Biometrika (2010), **97**, ?, *pp*. 1–9

Printed in Great Britain

A goodness-of-fit test of logistic regression models for case-control data with measurement errors

BY GANGGANG XU AND SUOJIN WANG

Department of Statistics, Texas A&M University, College Station, Texas 77843, USA

gang@stat.tamu.edu sjwang@stat.tamu.edu

SUMMARY

We study the problem of goodness-of-fit tests for logistic regression models for case-control data when some covariates are measured with errors. We first study the applicability of traditional test methods for this problem by simply ignoring measurement errors and show that in some scenarios they are still effective despite the inconsistency of the parameter estimators. We then develop a test procedure based on Zhang (2001) that can simultaneously test the validity of using logistic regression and correct the bias in parameter estimators for case-control data with nondifferential classical additive normal measurement error. Instead of using the information matrix considered by Zhang (2001), our test statistic uses a collection of preselected functions to reduce dimensionality. Simulation studies and an application are carried out to illustrate the usefulness of the test.

Some key words: Case-control study, Conditional score, Empirical likelihood, Logistic regression, Measurement error.

1. INTRODUCTION

Logistic regression has been an extensively used tool to analyze binary data collected from case-control studies in many research areas such as epidemiology, sociology, and biology, see, for example, Prentice & Pyke (1979). Therefore, testing the validity of using the logistic regression model in a case-control study is of great importance. Letting Y be the binary response variable and (Z, X) be covariates, we wish to test the following model:

$$\mathrm{pr}(Y = 1 \mid Z, X) = H(\beta_0^* + Z^{\mathrm{T}}\beta_Z + X^{\mathrm{T}}\beta_X), \tag{1}$$

where $H(x) = 1/\{1 + \exp(-x)\}$, β_0^* is a scalar parameter, β_Z and β_X are p_Z and p_X vectors. Various goodness-of-fit tests have been proposed for model (1), including parametric methods, see Su & Wei (1991) and Lin et al. (2002), nonparametric methods (Azzalini et al., 1989) and semi-parametric methods, see Qin & Zhang (1997), Zhang (2001) and Bondell (2007).

The process of data collection in case-control studies can be viewed as taking two independent samples $\{(Z_1^c, X_1^c), \ldots, (Z_{n_0}^c, X_{n_0}^c)\}$ and $\{(Z_1^t, X_1^t), \ldots, (Z_{n_1}^t, X_{n_1}^t)\}$ from the control population and the case population, respectively. Throughout the paper, we use $f(\omega_1, \omega_2 \mid Y = i)$ as the notation of conditional density function of random variable (ω_1, ω_2) given Y for $i = 0, 1$. It was shown in Qin & Zhang (1997) that, model (1) is equivalent to the following model:

$$h_1(Z, X)/h_0(Z, X) = \exp(\beta_0 + Z^{\mathrm{T}}\beta_Z + X^{\mathrm{T}}\beta_X), \tag{2}$$

where $h_i(Z, X) = f(Z, X \mid Y = i)$, $i = 0, 1$, $\beta_0 = \beta_0^* + \log\{(1 - \pi)/\pi\}$, and $\pi = \mathrm{pr}(Y = 1)$. Using model (2), Zhang (2001) proposed an information matrix test for model (1) with a test statistic having a closed form asymptotic distribution.

Another important issue arising in many case-control studies is that some covariates are measured with errors. Extensive work has been done in this area, see Carroll et al. (1993). In this paper, we assume two types of covariates: (i) Z is a vector of covariates measured without error; (ii) X is a vector of covariates measured with errors or covariates which are difficult to measure and a surrogate W is observed instead of the true X. Most existing methods focus only on correcting the bias in parameter estimation caused by ignoring the measurement error, see Stefanski & Carroll (1987), Carroll et al. (1993) and Carroll et al. (2006). However, when some covariates are measured with errors, the problem of validating the use of model (1) before applying these bias-correction techniques has not received the attention it deserves in the literature.

