# Saddlepoint Test in Measurement Error Models

Yanyuan MA and Elvezio RONCHETTI

We develop second-order hypothesis testing procedures in functional measurement error models for small or moderate sample sizes, where the classical first-order asymptotic analysis often fails to provide accurate results. In functional models no distributional assumptions are made on the unobservable covariates and this leads to semiparametric models. Our testing procedure is derived using saddlepoint techniques and is based on an empirical distribution estimation subject to the null hypothesis constraints, in combination with a set of estimating equations which avoid a distribution approximation. The validity of the method is proved in theorems for both simple and composite hypothesis tests, and is demonstrated through simulation and a farm size data analysis.

KEY WORDS: Empirical distribution function; Kullback–Leibler divergence; Relative error; Semiparametric estimation.

## 1. INTRODUCTION

Regression is arguably the most widely studied problem in statistics. Estimating and testing the parameters in the regression function thus has been well studied in the statistics literature and various approaches exist to best handle the situation at hand of a practitioner. However, when covariates are measured with error, the problem becomes more complex from a statistical point of view. A large body of literature in statistics, biostatistics, and econometrics has been devoted to the estimation and inference in regression problems such as linear or generalized linear models. For instance, Stefanski and Carroll (1987) derived unbiased estimating equations and the corresponding Wald tests for GLM with additive normal measurement error. For the same model, Buzas and Stefanski (1996) introduced estimating equations through instrumental variables. Finally, Hanfelt and Liang (1997) derived estimators and tests through an approximate likelihood. Excellent overviews are provided by the books Fuller (1987) and Carroll et al. (2006). In this article we consider the functional measurement error model approach by Tsiatis and Ma (2004) who constructed $\sqrt{n}$-consistent estimators for general regression measurement error models, where the estimators are obtained through solving an estimating equation of the form $\sum_{i=1}^{n} \boldsymbol{\psi}(\mathbf{A}_i; \boldsymbol{\beta}) = 0$ for $\boldsymbol{\beta}$, where $\mathbf{A}_1, \ldots, \mathbf{A}_n$ are $n$ independent identically distributed (iid) observations and $\boldsymbol{\psi}$ is the semiparametric efficient score derived under a possibly incorrect distribution of the unobserved variables. Explicit forms of $\boldsymbol{\psi}$ are available in important special cases, such as GLM with normal additive errors; see Appendix.

In spite of the availability of this estimation procedure, practical application remains an issue due to the sample size requirement. Although regularization or shrinkage type of methods have been proposed to handle small sample estimation (Fuller 1987, section 2.5), they only apply to very special models. In many situations, regression with errors in covariates are sufficiently complex that standard first order asymptotics provides reasonable approximations only for very large sample sizes. Simulation studies for these problems are typically carried out with sample sizes as large as 1000 in order to obtain a close approximation to root $n$ consistency for the estimator and 95% coverage for the corresponding confidence intervals, even when the estimation procedure is already optimal in the sense of Bickel et al. (1990). The difficulty is inherent to these models and an intuitive and informal understanding can be obtained by noting that in these models, the distribution of the unobservable latent variable in many situations can only be obtained at the $\log(n)$ rate, although being able to recover this distribution ensures the identifiability of such models.

A direct consequence on the restriction of sample size is the lack of accuracy of the subsequent testing procedure. Indeed the computation of $p$-values requires the knowledge of the distribution of the test statistic in its tails, which is harder to obtain and requires even larger sample sizes. For instance, the first-order optimal Wald test statistic is based on the estimator and its first-order mean and estimated variance, hence may not be sufficiently accurate for even moderate sample sizes in the measurement error model context.

In order to improve the accuracy of the testing procedure, we could consider two possible routes. The first one would be to compute adjustments to the asymptotic distributions of the test statistic by means of Edgeworth expansions, Bartlett corrections and related methods. However, these techniques would lead in general only to marginal improvements, in particular in terms of relative errors which is the meaningful measure in the tails of the distribution; see the remarks in Section 2. In this article we follow a second approach and we consider a new test for measurement error models, based on saddlepoint techniques. Saddlepoint approximations in statistics go back to the seminal article by Daniels (1954) and a rich literature has developed from there; cf., for instance, the books Field and Ronchetti (1990), Jensen (1995), and Butler (2007), and chapter 14 in Huber and Ronchetti (2009). These techniques provide extremely accurate approximations of tail probabilities of estimators and test statistics. Starting from the saddlepoint approximation of the distribution of $M$-estimators, Robinson, Ronchetti, and Young (2003) proposed a saddlepoint test statistic for testing hypotheses in general parametric models (see Appendix) which is asymptotically $\chi^2$-distributed as the three

Yanyuan Ma is Associate Professor, Department of Statistics, Texas A&M University, College Station, TX 77843-3143 (E-mail: *ma@stat.tamu.edu*). Elvezio Ronchetti is Professor, Research Center for Statistics and Department of Econometrics, University of Geneva, 1211 Geneva, Switzerland (E-mail: *elvezio.ronchetti@unige.ch*). The authors would like to thank the editor, the associate editor, two referees, and N. Wang for very constructive and helpful comments. The research of the first author was partially supported by a U.S. NSF grant DMS-0906341. The second author thanks the Department of Statistics, Texas A&M University for the hospitality during part of this research and the Swiss National Science Foundation Pro*Doc Program 114533 for partial financial support.


Journal of the American Statistical Association
March 2011, Vol. 106, No. 493, Theory and Methods
DOI: 10.1198/jasa.2011.tm10031

classical tests (Wald, score, likelihood ratio) but with a *relative error* of order $O(n^{-1})$. The test can be applied nonparametrically and recently Field, Robinson, and Ronchetti (2008) established its properties when testing a one-dimensional hypothesis. The main difference between the parametric and nonparametric setting is the availability of the distribution of the observations under the null hypothesis, which is used in constructing the test statistic. In the parametric setting, this distribution can be directly calculated under $H_0$ and using the estimated parameters, while in the nonparametric situation, one has to take an empirical distribution approach while satisfying the $H_0$ constraint. This can be obtained by means of the discrete distribution satisfying $H_0$ and closest in Kullback–Leibler divergence to the standard empirical distribution of the observations.

Regression with errors in the covariates falls into the category of semiparametric models, hence we need to take a combination of both, using the parametric model flavored $M$-estimator and the nonparametric model flavored empirical likelihood under constraints. By applying the empirical likelihood approach, we preserve a key advantage of the $M$-estimator in Tsiatis and Ma (2004), in that the distribution of the latent variable is never estimated.

The rest of the article is organized as follows. In Section 2, we describe the new testing procedure for the simple hypothesis case and establish its asymptotic distribution and relative error term under the null hypothesis. In Section 3, the procedure is extended to the composite hypothesis testing. We provide numerical results through simulations for linear and logistic models in Section 4 and illustrate the practical usefulness of the method in a data example in Section 5. Some discussion remarks are provided in Section 6. In the Appendix we give technical details and the conditions and proofs of the theorems.

## 2. SIMPLE HYPOTHESIS TESTING

Assume we have a regression model

$$Y = m(\mathbf{X}; \boldsymbol{\theta}) + \epsilon,$$

where $m$ is a known function up to the $d$-dimensional parameter $\boldsymbol{\theta}$ and $\epsilon$ is a random variable with conditional mean zero $E(\epsilon|\mathbf{X}) = 0$. Here and throughout the text, vectors are column vectors. In this section, we assume the distribution of $\epsilon$ is normal with known variance $\sigma^2$, although the following development does not rely on this assumption. In the situation when errors occur in the covariates, $\mathbf{X}$ is not observable and we observe instead $W$. Here we assume $\mathbf{W} = \mathbf{X} + \mathbf{U}$, where $\mathbf{U}$ is independent of $\mathbf{X}$ and has a known distribution. Note that the assumptions on $\mathbf{U}$ and $\epsilon$ are unnecessarily strong and are made here for the sole purpose of simplifying the simple hypothesis setting. These assumptions will be relaxed when we consider composite hypothesis testing in Section 3. We denote the iid observations $\mathbf{A}_i = (\mathbf{W}_i, Y_i)$, $i = 1, \ldots, n$.

In the simple hypothesis testing situation, we want to test $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$. When identifiable, root $n$ consistent estimators are known as the solution of a $d$-dimensional estimating equations of the form

$$\sum_{i=1}^{n} \boldsymbol{\psi}(\mathbf{a}_i; \boldsymbol{\theta}) = 0. \tag{1}$$

Here, the form of $\boldsymbol{\psi}(\mathbf{a}_i; \boldsymbol{\theta})$ is not explicit. We give the description on how to obtain $\boldsymbol{\psi}(\mathbf{a}_i; \boldsymbol{\theta})$ in the Appendix, and refer the readers to Tsiatis and Ma (2004) for more details on its derivation and properties. Notice however that our testing procedure is general and can be applied with other $\boldsymbol{\psi}$-functions.

