})^{-1}\mathbf{x}_{(h+1)}$.

For a given domain, the ULM is given by

[9] $\mathbf{Y}_{(n_i)} = \mathbf{X}_{(n_i)}\boldsymbol{\beta} + \mathbf{1}_{(n_i)}u_i + \mathbf{e}_{(n_i)}$,

where $\mathbf{1}_{(n_i)}$ is a vector of ones. Thus, by denoting $\mathbf{Y}_{(n_i)} = \mathbf{Y}_i$, $\mathbf{X}_{(n_i)} = \mathbf{X}_i$, and $\mathbf{e}_{(n_i)} = \mathbf{e}_i$ model [9] is seen to be a special case of model [1] with $q = 1$, and $\mathbf{G}_i = \sigma_u^2\mathbf{I}_{(n_i,n_i)}$. The covariance matrix under [9] is given by

[10] $\text{var}(\mathbf{Y}_{(n_i)}) = \mathbf{V}_{(n_i,n_i)} = \sigma_u^2\mathbf{1}_{(n_i)}\mathbf{1}_{(n_i)}^T + \sigma_e^2\mathbf{I}_{(n_i,n_i)}$

$$= \sigma_e^2\left(\mathbf{I}_{(n_i,n_i)} - n_i^{-1}\mathbf{1}_{(n_i)}\mathbf{1}_{(n_i)}^T\right) + (\sigma_e^2 + n_i\sigma_u^2)n_i^{-1}\mathbf{1}_{(n_i)}\mathbf{1}_{(n_i)}^T.$$

The ULM for all areas can be represented as

[11] $\mathbf{Y}_{(n)} = \mathbf{X}_{(n)}\boldsymbol{\beta} + \mathbf{H}_{(n)}\mathbf{u}_{(m)} + \mathbf{e}_{(n)}$,

where $n = \sum_{i=1}^{m} n_i$, $\mathbf{u}_{(m)}$ is defined below [5],

[12] $$\mathbf{H}_{(n)} = \begin{bmatrix} \mathbf{1}_{(n_1)} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{(n_2)} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{(n_m)} \end{bmatrix},$$

and

[13] $\text{var}(\mathbf{Y}_{(n)}) = \mathbf{V}_{(n,n)} = \sigma_u^2\mathbf{H}_{(n)}\mathbf{H}_{(n)}^T + \sigma_e^2\mathbf{I}_{(n,n)}$.

Substituting $\mathbf{V}_{(n,n)}$ from [13] into [3], the recursive residuals can be defined as in [4]. We denote them by $W_h^{ULM}$.

Under the definition of the recursive residuals in [6], $\hat{\boldsymbol{\beta}}_h$ is estimated using the first $h$ observations and it depends on the unknown variance component $A$, which appears in the numerator and the denominator of $W_h^{ALM}$. There are two potential problems. The first one concerns the precision of the estimator of $A$ when $h$ is small, $h > p$, and the second one is the asymptotic independence between $W_h^{ALM}$ and $W_{h'}^{ALM}$, for all $h \neq h'$ and $h, h' > p$. A third, related issue is the choice of the $h$ observations. As shown in Section 5, this choice is typically based on the values of the covariates rather than the domains to which the observations belong. It can then happen that only few domains are covered when $h$ is small, which can have undesirable effects on the variance component estimates.

In the next section, we propose formulas for recursive residuals that are easy to compute and establish their theoretical properties. Test statistics based on them are presented in Section 5.

## 4. Modified definitions of recursive residuals

Below we propose recursive residuals under ALM and ULM that do not require recursive estimates of the variance components.