2. TWO SPECIAL MEASUREMENT ERROR CASES

Throughout the paper, we assume nondifferential error structure that the surrogate W is independent of Y given (Z, X). With X unobserved, both models (1) and (2) are not directly testable. Define $g_i^*(Z, W) = f(Z, W \mid Y = i)$, $i = 0, 1$. If we ignore the measurement errors in covariates and use W instead of X to test model (1), we would test the model:

$$g_1^*(Z, W)/g_0^*(Z, W) = \exp(\beta_0' + Z^{\mathrm{T}}\beta_Z' + W^{\mathrm{T}}\beta_X'). \tag{3}$$

However, model (1) generally does not lead to model (3). Wang et al. (1997) showed that, under model (1) and the nondifferential error assumption, one has

$$g_1^*(Z, W)/g_0^*(Z, W) = \exp(\beta_0 + Z^{\mathrm{T}}\beta_Z)E\{\exp(X^{\mathrm{T}}\beta_X) \mid Z, W, Y = 0\}. \tag{4}$$

By (4), we see that unless the quantity $\log[E\{\exp(X^{\mathrm{T}}\beta_X) \mid Z, W, Y = 0\}]$ is linear in W and Z, we would end up with rejecting model (1) even if it is true, which leads to a test with inflated sizes. Below are two special cases where model (1) leads to model (3).

Example 1. As in Amstrong et al. (1989) and Kim & Jacquotte (1997), assuming $(X^{\mathrm{T}}, W^{\mathrm{T}})^{\mathrm{T}} \mid Y = 0$ is normally distributed with mean $(\mu_X^{\mathrm{T}}, \mu_W^{\mathrm{T}})^{\mathrm{T}}$ and variance $[(\Sigma_X, \Sigma_{XW})^{\mathrm{T}}, (\Sigma_{WX}, \Sigma_W)^{\mathrm{T}}]$ and is independent of Z, model (1) implies model (3) with $\beta_0' = \beta_0 + (\mu_X - \Sigma_{XW}\Sigma_W^{-1}\mu_W)^{\mathrm{T}}\beta_X + \frac{1}{2}\beta_X^{\mathrm{T}}(\Sigma_X - \Sigma_{XW}\Sigma_W^{-1}\Sigma_{WX})\beta_X$, $\beta_Z' = \beta_Z$ and $\beta_X' = \Sigma_W^{-1}\Sigma_{WX}\beta_X$. In fact, if $(Z, X, W) \mid Y = 0$ is normally distributed, this conclusion still holds.

Example 2. Another popular measurement error model called linear regression calibration model (Carroll et al., 2006) assumes that $X = \gamma_0 + \gamma_Z^{\mathrm{T}}Z + \gamma_W^{\mathrm{T}}W + U$, where U is independent of (Z, W) with $E(U \mid Z, W) = 0$ and γ_Z and γ_W are $p_Z \times p_X$ and $p_X \times p_X$ matrices, respectively. Under this measurement error structure, model (1) leads to model (3). In this case, $\beta_0' = \beta_0 + \gamma_0^{\mathrm{T}}\beta_X + \log[E\{\exp(U^{\mathrm{T}}\beta_X)\}]$, $\beta_Z' = \beta_Z + \gamma_Z\beta_X$ and $\beta_X' = \gamma_W\beta_X$. One well-known special case is the Berkson error assuming that $X = W + U$.

If our purpose is merely to conduct the goodness-of-fit test and ignore the bias in parameter estimation, traditional goodness of fit tests, such as Zhang (2001) and Lin et al. (2002), can still be used in the two examples above by ignoring the measurement errors in X, although they generally fail to produce consistent estimators of the original parameters $\Theta = (\beta_0, \beta_Z^{\mathrm{T}}, \beta_X^{\mathrm{T}})^{\mathrm{T}}$.