Consistency implies that $E\{\boldsymbol{\psi}(\mathbf{A}_i; \boldsymbol{\theta})\} = 0$, and this reduces to the constraints that $E\{\boldsymbol{\psi}(\mathbf{A}_i; \boldsymbol{\theta}_0)\} = 0$ under $H_0$. Denote by $F_0$ the true cumulative distribution function (cdf) of $\mathbf{A}_1, \ldots, \mathbf{A}_n$ under $H_0$ and by $\widehat{F}_0$ its empirical cdf. The conventional approach to obtain $\widehat{F}_0$ is to follow Owen (2001), which results in maximizing the empirical likelihood under various constraints. Denote the empirical likelihood without constraints $\widehat{F} = (\frac{1}{n}, \ldots, \frac{1}{n})$. This conventional approach is equivalent to minimizing the (forward) Kullback–Leibler divergence between $\widehat{F}$ and $F_0 = (\omega_1, \ldots, \omega_n)$, that is,

$$d_{\mathrm{KL}}(\widehat{F}, F_0) = \sum_{i=1}^{n} \frac{1}{n} \log\left[\frac{1/n}{\omega_i}\right] = -\frac{1}{n}\sum_{i=1}^{n} \log \omega_i - \log n.$$

This gives Owen's empirical log-likelihood ratio test statistic which is asymptotically $\chi^2$-distributed under $H_0$ with absolute error of order $O(n^{-1/2})$ as its classical counterpart. Similarly, in a saddlepoint test, we need to construct a suitable nonparametric estimation under $H_0$ in order to retain the second-order property of the saddlepoint test statistic. Indeed the idea is to apply the parametric saddlepoint test (see Appendix) by replacing the underlying distribution of the observations under $H_0$ by its nonparametric version. This nonparametric estimation corresponds to obtaining $\widehat{F}_0$ through minimizing the (backward) Kullback–Leibler divergence, that is,

$$d_{\mathrm{KL}}(F_0, \widehat{F}) = \sum_{i=1}^{n} \omega_i \log\left[\frac{\omega_i}{1/n}\right] = \sum_{i=1}^{n} \omega_i \log \omega_i + \log n,$$

subject to the constraints mentioned above. Specifically, we need to solve a constrained minimization problem

$$\min_{\omega_1, \ldots, \omega_n} \sum_{i=1}^{n} \omega_i \log\left[\frac{\omega_i}{1/n}\right],$$

subject to $\omega_i \geq 0$, $\sum_{i=1}^{n} \omega_i = 1$, and $\sum_{i=1}^{n} \omega_i \boldsymbol{\psi}(\mathbf{A}_i; \boldsymbol{\theta}_0) = 0$. The resulting $\widehat{F}_0$ is of the form

$$\widehat{F}_0(\mathbf{a}) = \sum_{i=1}^{n} e^{\boldsymbol{\psi}^T(\mathbf{a}_i; \boldsymbol{\theta}_0)\boldsymbol{\mu}} I(\mathbf{a}_i \leq \mathbf{a}) \Big/ \sum_{i=1}^{n} e^{\boldsymbol{\psi}^T(\mathbf{a}_i; \boldsymbol{\theta}_0)\boldsymbol{\mu}},$$

where $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta}_0) \in \mathcal{R}^d$ satisfies

$$\sum_{i=1}^{n} e^{\boldsymbol{\psi}^T(\mathbf{a}_i; \boldsymbol{\theta}_0)\boldsymbol{\mu}} \boldsymbol{\psi}(\mathbf{a}_i; \boldsymbol{\theta}_0) = 0.$$

This is a distribution with point mass at $\mathbf{a}_i$ with weights $e^{\boldsymbol{\psi}^T(\mathbf{a}_i; \boldsymbol{\theta}_0)\boldsymbol{\mu}} / \sum_{i=1}^{n} e^{\boldsymbol{\psi}^T(\mathbf{a}_i; \boldsymbol{\theta}_0)\boldsymbol{\mu}}$ for $i = 1, \ldots, n$. The advantage of this choice is that the resulting test will have relative error properties; see below.

Following the basic idea of the saddlepoint test (see Appendix) and using $\widehat{F}_0$ as the underlying distribution $F_0$, a second-order test statistic $2n\widehat{h}(\widehat{\boldsymbol{\theta}})$ can be formed as $2n\widehat{h}(\widehat{\boldsymbol{\theta}}) = 2n \sup_{\boldsymbol{\lambda}}\{-K_{\psi}(\boldsymbol{\lambda}; \widehat{\boldsymbol{\theta}})\}$ where $\widehat{\boldsymbol{\theta}}$ is the solution to (1), and

$$K_{\psi}(\boldsymbol{\lambda}; \widehat{\boldsymbol{\theta}}) = \log\left[\sum_{i=1}^{n} e^{\boldsymbol{\psi}^T(\mathbf{a}_i; \widehat{\boldsymbol{\theta}})\boldsymbol{\lambda} + \boldsymbol{\psi}^T(\mathbf{a}_i; \boldsymbol{\theta}_0)\boldsymbol{\mu}} \Big/ \sum_{i=1}^{n} e^{\boldsymbol{\psi}^T(\mathbf{a}_i; \boldsymbol{\theta}_0)\boldsymbol{\mu}}\right].$$

Equivalently, the test statistic can be written as

$$2n\widehat{h}(\widehat{\boldsymbol{\theta}}) = -2nK_\psi\{\boldsymbol{\lambda}(\widehat{\boldsymbol{\theta}}); \widehat{\boldsymbol{\theta}}\},$$

where $\boldsymbol{\lambda}(\widehat{\boldsymbol{\theta}})$ satisfies

$$\sum_{i=1}^{n} \boldsymbol{\psi}(\mathbf{a}_i; \widehat{\boldsymbol{\theta}}) e^{\boldsymbol{\psi}^T(\mathbf{a}_i;\widehat{\boldsymbol{\theta}})\boldsymbol{\lambda}(\widehat{\boldsymbol{\theta}}) + \boldsymbol{\psi}^T(\mathbf{a}_i;\boldsymbol{\theta}_0)\boldsymbol{\mu}(\boldsymbol{\theta}_0)} = 0.$$

The statistic $2n\widehat{h}(\widehat{\boldsymbol{\theta}})$ constructed above has the desired second-order property. This means that roughly speaking, when the sample size $n$ becomes moderate or large, the distribution of $2n\widehat{h}(\widehat{\boldsymbol{\theta}})$ is very close to a $\chi^2$, hence we can calculate $p$-value and proceed with the testing decision based on the $\chi^2$ distribution. We state this asymptotic property in the following theorem.

*Theorem 1.* Assume regularity conditions (C1), (C2), (C3) in the Appendix. Then, when $n \to \infty$, the $p$-value satisfies

$$\begin{aligned} p &= P_{H_0}\{2n\widehat{h}(\widehat{\boldsymbol{\theta}}) \geq 2n\widehat{h}(\widehat{\boldsymbol{\theta}}_{\text{obs}})\} \\ &= \{1 - Q_d(2n\widehat{h}(\widehat{\boldsymbol{\theta}}_{obs}))\}\{1 + O_p(n^{-1})\}, \end{aligned}$$

where $Q_d$ is the cdf of the $\chi^2$ distribution with $d$ degrees of freedom.

Let us make a few remarks on the theorem. The proof is deferred to the Appendix.

*Remark 1.* In Theorem 1, the $O_p(n^{-1})$ term is multiplied by $1 - Q_d(\cdot)$, that is, it is a relative error. Note that in the testing framework, $1 - Q_d(\cdot)$ is often a fairly small quantity, hence a relative error is practically more meaningful than an absolute error. For comparison, the error of the usual first-order test statistic (such as the Wald test) is absolute and of order $O_p(n^{-1/2})$.

*Remark 2.* In addition to the important relative error property explained in Remark 1, the advantage of the second-order test statistic is evident also from the contrast in the order of the error term, which would reduce a typical sample size of $n$ to that of roughly $\sqrt{n}$ in order to exhibit the same asymptotic behavior. In practical terms this means if we use the proposed testing procedure on a data with size 100, we can achieve the same level accuracy had we used a conventional first order test on a data with size 10,000. In fact, as mentioned in Remark 1, even after the square root sample size reduction, the second-order test is still superior to the usual first-order tests, because of its bounded relative error.

## 3. COMPOSITE HYPOTHESIS TESTING

Given the same model specification and observations as described in Section 2, suppose we now want to perform a test for only the second subvector of $\boldsymbol{\theta}$. We thus test a composite hypothesis $H_0 : \boldsymbol{\theta}_2 = \boldsymbol{\theta}_{20}$, where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T$, $\boldsymbol{\theta}_1 \in R^{d_1}, \boldsymbol{\theta}_2 \in R^{d_2}$. Note that we can now relax the assumption on $\epsilon$ and $U$. Instead of assuming their respective distributions are known, we assume that the distributions contain additional unknown parameters, while the model is still identifiable. Considering that these additional parameters are often not of main interest, we include these additional parameters in $\boldsymbol{\theta}_1$, the subvector on which we do not perform any test. For example, when duplicate measurements (Huang and Wang 2001) or instrumental variables (Buzas and Stefanski 1996) are used to estimate the variance of $U$, we can include the unknown variance elements into $\boldsymbol{\theta}_1$, and append the corresponding estimating equation for these variance elements into $\boldsymbol{\psi}$. The composite testing procedure we describe in the following is similar to the simple hypothesis in Section 2.

