TEST manuscript No.
(will be inserted by the editor)

Goodness-of-fit test with a robustness feature

Jiming Jiang · Mahmoud Torabi

Received: July 20, 2020 / Accepted: date

Abstract We develop a method originally proposed by R. A. Fisher into a general procedure, called tailoring, for deriving goodness-of-fit tests that are guaranteed to have a χ^2 asymptotic null distribution. The method has a robustness feature that it works correctly in testing a certain aspect of the model while some other aspect of the model may be misspecified. We apply the method to small area estimation. A connection, and difference, to the existing specification test is discussed. We evaluate performance of the tests both theoretically and empirically, and compare the performance with several existing methods. Our empirical results suggest that the proposed test is more accurate in size, and has either higher or similar power compared to the existing tests. The proposed test is also computationally less demanding than the specification test and other comparing methods. A real-data application is discussed.

Keywords Goodness-of-fit tests · Maple · Model diagnostics · Robustness · Small area estimation · Tailoring

Mathematics Subject Classification (2010) 62F05 · 62J05 · 62D99

J. Jiang
Department of Statistics, University of California, Davis, One Shields Avenue, Davis CA 95616, USA
Tel.: +1-530-752-1761
Fax: +1-530-752-7099
E-mail: jimjiang@ucdavis.edu

M. Torabi
Department of Community Health Sciences, University of Manitoba, 750 Bannatyne Ave, Winnipeg, Manitoba, R3E 0W3, Canada
Tel.: +1-204-272-3136
Fax: +1-204-789-3905
E-mail: Mahmoud.Torabi@umanitoba.ca

1 Introduction

Goodness-of-fit tests for mixed models, or mixed effects models, have received considerable attention in recent literature (e.g., Jiang 2001, Claeskens and Hart 2009, Dao and Jiang 2016). Such tests are relevant to many practical problems. For example, mixed effects models are extensively used in small area estimation (SAE; e.g., Rao and Molina 2015). Here the term small area typically refers to a population for which reliable statistics of interest cannot be produced based on direct sampling from the population due to certain limitations of the available data. Examples of small areas include a geographical region (e.g., a state, county, municipality), a demographic group (e.g., a specific age $\times$ sex $\times$ race group), a demographic group within a geographic region, etc. Statistical models, especially mixed effects models, have played key roles in improving small area estimates by borrowing strength from relevant sources. It is known, however, that in case of model misspecification, the traditional empirical best linear unbiased prediction (EBLUP) method may lose efficiency. See, for example, Jiang, Nguyen and Rao (2011). In case of model misspecification, an alternative method, known as observed best prediction (OBP), is shown to be more accurate than the EBLUP. On the other hand, when the underlying model is correctly specified, EBLUP is known to be more efficient than OBP (e.g., Jiang, Nguyen and Rao 2011, 2015). Therefore, it is important, in practice, to know whether or not the assumed model is appropriate in order to come up with a more efficient SAE strategy.

A standard assumption for mixed effects models in general (e.g., Jiang 2007), is that the random effects are normally distributed. This assumption has had substantial impact on many aspects of the inference. For example, estimation of the mean squared prediction errors of small area predictors is an important issue in SAE (e.g., Rao and Molina 2015). The well-known Prasad-Rao method (Prasad and Rao 1990) depends on the normality assumption and may not be accurate if the assumption fails (e.g., Lahiri and Rao 1995). Also, prediction interval obtained via parametric bootstrap methods (e.g., Chatterjee, Lahiri and Li 2008) depends heavily on the normality assumption. The normality assumption is even more critical for inference about generalized linear mixed models (GLMMs; e.g., Jiang 2007). See, for example, Jiang and Nguyen (2009). Although there are strategies that are less dependent on the normality assumption, those strategies are often less efficient than the normality-based method when the normality assumption actually holds, or approximately holds. Thus, it is important to check the validity of the normality assumption so that an appropriate, or more efficient, method can be used for the inference.

In the literature of mixed effects models, such problems as discussed above have to do with mixed model diagnostics; see, for example, Pierce (1982), sec. 2.4.1 of Jiang (2007), Claeskens and Hart (2009). Jiang (2001) proposed a χ^2-type goodness-of-fit test for linear mixed model (LMM) diagnostics, whose asymptotic null distribution is a weighted χ^2, where the weights are eigenvalues of some nonnegative definite matrix. Claeskens and Hart (2009) proposed

an alternative approach to the χ^2 test for checking the normality assumption in LMM. The authors considered a class of distributions that include the normal distribution as a reduced, special case. The test is based on the likelihood-ratio test (LRT) that compares the "estimated distribution" and the null distribution (i.e., normal). A model selection procedure via the information criteria is used to determine the larger class of distributions for the LRT. In particular, the asymptotic null distribution is in the form of the distribution of $\sup_{l\geq 1}\{2Q_l/l(l+3)\}$, where $Q_l = \sum_{q=1}^{l} \chi^2_{q+1}$, l is the order of polynomial, and $\chi^2_2, \chi^2_3, \ldots$ are independent such that χ^2_j has a χ^2 distribution with j degrees of freedom, $j \geq 2$.

The χ^2-type tests depend on the choice of cells, based on which the observed and expected cell frequencies are evaluated. As noted by Jiang and Nguyen (2009), performance of the χ^2 test is sensitive to the choice of the cells, and there is no "optimal choice" of such cells known in the literature. On the other hand, the Claeskens-Hart test depends on the choice of the information criterion. As is well known, there are different versions of the information criteria, such as AIC (Akaike 1973), BIC (Schwarz 1978), HQ (Hannan and Quinn 1979). The difference in the performance of the test by different information criteria is unclear. Furthermore, the weighted-χ^2 asymptotic null distribution of Jiang (2001) depends on eigenvalues of a certain matrix, whose expressions are complicated, and involve unknown parameters. These parameters need to be estimated in order to obtain the critical values of the tests. Due to such a complication, Jiang (2001) suggests to use a Monte-Carlo method to compute the critical value; but, by doing so, the usefulness of the asymptotic result may be undermined. Similarly, the asymptotic distribution of the Claeskens-Hart test is not simple and involves supreme of partial sums of χ^2 random variables.

It might be argued that, in today's computer era, having a χ^2 asymptotic distribution is, perhaps, not as important as in the past. However, there are, still, attractive features of the χ^2 limiting distribution that are worth pursuing. First, the χ^2 distribution corresponds to the right standardization–it is the "square" of the norm of a multivariate normal vector. In this regard, anything other than χ^2 leaves, at least, some room for improvement. In other words, if the limiting distribution is not a (central) χ^2, the test statistic has not been completely standardized. Note that, while there is only one way of a complete standardization, there are many, if not infinitely many, ways of incomplete standardization, so it may not be convincing why one way is preferred over the others. Second, having a computer-driven, non-analytic asymptotic distribution makes it difficult to study properties of the limiting distribution. For example, how does the reduction of complexity of the model under the null hypothesis play a role? It may not be easy to tell if all one gets are a bunch of numbers. A related issue is regarding direction of improvement. This may not be easy to see without a simple analytic expression for the asymptotic distribution.

In Section 2, we generalize a method initiated by Fisher (1922) in deriving goodness-of-fit tests (GoFTs) that are guaranteed to have asymptotic χ^2 null distributions. A robust feature of the proposed test is that it can be used to

test a certain aspect of the assumed model while another aspect of the model is misspecified. We also discuss a connection, and difference, between our GoFT and the specification test based on the generalized method of moments (GMM; e.g., Hall 2005) that is known in the econometrics literature.

In Section 3, we apply our generalized procedure to SAE to derive a goodness-of-fit test under the Fay-Herriot model (Fay and Herriot 1979). The test is developed under a predictive consideration that incorporates the special interest of SAE. In Section 4, we evaluate performance of the proposed GoFT via simulation studies. We compare our GoFT with several competing methods, including the specification test. The results show that our GoFT is more accurate in term of the size, and has higher or similar power compared to the competing methods. Our GoFT is also computationally less demanding than the specification test. A real data example is discussed in Section 5. Some concluding remarks and future directions are given in Section 6. Proofs and technical details are deferred to Appendix. Computer codes are provided as supplementary materials.

2 Tailoring

In this section, we describe a general approach to obtaining a test statistic that has an asymptotic χ^2 distribution under the null hypothesis. The original idea can be traced back to R. A. Fisher (1922), who used the method to obtain an asymptotic χ^2 distribution for Pearson's χ^2-test, when the so-called minimum chi-square estimator is used. However, Fisher did not put forward the method that he originated under a general framework, as we do here. Suppose that there is a sequence of s-dimensional random vectors, $B(\vartheta)$, which depend on a vector ϑ of unknown parameters with dimension r such that, when ϑ is the true parameter vector, one has $\mathrm{E}\{B(\vartheta)\} = 0$, $\mathrm{Var}\{B(\vartheta)\} = I_s$, and, as the sample size increases,

$$|B(\vartheta)|^2 \xrightarrow{\mathrm{d}} \chi^2_s, \tag{1}$$

where $|\cdot|$ denotes the Euclidean norm. However, because ϑ is unknown, one cannot use (1) for GoFT. What is typically done, such as in Pearson's χ^2-test, is to replace ϑ by an estimator, $\hat{\vartheta}$. Question is: what is $\hat{\vartheta}$? The ideal scenario would be that, after replacing ϑ by $\hat{\vartheta}$ in (1), one has a reduction of degrees of freedom (d.f.), which leads to

$$|B(\hat{\vartheta})|^2 \xrightarrow{\mathrm{d}} \chi^2_\nu, \tag{2}$$

where $\nu = s - r > 0$. This is the famous "subtract one degree of freedom for each parameter estimated" rule taught in many elementary statistics books (e.g., Rice 1995, p. 242). However, as is well known (e.g., Moore 1978), depending on what $\hat{\vartheta}$ is used, (2) may or may not hold, regardless of what degrees of freedom are actually involved. In fact, the only method that is known to achieve (2) without restriction on the distribution of the data is Fisher's

minimum χ^2 method. In a way, the method allows one to "cut-down" the d.f. of (1) by r, and thus convert an asymptotic χ^2_s to an asymptotic χ^2_ν. For such a reason, we have dubbed the method, under the more general setting below, *tailoring*. We develop the method with a heuristic derivation, with the rigorous justification given in the Appendix.