Instead of [6], we define the $h^{th}$ recursive residual as

[14] $\zeta_h^{ALM} = \dfrac{\widetilde{y}_{(h+1)} - \mathbf{x}_{(h+1)}^T \hat{\boldsymbol{\beta}}_h^{OLS}}{\sqrt{\text{var}(\widetilde{y}_{(h+1)} - \mathbf{x}_{(h+1)}^T \hat{\boldsymbol{\beta}}_h^{OLS})}}$,

where

[15] $\hat{\boldsymbol{\beta}}_h^{OLS} = (\mathbf{X}_{(h)}^T \mathbf{X}_{(h)})^{-1} \mathbf{X}_{(h)}^T \widetilde{\mathbf{Y}}_{(h)}$,

[16] $\text{var}(\widetilde{y}_{(h+1)} - \mathbf{x}_{(h+1)}^T \hat{\boldsymbol{\beta}}_h^{OLS}) = (\hat{A} + D_{h+1}) + \mathbf{x}_{(h+1)}^T (\mathbf{X}_{(h)}^T \mathbf{X}_{(h)})^{-1} \mathbf{X}_{(h)}^T \hat{\mathbf{Q}}_{(h,h)} \mathbf{X}_{(h)} (\mathbf{X}_{(h)}^T \mathbf{X}_{(h)})^{-1} \mathbf{x}_{(h+1)}$,

and $\hat{\mathbf{Q}}_{(h,h)}$ is obtained from $\mathbf{Q}_{(h,h)}$ by replacing $A$ with a consistent estimator $\hat{A}$ that is obtained using the direct estimates for all available domains. Note that $\zeta_h^{ALM}$ depends on $A$ only in its denominator, which represents the exact variance of $\{\widetilde{y}_{(h+1)} - \mathbf{x}_{(h+1)}^T \hat{\boldsymbol{\beta}}_h^{OLS}\}$. Thus, replacing $A$ by $\hat{A}$ implies that asymptotically $\zeta_h^{ALM} \sim N(0,1)$, and each $\zeta_h^{ALM}$ and $\zeta_{h'}^{ALM}$ ($h \neq h'$, $h,h' > p$) are asymptotically independent as $m \to \infty$.

For the ULM, we first transform the model and then apply the formula of McGilchrist and Sandland (1979. The transformed model is given by

[17] $\mathbf{KY}_{(n)} = \mathbf{KX}_{(n)} \boldsymbol{\beta} + \mathbf{KH}_{(n)} \mathbf{u} + \mathbf{Ke}_{(n)}$,

where

[18] $\mathbf{K} = \begin{bmatrix} \mathbf{K}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{K}_2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{K}_m \end{bmatrix}$,

[19] $\mathbf{K}_i = \mathbf{L}_{(n_i - 1)}$,

[20] $\mathbf{L}_{(n_i)} = \mathbf{I}_{(n_i, n_i)} - \mathbf{1}_{(n_i)} (\mathbf{1}_{(n_i)}^T \mathbf{1}_{(n_i)}) \mathbf{1}_{(n_i)}^T$.

This is an oft-used transformation. For instance, Fuller and Battese (1973) used it for testing normality of the observations under the ULM. Note that by [10] $\text{var}(\mathbf{K}_i \mathbf{Y}_{(n_i)}) = \sigma_e^2 \mathbf{K}_i \mathbf{K}_i^T$, which does not depend on $\sigma_u^2$. Also, note that $\mathbf{KH}_{(n)} = \mathbf{0}_{(n-m)}$ and hence

[21] $\text{var}(\mathbf{KY}_{(n)}) = \sigma_e^2 \mathbf{K}\mathbf{K}^T$.

Deriving the recursive residuals under [17] is achieved by replacing, in Section 2, $\mathbf{Y}_{(n)}$ by $\mathbf{KY}_{(n)}$, $\mathbf{X}_{(n)}$ by $\mathbf{KX}_{(n)}$, $\mathbf{e}_{(n)}$ by $\mathbf{Ke}_{(n)}$, and $\mathbf{V}_{(n)}$ by $\mathbf{K}\mathbf{K}^T$. After performing these replacements, equation [4] can be recomputed for model [17] where we denote by $\zeta_h^{TULM}$

the value of $W_h$ under the transformed ULM. Then, $\zeta_h^{TULM} \sim N(0, \sigma_e^2)$. Note that the proposed computations of the recursive residuals under model [17] do not require the estimation of $\sigma_u^2$.

## 5. Test statistics

The purpose of the previous section was to compute recursive residuals that possess the desirable properties of being independent and normally distributed with constant variance. This was achieved only asymptotically under the ALM and is satisfied under the transformed ULM. Testing a nonlinear covariate effect under the two models is accomplished after sorting the observations by the values of the variable of interest. This is apparent under the ALM. For the ULM, we argue that testing the nonlinear specification of a covariate is also possible under the transformed ULM. This is because the nonlinearity effect of any given covariate continues to hold also under the transformed version of that covariate, which can be detected using the recursive residuals $\zeta_h^{TULM}$.