We denote the estimating equation for $\boldsymbol{\theta}$ as

$$\sum_{i=1}^{n} \boldsymbol{\psi}(\mathbf{a}_i; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = 0. \tag{2}$$

The form of $\boldsymbol{\psi}(\mathbf{a}_i; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ follows the similar construction as in (1); the details are in the Appendix. Satisfying the consistency of $\boldsymbol{\psi}(\mathbf{A}_i; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, we construct the empirical cdf of $A_1, \ldots, A_n$ under $H_0$ through minimizing the backward Kullback–Leibler divergence $\sum_{i=1}^n \omega_i \log[\frac{\omega_i}{1/n}]$ subject to $\omega_i \geq 0$, $\sum_{i=1}^n \omega_i = 1$, and $\sum_{i=1}^n \omega_i \boldsymbol{\psi}(\mathbf{A}_i; \boldsymbol{\theta}_1, \boldsymbol{\theta}_{20}) = 0$ for some $\boldsymbol{\theta}_1$. Through a simple Lagrange multiplier and then maximizing once more with respect to $\boldsymbol{\theta}_1$ without constraint, we obtain the estimate

$$\widehat{F}_0(\mathbf{a}) = \sum_{i=1}^{n} e^{\boldsymbol{\psi}^T(\mathbf{a}_i;\boldsymbol{\theta}_1,\boldsymbol{\theta}_{20})\boldsymbol{\mu}} I(\mathbf{a}_i \leq \mathbf{a}) \Big/ \sum_{i=1}^{n} e^{\boldsymbol{\psi}^T(\mathbf{a}_i;\boldsymbol{\theta}_1,\boldsymbol{\theta}_{20})\boldsymbol{\mu}},$$

where $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta}_{20})$ and $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_1(\boldsymbol{\theta}_{20})$ satisfy

$$\sum_{i=1}^{n} e^{\boldsymbol{\psi}^T(\mathbf{a}_i;\boldsymbol{\theta}_1,\boldsymbol{\theta}_{20})\boldsymbol{\mu}} \boldsymbol{\psi}(\mathbf{a}_i; \boldsymbol{\theta}_1, \boldsymbol{\theta}_{20}) = 0,$$

$$\sum_{i=1}^{n} e^{\boldsymbol{\psi}^T(\mathbf{a}_i;\boldsymbol{\theta}_1,\boldsymbol{\theta}_{20})\boldsymbol{\mu}} \frac{\partial \boldsymbol{\psi}^T(\mathbf{a}_i; \boldsymbol{\theta}_1, \boldsymbol{\theta}_{20})}{\partial \boldsymbol{\theta}_1} \boldsymbol{\mu} = 0.$$

Similar to the simple hypothesis testing case, this is a discrete distribution with point mass at $\mathbf{a}_i$ with weights $e^{\boldsymbol{\psi}^T(\mathbf{a}_i;\boldsymbol{\theta}_1,\boldsymbol{\theta}_{20})\boldsymbol{\mu}} / \sum_{i=1}^n e^{\boldsymbol{\psi}^T(\mathbf{a}_i;\boldsymbol{\theta}_1,\boldsymbol{\theta}_{20})\boldsymbol{\mu}}$ for $i = 1, \ldots, n$.

Using $\widehat{F}_0$ as the distribution of $\mathbf{A}_i$'s under $H_0$, we can then construct the second-order test statistic $2n\widehat{h}(\widehat{\boldsymbol{\theta}}_2)$ from

$$2n\widehat{h}(\widehat{\boldsymbol{\theta}}_2) = 2n \inf_{\boldsymbol{\gamma}} \sup_{\boldsymbol{\lambda}} \{-K_\psi(\boldsymbol{\lambda}; \boldsymbol{\gamma}, \widehat{\boldsymbol{\theta}}_2)\},$$

where

$$K_\psi(\boldsymbol{\lambda}; \boldsymbol{\gamma}, \widehat{\boldsymbol{\theta}}_2) = \log\left[\sum_{i=1}^{n} e^{\boldsymbol{\psi}^T(\mathbf{a}_i;\boldsymbol{\gamma},\widehat{\boldsymbol{\theta}}_2)\boldsymbol{\lambda} + \boldsymbol{\psi}^T\{\mathbf{a}_i;\boldsymbol{\theta}_1(\boldsymbol{\theta}_{20}),\boldsymbol{\theta}_{20}\}\boldsymbol{\mu}(\boldsymbol{\theta}_{20})} \Big/ \sum_{i=1}^{n} e^{\boldsymbol{\psi}^T\{\mathbf{a}_i;\boldsymbol{\theta}_1(\boldsymbol{\theta}_{20}),\boldsymbol{\theta}_{20}\}\boldsymbol{\mu}(\boldsymbol{\theta}_{20})}\right],$$

and $\widehat{\boldsymbol{\theta}}_2$ contains the last $d_2$ components of $\widehat{\boldsymbol{\theta}}$ which solves (2). Simple calculations show that the test statistic can be written as

$$2n\widehat{h}(\widehat{\boldsymbol{\theta}}_2) = -2nK_\psi\{\boldsymbol{\lambda}(\widehat{\boldsymbol{\theta}}_2); \boldsymbol{\gamma}(\widehat{\boldsymbol{\theta}}_2), \widehat{\boldsymbol{\theta}}_2\},$$

where $\boldsymbol{\lambda} = \boldsymbol{\lambda}(\widehat{\boldsymbol{\theta}}_2)$, $\boldsymbol{\gamma} = \boldsymbol{\gamma}(\widehat{\boldsymbol{\theta}}_2)$ satisfies

$$\sum_{i=1}^{n} \omega_i e^{\boldsymbol{\psi}^T(\mathbf{a}_i;\boldsymbol{\gamma},\widehat{\boldsymbol{\theta}}_2)\boldsymbol{\lambda}} \boldsymbol{\psi}(\mathbf{a}_i; \boldsymbol{\gamma}, \widehat{\boldsymbol{\theta}}_2) = 0,$$

$$\sum_{i=1}^{n} \omega_i e^{\boldsymbol{\psi}^T(\mathbf{a}_i;\boldsymbol{\gamma},\widehat{\boldsymbol{\theta}}_2)\boldsymbol{\lambda}} \frac{\partial \boldsymbol{\psi}^T(\mathbf{a}_i; \boldsymbol{\gamma}, \widehat{\boldsymbol{\theta}}_2)}{\partial \boldsymbol{\gamma}} \boldsymbol{\lambda} = 0,$$

and $\omega_i = e^{\boldsymbol{\psi}^T\{\mathbf{a}_i;\boldsymbol{\theta}_1(\boldsymbol{\theta}_{20}),\boldsymbol{\theta}_{20}\}\boldsymbol{\mu}(\boldsymbol{\theta}_{20})} / \sum_{i=1}^n e^{\boldsymbol{\psi}^T\{\mathbf{a}_i;\boldsymbol{\theta}_1(\boldsymbol{\theta}_{20}),\boldsymbol{\theta}_{20}\}\boldsymbol{\mu}(\boldsymbol{\theta}_{20})}$.

The construction of the test statistic $2n\widehat{h}(\widehat{\boldsymbol{\theta}}_2)$ is based on the natural extension of the idea used in the simple hypothesis testing, with the obvious adaption to the fact that no properties are hypothesized about $\boldsymbol{\theta}_1$ under $H_0$. The only part that is not automatic is the minimization over the first component, denoted $\boldsymbol{\gamma}$, in the minimax procedure. The intuition of this operation can be gained as follows. For any fixed $\boldsymbol{\gamma}$, the test statistic would have been $-2n\sup_{\boldsymbol{\lambda}} K_\psi(\boldsymbol{\lambda}; \boldsymbol{\gamma}, \widehat{\boldsymbol{\theta}}_2)$, which describes the "rarity" of the observation under the $H_0$ model for that specific $\boldsymbol{\gamma}$. However, since the real null hypothesis does not impose any requirement on $\boldsymbol{\gamma}$, the $H_0$ we have at hand is much less restrictive, hence as long as we can find a "good" $\boldsymbol{\gamma}$ that constitutes a "reasonable" value of the test statistic, we should not reject $H_0$ even if all other values of $\boldsymbol{\gamma}$ would yield very "rare" values of the corresponding test statistic. Since the target test statistic is to follow a $\chi^2$ distribution, it is thus natural to select the minimum value of $-2n\sup_{\boldsymbol{\lambda}} K_\psi(\boldsymbol{\lambda}; \boldsymbol{\gamma}, \widehat{\boldsymbol{\theta}}_2)$ across all $\boldsymbol{\gamma}$ as the final test statistic. Thus when and only when the "most favorable" test statistic indicates a "unlikeliness" of the $H_0$, it would indicate a rejection of $H_0$ no matter what the first component $\boldsymbol{\theta}_1$ (or $\boldsymbol{\gamma}$) would be. The choice of taking the minimization can actually be shown to be the only right way of generalizing a simple hypothesis testing procedure to a composite one through considering $p$-values, see Gatto (2006, p. 286).

The rational of the testing procedure is almost identical as in the simple hypothesis case. Consider drawing a sample $\mathbf{A}_1^*, \ldots, \mathbf{A}_n^*$ from $\widehat{F}_0(\mathbf{a})$. Conceptually, if $F_0$ were known, we could construct the ideal test statistic $h(\widehat{\boldsymbol{\theta}}_2)$ from

$$h(\widehat{\boldsymbol{\theta}}_2) = \inf_{\boldsymbol{\gamma}} \sup_{\boldsymbol{\lambda}} \{-K_\psi(\boldsymbol{\lambda}; \boldsymbol{\gamma}, \widehat{\boldsymbol{\theta}}_2)\},$$

where

$$K_\psi(\boldsymbol{\lambda}; \boldsymbol{\gamma}, \widehat{\boldsymbol{\theta}}_2) = \log E_{F_0}\{e^{\boldsymbol{\psi}^T(\mathbf{A}_i; \boldsymbol{\gamma}, \widehat{\boldsymbol{\theta}}_2)\boldsymbol{\lambda}}\}.$$

This yields the resulting $h$ function as

$$h(\widehat{\boldsymbol{\theta}}_2) = -K_\psi\{\boldsymbol{\lambda}(\widehat{\boldsymbol{\theta}}_2); \boldsymbol{\gamma}(\widehat{\boldsymbol{\theta}}_2), \widehat{\boldsymbol{\theta}}_2\},$$

where $\boldsymbol{\lambda}(\widehat{\boldsymbol{\theta}}_2), \boldsymbol{\gamma}(\widehat{\boldsymbol{\theta}}_2)$ satisfy

$$E_{F_0}\{e^{\boldsymbol{\psi}^T(\mathbf{A}_i; \boldsymbol{\gamma}, \widehat{\boldsymbol{\theta}}_2)\boldsymbol{\lambda}} \boldsymbol{\psi}(\mathbf{A}_i; \boldsymbol{\gamma}, \widehat{\boldsymbol{\theta}}_2)\} = 0,$$

$$E_{F_0}\left\{e^{\boldsymbol{\psi}^T(\mathbf{A}_i; \boldsymbol{\gamma}, \widehat{\boldsymbol{\theta}}_2)\boldsymbol{\lambda}} \frac{\partial \boldsymbol{\psi}^T(\mathbf{A}_i; \boldsymbol{\gamma}, \widehat{\boldsymbol{\theta}}_2)}{\partial \boldsymbol{\gamma}} \boldsymbol{\lambda}\right\} = 0.$$

The following theorem provides the asymptotic distribution of the test statistic under $H_0 : \boldsymbol{\theta}_2 = \boldsymbol{\theta}_{20}$ in the composite hypothesis case.