The "right" estimator of ϑ for tailoring is supposed to be the solution to an estimating equation of the following form:

$$C(\vartheta) \equiv A(\vartheta)B(\vartheta) = 0, \tag{3}$$

where $A(\vartheta)$ is an $r \times s$ non-random matrix that plays the role of tailoring the s-dimensional vector, $B(\vartheta)$, to the r-dimensional vector, $C(\vartheta)$. The specification of A will become clear at the end of the derivation. Throughout the derivation, ϑ denotes the true parameter vector. For notation simplicity, we use A for $A(\vartheta)$, $\hat{A}$ for $A(\hat{\vartheta})$, etc, where $\hat{\vartheta}$ is the solution of (3). Under regularity conditions, one has the following expansions, which can be derived from the Taylor series expansion and large-sample theory (e.g., Jiang 2010):

$$\hat{\vartheta} - \vartheta \approx -\left\{\mathrm{E}_\vartheta\left(\frac{\partial C}{\partial \vartheta'}\right)\right\}^{-1} C, \tag{4}$$

$$\hat{B} \approx B - \mathrm{E}_\vartheta\left(\frac{\partial B}{\partial \vartheta'}\right)\left\{\mathrm{E}_\vartheta\left(\frac{\partial C}{\partial \vartheta'}\right)\right\}^{-1} C. \tag{5}$$

Because $\mathrm{E}_\vartheta\{B(\vartheta)\} = 0$ [see above (1)], one has

$$\mathrm{E}_\vartheta\left(\frac{\partial C}{\partial \vartheta'}\right) = A\mathrm{E}_\vartheta\left(\frac{\partial B}{\partial \vartheta'}\right). \tag{6}$$

Combining (5) and (6), we get

$$\hat{B} \approx \{I_s - U(AU)^{-1}A\}B, \tag{7}$$

where $U = \mathrm{E}_\vartheta(\partial B/\partial \vartheta')$. We assume that A is chosen such that

$$U(AU)^{-1}A \text{ is symmetric.} \tag{8}$$

Then, it is easy to verify that $I_s - U(AU)^{-1}A$ is symmetric and idempotent. If we further assume that the following limit exists:

$$I_s - U(AU)^{-1}A \longrightarrow P, \tag{9}$$

then P is also symmetric and idempotent. Thus, assuming that $B \xrightarrow{\mathrm{d}} N(0, I_s)$, which is typically the argument leading to (1), one has, by (7), $\hat{B} \xrightarrow{\mathrm{d}} N(0, P)$, hence (e.g., Searle 1971, p. 58) $|\hat{B}|^2 \xrightarrow{\mathrm{d}} \chi^2_\nu$, where $\nu = \mathrm{tr}(P) = s - r$. This is exactly (2).

It remains to answer one last question: Is there such a non-random matrix $A = A(\vartheta)$ that satisfies (8) and (9)? We show that, not only the answer is

yes, there is an optimal one. Let $A = N^{-1}U'W$, where W is a symmetric, non-random matrix to be determined, and N is a normalizing constant that depends on the sample size. By (4) and the fact that $\mathrm{Var}_\vartheta(B) = I_s$ [see above (1)], we have

$$\mathrm{var}_\vartheta(\hat{\vartheta}) \approx (U'WU)^{-1}U'W^2U(U'WU)^{-1} \geq (U'U)^{-1}, \tag{10}$$

by, for example, Lemma 5.1 of Jiang (2010). The equality on the right side of (10) holds when $W = I_s$, giving the optimal A:

$$A = A(\vartheta) = \frac{U'}{N} = \frac{1}{N}\mathrm{E}_\vartheta\left(\frac{\partial B'}{\partial \vartheta}\right). \tag{11}$$

The A given by (11) clearly satisfies (8) [which is $U(U'U)^{-1}U'$]. It will be seen in the next section that, with $N = m$, (9) is expected to be satisfied. It should be noted that the solution to (3), $\hat{\vartheta}$, does not depend on the choice of N.

Remark 1. A basic assumption for the tailoring method to work is that $\mathrm{E}\{B(\vartheta)\} = 0$ when ϑ is the true parameter vector. However, from the proof of the result (see Appendix) it is seen that the condition "ϑ is the true parameter vector" is not critical. For example, in case there is a model misspecification, a "true parameter vector" may not exist. Nevertheless, what is important is that there is some parameter vector, ϑ, which is not necessarily the true parameter vector, such that the equation

$$A(\vartheta)\mathrm{E}\{B(\vartheta)\} = 0 \tag{12}$$

holds. This equation holds, of course, when ϑ is the true parameter vector, but it can also hold when the true parameter vector does not exist, such as under model misspecification. In fact, in the latter case, one may define the "true parameter vector" as the unique ϑ, assumed exist, that satisfies (12). Note that the number of equations in (12) is the same as the dimension of ϑ; thus, one expect that a solution exists and is unique, under some regularity conditions. To see that (12) is the key, note that under (12), (3) is equivalent to $A(\vartheta)[B(\vartheta) - \mathrm{E}\{B(\vartheta)\}] = 0$, where the expectation is with respect to the true underlying distribution. It follows that one can replace $B(\vartheta)$ by $B(\vartheta) - \mathrm{E}\{B(\vartheta)\}$, which has mean zero, and all of the arguments in the proof go through. This property has given tailoring some unexpected robustness feature, that is, it can work correctly in spite of some model misspecification. We illustrate more specifically in the next section.

Remark 2. There is a connection between the tailoring method and the specification test (ST) based on GMM (e.g., Hall 2005). However, there is also a difference. The difference is that ST is equivalent to (3) with $A(\vartheta)$ given by (11) without the expectation, but for tailoring the expectation is taken first before using it in (3). One may compare this difference to that between the observed Fisher information and expected one in maximum likelihood (ML) estimation (Efron and Hinkley 1978). Although it may be argued that, asymptotically, this difference may be of lower order—in fact, ST may also be viewed as an extension of the original idea of Fisher (1922), and it has the same

asymptotic null distribution as tailoring — , finite-sample performance may differ. We demonstrate this difference in our simulation study in Section 4. Furthermore, because, after taking the expectation, some terms in $\partial B'/\partial\vartheta$ in (11) may vanish, the form of $A(\vartheta)$ in (3) may be substantially simplified. One may, again, compare this to Fisher scoring in ML. For example, McCullagh and Nelder (1989, p. 42) developed the well-known GLM algorithm and noted that it often simplifies the numerical computation of the ML estimator. In our simulation study, we have also observed that tailoring is computationally less demanding than ST, apparently also due to the simplification of taking the expectation. See Section 4 for more detail.

Remark 3. The asymptotic covariance matrix of $\hat{\vartheta}$, that is, the left side of (10), has a "sandwich" expression, which is similar to the well-known sandwich estimator of the (asymptotic) covariance matrix of a GEE (generalized estimating equations) estimator. See, for example, Kauermann and Carroll (2001), who studied impact of the sandwich estimator in terms of relative efficiency and coverage probability of the resulting confidence interval. The sandwich estimator provides robust estimation of the variation of the GEE estimator when the variance-covariance structure of the data is misspecified. The robustness feature of tailoring is, in a way, more general in that it is not necessarily with respect to misspecification of the variance-covariance structure. For example, in the next section we are mainly concerned with misspecification of the mean function.

3 Applying tailoring to SAE

As noted, we intend to develop a GoFT that takes into account the special interests in SAE problems. The development is based on an appropriate objective function in conjunction with the tailoring method. We shall focus on area-level models (Fay and Herriot 1979); extension of the method to other types of SAE models, such as the nested-error regression model (Battese, Fuller and Harter 1988), is fairly straightforward.

There are different aspects of the model that are subject to model checking. Although the focus here is on testing for the normality assumption of the random effects, the method can be easily extended to testing other aspects of the assumed model. As noted (see Remark 1 of the previous section), the proposed test has a robustness feature that it can be used to test one aspect of the model assumption, here normality, while other aspects of the model, for example, the mean function, may be misspecified.

The Fay-Herriot (FH) model may be expressed as that (i) $(y_i, \theta_i), i = 1, \ldots, m$ are independent; (ii) $y_i|\theta_i \sim N(\theta_i, D_i)$; and (iii) $\theta_i \sim N(x_i'\beta, \sigma^2)$. Here, y_i is the direct estimate from the ith area, θ_i is the small area mean, x_i is a vector of observed covariates, β is a vector of unknown parameters, σ^2 is an unknown variance, and D_i is a sampling variance that is assumed known. The normality assumption has to do with (iii). The reason that this is not an issue with (ii) is because, in practice, y_i is typically a sample summary

such as a sample mean or proportion; as a result, the normality assumption in (ii) often holds approximately due to the central limit theorem (CLT). However, there is no obvious reason to believe that the CLT should hold for (iii). Thus, we consider a broader class of distributions, namely, the skewed normal distribution (SN; Azzalini and Capitanio 2014), which includes the normal distribution as a special case. Under the SN distribution, (iii) is replaced by (iii) $\theta_i \sim \text{SN}(x_i'\beta, \sigma^2, \alpha)$, which denotes the SN distribution with mean $x_i'\beta$, variance σ^2, and skewness parameter α (see below), noting that $\alpha = 0$ leads to the normal distribution. We denote the model parameters as $\psi = (\beta', \sigma^2, \alpha)'$.

Suppose that, under the null hypothesis, there is a reduction in the dimension of the parameter vector such that $\gamma = \gamma_0$ under the null hypothesis, where γ is a sub-vector of ψ and γ_0 is known. Let ϑ denote the vector of parameters in ψ other than γ. In this section, notation such as E_ϑ, etc. will be understood as expectation, etc. under the null hypothesis.

The LRT is often used in the context of goodness-of-fit. However, because, in SAE, the primary interest is prediction of mixed effects (e.g., Jiang 2007, Rao and Molina 2015), it is reasonable to develop something that is closely related to the predictive interest. To motivate something that is in a similar spirit of LRT, but takes into account the SAE interest, let us consider the problem from a "Bayesian" perspective.