3. CLASSICAL MEASUREMENT ERROR

3·1. *Model specification and parameter estimation*

In general, blindly applying a goodness of fit test to the observed data (Y, Z, W) without adjusting for errors in X would lead to unreliable test and estimation results. Here we propose an approach that can simultaneously test the validity of using logistic regression for case-control data and correct the estimation bias under the nondifferential classical additive error model:

$$W = X + U, \tag{5}$$

where $U \sim N(0, \Sigma_U)$ with Σ_U known. Here X is not necessarily normally distributed, but the moment generating function of $X \mid Z, W, Y = 0$ is assumed to exist at β_X due to (4). As in Stefanski & Carroll (1987), we define a pseudo-value for X as $\Delta = W + Y\Sigma_U\beta_X$, which is originally proposed in a prospective perspective where the sample is taken from the whole population. In the case-control study context, it basically tells us that for the control group, we would use W as X but for the case group, a shift from the observed data W is taken to generate a new pseudo covariate value. Define $g_i(Z, \Delta) = f(Z, \Delta \mid Y = i)$, $i = 0, 1$. As is shown in the supplementary material, under error structure (5), model (1) leads to the following equation

$$g_1(Z, \Delta)/g_0(Z, \Delta) = w(Z, \Delta, \Theta), \tag{6}$$

where $\Theta = (\beta_0, \beta_Z^{\mathrm{T}}, \beta_X^{\mathrm{T}})^{\mathrm{T}}$ and $w(Z, \Delta, \Theta) = \exp(\beta_0 + Z^{\mathrm{T}}\beta_Z + \Delta^{\mathrm{T}}\beta_X - \beta_X^{\mathrm{T}}\Sigma_U\beta_X/2)$.

Suppose that two independent samples are collected: $\{(Z_1^c, \Delta_1^c), \ldots, (Z_{n_0}^c, \Delta_{n_0}^c)\}$ with density $g_0(Z, \Delta)$, and $\{(Z_1^t, \Delta_1^t), \ldots, (Z_{n_1}^t, \Delta_{n_1}^t)\}$ with density $g_1(Z, \Delta)$. Denote $T = (Z^{\mathrm{T}}, \Delta^{\mathrm{T}})^{\mathrm{T}}$ and $\{T_1, \ldots, T_n\}$ as the combined sample of the controls and then the cases with $n = n_0 + n_1$. Let $G_0(T)$ be the probability measure corresponding to the baseline density $g_0(T)$. To estimate Θ and g_0, Qin & Zhang (1997) proposed to minimize the empirical likelihood function

$$L(\Theta, G_0) = \prod_{i=1}^{n_0} dG_0(Z_i^c, \Delta_i^c) \prod_{j=1}^{n_1} w(Z_j^t, \Delta_j^t, \Theta) dG_0(Z_j^t, \Delta_j^t) = \prod_{i=1}^{n} p_i \prod_{j=1}^{n_1} w(Z_j^t, \Delta_j^t, \Theta), \tag{7}$$

where $p_i = dG_0(T_i)$, $i = 1, \ldots, n$. Write $w_i = w(T_i, \Theta)$. Qin & Zhang (1997) showed that, for fixed Θ, maximizing (7) with respect to p_i's, under constraints $\sum_{i=1}^n p_i = 1$, $p_i \geq 0$ and $\sum_{i=1}^n p_i\{w_i - 1\} = 0$, yields $p_i = n_0^{-1}\{1 + \rho_n w_i\}$ with $\rho_n = n_1/n_0 \to \rho$ as $n \to \infty$. Pretending that each $\Delta_i = W_i + Y_i\Sigma_U\beta_X$ is observed, after plugging $p_i = n_0^{-1}\{1 + \rho_n w_i\}$ to (7), it can be shown that maximizing (7) with respect to Θ is equivalent to solving $\sum_{i=1}^n \phi_i(\Theta) = 0$, where

$$\phi_i(\Theta) = \frac{I(i > n_0) - I(i \leq n_0)\rho_n w_i}{1 + \rho_n w_i} \begin{pmatrix} 1 \\ T_i \end{pmatrix}.$$

From now on, we consider Δ_i in $\phi_i(\Theta)$ as a function of β_X.