*Theorem 2.* Assume the regularity conditions (C1), (C2), (C3) in the Appendix. Then, when $n \to \infty$, the $p$-value satisfies

$$\begin{aligned} p &= P_{H_0}\{2n\widehat{h}(\widehat{\boldsymbol{\theta}}_2) \geq 2n\widehat{h}(\widehat{\boldsymbol{\theta}}_{2,\text{obs}})\} \\ &= \{1 - Q_{d_2}(2n\widehat{h}(\widehat{\boldsymbol{\theta}}_{2,\text{obs}}))\}\{1 + O_p(n^{-1})\}, \end{aligned}$$

where $Q_{d_2}$ is the cdf of the $\chi^2$ distribution with $d_2$ degrees of freedom.

## 4. SIMULATION STUDY

We perform several simulation studies to demonstrate the performance of the second-order test in comparison to the traditional first-order test. To focus on the relative performance of the tests only, we use the same estimation procedure for both the first and the second-order tests. For simplicity, we considered only linear and logistic models with normal additive error, hence the estimator by Tsiatis and Ma (2004) reduces to the conditional score estimator by Stefanski and Carroll (1987), and this is what we implement. The first-order test is carried out using a Wald test most times. For situations where alternative testing procedures such as the adjusted test by Fuller (1987, section 2.5) or the quasi-likelihood based test by Hanfelt and Liang (1997) are applicable, we also included these for comparison.

We first illustrate the case of a simple hypothesis. We generated datasets from three regression models. The first one (model 1) is a linear regression model $Y = Z\beta_1 + X\beta_2 + \epsilon$, where $X$ is measured with error while $Z$ is without error. The second one (model 2) has an additional interaction term $Y = Z\beta_1 + X\beta_2 + XZ\beta_3 + \epsilon$, and the third one is a logistic model where the logit function of the conditional mean is $\text{logit}\{E(Y|X)\} = Z\beta_1 + X\beta_2$. In all the models, $Z$ and $X$ are generated from various uniform distributions while the measurement error on $X$ is generated from a normal distribution to achieve various reliability ratios. In the first two models, the model error $\epsilon$ is generated from a standard normal distribution. The true parameter value for $\beta$ is zero, which corresponds to no covariate effect, and we test $H_0 : \beta = 0$. We performed 100,000 simulations for various sample sizes $n = 50, 100, 150, 200, 250, 300$, and we report the most representative results for $n = 50, 100$. The Monte Carlo standard errors vary from 0.0001 for the nominal level 0.01 to 0.0004 for the nominal level 0.2. The results are summarized in Table 1. We can see that the second-order test performs consistently better than the Wald test and the adjusted test, while it is usually competitive or superior than the quasi-likelihood test, especially when the sample size is small.

To demonstrate the composite hypothesis testing performance, we test $\beta_2 = 0$ in model 1 (with corresponding coefficient of determination $R^2 = 1/13$) and model 3, and $\beta_3 = 0$ (with corresponding $R^2 = 17/29$) as well as $\beta_2 = \beta_3 = 0$ (with corresponding $R^2 = 1/13$) in model 2. These correspond to no covariate effect from $X$, no covariate effect from the intersection term and no covariate effect from $X$ and $XZ$. All the other unconstrained unknown parameters, including the parameters we are not testing and the measurement error variances are considered as nuisance parameters. Here, measurement error variances are estimated using repeated measurements. Table 2 contains the results of the three sets of composite tests for various sample sizes and various reliability ratios. Similarly, the second-order test often has superior performance especially when the sample size is small and the test level is far in the tails.

We also experimented with several situations when our model assumption is violated, including generating nonnormal measurement errors or model errors in the first two models. The second-order testing procedure demonstrated a certain robustness property. This can be understood as a result of the

Table 1. Level of the simple test

| | Reliability ratio 0.3 | | | | Reliability ratio 0.5 | | | | Reliability ratio 0.9 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Nominal | 0.01 | 0.05 | 0.1 | 0.2 | 0.01 | 0.05 | 0.1 | 0.2 | 0.01 | 0.05 | 0.1 | 0.2 |
| | | | | | | Model 1, $n = 50$ | | | | | | |
| 2nd | 0.0167 | 0.0697 | 0.1283 | 0.2356 | 0.0138 | 0.0640 | 0.1217 | 0.2303 | 0.0076 | 0.0410 | 0.0885 | 0.1929 |
| Wald | 0.0275 | 0.0862 | 0.1462 | 0.2518 | 0.0278 | 0.0872 | 0.1472 | 0.2524 | 0.0286 | 0.0883 | 0.1493 | 0.2553 |
| Adj | 0.0282 | 0.0877 | 0.1478 | 0.2531 | 0.0297 | 0.0897 | 0.1497 | 0.2536 | 0.0326 | 0.0933 | 0.1526 | 0.2551 |
| QL | 0.0054 | 0.0440 | 0.0882 | 0.1991 | 0.0051 | 0.0440 | 0.0882 | 0.2002 | 0.0059 | 0.0459 | 0.0895 | 0.2023 |
| | | | | | | Model 1, $n = 100$ | | | | | | |
| 2nd | 0.0133 | 0.0598 | 0.1145 | 0.2185 | 0.0122 | 0.0577 | 0.1116 | 0.2164 | 0.0074 | 0.0450 | 0.0954 | 0.2007 |
| Wald | 0.0178 | 0.0679 | 0.1227 | 0.2271 | 0.0182 | 0.0683 | 0.1231 | 0.2279 | 0.0185 | 0.0693 | 0.1235 | 0.2293 |
| Adj | 0.0181 | 0.0687 | 0.1245 | 0.2284 | 0.0192 | 0.0696 | 0.1251 | 0.2292 | 0.0216 | 0.0730 | 0.1274 | 0.2310 |
| QL | 0.0146 | 0.0554 | 0.0960 | 0.1915 | 0.0148 | 0.0553 | 0.0954 | 0.1910 | 0.0152 | 0.0553 | 0.0960 | 0.1915 |
| | | | | | | Model 2, $n = 50$ | | | | | | |
| 2nd | 0.0133 | 0.0644 | 0.1269 | 0.2463 | 0.0091 | 0.0491 | 0.1040 | 0.2191 | 0.0071 | 0.0391 | 0.0856 | 0.1911 |
| Wald | 0.0530 | 0.1302 | 0.1993 | 0.3135 | 0.0550 | 0.1338 | 0.2038 | 0.3176 | 0.0568 | 0.1385 | 0.2071 | 0.3208 |
| QL | 0.0050 | 0.0365 | 0.0826 | 0.1638 | 0.0054 | 0.0355 | 0.0817 | 0.1615 | 0.0057 | 0.0351 | 0.0819 | 0.1608 |
| | | | | | | Model 2, $n = 100$ | | | | | | |
| 2nd | 0.0121 | 0.0609 | 0.1205 | 0.2289 | 0.0089 | 0.0525 | 0.1094 | 0.2192 | 0.0070 | 0.0444 | 0.0950 | 0.2026 |
| Wald | 0.0276 | 0.0887 | 0.1499 | 0.2587 | 0.0284 | 0.0913 | 0.1527 | 0.2613 | 0.0297 | 0.0911 | 0.1521 | 0.2616 |
| QL | 0.0057 | 0.0398 | 0.0906 | 0.1800 | 0.0057 | 0.0388 | 0.0907 | 0.1778 | 0.0063 | 0.0393 | 0.0919 | 0.1754 |
| | | | | | | Model 3, $n = 50$ | | | | | | |
| 2nd | 0.0071 | 0.0432 | 0.0927 | 0.1962 | 0.0070 | 0.0428 | 0.0924 | 0.1965 | 0.0059 | 0.0402 | 0.0899 | 0.1930 |
| Wald | 0.0052 | 0.0387 | 0.0869 | 0.1906 | 0.0052 | 0.0388 | 0.0872 | 0.1906 | 0.0053 | 0.0387 | 0.0879 | 0.1910 |
| QL | 0.0065 | 0.0321 | 0.0785 | 0.1698 | 0.0136 | 0.0454 | 0.1054 | 0.1962 | 0.0064 | 0.0321 | 0.0807 | 0.1669 |
| | | | | | | Model 3, $n = 100$ | | | | | | |
| 2nd | 0.0079 | 0.0464 | 0.0962 | 0.1981 | 0.0077 | 0.0461 | 0.0962 | 0.1978 | 0.0072 | 0.0452 | 0.0949 | 0.1970 |
| Wald | 0.0069 | 0.0439 | 0.0935 | 0.1951 | 0.0069 | 0.0441 | 0.0933 | 0.1952 | 0.0071 | 0.0444 | 0.0937 | 0.1956 |
| QL | 0.0066 | 0.0319 | 0.0807 | 0.1707 | 0.0069 | 0.0317 | 0.0806 | 0.1708 | 0.0063 | 0.0322 | 0.0814 | 0.1681 |

robustness of the estimating equation, since the validity of the second-order test requires only the consistency of the estimating equation. Finally, we performed an empirical power study of the second-order test in comparison with the Wald test (Figure 1). Overall, the two tests are comparable at the practically useful power numbers (we reported 70% and 90% at level 5%), with neither test clearly dominating the other.