If the correct functional form of a given covariate is not represented in the fitted model, then this misspecification causes the recursive residuals to become systematically positive or negative over a wide range of values of the misspecified covariate. This is clearly the case when the misspecified covariates are nonlinearly related to the covariates in the mean function of the working model. Examples of commonly exhibited nonlinear covariate effects in practice are given in Section 6. Thus, the sum of the recursive residuals tends to depart from zero and hence can be used to quantify this type of misspecification. The proposed test statistics for the ALM and ULM are given below.

For the ALM, the test statistic is benefiting from the independence between the sample mean and the sample variance of the asymptotically independent recursive residuals $\zeta_h^{ALM}$ defined in [14]. The test statistic is given by

$$[22]\quad T_{ALM} = \frac{(m-p)^{-1/2} \sum_{h=p+1}^{m} \zeta_h^{ALM}}{\sqrt{(m-p-1)^{-1} \sum_{h=p+1}^{m} (\zeta_h^{ALM} - \zeta_*^{ALM})^2}},$$

where $\zeta_*^{ALM} = (m-p)^{-1} \sum_{h=p+1}^{m} \zeta_h^{ALM}$. Under the null hypothesis that the ALM is correctly specified, $T_{ALM} \sim t_{m-p-1}$ as $m \to \infty$. When misspecification is present, the magnitude of $\zeta_*^{ALM}$ increases in absolute value and the numerator of [22] increases leading to a bigger value of $T_{ALM}$.

For the ULM, the test procedures begin by applying the transformation as given in Section 4. The number of observations under the transformed ULM is reduced to $n_* = \sum_{i=1}^{m} (n_i - 1) = n - m$ because the maximum number of linearly independent rows in $\mathbf{L}_{(n_i)}$, which represents an orthogonal projection on the complement space spanned by the

unit vector $\mathbf{1}_{(n_i)}$, is $n_i - 1$ rows. Note that the transformation in [17] results in deleting the intercept term and hence the number of covariates under the transformed model reduces to $p - \lambda$ where $\lambda$ equals one when the original model involves an intercept and equals zero otherwise. Moreover, when $\lambda$ equals one, the first column in $\mathbf{KX}_{(n)}$ will be the zero vector and should be removed from this matrix when the recursive residuals $\zeta_h^{TULM}$, $h = (p - \lambda + 1), ..., n_*$, are computed.

Sorting the data comes next. Note that each observation represents under the transformed model the original observation under the ULM minus its domain mean value. Thus, data sorting can be achieved by sorting the observations under [17] by the corresponding values of the original covariate under test. A better performance of this test can be expected when the latter sorting is performed regardless to which domain each observation belongs. The prescribed data sorting is reasonable in this case because the objective of the test is to detect a nonlinear misspecification in the functional relationship between response and the explanatory variables.

The proposed test statistic is given by

[23] $$T_{TULM} = \frac{(n_* - p + \lambda)^{-1/2} \sum_{h=p-\lambda+1}^{n_*} \zeta_h^{TULM}}{\sqrt{(n_* - p + \lambda - 1)^{-1} \sum_{h=p-\lambda+1}^{n_*} (\zeta_h^{TULM} - \zeta_*^{TULM})^2}},$$

where $\zeta_*^{TULM} = (n_* - p + \lambda)^{-1} \sum_{h=p-\lambda+1}^{n_*} \zeta_h^{TULM}$. Under the null hypothesis that the ULM is correctly specified (hence the transformed ULM), $T_{TULM} \sim t_{n_* - p + / - 1}$ as $n_* \to \infty$.

## 6. Simulation studies

The performance of the proposed test statistics $T_{ALM}$ and $T_{TULM}$ has been evaluated by running separate simulation studies to assess the size and power of each test. In order to evaluate the size of the tests, the sample data were generated under the correct model and the recursive parameter estimates and residuals were computed under the same model. Then, each test statistic was computed under the corresponding correct model. The process was repeated $r = 2000$ times. The size of the test was calculated by computing the proportion of times that it rejects the null hypothesis that the fitted model is correctly specified. The test rejects the null hypothesis when the value of the test statistic exceeds the chosen critical value that is determined by the *t*-distribution, as indicated under [22] and [23].