Lemma 1 below shows the consistency of the estimator $\widehat{\Theta}$, which solves $\sum_{i=1}^n \phi_i(\Theta) = 0$, is parallel to Zhang's (2001) point estimator, and reduces to his estimator when $\Sigma_U = 0$. Lemma 1's proof is a standard one for *M*-estimators. It suffices to show $\sum_{i=1}^n E\{\phi_i(\Theta_0)\} = 0$, which is easy using (6). For details of the proof and regularity conditions, please refer to Huber (1967).

LEMMA 1. *Let Θ_0 be the true values of the parameters. If model (1) is true, then under some regularity conditions,*

$$n^{1/2}(\widehat{\Theta} - \Theta_0) \to N\left(0, B^{-1}AB^{-\mathrm{T}}\right) \text{ in distribution},$$

where $B = -\lim_{n\to\infty} n^{-1}\sum_{i=1}^n E\left\{\partial\phi_i(\Theta_0)/\partial\Theta^{\mathrm{T}}\right\}$, $A = \lim_{n\to\infty} n^{-1}\sum_{i=1}^n \mathrm{cov}\left\{\phi_i(\Theta_0)\right\}$.

3·2. *Test statistic construction*

Let $J(T) = (1, T^{\mathrm{T}})^{\mathrm{T}}$, $\xi(\Theta) = (0, 0_{1\times p_Z}, \beta_X^{\mathrm{T}}\Sigma_U)^{\mathrm{T}}$, and $\Gamma = \mathrm{diag}\{0, 0_{p_Z\times p_Z}, \Sigma_U\}$. Noting that $\partial\Delta_i/\partial\beta_X^{\mathrm{T}} = \Sigma_U$ for $i = n_0 + 1, \ldots, n$, straightforward algebra yields

$$\frac{\partial\phi_i(\Theta)}{\partial\Theta^{\mathrm{T}}} = -\frac{\rho_n w_i}{\{1 + \rho_n w_i\}^2} J(T_i)J(T_i)^{\mathrm{T}} + \frac{I(i > n_0)}{1 + \rho_n w_i}\Gamma + \frac{\rho_n w_i I(i \leq n_0)}{\{1 + \rho_n w_i\}^2} J(T_i)\xi(\Theta)^{\mathrm{T}}.$$

For simplicity, write $w_0(T) = w(T, \Theta_0)$ and $\xi_0 = \xi(\Theta_0)$. Then define $D = \int \{1 + \rho w_0(T)\}^{-1} w_0(T) J(T) J(T)^{\mathrm{T}}\, dG_0(T)$ and $D^* = \int \{1 + \rho w_0(T)\}^{-2} w_0(T) J(T) \xi_0^{\mathrm{T}}\, dG_0(T) + \int \{1 + \rho w_0(T)\}^{-1} w_0(T) \Gamma\, dG_0(T)$. Using (6), one has $B = (1 + \rho)^{-1} \rho (D - D^*)$, $A = (1 + \rho)^{-1} \rho D - \rho D_1 D_1^{\mathrm{T}}$, where D_1 is the first column of matrix D. The term D^* is introduced by the errors in X. When X is measured without error, we have $\Sigma_U = 0_{p_X \times p_X}$ and $D^* = 0_{p \times p}$ with $p = 1 + p_Z + p_X$. Thus B and A reduce to the corresponding terms in Qin & Zhang (1997).

We construct our test statistic in the spirit of Zhang (2001). Let $\mathcal{G} = \{f_k(T) : \mathbb{R}^{p_Z + p_X} \to \mathbb{R}^1, k = 1, \ldots, K\}$ be a collection of appropriately defined functions. In practice, it would be generally sufficient to consider continuous functions. For each f_k, define

$$q_k(\Theta) = \frac{1}{n} \sum_{i=1}^{n} \frac{\rho_n w_i - \{I(i > n_0) + I(i \le n_0) \rho_n^2 w_i\}}{\{1 + \rho_n w_i\}^2} f_k(T_i)$$

for $k = 1, \ldots, K$. It is straightforward to show that if model (1) is true, we have $E\{q_k(\Theta_0)\} = 0$ for any function f_k due to (6). This motivates us to construct the following statistic

$$\widehat{Q}_n = Q_n(\widehat{\Theta}) = (q_1(\widehat{\Theta}), \ldots, q_K(\widehat{\Theta}))^{\mathrm{T}}.$$