## 5. DATA EXAMPLE

To illustrate the method, we consider a farm size dataset, originally given in Fuller (1987, p. 201). The data contains information on farm sizes ($Y$), farmers' experience ($X_1$) and education ($X_2$), and the interest is to study the relation of the farm size to the two covariates. Both experience and education are measured with error, and the error variances are known to be 0.2013, 0.1808 respectively. In addition, the error variance on measuring the farm size is also known to be 0.0997. A model considered in Fuller (1987) has a linear form $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$, and based on the normality assumption of both $X_1$ and $X_2$, it is concluded that farm size is positively correlated with experience and education.

We relaxed the normality assumption of the unobservable covariates and performed the analysis on the original dataset with 176 observations. For this moderate sample size, both first-order and second-order analysis came to the same conclusion of not rejecting the zero slope hypotheses. We report the estimates, the standard error and the $p$-values of the two tests in Table 3. Note that despite the same conclusion of the two tests, the $p$-value about experience from the two tests are quite different, with the $p$-value of the Wald test much smaller than that of the second-order test. This raises the suspicion that the small $p$-value of the Wald test could be artifactual due to its finite sample performance under moderate sample sizes.

We thus considered a random subsample of size 50 to reflect further a small sample scenario. The scatter plot of the data set is in Figure 2. Using the methods described above, the estimate of the intercept is 5.26, while the slopes are $-0.46$ for experience and 0.27 for education. Using first order asymptotics, the standard deviation associated with the two slope values are 0.15 and 0.10, indicating significance of both covariates. In fact, using two separate Wald tests for zero coefficient of each covariate, we find the $p$-values to be 0.0024 and 0.0102, which would indicate that farm size is negatively correlated with experience and positively correlated with education.

However, considering that the measurement error is quite large in the data, and the sample size is rather small, it is more appropriate to use a second-order analysis. The result of this analysis indeed produced different results, with the corresponding $p$-values to be 0.1905 and 0.5446 respectively. Contrary to the first-order result, the conclusion here is that both covariates are not statistically significantly linked to farm size, a sample size 50 is simply too small to allows us to understand the relation between farm size and the education, experience of the

Table 2. Level of the composite test

| | Reliability ratio 0.3 | | | | Reliability ratio 0.5 | | | | Reliability ratio 0.9 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Nominal | 0.01 | 0.05 | 0.1 | 0.2 | 0.01 | 0.05 | 0.1 | 0.2 | 0.01 | 0.05 | 0.1 | 0.2 |
| | | | | | Model 1, $n = 50$ | | | | | | | |
| 2nd | 0.0143 | 0.0612 | 0.1163 | 0.2215 | 0.0104 | 0.0559 | 0.1116 | 0.2181 | 0.0053 | 0.0408 | 0.0936 | 0.2028 |
| Wald | 0.0189 | 0.0685 | 0.1234 | 0.2266 | 0.0196 | 0.0694 | 0.1247 | 0.2280 | 0.0196 | 0.0698 | 0.1248 | 0.2277 |
| Adj | 0.0232 | 0.0785 | 0.1366 | 0.2401 | 0.0243 | 0.0788 | 0.1353 | 0.2378 | 0.0254 | 0.0790 | 0.1349 | 0.2366 |
| QL | 0.0147 | 0.0629 | 0.1181 | 0.2248 | 0.0209 | 0.0774 | 0.1384 | 0.2494 | 0.0171 | 0.0674 | 0.1242 | 0.2317 |
| | | | | | Model 1, $n = 100$ | | | | | | | |
| 2nd | 0.0125 | 0.0565 | 0.1074 | 0.2092 | 0.0110 | 0.0541 | 0.1049 | 0.2082 | 0.0074 | 0.0474 | 0.0992 | 0.2034 |
| Wald | 0.0144 | 0.0597 | 0.1107 | 0.2122 | 0.0148 | 0.0604 | 0.1115 | 0.2130 | 0.0146 | 0.0603 | 0.1124 | 0.2137 |
| Adj | 0.0189 | 0.0687 | 0.1239 | 0.2287 | 0.0187 | 0.0697 | 0.1232 | 0.2266 | 0.0213 | 0.0742 | 0.1288 | 0.2320 |
| QL | 0.0138 | 0.0601 | 0.1137 | 0.2179 | 0.0197 | 0.0740 | 0.1344 | 0.2432 | 0.0149 | 0.0635 | 0.1186 | 0.2237 |
| | | | | | Model 2, $H_0 : \beta_3 = 0$, $n = 50$ | | | | | | | |
| 2nd | 0.0095 | 0.0595 | 0.1205 | 0.2316 | 0.0080 | 0.0554 | 0.1162 | 0.2292 | 0.0160 | 0.0679 | 0.1284 | 0.2396 |
| Wald | 0.0309 | 0.0921 | 0.1512 | 0.2564 | 0.0532 | 0.1280 | 0.1926 | 0.2998 | 0.2802 | 0.3904 | 0.4627 | 0.5583 |
| | | | | | Model 2, $H_0 : \beta_3 = 0$, $n = 100$ | | | | | | | |
| 2nd | 0.0117 | 0.0571 | 0.1132 | 0.2187 | 0.0109 | 0.0565 | 0.1116 | 0.2174 | 0.0149 | 0.0638 | 0.1195 | 0.2232 |
| Wald | 0.0203 | 0.0720 | 0.1286 | 0.2325 | 0.0347 | 0.0990 | 0.1633 | 0.2707 | 0.2371 | 0.3560 | 0.4320 | 0.5347 |
| | | | | | Model 2, $H_0 : \beta_2 = \beta_3 = 0$, $n = 50$ | | | | | | | |
| 2nd | 0.0093 | 0.0577 | 0.1202 | 0.2387 | 0.0071 | 0.0498 | 0.1092 | 0.2283 | 0.0027 | 0.0262 | 0.0693 | 0.1754 |
| Wald | 0.0420 | 0.1114 | 0.1764 | 0.2857 | 0.0444 | 0.1142 | 0.1791 | 0.2890 | 0.0585 | 0.1320 | 0.1980 | 0.3069 |
| | | | | | Model 2, $H_0 : \beta_2 = \beta_3 = 0$, $n = 100$ | | | | | | | |
| 2nd | 0.0116 | 0.0590 | 0.1150 | 0.2244 | 0.0100 | 0.0552 | 0.1107 | 0.2203 | 0.0045 | 0.0380 | 0.0884 | 0.1974 |
| Wald | 0.0245 | 0.0802 | 0.1370 | 0.2447 | 0.0252 | 0.0813 | 0.1393 | 0.2456 | 0.0291 | 0.0868 | 0.1462 | 0.2524 |
| | | | | | Model 3, $n = 50$ | | | | | | | |
| 2nd | 0.0085 | 0.0497 | 0.1023 | 0.2038 | 0.0084 | 0.0494 | 0.1021 | 0.2040 | 0.0084 | 0.0493 | 0.1019 | 0.2052 |
| Wald | 0.0074 | 0.0471 | 0.0997 | 0.2014 | 0.0074 | 0.0472 | 0.0999 | 0.2019 | 0.0073 | 0.0472 | 0.0996 | 0.2032 |
| QL | 0.0149 | 0.0603 | 0.1121 | 0.2139 | 0.0155 | 0.0611 | 0.1146 | 0.2161 | 0.0158 | 0.0626 | 0.1156 | 0.2192 |
| | | | | | Model 3, $n = 100$ | | | | | | | |
| 2nd | 0.0099 | 0.0499 | 0.1006 | 0.2020 | 0.0099 | 0.0501 | 0.1006 | 0.2018 | 0.0101 | 0.0501 | 0.1004 | 0.2017 |
| Wald | 0.0092 | 0.0486 | 0.0992 | 0.2008 | 0.0092 | 0.0490 | 0.0994 | 0.2006 | 0.0094 | 0.0488 | 0.0992 | 0.2004 |
| QL | 0.0122 | 0.0544 | 0.1051 | 0.2071 | 0.0127 | 0.0558 | 0.1070 | 0.2098 | 0.0133 | 0.0569 | 0.1087 | 0.2112 |

farmers. The subsample analysis results are summarized in Table 3.

The above analysis indicates that the conclusion we obtained through the Wald tests in the subsample analysis is a result of sampling randomness and it does not reflect the general relation between the random variables. The advantage of the second-order analysis is that even when working with a small subsample, we still will not be misled to false conclusion.

To see the effect of the small sample size and how often it affects the conclusion, we repeated the above subsample analysis on 100 randomly generated subsamples, and find that 14 of the 100 times, the first-order test would either detect the education or the experience to be a significant predictor of the farm size, while the second-order test consistently concludes nonsignificance for all the 100 subsamples. Extending this observation, we can understand the large difference between the $p$-values on experience of the Wald test and the second order test as a small-moderate sample size effect.

## 6. DISCUSSION

We have developed a second-order test under the measurement error model framework that can be used whenever an estimating equation is available. In order to provide an explicit expression for the score function $\boldsymbol{\psi}$, we have assumed in our case the error distribution to be normal, but we have found that the procedure is robust to this assumption. In addition, the method is applicable in more general situations than the measurement error models. Consider for instance the situation where we have observations $A_1, \ldots, A_n$ from a semiparametric model with probability density function (pdf) $f(A; \boldsymbol{\theta}, \eta)$, where the parameter $\boldsymbol{\theta} \in \mathbb{R}^d$ and $\eta$ is an infinite dimensional nuisance parameter. In measurement error model, $\eta$ corresponds to the distribution of the unobservable covariate; in general, it can be any other nuisance parameter. If in this class of semiparametric models, there exists an estimating equation $\boldsymbol{\psi}(\mathbf{A}_i; \boldsymbol{\theta}) \in \mathbb{R}^d$ so that $E\{\boldsymbol{\psi}(\mathbf{A}_i; \boldsymbol{\theta})\} = 0$, then our second-order testing method is applicable. This implies that the testing procedure can be applied to restricted moment models, mixture models, generalized linear latent variable models, etc.