### 6.1 Area-level model

We generated the direct estimates under the ALM as

[24] $$\widetilde{y}_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + u_i + e_i$$

where $i = 1,...,110$ , $\beta_0 = 1$ , $\beta_1 = 1$ , $\beta_2 = 3$ , $\beta_3 = 1$ , $\beta_4 = 1$ , $x_1 \sim U(1,9)$ , $x_2 \sim U(0.1,3)$ , $x_3 \sim N(2.0.2)$ , and $x_4 \sim N(2.0.2)$ . The random effects were generated as $u_i \sim N(0, A)$ , $A = 1$ and the sampling errors as $e_i \sim N(0, D_i)$ where $D_i \in [0.5,1.5]$. Thus, the left-hand side of [24] is generated by adding $u_i$ and $e_i$ to the mean function of the model. This process was replicated $r = 2000$ times.

In order to assess the power of the test, we generate the data from four models given by

[25] $\tilde{y}_i = \beta_0 + \beta_1 x_{1i} + \beta_2 \ln x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + u_i + e_i$,

[26] $\tilde{y}_i = \beta_0 + \beta_1 x_{1i} + 0.15 x_{1i}^2 + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + u_i + e_i$,

[27] $\tilde{y}_i = \beta_0 + \beta_1 x_{1i} + 0.15 x_{1i}^2 + \beta_2 \ln x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + u_i + e_i$,

[28] $\tilde{y}_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + 2x_{5i} + u_i + e_i$,

where $x_5 \sim N(\ln x_2, 1)$ . We proceeded by fitting model [24].

The four models [25-28] are examples of model misspecification where a nonlinear effect of one (or more) of the covariates is not accounted for under the fitted model [24]. This kind of misspecification is frequently recognized in practice, as commented in Pan and Lin (2005). To ensure a reasonable power of the test, the observations are sorted in advance by the values of the covariate that is being tested for misspecification. Extension of this paradigm is obtained by assuming that more than one covariate is misspecified under the assumed model. In this case, we propose sorting the observations by the fitted values before computing the test statistic. The empirical size and power of the test are summarized in Table 1.

Table 1. Proportions of rejection of the null model [24]

| Nominal Level | Misspecified Model | | | | |
|---|---|---|---|---|---|
| | Correct Model | Model [25] | Model [26] | Model [27] | Model [28] |
| 5% | 4.98 | 100 | 97.1 | 100 (73.6) | 96.4 |
| 2.5% | 2.51 | 99.9 | 94.5 | 99.8 (58.7) | 89.4 |

The nominal levels in Table 1 refer to the probability that the proposed test statistic exceeds the 95 and 97.5 percentiles of the *t*-distribution with $m - p - 1 = 104$ degrees of freedom, where $p = 5$ . The results in the table indicate that the test possesses the correct size when the fitted model is the correct model. Results on the power of the test are summarized in the remaining columns of the table. When $\ln x_2$ is the correct form of the second covariate in [24], sorting the observations by the values of this variable before computing $T_{ALM}$ yields extremely high powers whereby the test rejects the null hypothesis in 100% of the times at the 5% nominal level and 99.9% of the times at the 2.5% nominal level.

Fitting model [24] when [26] is the correct model, i.e. ignoring $x_1^2$, yields a power of 97.1% and 94.5% at the 5% and 2.5% nominal levels, respectively, where in this case the observations are sorted by the values of $x_1$. When model [27] was used for generating the data, we ran the test after sorting the observations by the values of $x_2$. The power of the test when sorting the observations by the fitted values is also given in Table 1 between the brackets. Fitting model [24] instead of [27] yields good powers under the two sorting methods. However, soring by the values of a single misspecified variable produces a significantly higher power than sorting by the values of the linear combinations of all the variables (i.e. the fitted values).