In general, let θ be a vector of unobserved quantities that one wishes to predict (e.g., small area mean θ_i in FH model), ψ be the vector of parameters involved in either $f(y|\theta)$ or $f(\theta)$, and y a vector of observations. The likelihood function may be viewed, using a Bayesian term, as a marginal likelihood with the distribution of θ, $f(\theta)$, treated as *a prior*, that is,

$$f(y|\psi) = \int f(y|\theta, \psi) f(\theta|\psi) d\theta. \tag{13}$$

The likelihood is used for estimation of fixed parameters, which are associated with either $f(\theta)$ or $f(y|\theta)$ or both. To come up with a predictive version of the likelihood, we may simply replace the prior in (13) by its "posterior", that is, the conditional pdf of θ given y, $f(\theta|y)$. With this replacement, we obtain

$$f(y|y, \psi) \equiv \int f(y|\theta, \psi) f(\theta|y, \psi) d\theta. \tag{14}$$

We call (14) the *predictive likelihood*, or PL. The reason is that, if parameter estimation is of primary interest, one uses the prior, $f(\theta)$, to obtain the (marginal) likelihood (13). Now, because we replace $f(\theta)$ by $f(\theta|y)$, which is the main outcome for the prediction of θ, and then go through the same operation, the output (14) should be called a predictive likelihood. It should be noted that the predictive likelihood is not necessarily a likelihood, as it does not always possess some of the well-known properties of the likelihood. However, we can, at least, adjust the score equation of the PL to make it unbiased.

The adjusted PL score is given by

$$s_{\rm a}(\psi) = \frac{\partial}{\partial\psi}\log f(y|y,\psi) - {\rm E}_\psi\left\{\frac{\partial}{\partial\psi}\log f(y|y,\psi)\right\}. \tag{15}$$

We call the estimator of ψ obtained by solving the adjusted PL score equation, $s_{\rm a}(\psi)=0$, or, equivalently, the following equation:

$$\frac{\partial}{\partial\psi}\log f(y|y,\psi) = {\rm E}_\psi\left\{\frac{\partial}{\partial\psi}\log f(y|y,\psi)\right\} \tag{16}$$

maximum adjusted PL estimator, or Maple, in view of its analogy to the MLE.

Under the FH model, it is easy to show that $f(\theta|y,\psi)=\prod_{i=1}^m f(\theta_i|y_i,\psi)$ where $\theta=(\theta_1,...,\theta_m)$. Thus, we have

$$f(y|y,\psi) = \int\prod_{i=1}^m f(y_i|\theta_i,\psi)f(\theta_i|y_i,\psi)d\theta = \prod_{i=1}^m\int f(y_i|\theta_i,\psi)f(\theta_i|y_i,\psi)d\theta_i = \prod_{i=1}^m f(y_i|y_i,\psi), \tag{17}$$

where $f(y_i|y_i,\psi) = \int f(y_i|\theta_i,\psi)f(\theta_i|y_i,\psi)d\theta_i$. Without the null hypothesis that the random effects are normal, that is, $\alpha=0$, we have $y_i|\theta_i\sim N(\theta_i,D_i)$ and $\theta_i\sim{\rm SN}(x_i'\beta,\sigma^2,\alpha)$. It is then shown in the Appendix that

$$f(y_i|y_i,\psi) = \frac{1}{\sqrt{D_i(1+B_i)}}\phi\left[\frac{y_i-x_i'\beta}{\sqrt{D_i(1+B_i)}/(1-B_i)}\right]\frac{\Phi[\alpha_{2i}(y_i-x_i'\beta)]}{\Phi[\alpha_{3i}(y_i-x_i'\beta)]} \tag{18}$$

where $\alpha_{si}=(4-s)\sigma\alpha/\sqrt{\{(4-s)\sigma^2+D_i\}\{(4-s)\sigma^2+(1+\alpha^2)D_i\}}$, $s=2,3$, $B_i=\sigma^2/(\sigma^2+D_i)$, and $\Phi(\cdot),\phi(\cdot)$ denote the cdf, pdf of $N(0,1)$, respectively. Note that, when $\alpha=0$, (18) reduces to that under normality. Also note that $f(y_i|y_i,\psi)\neq f(y_i,\psi)$.

By (17), the PL can be expressed as $\prod_{i=1}^m f(y_i|y_i,\psi)$. To test ${\rm H}_0:\alpha=0$, let

$$b_i(y_i,\vartheta) = \{(\partial/\partial\psi)\log f(y_i|y_i,\psi)\}|_{\alpha=0} - E[\{(\partial/\partial\psi)\log f(y_i|y_i,\psi)\}|_{\alpha=0}].$$

One can derive the adjusted PL equation, (16), as follows:

$$\begin{aligned}
&\left.\frac{\partial\log f(y_i|y_i,\psi)}{\partial\beta}\right|_{\alpha=0} - {\rm E}\left\{\left.\frac{\partial\log f(y_i|y_i,\psi)}{\partial\beta}\right|_{\alpha=0}\right\} = a_i(\sigma^2)x_i(y_i-x_i'\beta),\\
&\left.\frac{\partial\log f(y_i|y_i,\psi)}{\partial\sigma^2}\right|_{\alpha=0} - {\rm E}\left\{\left.\frac{\partial\log f(y_i|y_i,\psi)}{\partial\sigma^2}\right|_{\alpha=0}\right\} = b_i(\sigma^2)(y_i-x_i'\beta)^2 - c_i(\sigma^2),\\
&\left.\frac{\partial\log f(y_i|y_i,\psi)}{\partial\alpha}\right|_{\alpha=0} - {\rm E}\left\{\left.\frac{\partial\log f(y_i|y_i,\psi)}{\partial\alpha}\right|_{\alpha=0}\right\} = d_i(\sigma^2)(y_i-x_i'\beta),
\end{aligned}$$

where $a_i(\sigma^2)=(1-B_i)^2/D_i(1+B_i)$, $b_i(\sigma^2)=(1-B_i)^3(3+B_i)/2D_i^2(1+B_i)^2$, $c_i(\sigma^2)=(1-B_i)^2(3+B_i)/2D_i(1+B_i)^2$, and $d_i(\sigma^2)=\sqrt{2/\pi}\sigma(1-B_i)^2/D_i(1+B_i)$.

Let ϑ denote the true ϑ. If the model is correctly specified under the null hypothesis, then, under the null hypothesis, $\sum_{i=1}^m b_i(y_i, \vartheta)$ is a sum of independent random vectors with mean zero. On the other hand, if there is some misspecification in the mean function that the true β, hence the true ϑ, does not exist (under the null hypothesis), we again define the "true ϑ" as the unique solution to (12). Then, all of the arguments in the derivation of Section 2 go through by replacing $\sum_{i=1}^m b_i(y_i, \vartheta)$ with $\sum_{i=1}^m [b_i(y_i, \vartheta) - \mathrm{E}\{b_i(y_i, \vartheta)\}]$. Furthermore, we have $V_b(\vartheta) = \mathrm{Var}_\vartheta\{\sum_{i=1}^m b_i(y_i, \vartheta)\} = \sum_{i=1}^m \mathrm{Var}_\vartheta\{b_i(y_i, \vartheta)\}$, where

$$\mathrm{Var}_\vartheta\{b_i(y_i, \vartheta)\} = \begin{bmatrix} g_i(\sigma^2)x_i x_i' & 0_p & g_i(\sigma^2)(x_i\sqrt{2/\pi}\sigma) \\ 0_p' & h_i(\sigma^2) & 0 \\ g_i(\sigma^2)(x_i'\sqrt{2/\pi}\sigma) & 0 & g_i(\sigma^2)(2\sigma^2/\pi) \end{bmatrix}$$

with $g_i(\sigma^2) = (1 - B_i)^3/D_i(1 + B_i)^2$ and $h_i(\sigma^2) = (1 - B_i)^4(3 + B_i)^2/2D_i^2(1 + B_i)^4$.

Thus, if we let $B(\vartheta) = V_b^{-1/2}(\vartheta)\sum_{i=1}^m b_i(y_i, \vartheta)$, we have $B(\vartheta) \xrightarrow{\mathrm{d}} N(0, I_s)$, where $s = \dim(\psi) = 3$. It follows that (1) holds. Because $r = \dim(\vartheta) = 2 < s$, the tailoring method applies to yield (2) with $\nu = s - r = 1$. In particular, we have

$$A(\vartheta) = \frac{1}{m}\left\{\sum_{i=1}^m \mathrm{E}_\vartheta\left(\frac{\partial b_i'}{\partial \vartheta}\right)\right\} V_b^{-1/2}(\vartheta),$$

where b_i is defined above and

$$\mathrm{E}_\vartheta\left(\frac{\partial b_i'}{\partial \vartheta}\right) = -\begin{bmatrix} a_i(\sigma^2)x_i x_i' & 0_p & d_i(\sigma^2)x_i \\ 0_p' & b_i(\sigma^2) & 0 \end{bmatrix}.$$

This gives $A(\vartheta)$ for solving the tailoring equation (3).