## APPENDIX

### Description of $\boldsymbol{\psi}(\mathbf{a}_i; \boldsymbol{\theta})$ in (1)

First, we denote the score function $\partial \log p_{Y|X}(y|\mathbf{x}; \boldsymbol{\theta})/\partial \boldsymbol{\theta}$ as $\mathbf{S}_\theta^F(\mathbf{X}_i, Y_i; \boldsymbol{\theta})$. Specifically,

$$\mathbf{S}_\theta^F(\mathbf{X}_i, Y_i; \boldsymbol{\theta}) = \{Y_i - m(\mathbf{X}_i; \boldsymbol{\theta})\} m'_\theta(\mathbf{X}; \boldsymbol{\theta})/\sigma^2.$$

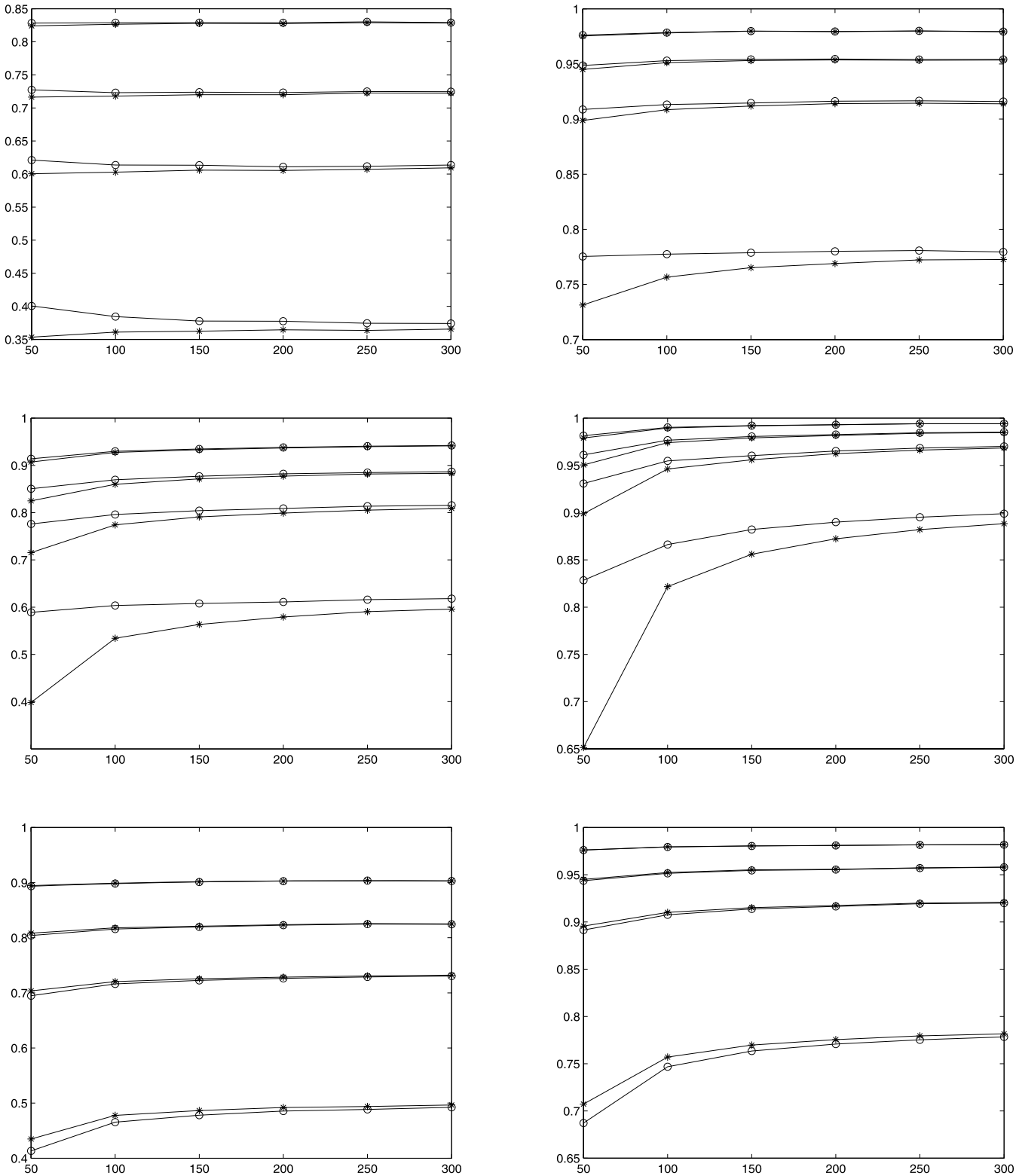


Figure 1. Power study of the second-order test (indicated with $*$) and the Wald test (indicated with $\circ$) for model 1 (first row), model 2 (second row), and model 3 (last row) at approximately 70% (left) and 90% (right) power for a level 5% test. Each plot contains four groups of curves, indicating the power of the test if the nominal Type I error is respectively 0.01, 0.05, 0.1, and 0.2.

Table 3. Estimates (est), standard deviation (sd) and $p$-values of the farm size data. The last block is the number of $p$-values smaller than 0.05 in the 100 subsample analysis

| | Intercept | Education | Experience |
|---|---|---|---|
| | | Complete data | |
| est | 5.8665 | 0.0607 | −0.2551 |
| sd | 0.6459 | 0.1777 | 0.1495 |
| $p$-value (Wald) | – | 0.7326 | 0.0878 |
| $p$-value (2nd order) | – | 0.7463 | 0.7264 |
| | | A subsample of size $n = 50$ | |
| est | 5.2553 | 0.2659 | −0.4585 |
| sd | 0.4727 | 0.1036 | 0.1508 |
| $p$-value (Wald) | – | 0.0102 | 0.0024 |
| $p$-value (2nd order) | – | 0.5446 | 0.1905 |
| | | 100 subsamples of size $n = 50$ | |
| Wald | – | 2 | 12 |
| 2nd order | – | 0 | 0 |

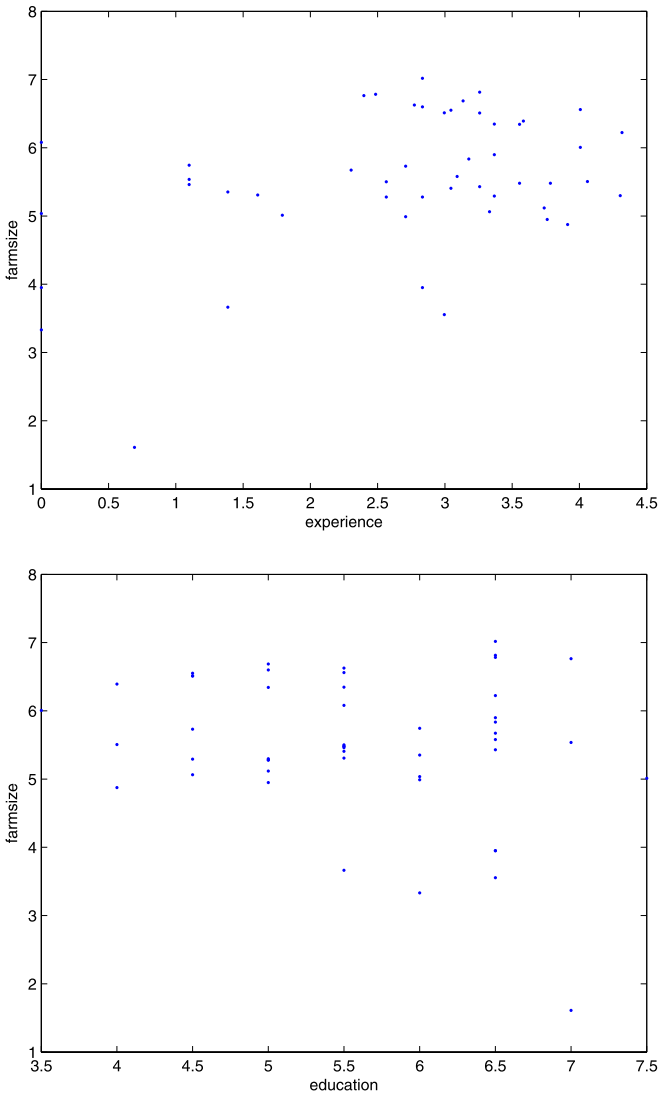


Figure 2. The farm size data. Left: farm size ($Y$ label) and experience ($X$ label). Right: farm size ($Y$ label) and education ($X$ label). The online version of this figure is in color.

We now adopt a pdf for $\mathbf{X}$, $f_X^*$, and use $E^*$ to denote expectations calculated using $f_X^*$. Note that $f_X^*$ does not need to be the true distribution of $\mathbf{X}$. We then solve for $\boldsymbol{\alpha}(\mathbf{X})$, which is a function that satisfies

$$E\big[E^*\{\boldsymbol{\alpha}(\mathbf{X})|\mathbf{W}, Y\}|\mathbf{X}\big] = E\big[E^*\{\mathbf{S}_\theta^F(\mathbf{X}_i, Y_i; \boldsymbol{\theta})|\mathbf{W}, Y\}|\mathbf{X}\big].$$

Finally, we form $\boldsymbol{\psi}(\mathbf{a}_i; \boldsymbol{\theta})$ through

$$\boldsymbol{\psi}(\mathbf{a}_i; \boldsymbol{\theta}) = E^*\{S_\theta^F(\mathbf{X}_i, Y_i; \boldsymbol{\theta})|\mathbf{W} = \mathbf{w}_i, Y = y_i\} - E^*\{\boldsymbol{\alpha}(\mathbf{X})|\mathbf{W} = \mathbf{w}_i, Y = y_i\}$$

for $i = 1, \ldots, n$.