Unlike the previous examples of model misspecification, when $x_5$ is an important variable in model [28] but is absent in model [24], sorting the observations by the values of $x_2$ (that is nonlinearly related to $x_5$) is necessary to obtain a high power of the test. This indicates that rejecting the null hypothesis may not be because of incorrectly specifying $x_2$ but rather omitting $x_5$, which is an important covariate in the true model and nonlinearly related to $x_2$. Under this scenario, the test captures the incorrect specification 96.4% and 89.4% of the times at the corresponding 5% and 2.5% nominal levels. Although the test is capable of detecting this example of model misspecification, it would be challenging to improve the mean function of the model if $x_5$ is unknown.

**6.2 Unit-level model**

In this part, the performance of the proposed test statistic in [23] is assessed with respect to the transformed version of the ULM using the transformation matrix in [20]. Under the null hypothesis that the fitted model is correctly specified, $T_{TULM}$ is compared to the critical values of the *t*-distribution with $n-m-p$ degrees of freedom to obtain the size of the test. We generated data from the following ULM

[29] $y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \beta_4 x_{4ij} + u_i + e_{ij}$,

where $i = 1,...,20$, $j = 1,...,7$, for all $i$, $\beta_0 = 1$, $\beta_1 = 1$, $\beta_2 = 3$, $\beta_3 = 1$, $\beta_4 = 1$, $x_1 \sim U(1,9)$, $x_2 \sim U(0.1,3)$, $x_3 \sim N(2,0.2)$, $x_4 \sim N(2,0.2)$, $u_i \sim N(0,\sigma_u^2)$, $e_{ij} \sim N(0,\sigma_e^2)$, $\sigma_u^2 = 1$, and $\sigma_e^2 = 1$.

The test statistic was computed under the transformed ULM that can be represented as

[30] $y_{ij}^* = \beta_1 x_{1ij}^* + \beta_2 x_{2ij}^* + \beta_3 x_{3ij}^* + \beta_4 x_{4ij}^* + e_{ij}^*$,

where $y_{ij}^* = y_{ij} - \sum_{j=1}^{7} y_{ij}/7$, $x_{hij}^* = x_{hij} - \sum_{j=1}^{7} x_{hij}/7$, $h = 1,...,4$, and $e_{ij}^* = e_{ij} - \sum_{j=1}^{7} e_{ij}/7$.

Note that only $n_i - 1 = 6$ observations per domain will be used under [30] in order to compute $T_{TULM}$ and thus the number of observations becomes $n_* = n - m = 140 - 20 = 120$. Corresponding to the 5% and 2.5% nominal levels, the

size of the test is obtained by finding the proportion of times (out of $r = 2000$ replicates) that $|T_{TULM}| > t_{\alpha/2,116}$ where $\alpha$ denotes the nominal level.

In order to assess the power of the test, we generate the unit-level data from the following models

[31] $y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 \ln x_{2ij} + \beta_3 x_{3ij} + \beta_4 x_{4ij} + u_i + e_{ij}$,

[32] $y_{ij} = \beta_0 + \beta_1 x_{1ij} + 0.15 x_{1ij}^2 + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \beta_4 x_{4ij} + u_i + e_{ij}$,

[33] $y_{ij} = \beta_0 + \beta_1 x_{1ij} + 0.15 x_{1ij}^2 + \beta_2 \ln x_{2ij} + \beta_3 x_{3ij} + \beta_4 x_{4ij} + u_i + e_{ij}$,

[34] $y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \beta_4 x_{4ij} + 2x_{5ij} + u_i + e_{ij}$,

where $x_5 \sim N(\ln x_2, 1)$, but fitted model [30] to compute the test statistic. In doing so, we emphasize the need to sort the observations by the values of the covariate under test, regardless of the domain membership of these observations. When more than one covariate is suspected of being misspecified under the ULM, we sort the observations by the fitted values obtained under [29]. The empirical power of the test under the above four scenarios is summarized in Table 2. The results when the observations are sorted by the fitted values are recorded between the brackets therein.