4 Simulation study

We carry out a simulation study to evaluate performance of the tailoring methods based on Maple, described in the previous section, and compare it with existing methods. Specifically, we compare our method with those of Pierce (1982), Jiang (2001), Claeskens and Hart (2009), and ST based on GMM (e.g., Hall 2005). For Pierce (1982), the test statistic under the FH model for $\mathrm{H}_0 : \alpha = 0$ is given by $\hat{F} \equiv m\hat{T}_m^2/V$, where

$$\hat{T}_m = \frac{1}{m}\sum_{i=1}^m \frac{\sqrt{D_i}(y_i - x_i'\hat{\beta})}{\hat{\sigma}^2 + D_i},$$

$$V = \frac{1}{m}\sum_{i=1}^m \frac{D_i}{D_i + \hat{\sigma}^2} - mP\{\mathrm{var}(\hat{\psi} - \psi)\}P'$$

with $P = \lim \mathrm{E}(\partial T_m/\partial\psi)$. The asymptotic null distribution of the test statistic is χ^2_1. In the current case, it can be shown that

$$P = -\lim\left\{\frac{1}{m}\begin{bmatrix}\sum_{i=1}^m \sqrt{D_i}x_i'/(D_i+\sigma^2)\\ 0\end{bmatrix}\right\}.$$

In the case of Jiang (2001), one has the test statistic

$$\hat{\chi}^2_{\mathrm{J}} = \frac{1}{m}\sum_{k=1}^{K}\{N_k - p_k(\hat{\psi})\}^2,$$

where $N_k = \sum_{i=1}^m 1_{(y_i\in C_k)} = \#\{1 \le i \le m : y_i \in C_k\}$, and $p_k(\psi) = \sum_{i=1}^m P_\psi(y_i \in C_k) = \sum_{i=1}^m p_{ik}(\psi)$. More specifically, the cells, C_k, $1 \le k \le K$ are defined as follows: $C_1 = (-\infty, c_1], C_k = (c_{k-1}, c_k], 2 \le k \le K-1$, and $C_K = (c_{K-1}, \infty)$. Regarding the choice of K and c_k's, by Jiang (2001), we may choose $K = \max(p+2, [m^{1/5}])$, where p is the dimension of β. Once K is chosen, the c_k's are chosen so that there are equal number of y_i's within each $C_k, 1 \le k \le K$. It then follows that $N_k = m/K, 1 \le k \le K$. Finally, the $p_{ik}(\psi)$ have the following expressions:

$$\begin{aligned}
p_{i1}(\psi) &= \Phi\left(\frac{c_1 - x_i'\beta}{\sqrt{\sigma^2+D_i}}\right),\\
p_{ik}(\psi) &= \Phi\left(\frac{c_k - x_i'\beta}{\sqrt{\sigma^2+D_i}}\right) - \Phi\left(\frac{c_{k-1} - x_i'\beta}{\sqrt{\sigma^2+D_i}}\right), \quad 2 \le k \le K-1,\\
p_{iK}(\psi) &= 1 - \Phi\left(\frac{c_{K-1} - x_i'\beta}{\sqrt{\sigma^2+D_i}}\right).
\end{aligned}$$

We then use a Monte-Carlo method (e.g., bootstrapping) to compute the critical values, as suggested by Jiang (2001).

In the case of Claeskens and Hart (2009), one uses the test statistic

$$\hat{\chi}^2_{\mathrm{CH}} = \max_{1\le l\le M}\frac{2\{\log L_l - \log L_{M=0}\}}{l(l+3)/2},$$

where $\log L$ is the log-likelihood and M is the order of polynomial which plays the role of a smoothing parameter. The test is based on the LRT which compares the estimated distribution ($M > 0$) and the null distribution ($M = 0$; i.e., normal). Similar to the Jiang (2001), one needs to use replications from the test statistic above to approximate the critical values. We consider $M = 2$ in our simulation study.

As noted (see Remark 2 in Section 2), the ST is simply tailoring with the expectation sign in (11) removed. In the case of Maple, to obtain the A corresponding to ST, we have $\partial b_i'/\partial\vartheta = (a_{ist})_{1\le s\le 2, 1\le t\le 3}$, where

$$\begin{aligned}
a_{i11} &= -a_i(\sigma^2)x_i x_i',\\
a_{i12} &= -2b_i(\sigma^2)x_i(y_i - x_i'\beta),
\end{aligned}$$

$$a_{i13} = -d_i(\sigma^2)x_i,$$

$$a_{i21} = -2b_i(\sigma^2)x_i'(y_i - x_i'\beta),$$

$$a_{i22} = \{\partial b_i(\sigma^2)/\partial\sigma^2\}(y_i - x_i'\beta)^2 - \{\partial c_i(\sigma^2)/\partial\sigma^2\},$$

$$a_{i23} = \{\partial d_i(\sigma^2)/\partial\sigma^2\}(y_i - x_i'\beta),$$

with

$$\frac{\partial b_i(\sigma^2)}{\partial\sigma^2} = -\frac{2b_i(\sigma^2)(1-B_i)(3+B_i)}{2D_i(1+B_i)} - \frac{a_i(\sigma^2)\{2(1+B_i)^2 + (1-B_i)(3+B_i)\}}{2(2\sigma^2+D_i)^2},$$

$$\frac{\partial c_i(\sigma^2)}{\partial\sigma^2} = -\frac{2b_i(\sigma^2)(3+B_i)}{2(1+B_i)} - \frac{a_i(\sigma^2)D_i}{(2A+D_i)^2},$$

and $\partial d_i(\sigma^2)/\partial\sigma^2 = -2b_i(\sigma^2)\sqrt{2/\pi}\sigma + a_i(\sigma^2)/\sqrt{2\pi}\sigma$.

To evaluate and compare performance of the aforementioned methods, let $\hat{B}^2_{\rm PL}$, $\hat{F}$, $\hat{\chi}^2_{\rm J}$, $\hat{\chi}^2_{\rm CH}$, and $\hat{B}^2_{\rm ST}$ represent the test statistics for tailoring/Maple, Pierce (1982), Jiang (2001), Claeskens-Hart (2009), and ST, respectively [for notation simplicity we write $|B(\hat{\vartheta})|^2$ as $\hat{B}^2$]. We consider two situations where the assumed model is either correct or misspecified. The assumed model is a FH model:

$$y_i = \beta_1 x_i + v_i + e_i, \quad i = 1, \ldots, m;$$

however, the data are generated under the following FH model:

$$y_i = \beta_1 x_i + v_i + e_i, \quad 1 \le i \le n,$$
$$y_i = \beta_2 x_i + v_i + e_i, \quad n+1 \le i \le m,$$

where $m = 2n, D_i = D_{i1}$ for $1 \le i \le n$ and $D_i = D_{i2}$ for $n+1 \le i \le m$. We choose $\sigma^2 = 10$, noting that $v_i \sim SN(0, \sigma^2, \alpha)$ and $e_i \sim N(0, D_i)$. The D_{i1} are generated from the uniform distribution between 3.5 and 4.5. There are two scenarios for D_{i2}, one generated from $U(3.5, 4.5)$ and the other from $U(0.5, 1.5)$. Let $\beta_1 = 1$, and $\beta_2 = 1$ or 3; and the true value of α is 0 under the null hypothesis, and 0.5 under the alternative. The x_i's are generated from the uniform distribution between 0 and 1, and fixed during the simulation study. Note also that the D_i's are fixed during the simulation study. It is seen that, when $\beta_1 \ne \beta_2$, the underlying model is misspecified.

We consider testing $H_0 : \alpha = 0$ with three different levels of significance, 0.01, 0.05, 0.10, and four different sample sizes, $m = 50, 100, 200$, and 500. We run $R = 5,000$ simulations to calculate $\hat{B}^2_{\rm PL}, \hat{F}, \hat{\chi}^2_{\rm J}$, $\hat{\chi}^2_{\rm CH}$, and $\hat{B}^2_{\rm ST}$. In particular, we generate response variable $y_i^{(r_1)} = \beta_1 x_i + v_i^{(r_1)} + e_i^{(r_1)}, (1 \le i \le n; r_1 = 1, ..., R)$, and $y_i^{(r_1)} = \beta_2 x_i + v_i^{(r_1)} + e_i^{(r_1)}, (n+1 \le i \le m; r_1 = 1, ..., R)$, where $v_i^{(r_1)} \sim SN(0, \sigma^2, \alpha = 0)$ and $e_i^{(r_1)} \sim N(0, D_{i1})$ for $1 \le i \le n$ and $e_i^{(r_1)} \sim N(0, D_{i2})$ for $n+1 \le i \le m$. For each simulated dataset, we estimate σ^2 and β_1 for $\hat{B}^2_{\rm PL}$, $\hat{F}$, $\hat{\chi}^2_{\rm J}$, $\hat{\chi}^2_{\rm CH}$, and $\hat{B}^2_{\rm ST}$, where $r = 2$ and $s = 3$. Note that for

$\hat{B}^2_{\rm PL}$, we use tailoring to estimate the model parameters; we use the Prasad-Rao approach for $\hat{F}$ and $\hat{\chi}^2_{\rm J}$ as it is computationally faster, the MLE for $\hat{\chi}^2_{\rm CH}$, and GMM for $\hat{B}^2_{\rm ST}$.

Also, we use 1000 replications to obtain the critical values, in each simulation run, for $\hat{\chi}^2_{\rm J}$, and 100 replication run for $\hat{\chi}^2_{\rm CH}$ (due to the fact that the latter is computationally more intensive). To obtain the sizes of the tests, we count the number of times (out of R) that $\hat{B}^{2(r_1)}_{\rm PL}$, $\hat{F}^{(r_1)}$, and $\hat{B}^{2(r_1)}_{\rm ST}$ exceed the critical values for the three different levels of significance, namely, $\chi^2_{(0.01)}(1) = 6.63$, $\chi^2_{(0.05)}(1) = 3.84$, $\chi^2_{(0.10)}(1) = 2.70$, and divide those numbers by R. In the cases of Jiang (2001) and Claeskens and Hart (2009), $\hat{\chi}^{2(r_1)}_{\rm J}$ and $\hat{\chi}^{2(r_1)}_{\rm CH}$ are compared with their corresponding critical values obtained using the bootstrap approach~~es~~ (i.e., 1000 replications for Jiang test and 100 replications for CH test under the null hypothesis), in each simulation run. The powers of the tests are obtained the same way, the only difference being that the sizes are computed when the data are generated under the null hypothesis $\alpha = 0$, while the powers under the alternative of $\alpha = 0.5$.

The empirical size and power for different levels of significance, different scenarios, and different methods are reported in Tables 1–3. It seems that, with the increasing sample size (m) and for all three different levels of significance, $\hat{B}^2_{\rm PL}$ and $\hat{\chi}^2_{\rm J}$ have approximately the right size under different scenarios. However, in the case of $\hat{F}$, the test does not seem to have the right size if there are misspecifications in the mean and significant change in the range of the sampling variances for the small areas. The size also does not seem to improve for $\hat{\chi}^2_{\rm CH}$ with increasing sample size. As for $\hat{B}^2_{\rm ST}$, it seems that the test does not have the right size until $m = 500$. Regarding the power, $\hat{B}^2_{\rm PL}$ seems to perform very well under all scenarios. The power performance of $\hat{\chi}^2_{\rm J}$ seems to be poor compared to $\hat{B}^2_{\rm PL}$, while the power performance of $\hat{F}$ and $\hat{B}^2_{\rm ST}$ is similar to that of $\hat{B}^2_{\rm PL}$. It appears that the power of $\hat{\chi}^2_{\rm CH}$ does not also improve with increasing sample size.