In Tsiatis and Ma (2004), it is shown that the above $\boldsymbol{\psi}$ is guaranteed to yield a consistent estimating equation. Under special cases such as generalized linear model and normal additive errors, simplification occurs and an explicit form of $\boldsymbol{\psi}$ becomes available (Ma and Tsiatis 2006). As examples, in model 1 of the simulation, $\boldsymbol{\psi}(\mathbf{a}_i; \boldsymbol{\theta}) = (Z_i, \Delta_i)^T\{Y_i - (Z_i\beta_1 + \Delta_i\beta_2)/(1 + \beta_2^2 r)\}$, where $r = \sigma_u^2\sigma_\epsilon^{-2}$, $\sigma_\epsilon^2$ and $\sigma_u^2$ are respectively the variances of the model error and the measurement error, and $\Delta_i = W_i + Y_i\beta_2 r$. In model 2 of the simulation, $\boldsymbol{\psi}(\mathbf{a}_i; \boldsymbol{\theta}) = (Z_i, \Delta_i, Z_i\Delta_i)^T[Y_i - (Z_i\beta_1 + \Delta_i(\beta_2 + \beta_3 Z_i))\}/\{1 + (\beta_2 + \beta_3 Z_i)^2 r\}]$, where $r = \sigma_u^2\sigma_\epsilon^{-2}$, $\sigma_\epsilon^2$ and $\sigma_u^2$ are respectively the variances of the model error and the measurement error, and $\Delta_i = W_i + Y_i(\beta_2 + \beta_3 Z_i)r$. In model 3 of the simulation, $\boldsymbol{\psi}(\mathbf{a}_i; \boldsymbol{\theta}) = \{Z_i, \Delta_i + (d_i - 1)\sigma_u^2\beta_2\}^T(Y_i - d_i)$, where $\Delta_i = W_i + Y_1\beta_2\sigma_u^2$ and $d_i = (1 + e^{-\beta_1 Z_i - \beta_2\Delta_i + \beta_2^2\sigma_u^2/2})^{-1}$. In the data example, $\boldsymbol{\psi}(\mathbf{a}_i; \boldsymbol{\theta}) = (1, \Delta_i)^T\{Y_i - (\beta_0 + \beta_1\Delta_{i1} + \beta_2\Delta_{i2})/(1 + \beta_1^2 r_1 + \beta_2^2 r_2)\}$, where $\Delta_i = (W_{i1}, W_{i2})^T + Y_i(r_1\beta_1, r_2\beta_2)^T$, $r_1 = \sigma_{u_1}^2\sigma_\epsilon^{-2}$, $r_2 = \sigma_{u_2}^2\sigma_\epsilon^{-2}$, and $\sigma_\epsilon^2$, $\sigma_{u_1}^2$ and $\sigma_{u_2}^2$ are respectively the variances of the model error and the measurement error on $X_{i1}$ and on $X_{i2}$.

### Parametric Saddlepoint Test

For completeness, we summarize here the basic idea of the saddlepoint test for $M$-estimators as introduced by Robinson, Ronchetti, and Young (2003). For simplicity we consider the simple hypothesis situation as described in Section 2, (1). The saddlepoint test statistic is defined by

$$h(\hat{\boldsymbol{\theta}}) = \sup_{\boldsymbol{\lambda}}\{-K_\psi(\boldsymbol{\lambda}; \hat{\boldsymbol{\theta}})\}, \tag{A.1}$$

where $\hat{\boldsymbol{\theta}}$ is the solution of (1) and

$$K_\psi(\boldsymbol{\lambda}; \boldsymbol{\theta}) = \log E_{F_{\theta_0}}\big[e^{\boldsymbol{\lambda}^T\boldsymbol{\psi}(\mathbf{A}_i; \boldsymbol{\theta})}\big]$$

is the cumulant generating function of the score function $\boldsymbol{\psi}(\mathbf{A}_i; \boldsymbol{\theta})$.

Under $H_0$, $2nh(\hat{\boldsymbol{\theta}}) \xrightarrow{D} \chi_d^2$, with relative error of order $n^{-1}$.

In the composite hypothesis case, the nuisance parameters are marginalized by taking the inf over the nuisance parameters in (A.1).

This test statistic can be viewed as the Legendre transform of the score function $\boldsymbol{\psi}$ and is first-order equivalent to the three classical tests (Wald, score, and likelihood-ratio) but exhibits better second-order properties; see Remarks 1 and 2 in Section 2.

The test requires the existence of the cumulant generating function of the score function $\boldsymbol{\psi}(\mathbf{A}_i; \boldsymbol{\theta})$ computed with respect to the distribution of the observations under the null hypothesis. If $\boldsymbol{\psi}$ is bounded with respect to $\mathbf{A}_i$, this condition is always satisfied. If $\boldsymbol{\psi}$ is linear, then the underlying distribution must have exponential tails; cf. the discussion in Huber and Ronchetti (2009, remark in section 14.3).

### Regularity Conditions for Theorems 1 and 2

For completeness we list here the regularity conditions needed in the theorems as in Field, Robinson, and Ronchetti (2008).

(C1) $E\{\partial\boldsymbol{\psi}(\mathbf{A}; \boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}^T\}$ is an invertible matrix.

This condition is a standard requirement for regular asymptotically linear estimators and is satisfied in all the problems we consider. It ensures that there is a compact local neighborhood $B$ where $\boldsymbol{\theta}_0$ is an interior point, as long as $E\{\partial\boldsymbol{\psi}(\mathbf{A};\boldsymbol{\theta})/\partial\boldsymbol{\theta}^T\}$ is continuous with respect to $\boldsymbol{\theta}$, then $\boldsymbol{\theta}_0$ is the unique solution to $E\{\boldsymbol{\psi}(\mathbf{A};\boldsymbol{\theta})\}=0$ in $B$.

(C2) All the components in $\boldsymbol{\psi}(\mathbf{a};\boldsymbol{\theta})$ and their first four derivatives with respect to $\boldsymbol{\theta}$ exist and are bounded and continuous.

The third condition is a technical smoothness condition which is required to apply an Edgeworth expansion to the random vector $\mathbf{U}^\tau$ defined below and it is used in the classical saddlepoint analysis. Define $\mathbf{U}_{j\theta}$ to be the concatenation of $\mathbf{L}_{j\theta}, \mathbf{V}_{j\theta}$, and $\mathbf{Q}_{j\theta}$, where $\mathbf{L}_{j\theta} = \boldsymbol{\psi}(\mathbf{A}_j;\boldsymbol{\theta})$, $\mathbf{V}_{j\theta}$ is the vector formed by the elements of $\partial\boldsymbol{\psi}(\mathbf{A}_j;\boldsymbol{\theta})/\partial\boldsymbol{\theta}^T$ and $\boldsymbol{\psi}(\mathbf{A}_j;\boldsymbol{\theta})\boldsymbol{\psi}^T(\mathbf{A}_j;\boldsymbol{\theta})$, and $\mathbf{Q}_{j\theta}$ is formed by the the elements of $\partial\mathbf{V}_{j\theta}/\partial\boldsymbol{\theta}^T$. The dimension of $\mathbf{L}_{j,\theta}$ is $d$. Denote the dimension of $\mathbf{V}_{j\theta}$ and $\mathbf{Q}_{j\theta}$ to be $p$ and $r$. Define

$$\kappa(\boldsymbol{\tau};\boldsymbol{\theta}) = \log E\exp\{\boldsymbol{\tau}^T\boldsymbol{\psi}(\mathbf{A};\boldsymbol{\theta})\},$$

where $E$ is calculated under the true parameter value $\boldsymbol{\theta}_0$. Let $\boldsymbol{\tau}(\boldsymbol{\theta})$ be the solution to

$$\frac{\partial\kappa(\boldsymbol{\tau};\boldsymbol{\theta})}{\partial\boldsymbol{\tau}} = 0.$$

Denote the distribution function of $\mathbf{U}_{j\theta}$ to be $F_U$. Let $\mathbf{U}^\tau$ be a random vector with distribution function

$$F^\tau(\mathbf{l},\mathbf{v},\mathbf{q}) = \int_{(\mathbf{l}',\mathbf{v}',\mathbf{q}')\le(\mathbf{l},\mathbf{v},\mathbf{q})} e^{\boldsymbol{\tau}(\boldsymbol{\theta})^T\mathbf{l}'-\kappa\{\boldsymbol{\tau}(\boldsymbol{\theta});\boldsymbol{\theta}\}}\,dF_U(\mathbf{l}',\mathbf{v}',\mathbf{q}').$$

Denote further $\boldsymbol{\Sigma}_\tau = \operatorname{cov}\mathbf{U}^\tau$, $\boldsymbol{\mu}_\tau = E\mathbf{U}^\tau$. And let $\phi_\tau(\boldsymbol{\xi}) = Ee^{i\boldsymbol{\xi}^T\mathbf{U}^\tau}$.

(C3) There exist positive constants $c$, $C$, and $\rho$, such that $c < \det\boldsymbol{\Sigma}_\tau^{1/2} < C$, and $|\phi_\tau(\boldsymbol{\xi})| < 1-\rho$ for all $c < |\boldsymbol{\xi}| < Cn^{(d+p+r+1)/2}$.