Table 2. Proportions of rejection of the null model [30]

| Nominal Level | Misspecified Model | | | | |
|---|---|---|---|---|---|
| | Correct Model | Model [31] | Model [32] | Model [33] | Model [34] |
| 5% | 5.00 | 98.8 | 98.5 | 97.0 (96.9) | 99.9 |
| 2.5% | 2.50 | 97.6 | 96.3 | 93.0 (93.2) | 99.2 |

By Table 2, the empirical size of the test when the model is correctly specified equals to the corresponding nominal level. In addition, the test performs well in terms of its power under the four misspecified functional forms in [31-34] (i.e., missing $\ln x_2$, $x_1^2$, $x_1^2$ & $\ln x_2$, and $x_5$) as shown in Table 2. The power of the test when both the covariates $x_1$ and $x_2$ are misspecified (model [33]) was obtained after sorting the observations by the values of $x_2$.

A result that worth further mentioning is that sorting the observations by the fitted values under model [29] yields similar powers to the case of sorting the observations by the values of $x_2$. This may be due to the dominance of $x_2$ with $\beta_2 = 3$ on the fitted values such that both ways of sorting yield similar powers of

the test. Further study is needed to find sorting methods that would maintain higher powers.

## 7. Conclusion and further work

This paper develops new goodness of fit test statistics that utilize the concept of recursive residuals, with application to two models in common use for small area estimation. We propose simple formulas of the recursive residuals, establish their theoretical properties and demonstrate their usefulness via simulation studies.

Some words of caution seem warranted. Under the ALM, the independence of the recursive residuals depends on there being a sufficiently large number of domains and the normality of the random effects. Failure to satisfy one of these two conditions may influence the convergence of the test statistic $T_{ALM}$ to the $t$-distribution. In such cases, one may consider the use of bootstrap samples for approximating the distribution of the test statistic.

Under the ULM, a domain needs to have sample size greater than one in order to be included in the transformed model. A domain may also become ineffective when the observations within the domain are all the same. The effective sample size is reduced under the transformed model. For example, the sample size that is used to compute $T_{TULM}$ is $n_* = n/2$ when each domain contains two distinct observations. Extension of the transformation approach to models with more complicated covariance structures seems possible. However, one needs to be careful about the choice of the transformation matrix and the method of estimating the covariance matrix of the residual errors. An exact linear dependence between the columns of $\mathbf{X}$ and $\mathbf{Z}$ can, for example, affect the covariate to be tested when the ULM is transformed as shown in [17].

## Acknowledgements

This research has been partially supported by the Southampton Statistical Sciences Research Institute (S3RI), University of Southampton.

## References

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

Brown, R. L., Durbin, J. and Evans, J. M. (1975). Techniques of testing the constancy of regression relationships over time (with discussion). Journal of Royal Statistical Society B, 37, 149-192.

Crainiceanu and Ruppert (2005). Exact likelihood ratio tests for penalised splines. Biometrika, 92, 91-103.

Fay, R. E. and Herriot, R. A. (1979). Estimation of income for small places: an application of James-Stein procedure to census data. Journal of American Statistical Association, 74, 269-277.

Fuller, W. A. and Battese, G. E. (1973). Transformations for estimation of linear models with nested-error structure. Journal of American Statistical Association, 68, 626-632.

Harvey, A. C. and Collier, P. (1977). Testing for functional misspecification in regression analysis. Journal of Econometrics, 6, 103-119.

Haslett, J. and Haslett, S. J. (2007). The three basic types of residuals for a linear model. International Statistical Review, 75, 1-24.

Jiang, J. and Lahiri, P. (2006). Mixed model prediction and small area estimation. Test, 15, 1-96.

Jiang, J., Nguyen, T. and Rao, J. S. (2011). Best predictive small area estimation. Journal of American Statistical Association, 106, 732-745.

Kianifard, F. and Swallow, W.H. (1996). A review of the development and application of recursive residuals in linear models. Journal of American Statistical Association, 91, 391-400.

McGilchrist, C. A. and Sandland, R. L. (1979). Recursive estimation of the general linear model with dependent errors. Journal of Royal Statistical Society B, 41, 65-68.

Pan, Z. and Lin, D. Y. (2005). Goodness-of-fit methods for generalized linear mixed models. Biometrics, 61, 1000-1009.

Rao, J. N. K. (2003). Small Area Estimation. New Jersey: John Wiley & Sons, Inc.

Zhang, D. and Lin, X. (2003). Hypothesis testing in semiparametric additive mixed models. Biostatistics 4, 57–74.