Note, in particular, that $\hat{B}^2_{\rm ST}$ performs poorly in size unless $m = 500$. To further investigate the possible reason for this, we provide in Table 4 median estimates of σ^2 over the simulation runs R under different sample sizes and scenarios, for PL (tailoring) and ST (GMM). It is seen that the estimate of σ^2 by GMM performs poorly until $m = 500$.

It has also been observed that PL (tailoring) is computationally much less demanding than ST (GMM). For example, the rate of convergence for the parameter estimates GMM/ST, in terms of the number of iterations needed for the Newton-Raphson procedure to achieve a given level of accuracy (the larger m the slower convergence), was much lower than that for the corresponding Maple/tailoring method.

Table 1 Size (Power) under different sample sizes and scenarios—level of significance equal to 0.01

D_{i2}	β_2	m	$\hat{B}^2_{\rm PL}$	$\hat{F}$	$\hat{\chi}^2_{\rm J}$	$\hat{\chi}^2_{\rm CH}$	$\hat{B}^2_{\rm ST}$
$U(3.4, 4.5)$	1	50	0.009 (0.660)	0.009 (0.848)	0.009 (0.032)	0.014 (0.818)	0.002 (0.490)
		100	0.006 (0.894)	0.006 (0.964)	0.010 (0.019)	0.010 (0.763)	0.000 (0.899)
		200	0.010 (0.992)	0.010 (0.999)	0.012 (0.008)	0.003 (0.696)	0.000 (0.985)
		500	0.011 (1.000)	0.011 (1.000)	0.011 (0.0003)	0.002 (0.659)	0.008 (1.000)
	3	50	0.010 (0.647)	0.009 (0.841)	0.014 (0.021)	0.012 (0.834)	0.003 (0.474)
		100	0.006 (0.882)	0.006 (0.962)	0.013 (0.013)	0.014 (0.821)	0.000 (0.889)
		200	0.009 (0.991)	0.009 (0.998)	0.010 (0.004)	0.000 (0.698)	0.000 (0.983)
		500	0.010 (1.000)	0.010 (1.000)	0.012 (0.0003)	0.004 (0.667)	0.004 (1.000)
$U(0.5, 1.5)$	1	50	0.008 (0.577)	0.010 (0.773)	0.010 (0.009)	0.020 (0.795)	0.000 (0.579)
		100	0.009 (0.821)	0.007 (0.949)	0.013 (0.013)	0.020 (0.795)	0.000 (0.824)
		200	0.009 (0.975)	0.009 (0.997)	0.011 (0.005)	0.000 (0.670)	0.000 (0.975)
		500	0.010 (1.000)	0.008 (1.000)	0.011 (0.0004)	0.001 (0.673)	0.000 (1.000)
	3	50	0.007 (0.574)	0.024 (0.665)	0.015 (0.033)	0.022 (0.804)	0.000 (0.570)
		100	0.009 (0.803)	0.033 (0.893)	0.016 (0.019)	0.013 (0.811)	0.000 (0.805)
		200	0.008 (0.971)	0.075 (0.982)	0.015 (0.003)	0.001 (0.710)	0.000 (0.972)
		500	0.009 (1.000)	0.193 (1.000)	0.011 (0.0004)	0.002 (0.826)	0.000 (1.000)

Table 2 Size (Power) under different sample sizes and scenarios—level of significance equal to 0.05

D_{i2}	β_2	m	$\hat{B}^2_{\rm PL}$	$\hat{F}$	$\hat{\chi}^2_{\rm J}$	$\hat{\chi}^2_{\rm CH}$	$\hat{B}^2_{\rm ST}$
$U(3.4, 4.5)$	1	50	0.050 (0.855)	0.048 (0.896)	0.049 (0.023)	0.068 (0.818)	0.014 (0.634)
		100	0.048 (0.966)	0.047 (0.979)	0.051 (0.017)	0.039 (0.770)	0.000 (0.969)
		200	0.053 (0.998)	0.053 (0.999)	0.051 (0.004)	0.011 (0.696)	0.001 (0.991)
		500	0.050 (1.000)	0.050 (1.000)	0.048 (0.0002)	0.013 (0.659)	0.054 (1.000)
	3	50	0.048 (0.849)	0.045 (0.889)	0.053 (0.021)	0.057 (0.837)	0.016 (0.629)
		100	0.046 (0.963)	0.045 (0.979)	0.060 (0.013)	0.045 (0.825)	0.000 (0.967)
		200	0.051 (0.998)	0.051 (0.999)	0.053 (0.003)	0.002 (0.701)	0.000 (0.990)
		500	0.048 (1.000)	0.046 (1.000)	0.049 (0.000)	0.010 (0.667)	0.046 (1.000)
$U(0.5, 1.5)$	1	50	0.051 (0.799)	0.052 (0.849)	0.051 (0.025)	0.055 (0.798)	0.000 (0.813)
		100	0.046 (0.936)	0.049 (0.967)	0.052 (0.017)	0.055 (0.798)	0.000 (0.932)
		200	0.053 (0.994)	0.049 (0.999)	0.048 (0.006)	0.014 (0.677)	0.000 (0.993)
		500	0.049 (1.000)	0.049 (1.000)	0.048 (0.000)	0.010 (0.673)	0.033 (1.000)
	3	50	0.048 (0.792)	0.092 (0.772)	0.061 (0.021)	0.061 (0.805)	0.000 (0.802)
		100	0.046 (0.928)	0.122 (0.935)	0.074 (0.011)	0.050 (0.816)	0.000 (0.924)
		200	0.051 (0.993)	0.209 (0.993)	0.070 (0.003)	0.008 (0.715)	0.000 (0.992)
		500	0.047 (1.000)	0.415 (1.000)	0.081 (0.000)	0.013 (0.829)	0.031 (1.000)

5 Median income data

We discuss two applications of the tailoring method regarding the median income data of four-person families at the state level in the USA (Ghosh, Nangia and Kim, 1996). The first application has to do for choosing an appropriate model; the second is about checking the normality assumption. The data has been analyzed by several researchers using different set-ups. In this analysis, the response variable y_i is the four-person median income from the sample survey at state i in year 1989, and x_i is the census four-person median income at state i in year 1979 ($i = 1, ..., m = 51$).

Table 3 Size (Power) under different sample sizes and scenarios—level of significance equal to 0.10

D_{i2}	β_2	m	$\hat{B}^2_{\mathrm{PL}}$	$\hat{F}$	$\hat{\chi}^2_{\mathrm{J}}$	$\hat{\chi}^2_{\mathrm{CH}}$	$\hat{B}^2_{\mathrm{ST}}$
$U(3.4, 4.5)$	1	50	0.101 (0.917)	0.098 (0.919)	0.094 (0.032)	0.122 (0.819)	0.035 (0.676)
		100	0.098 (0.984)	0.095 (0.986)	0.103 (0.018)	0.085 (0.770)	0.000 (0.987)
		200	0.104 (0.999)	0.104 (0.999)	0.095 (0.004)	0.024 (0.696)	0.013 (0.992)
		500	0.094 (1.000)	0.095 (1.000)	0.099 (0.0004)	0.030 (0.659)	0.097 (1.000)
	3	50	0.096 (0.913)	0.094 (0.913)	0.106 (0.031)	0.111 (0.837)	0.034 (0.674)
		100	0.097 (0.982)	0.092 (0.984)	0.113 (0.014)	0.097 (0.825)	0.000 (0.985)
		200	0.101 (0.999)	0.100 (0.999)	0.104 (0.007)	0.016 (0.702)	0.012 (0.992)
		500	0.094 (1.000)	0.094 (1.000)	0.102 (0.0004)	0.000 (0.667)	0.093 (1.000)
$U(0.5, 1.5)$	1	50	0.103 (0.886)	0.102 (0.885)	0.095 (0.044)	0.121 (0.799)	0.000 (0.891)
		100	0.096 (0.972)	0.100 (0.977)	0.096 (0.021)	0.121 (0.799)	0.000 (0.970)
		200	0.110 (0.997)	0.102 (0.999)	0.099 (0.003)	0.037 (0.681)	0.000 (0.998)
		500	0.099 (1.000)	0.098 (1.000)	0.098 (0.0003)	0.021 (0.681)	0.076 (1.000)
	3	50	0.097 (0.879)	0.171 (0.812)	0.121 (0.038)	0.114 (0.806)	0.000 (0.883)
		100	0.097 (0.965)	0.209 (0.953)	0.138 (0.018)	0.094 (0.817)	0.000 (0.963)
		200	0.105 (0.996)	0.316 (0.996)	0.132 (0.004)	0.022 (0.716)	0.001 (0.997)
		500	0.097 (1.000)	0.535 (1.000)	0.120 (0.000)	0.023 (0.829)	0.075 (1.000)

Table 4 Median estimate of σ^2 for PL (tailoring) and ST (GMM) methods under different sample sizes and scenarios

D_{i2}	β_2	m	PL (tailoring)	ST (GMM)
$U(3.4, 4.5)$	1	50	10.00	1237e+5
		100	10.00	1470.00
		200	10.04	11.00
		500	10.01	10.49
	3	50	11.00	1255e+5
		100	10.00	585.30
		200	10.39	12.00
		500	10.33	10.83
$U(0.5, 1.5)$	1	50	10.00	1661e+4
		100	10.00	1038e+3
		200	10.02	12.00
		500	10.00	10.53
	3	50	11.00	1589e+4
		100	10.00	980100
		200	10.36	12.00
		500	10.33	10.83

5.1 Choosing an appropriate model

An inspection of the scatter plot (Figure 1) suggests that a quadratic model may fit the data well. As a starting point, we test whether a quadratic mixed model rather than linear mixed model fits the data well. That is,

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + v_i + e_i, i = 1, ..., m = 51, \tag{19}$$

where the v_i's are state-specific random effects and e_i's are sampling errors. It is assumed that v_i and e_i are independent with $v_i \sim N(0, \sigma^2)$ and $e_i \sim N(0, D_i)$ with known D_i. We now test $H_0 : \beta_2 = 0$ vs $H_1 : \beta_2 \neq 0$. In the case of Maple/tailoring approach, the model parameter estimates are $\hat{\beta}_0 = 4503.2, \hat{\beta}_1 = 1.60, \hat{\sigma}^2 = 1.9 \times 10^7$ which result in rejecting H_0 as $\hat{B}^2_{PL} = 5.36 > 3.84 [= \chi^2_{0.05}(1)]$. Thus, the test suggests that the linear model is inappropriate.