### Proof of Theorem 1

We use results from the existing literature on bootstrap for second-order tests. Let $\mathbf{A}_1^*,\ldots,\mathbf{A}_n^*$ be a sample drawn from $\widehat{F}_0$. To make the notation more precise, assume $\widehat{\boldsymbol{\theta}}$ solves $\sum_{i=1}^n\boldsymbol{\psi}(\mathbf{a}_i;\widehat{\boldsymbol{\theta}})=0$, $\widehat{\boldsymbol{\Theta}}$ solves $\sum_{i=1}^n\boldsymbol{\psi}(\mathbf{A}_i;\widehat{\boldsymbol{\Theta}})=0$, and $\widehat{\boldsymbol{\Theta}}^*$ solves $\sum_{i=1}^n\boldsymbol{\psi}(\mathbf{A}_i^*;\widehat{\boldsymbol{\Theta}}^*)=0$. In the bootstrap world, when the sample space is $\{\mathbf{A}_1,\ldots,\mathbf{A}_n\}$, under $H_0$, the true distribution is $\widehat{F}_0$, and the ideal test statistic in this world is constructed exactly through $2n\widehat{h}(\cdot)$, where "$\cdot$" will be replaced with the estimated value from the corresponding estimating equations. Then, under the above conditions (C1), (C2), (C3), using the result in Field, Robinson, and Ronchetti (2008) there exists a saddlepoint approximation to the distribution of $\widehat{\boldsymbol{\Theta}}^*$ in the bootstrap case (in which the underlying distribution does not have a density). Therefore, equation (1.6) in Robinson, Ronchetti, and Young (2003) can be applied to get the $p$-value $p^*\{2n\widehat{h}(\widehat{\boldsymbol{\theta}})\}$ in the bootstrap world for a fixed $\widehat{h}(\widehat{\boldsymbol{\theta}})$

$$\begin{aligned}p^*\{2n\widehat{h}(\widehat{\boldsymbol{\theta}})\} &= P^*_{H_0}\{2n\widehat{h}(\widehat{\boldsymbol{\Theta}}^*)\ge 2n\widehat{h}(\widehat{\boldsymbol{\theta}})\}\\ &= \{1-Q_d(n\widehat{u}^2)\}\big\{1+O((1+n\widehat{u}^2)/n)\big\},\end{aligned}$$

where $\widehat{u}=\sqrt{2\widehat{h}(\widehat{\boldsymbol{\theta}})}$. Since $\widehat{h}(\widehat{\boldsymbol{\theta}})=O_p(n^{-1})$, this leads to

$$p^*\{2n\widehat{h}(\widehat{\boldsymbol{\theta}})\} = \{1-Q_d(n\widehat{u}^2)\}\{1+O(n^{-1})\}.$$

Moreover, from Field, Robinson, and Ronchetti (2008, theorem 3), the $p$-value in the original world, $p\{2n\widehat{h}(\widehat{\boldsymbol{\theta}})\}$, is related to $p^*\{2n\widehat{h}(\widehat{\boldsymbol{\theta}})\}$ through

$$\begin{aligned}p\{2n\widehat{h}(\widehat{\boldsymbol{\theta}})\} &= p^*\{2n\widehat{h}(\widehat{\boldsymbol{\theta}})\}\big[1+O_p\{\sqrt{n}\widehat{h}(\widehat{\boldsymbol{\theta}})^3\vee n^{-1}\}\big]\\ &= p^*\{2n\widehat{h}(\widehat{\boldsymbol{\theta}})\}\{1+O(n^{-1})\}\end{aligned}$$

using again that $\widehat{h}(\widehat{\boldsymbol{\theta}})=O_p(n^{-1})$. Hence the $p$-value of the observed sample is

$$\begin{aligned}p\{2n\widehat{h}(\widehat{\boldsymbol{\theta}})\} &= P_{H_0}\{2n\widehat{h}(\widehat{\boldsymbol{\Theta}})\ge 2n\widehat{h}(\widehat{\boldsymbol{\theta}})\}\\ &= \big\{1-Q_d(2n\widehat{h}(\widehat{\boldsymbol{\theta}}))\big\}\{1+O_p(n^{-1})\}\end{aligned}$$

and this proves the theorem.

### Description of $\boldsymbol{\psi}(\mathbf{a}_i;\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$ in (2)

Assume the first two components of $\boldsymbol{\theta}_1$ are the standard deviation of $\epsilon$, $\sigma_\epsilon$, and that of $U$, $\sigma_U$. Then the score function has the first two components

$$\begin{bmatrix}\{Y_i-m(X_i;\boldsymbol{\theta})\}^2/\sigma_\epsilon^3-1/\sigma_\epsilon\\ (W_i-X_i)^2/\sigma_U^3-1/\sigma_U\end{bmatrix}.$$

The remaining components of the score function are of the same form as the score function used in constructing $\boldsymbol{\psi}(\mathbf{a}_1;\boldsymbol{\theta})$ in (1).

Using the new score function, following the same procedure as before, one can construct the $\boldsymbol{\psi}(\mathbf{a}_1;\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$ in (2).

### Proof of Theorem 2

This proof follows the same lines as that of Theorem 1. Assume $\widehat{\boldsymbol{\theta}}_2$ solves $\sum_{i=1}^n\boldsymbol{\psi}(\mathbf{a}_i;\widehat{\boldsymbol{\theta}}_2)=0$, $\widehat{\boldsymbol{\Theta}}_2$ solves $\sum_{i=1}^n\boldsymbol{\psi}(\mathbf{A}_i;\mathbf{V},\widehat{\boldsymbol{\Theta}}_2)=0$, and $\widehat{\boldsymbol{\Theta}}_2^*$ solves $\sum_{i=1}^n\boldsymbol{\psi}(\mathbf{A}_i^*;\mathbf{V}^*,\widehat{\boldsymbol{\Theta}}_2^*)=0$. In the bootstrap world, when the sample space is $\{\mathbf{A}_1,\ldots,\mathbf{A}_n\}$, under $H_0$, the true distribution is $\widehat{F}_0$, and the ideal test statistic in this world is constructed exactly using $\widehat{h}$. Then, by the same arguments as in the proof of Theorem 1 and by noticing that the additional condition (A2) in Robinson, Ronchetti, and Young (2003, p. 1160), is clearly satisfied in our case, we obtain

$$\begin{aligned}p^*\{2n\widehat{h}(\widehat{\boldsymbol{\theta}}_2)\} &= P^*_{H_0}\{2n\widehat{h}(\widehat{\boldsymbol{\Theta}}_2^*)\ge 2n\widehat{h}(\widehat{\boldsymbol{\theta}}_2)\}\\ &= \big\{1-Q_{d_2}(n\widehat{u}^2)\big\}\{1+O(n^{-1})\},\end{aligned}$$

where $\widehat{u}=\sqrt{2\widehat{h}(\widehat{\boldsymbol{\theta}}_2)}$ and

$$p\{2n\widehat{h}(\widehat{\boldsymbol{\theta}}_2)\} = p^*\{2n\widehat{h}(\widehat{\boldsymbol{\theta}}_2)\}\{1+O(n^{-1})\}.$$

This proves the theorem.

*[Received January 2010. Revised September 2010.]*

## REFERENCES


Bickel, P., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1990), *Efficient and Adaptive Inference in Semiparametric Models*, New York: Springer. [147]

Butler, R. W. (2007), *Saddlepoint Approximations With Applications*, Cambridge: Cambridge University Press. [147]

Buzas, J. S., and Stefanski, L. A. (1996), "Instrumental Variable Estimation in Generalized Linear Measurement Error Models," *Journal of the American Statistical Association*, 91, 999–1006. [147,149]

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. (2006), *Measurement Error in Nonlinear Models: A Modern Perspective* (2nd ed.), London: CRC Press. [147]

Daniels, H. E. (1954), "Saddlepoint Approximations in Statistics," *Annals of Mathematical Statistics*, 25, 631–650. [147]

Field, C., and Ronchetti, E. (1990), *Small Sample Asymptotics. Institute of Mathematical Statistics Lecture Notes—Monograph Series*, Hayward, CA: Institute of Mathematical Statistics. [147]

Field, C., Robinson, J., and Ronchetti, E. (2008), "Saddlepoint Approximations for Multivariate *M*-Estimators With Applications to Bootstrap Accuracy," *Annals of the Institute of Statistical Mathematics*, 60, 205–224. [148,154,155]

Fuller, W. A. (1987), *Measurement Error Models*, New York: Wiley. [147,150,151]

Gatto, R. (2006), "A Bootstrap Test for Circular Data," *Cummunications in Statistics*, 35, 281–292. [150]

Hanfelt, J. J., and Liang, K. Y. (1997), "Approximate Likelihoods for Generalized Linear Errors-in-Variables Models," *Journal of the Royal Statistical Society, Ser. B*, 59, 627–637. [147,150]

Huang, Y., and Wang, C. Y. (2001), "Consistent Functional Methods for Logistic Regression With Errors in Covariates," *Journal of the American Statistical Association*, 95, 1209–1219. [149]

Huber, P. J., and Ronchetti, E. (2009), *Robust Statistics* (2nd ed.), New York: Wiley. [147,154]

Jensen, J. L. (1995), *Saddlepoint Approximations*, London: Oxford University Press. [147]

Ma, Y., and Tsiatis, A. A. (2006), "Closed Form Semiparametric Estimators for Measurement Error Models," *Statistica Sinica*, 16, 183–193. [154]

Owen, A. (2001), *Empirical Likelihood*, New York: Chapman & Hall/CRC. [148]

Robinson, J., Ronchetti, E., and Young, G. A. (2003), "Saddlepoint Approximations and Tests Based on Multivariate *M*-Estimates," *The Annals of Statistics*, 31, 1154–1169. [147,154,155]

Stefanski, L. A., and Carroll, R. J. (1987), "Conditional Scores and Optimal Scores for Generalized Linear Measurement-Error Models," *Biometrika*, 74, 703–716. [147,150]

Tsiatis, A. A., and Ma, Y. (2004), "Locally Efficient Semiparametric Estimators for Functional Measurement Error Models," *Biometrika*, 91, 835–848. [147, 148,150,154]