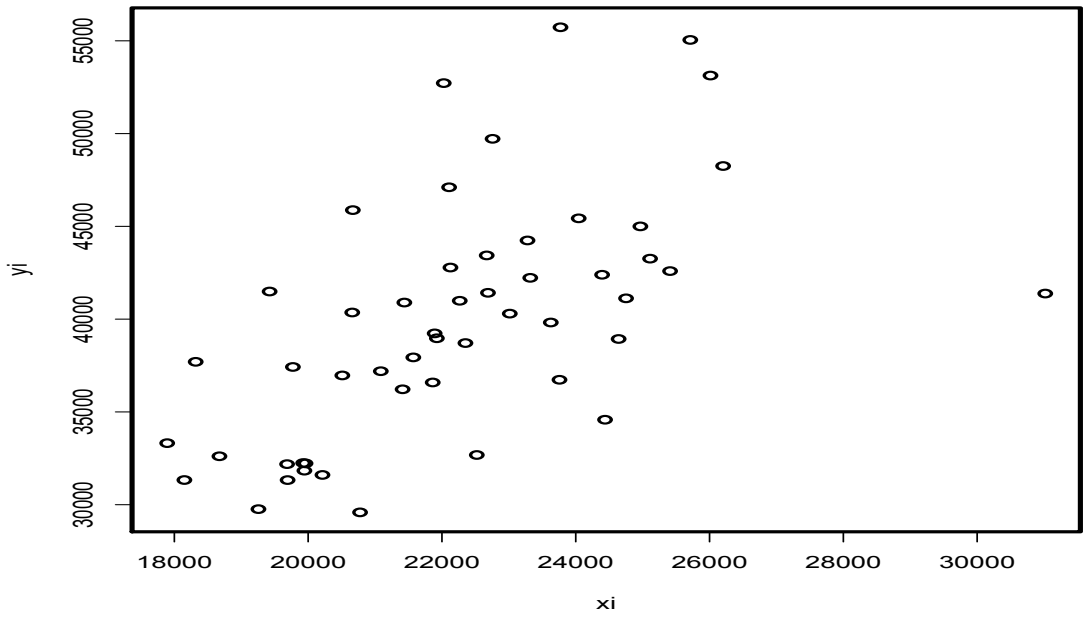

Fig. 1 Plot of survey income in 1989 (y) vs family median income in 1979 (x)

Based on the above result, we can also evaluate the cubic model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + v_i + e_i, \tag{20}$$

or a quadratic-outlying (Q-O) model (due to the point at the right corner of the scatterplot of y_i vs x_i; see Jiang, Nguyen and Rao 2011), as

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 1_{(x_i > 30000)} + v_i + e_i. \tag{21}$$

In the case of Maple/tailoring approach, the model parameter estimates are $\hat{\beta}_0 = -72375.1, \hat{\beta}_1 = 845.1, \hat{\beta}_2 = -1.50, \hat{\sigma}^2 = 16982730$, which cannot reject H_0 as $\hat{B}^2_{PL} = 1.19 < 3.84\ [= \chi^2_{0.05}(1)]$. Thus, the test confirms the quadratic model is an appropriate model for this data rather than the cubic model.

Finally, we consider the Q-O model. To evaluate the Q-O model, we use Maple in conjunction with tailoring to test $H_0 : \beta_3 = 0$ in model (21). In the case of Maple/tailoring approach, the model parameter estimates are $\hat{\beta}_0 = -72375.1, \hat{\beta}_1 = 849.1, \hat{\beta}_2 = -1.50, \hat{\sigma}^2 = 16982730$, which cannot reject the H_0 as $\hat{B}^2_{PL} = 1.31 < 3.84\ [= \chi^2_{0.05}(1)]$. Thus, the test confirms that the quadratic model as an appropriate model for this data rather than the Q-O model.

Overall, we conclude that the quadratic model is a good fit for the data.

It should be noted that we also applied the methods of Pierce (1982), Jiang (2001), and Claeskens and Hart (2009) to this data. None of these tests were able to reject the linear model null hypothesis. This seems to be consistent with the pattern observed in our simulation study in Section 4 that these tests appear to have lower power than our tests.

5.2 Checking the normality assumption

It is known that income data are typically not normal. In this application, our goal is to check the normality assumption for median incomes of four-person families at the state level in the USA (Ghosh, Nangia and Kim 1996). Following Section 5.1, we consider the quadratic model (19).

We consider testing $\mathrm{H}_0 : \alpha = 0$ vs $\mathrm{H}_1 : \alpha \neq 0$. First, we apply the Maple approach in conjunction with tailoring. The parameter estimates are $\hat{\beta}_1 = 2.07, \hat{\beta}_2 = -1.2 \times 10^{-5}, \hat{\sigma}^2 = 1.9 \times 10^7$, which result in $\hat{B}^2_{\mathrm{PL}} = 4.67 > 2.70 = \chi^2_1(0.90)$, rejecting H_0 at the 10% significance level.

Next, we apply the ST method to this data. In this case, the GMM estimates are $\hat{\beta}_1 = 2.07, \hat{\beta}_2 = -1.3 \times 10^{-5}, \hat{\sigma}^2 = 1.9 \times 10^7$, which result $\hat{B}^2_{\mathrm{ST}} = 4.93 > 2.70$, also rejecting H_0 at the 10% significance level.

Note that, although the values of tailoring and GMM estimates are very close, the GMM equation for ST is more complicated than the tailoring one due to not taking the expected value of the A matrix (see Remark 2 in Section 2). This may explain the poor performance of ST in terms of the size (when m is moderate) and computational efficiency in our simulation study reported in Section 4.

We also applied the methods of Pierce (1982), Jiang (2001), and Claeskens and Hart (2009) to test the hypothesis. None of these tests were able to reject the normality assumption. For the latter two methods, this seems to be consistent with the pattern observed in our simulation study in Section 4 that these tests appear to be less powerful.

6 Discussion

There are multiple ways of checking for goodness-of-fit. A main reason that PL is considered in the context of SAE is due to an intuitive fact that it gets the predictive distribution of θ, the vector of small area means, involved in the process. To illustrate with a simple example, consider the following James-Stein example. Suppose that $y_i = \theta_i + e_i$, $i = 1, \ldots, m$, where $\theta_i, e_i, i = 1, \ldots, m$ are independent such that $\theta_i \sim N(\mu, A)$, $e_i \sim N(0, 1)$. The model is a special case of the Fay-Herriot model. From a Bayesian perspective, θ_i has a prior distribution, which is normal with mean μ and variance A. However, the predictive distribution of θ_i, given the data $y = (y_i)_{1 \leq i \leq m}$, is normal with mean $w\mu + (1 - w)y_i$ and variance wA, where $w = (A + 1)^{-1}$. It is clear that the data has an impact on understanding the distribution of θ, going from

the prior distribution to the predictive distribution. This is what we want to check with our goodness-of-fit test. In contrast, the traditional likelihood, which corresponds to using the prior distribution of θ instead of the predictive distribution [compare (13) and (14)]. The data has no impact on the prior distribution; in other words, the prior distribution is not sensitive to how one predicts the distribution of θ using the data. Therefore, intuitively, the likelihood based method has little to do with the main interest of SAE, that is, the prediction of θ.

The next question is how this intuition makes a difference. This has to do with the main objective of using a statistical model. Is the model used for interpretation or prediction? If the model is used for interpretation, then perhaps one can ignore a few outliers because, here, the focus is the main trend, or big picture. However, if the main objective is prediction, the outliers may not be ignored. In practice, ignoring a few outliers can result in the cost of millions of dollars, if not billions of dollars. The objective is taken seriously, for example, in our real-data example (see Section 5.1). Here, a single point on the right side appears to be an outlier. If one uses a non-predictive goodness-of-fit test, such as Pierce (1982), Jiang (2001), and Claeskens and Hart (2009), none of these tests have rejected the linear model null hypothesis (see the last paragraph of Section 5.1). This suggests that these tests tend to look at the big picture, and therefore ignore the "outlier". On the other hand, our predictive-based test, that is, PL/tailoring, is able to reject the null hypothesis. This means that PL is taking the "outlier" more seriously by considering its potential impact on the prediction.

Regarding extension of the proposed method to other SAE models, note that tailoring is a general method that can be implemented as long as one has a base function, $B(\vartheta)$, in hand that satisfies certain conditions (see the first paragraph of Section 2). For example, to extend our GoFT to the nested-error regression (NER; Battese, Fuller and Harter 1988) model, we need to (I) set up a framework for GoFT; and (II) construct an appropriate base function. We discuss these two parts below. (I) An NER model can be expressed as $y_{ij} = x'_{ij}\beta + v_i + e_{ij}$, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$, where y_{ij} is the outcome variable, x_{ij} is a known vector of auxiliary variables, β is a vector of fixed effects, v_i is an area-specific random effect, and e_{ij} is an error. The standard assumptions are that the random effects and errors are independent with (i) $v_i \sim N(0, \sigma_v^2)$ and (ii) $e_{ij} \sim N(0, \sigma_e^2)$. Note that, unlike in the Fay-Herriot model, here, it may not be reasonable to assume that the distribution of e_{ij} is normal, because the central limit theorem (CLT) may not apply. Thus, to set up the GoFT framework, we may replace both (i) and (ii) by the skewed normal distribution family. (II) The general principle of PL (see the middle part of Section 3) still applies to this case. We just need to derive the resulting base function following the general steps and, with the base function, the resulting GoFT by the tailoring method. We will develop the details, and study performance of the resulting test in our future work.

Appendix

This appendix provides justification for the heuristic derivation given in Section 2 regarding the asymptotic null distribution as well as other details.

A.1 Notation and regularity conditions

Let ϑ_0 denote the true ϑ; $\|M\| = \sqrt{\lambda_{\max}(M'M)}$ the spectral norm of matrix M, where $\lambda_{\max}$ denotes the largest eigenvalue; $\lambda_{\min}$ the smallest eigenvalue; and $|v| = \sqrt{v'v}$ the Euclidean norm of vector v. Note that the matrices $A(\vartheta), B(\vartheta)$, etc. in Section 2 depend on the sample size, m, although the dependence will be implicit in notation.

Suppose that $B(\vartheta)$ in (1) can be expressed as $B(\vartheta) = V^{-1/2}(\vartheta)\sum_{i=1}^{m} b_i(Y_i, \vartheta)$, where $Y_1, \ldots, Y_m$ are independent random vectors, $\mathrm{E}_\vartheta\{b_i(Y_i, \vartheta)\} = 0$, $\mathrm{Var}_\vartheta\{b_i(Y_i, \vartheta)\}$ exists, and $V(\vartheta) = \sum_{i=1}^{m} \mathrm{Var}_\vartheta\{b_i(Y_i, \vartheta)\}$ is nonsingular. Then, with $N = m$ and $A(\vartheta)$ given by (11), the $C(\vartheta)$ in (3) can be expressed as

$$C(\vartheta) = \frac{1}{m}\sum_{i=1}^{m} \mathrm{E}_\vartheta \left(\frac{\partial B'}{\partial \vartheta}\right) V^{-1/2}(\vartheta) b_i(Y_i, \vartheta) = \frac{1}{m}\sum_{i=1}^{m} c_{m,i}(\vartheta)$$

with $c_{m,i}(\vartheta)$ defined in an obvious way. Denote $\Delta C(\vartheta) = (\partial/\partial\vartheta')C(\vartheta)$. We assume the following regularity conditions as $m \to \infty$:

A1. The parameter space of ϑ, Θ, is an open subset of R^r, and $c_{m,i}$ is continuously differentiable with respect to ϑ for each $1 \le i \le m$.

A2. With probability tending to one, $\Delta C(\vartheta_0)$ is nonsingular.

A3. $D(\vartheta) = \lim_{m\to\infty} \mathrm{E}\{\Delta C(\vartheta)\}$ exists, and there is a constant $\delta > 0$ such that

$$\sup_{|\vartheta - \vartheta_0| < \delta} \|\Delta C(\vartheta) - D(\vartheta)\| \xrightarrow{\mathrm{P}} 0.$$

A.2 Asymptotic behavior of $\hat{\vartheta}$

In this subsection, we state a result regarding existence, uniqueness, and consistency of $\hat{\vartheta}$, the solution to the tailoring equation (3) which is our estimator of ϑ used in (2). The proof is very similar to that of Theorem 2 of Foutz (1977), and therefore is omitted.

Lemma 1. Under assumptions *A1–A3*, there exists a sequence of estimators, $\hat{\vartheta}$, such that $C(\hat{\vartheta}) = 0$ with probability tending to one, and $\hat{\vartheta} \xrightarrow{\mathrm{P}} \vartheta_0$. Furthermore, if $\tilde{\vartheta}$ also satisfies the above, then $\mathrm{P}(\tilde{\vartheta} = \hat{\vartheta}) \to 1$ as $m \to \infty$.

A.3 Asymptotic distribution of $B(\hat{\vartheta})$

We assume the following additional regularity conditions:

A4. there is a full rank matrix, Q, such that

$$\frac{1}{\sqrt{m}}\mathrm{E}_{\vartheta_0}\left(\frac{\partial B}{\partial \vartheta'}\right) \longrightarrow Q \text{ and } \frac{1}{\sqrt{m}}\left\{\frac{\partial B}{\partial \vartheta'} - \mathrm{E}_{\vartheta_0}\left(\frac{\partial B}{\partial \vartheta'}\right)\right\} \xrightarrow{\mathrm{P}} 0,$$

where $\partial B/\partial \vartheta'$ is evaluated at ϑ_0.

A5. There is a compact subspace of $\Theta_{\mathrm{c}} \subset \Theta$ that contains ϑ_0 as an interior point such that the $\sup_{\vartheta\in\Theta_{\mathrm{c}}} \|\cdot\|$ of $V(\vartheta)/m$ and of its up to second partial derivatives (with respect to ϑ) are bounded, and $\liminf[\inf_{\vartheta\in\Theta_{\mathrm{c}}} \lambda_{\min}\{V(\vartheta)/m\}] > 0$.

A6. For the same Θ_{c}, the $\sup_{\vartheta\in\Theta_{\mathrm{c}}} \|\cdot\|$ of $m^{-1}\sum_{i=1}^{m} \mathrm{E}_{\vartheta}(\partial b_i/\partial\theta')$ and of its up to second partial derivatives (with respect to ϑ) are bounded; and the $\sup_{\vartheta\in\Theta_{\mathrm{c}}} \|\cdot\|$ of $m^{-1}\sum_{i=1}^{m} b_i$ and its up to second partial derivatives with respect to ϑ are bounded in probability.

A7. $\forall \epsilon > 0$, $\max_{1\le i\le m} \mathrm{E}_{\vartheta_0}\{b_i^2 1_{(|b_i|>\epsilon m)}\} \to 0$ as $m \to \infty$, where $b_i = b_i(Y_i, \vartheta_0)$.

Theorem 1. Let $\hat{\vartheta}$ denote the estimator in Lemma 1. Under assumptions *A1*–*A7*, we have $B(\hat{\vartheta}) \xrightarrow{\mathrm{d}} N(0, P)$, where $P = I_s - Q(Q'Q)^{-1}Q'$ is idempotent with rank $\nu = s - r$.

Proof: First, by assumptions *A5*, *A7*, and the central limit theorem for an array of independent random variables (e.g., Theorem 6.12 of Jiang 2010), it follows that

$$B(\vartheta_0) \xrightarrow{\mathrm{d}} N(0, I_s). \tag{A.1}$$

Next, by the Taylor series expansion, we have

$$0 = C(\hat{\vartheta}) = C(\vartheta_0) + \frac{\partial C}{\partial \vartheta'}(\hat{\vartheta} - \vartheta_0) + \frac{1}{2}\left[(\hat{\vartheta} - \vartheta_0)'\frac{\partial^2 C_{(k)}}{\partial\vartheta\partial\vartheta'}\right]_{1\le k\le r}(\hat{\vartheta} - \vartheta_0), \tag{A.2}$$

where $\partial C/\partial\vartheta'$ is evaluated at ϑ_0, and $\partial^2 C_{(k)}/\partial\vartheta\partial\vartheta'$ denotes the kth component of C evaluated at some $\vartheta_{(k)}$ that lies between ϑ_0 and $\hat{\vartheta}$. By assumptions *A4*–*A6*, it follows that

$$\frac{\partial C}{\partial \vartheta'} = Q'Q + o_{\mathrm{P}}(1). \tag{A.3}$$

Similarly, by assumptions *A5* and *A6*, it can be shown that

$$\frac{\partial^2 C_{(k)}}{\partial\vartheta\partial\vartheta'} = O_{\mathrm{P}}(1), \quad 1 \le k \le r. \tag{A.4}$$

By (A.2)–(A.4), and Lemma 1, we have $0 = C(\vartheta_0) + \{Q'Q + o_{\mathrm{P}}(1)\}(\hat{\vartheta} - \vartheta_0)$, or

$$\hat{\vartheta} - \vartheta_0 = -\{Q'Q + o_{\mathrm{P}}(1)\}^{-1}C(\vartheta_0). \tag{A.5}$$

On the other hand, again by the Taylor series expansion, we have

$$B(\hat{\vartheta}) = B(\vartheta_0) + \frac{\partial B}{\partial \vartheta'}(\hat{\vartheta} - \vartheta_0) + \frac{1}{2}\left[(\hat{\vartheta} - \vartheta_0)\frac{\partial^2 B_{(k)}}{\partial \vartheta \partial \vartheta'}\right]_{1 \le k \le s}(\hat{\vartheta} - \vartheta_0), \quad \text{(A.6)}$$

where $\partial B/\partial \vartheta'$ is evaluated at ϑ_0, and $\partial^2 B_{(k)}/\partial\vartheta\partial\vartheta'$ denotes the kth component of B evaluated at some $\vartheta_{(k)}$ that lies between ϑ_0 and $\hat{\vartheta}$. By assumption *A4* and (A.1), it is easy to see that $C(\vartheta_0) = O_{\mathrm{P}}(m^{-1/2})$. It follows by (A.5) that $\hat{\vartheta} - \vartheta_0 = O_{\mathrm{P}}(m^{-1/2})$. Therefore, by assumptions *A5* and *A6*, it can be shown that the last term on the right side of (A.6) is $o_{\mathrm{P}}(1)$. Furthermore, by assumption *A4*, we have

$$\sqrt{m}A(\vartheta_0) \longrightarrow Q', \quad \frac{1}{\sqrt{m}}\frac{\partial B}{\partial \vartheta'} \xrightarrow{\mathrm{P}} Q. \quad \text{(A.7)}$$

Combining (A.5)–(A.7), we have

$$\begin{aligned} B(\hat{\vartheta}) &= B(\vartheta_0) - \left(\frac{1}{\sqrt{m}}\frac{\partial B}{\partial \vartheta'}\right)\{Q'Q + o_{\mathrm{P}}(1)\}^{-1}\sqrt{m}A(\vartheta_0)B(\vartheta_0) \\ &= \{I_s - Q(Q'Q)^{-1}Q'\}B(\vartheta_0) + o_{\mathrm{P}}(1) \\ &\xrightarrow{\mathrm{d}} N(0, P), \end{aligned}$$

and $P = I_s - Q(Q'Q)^{-1}Q'$ is idempotent with $\mathrm{rank}(P) = s - r$.

Corollary 1. Under the conditions of Theorem 1, we have $|B(\hat{\vartheta})|^2 \xrightarrow{\mathrm{d}} \chi^2_{s-r}$.

A.4 Derivation of (18)

We have $f(\theta_i|y_i) = f(y_i|\theta_i)f(\theta_i)/\int f(y_i|\theta_i)f(\theta_i)d\theta_i = I_{i1}/I_{i2}$. For I_{i1}, we have

$$I_{i1} = \frac{2}{\sigma\sqrt{D_i}}\phi\left(\frac{y_i - \theta_i}{\sqrt{D_i}}\right)\phi\left(\frac{\theta_i - x_i'\beta}{\sigma}\right)\Phi\left(\alpha\frac{\theta_i - x_i'\beta}{\sigma}\right).$$

Next, we can show, after some simplification, that

$$\frac{(y_i - \theta_i)^2}{D_i} + \frac{(\theta_i - x_i'\beta)^2}{\sigma^2} = \frac{(\theta_i - \mu_i)^2}{\sigma_i^2} + \frac{(y_i - x_i'\beta)^2}{\sigma^2 + D_i},$$

where $\mu_i = (D_i x_i'\beta + \sigma^2 y_i)/(\sigma^2 + D_i)$ and $\sigma_i^2 = B_i D_i$ with $B_i = \sigma^2/(\sigma^2 + D_i)$. It follows that I_{i1} can be expressed as

$$I_{i1} = \frac{2}{\sigma\sqrt{D_i}}\phi\left(\frac{\theta_i - \mu_i}{\sigma_i}\right)\phi\left(\frac{y_i - x_i'\beta}{\sqrt{\sigma^2 + D_i}}\right)\Phi\left(\alpha\sqrt{1 - B_i}\frac{\theta_i - \mu_i}{\sigma_i} + \alpha\sqrt{B_i}\frac{y_i - x_i'\beta}{\sqrt{\sigma^2 + D_i}}\right).$$

Thus, we obtain the expression $f(\theta_i|y_i) =$

$$\left\{\sigma_i\Phi\left(\frac{\alpha\sqrt{B_i}d_i}{\sqrt{1 + \alpha^2(1 - B_i)}}\right)\right\}^{-1}\phi\left(\frac{\theta_i - \mu_i}{\sigma_i}\right)\Phi\left(\alpha\sqrt{1 - B_i}\frac{\theta_i - \mu_i}{\sigma_i} + \alpha\sqrt{B_i}d_i\right),$$

where $d_i = (y_i - x_i'\beta)/\sqrt{\sigma^2 + D_i}$, and we have used the following fact:

Azzalini and Capitanio (2014): Let $\phi(\cdot)$ and $\Phi(\cdot)$ denote pdf and cdf of the standard normal distribution. Then for all constants $a, b \in R$ and real value u, we have $\int_{-\infty}^{+\infty} \phi(u)\Phi(a + bu)du = \Phi(a/\sqrt{1 + b^2})$.

To calculate $f(y_i|y_i)$, we have

$$f(y_i|y_i) = \left\{\sigma_i\Phi\left(\frac{\alpha\sqrt{B_i}d_i}{\sqrt{1+\alpha^2(1-B_i)}}\right)\right\}^{-1}\frac{1}{\sqrt{D_i}}$$
$$\times\int_{-\infty}^{+\infty}\phi\left(\frac{y_i-\theta_i}{\sqrt{D_i}}\right)\phi\left(\frac{\theta_i-\mu_i}{\sigma_i}\right)\Phi\left(\alpha\sqrt{1-B_i}\frac{\theta_i-\mu_i}{\sigma_i}+\alpha\sqrt{B_i}d_i\right)d\theta_i.$$

It can now be shown, after some simplification, that

$$\frac{(y_i-\theta_i)^2}{D_i}+\frac{(\theta_i-\mu_i)^2}{\sigma_i^2}=\frac{(\theta_i-\mu_{\theta i})^2}{\sigma_{\theta i}^2}+\frac{(y_i-\mu_i)^2}{\sigma_i^2+D_i},$$

where $\mu_{\theta i} = (D_i\mu_i + \sigma_i^2 y_i)/(\sigma_i^2 + D_i)$ and $\sigma_{\theta i}^2 = \sigma_i^2 D_i/(\sigma_i^2 + D_i)$. Thus, we have

$$f(y_i|y_i) = \frac{1}{\sqrt{D_i}}\phi\left(\frac{y_i-\mu_i}{\sqrt{\sigma_i^2+D_i}}\right)\left\{\sigma_i\Phi\left(\frac{\alpha\sqrt{B_i}d_i}{\sqrt{1+\alpha^2(1-B_i)}}\right)\right\}^{-1}$$
$$\times\int_{-\infty}^{+\infty}\phi\left(\frac{\theta_i-\mu_{\theta i}}{\sigma_{\theta i}}\right)\Phi\left(\alpha\sqrt{\frac{D_i}{2\sigma^2+D_i}}\frac{\theta_i-\mu_{\theta i}}{\sigma_{\theta i}}+\alpha f_i\right)d\theta_i$$
$$=\left\{\sigma_i\Phi\left(\frac{\alpha\sqrt{B_i}d_i}{\sqrt{1+\alpha^2(1-B_i)}}\right)\right\}^{-1}\frac{\sigma_{\theta i}}{\sqrt{D_i}}\phi\left(\frac{y_i-\mu_i}{\sqrt{\sigma_i^2+D_i}}\right)$$
$$\times\Phi\left(\frac{\alpha f_i}{\sqrt{1+\alpha^2 D_i/(2\sigma^2+D_i)}}\right),$$

where $f_i = \{2\sigma/(2\sigma^2 + D_i)\}(y_i - x_i'\beta)$. From here it is easy to derive (18).

Supplementary Materials

The supplementary materials contain R codes and corresponding "readme" files for the simulation and application conducted in this work.

Acknowledgements

Constructive comments and suggestions of two referees, which led to an improved version of this article, are greatly appreciated. The research of Jiming Jiang is supported by the NSF grants SES-1121794 and DMS-1713120. The research of Mahmoud Torabi is supported by a grant from the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

1. Akaike H (1956) Information theory as an extension of the maximum likelihood principle, in *S*econd International Symposium on Information Theory (BN Petrov and F Csaki eds.), 267–281, Akademiai Kiado, Budapest
2. Azzalini A, Capitanio A (2014) The skew-normal and related families. Cambridge University Press, New York
3. Battese GE, Fuller WA, Harter R M (1988) An error-components model for prediction of county crop areas using survey and satellite data. J Am Stat Assoc 80: 28–36
4. Chatterjee S, Lahiri P, Li H (2008) Parametric bootstrap approximation to the distribution of EBLUP, and related prediction intervals in linear mixed models. Ann Stat 36:1221–1245
5. Claeskens G, Hart J D (2009) Goodness-of-fit tests in mixed models (with discussion). TEST 18:213–239
6. Dao C, Jiang J (2016) A modified Pearson's χ^2 test with application to generalized linear mixed model diagnostics. Ann Math Sci Appl 1:195–215
7. Datta GS, Hall P, Mandal A (2011) Model selection by testing for the presence of small-area effects, and application to area-level data. J Am Stat Assoc 106:362–374
8. Efron B, Hinkley DV (1978) Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information. Biometrika 65:457–487
9. Fay R E, Herriot RA (1979) Estimates of income for small places: an application of James-Stein procedures to census data. J Am Stat Assoc 74:269–277
10. Fisher RA (1922) On the interpretation of chi-square from contingency tables, and the calculation of P. J Roy Stat Soc 85:87–94
11. Foutz RV (1977) On the unique consistent solution to the likelihood equation. J Am Stat Assoc 72:147–148
12. Ganesh N (2009) Simultaneous credible intervals for small area estimation problems. J Mult Anal 100:1610–1621
13. Ghosh M, Nangia N, Kim D (1996) Estimation of median income of four-person families: A Bayesian time series approach. J Am Stat Assoc 91:1423–1431
14. Hall AR (2005) Generalized method of moments (Advanced Texts in Econometrics). Oxford University Press, Oxford
15. Hannan EJ, Quinn BG (1979) The determination of the order of an autoregression. J Roy Stat Soc B 41:190–195
16. Jiang J (2001) Goodness-of-fit tests for mixed model diagnostics. Ann Stat 29:1137–1164
17. Jiang J (2007) Linear and generalized linear mixed models and their applications. Springer, New York
18. Jiang J (2010) Large sample techniques for Statistics. Springer, New York
19. Jiang J, Nguyen T (2009) Comments on: Goodness-of-fit tests in mixed models by G Claeskens and JD Hart. TEST 18:248–255
20. Jiang J, Nguyen T, Rao JS (2010) Fence method for nonparametric small area estimation. Surv Method 36:3–11
21. Jiang J, Nguyen T, Rao JS (2011) Best predictive small area estimation. J Am Stat Assoc 106:732–745
22. Jiang J, Nguyen T, Rao JS (2015) Observed best prediction via nested-error regression with potentially misspecified mean and variance. Surv Method 41:37–55
23. Kauermann G, Carroll RJ (2001) A note on the efficiency of sandwich covariance matrix estimation. J Am Stat Assoc 96:1387–1396
24. Lahiri P, Rao JNK (1995) Robust estimation of mean squared error of small area estimators. J Am Stat Assoc 90:758–766
25. McCullagh P, Nelder JA (1989) Generalized linear models, 2nd edn. Chapman and Hall, London
26. Moore DS (1978) Chi-square tests, in *S*tudies in Statistics (RV Hogg, edn.), Mathematical Society of America, Providence, RI
27. Morris CN, Christiansen CL (1995) Hierarchical models for ranking and for identifying extremes with applications, in *B*ayes Statistics 5, Oxford: Oxford University Press
28. Pierce D (1982) The asymptotic effect of substituting estimators for parameters in certain types of statistics. Ann Stat 10:475–478

29. Prasad NGN, Rao JNK (1990) The estimation of mean squared error of small-area estimators. J Am Stat Assoc 85:163–171
30. Rao JNK, Molina I (2015) Small area estimation, 2nd edn. Wiley, New York
31. Rice JA (1995) Mathematical Statistics and data analysis, 2nd edn. Duxbury Press, Belmont, CA
32. Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464
33. Searle SR (1971) Linear models, Wiley, New York