For $k, l = 1, \ldots, K$, define function $S_{f_k}(T) = (0, 0_{1 \times p_Z}, \partial f_k(T)/\partial \Delta^{\mathrm{T}} \times \Sigma_U)^{\mathrm{T}}$ and

$$b_k = \frac{\rho}{1 + \rho} \int \frac{w_0(T)\{1 - \rho w_0(T)\}}{\{1 + \rho w_0(T)\}^2} f_k(T) J(T)\, dG_0(T),$$

$$b_k^* = \frac{\rho}{1 + \rho} \int \left[\frac{w_0(T)\{1 - 3\rho w_0(T)\}}{\{1 + \rho w_0(T)\}^3} f_k(T) \xi_0 + \frac{w_0(T)\{1 - \rho w_0(T)\}}{\{1 + \rho w_0(T)\}^2} S_{f_k}(T) \right] dG_0(T),$$

$$C_{kl} = \frac{\rho}{1 + \rho} \int \frac{w_0(T)\{1 - \rho w_0(T)\}^2}{\{1 + \rho w_0(T)\}^3} f_k(T) f_l(T)\, dG_0(T), \quad F_k = (b_k - b_k^*)^{\mathrm{T}} B^{-1}.$$

The next theorem gives the asymptotic property of $\widehat{Q}_n$. Its proof is given in the Appendix.

THEOREM 1. *Under some regularity conditions, if model (1) is true, then as $n \to \infty$*

$$n^{1/2} \widehat{Q}_n \to N(0, \Sigma) \textit{ in distribution,}$$

where the (k, l)th entry of Σ is $\sigma_{kl} = C_{kl} - F_k b_l - F_l b_k + (1 + \rho)^{-1} \rho F_k D F_l^{\mathrm{T}}$, $k, l = 1, \ldots, K$.

Theorem 1 is a generalization of Theorem 1 in Zhang (2001) in two aspects. First, the collection of functions used in Zhang (2001) is upper triangular elements of $J(T)J(T)^{\mathrm{T}}$ and the number of functions used is $K = (1 + p_Z + p_X)(2 + p_Z + p_X)$. Our method does not have specific form restrictions of f_k and the number of functions K is flexible. Second, when there is no measurement error, the covariance matrix reduces to the same form as in Zhang (2001).

To construct a Wald-type statistic, a consistent estimator of Σ is needed. We use the same empirical version of Σ, denoted by $\widehat{\Sigma}$, as in Zhang (2001) by replacing Θ_0 with $\widehat{\Theta}$ and G_0 with $\widehat{G}_0(t) = n_0^{-1} \sum_{i=1}^{n} \{1 + \rho_n w(\widehat{T}_i, \widehat{\Theta})\}^{-1} I(\widehat{T}_i \le t)$, where $\widehat{T}_i = (Z_i^{\mathrm{T}}, \widehat{\Delta}_i^{\mathrm{T}})^{\mathrm{T}}$ and $\widehat{\Delta}_i = W_i + Y_i \Sigma_U \hat{\beta}_X$ for $i = n_0 + 1, \ldots, n$. Using a result in Qin & Zhang (1997) and by Lemma 1, it is readily shown that $\widehat{G}_0(t)$ is consistent for $G_0(t)$ under model (1). With a proper collection of functions such that Σ is invertible, by Theorem 1, the following test statistic

$$M_n = n \widehat{Q}_n^{\mathrm{T}} \widehat{\Sigma}^{-1} \widehat{Q}_n$$

has an asymptotic χ^2_K distribution under model (1) as $n \to \infty$. However, in the case of $\widehat{\Sigma}$ being singular, to maximize the numerical stability, we find the generalized inverse of the correlation matrix $\widehat{R}$ from $\widehat{\Sigma}$, denoted by $\widehat{R}^{+}$. Let $L = \mathrm{diag}\{\hat{\sigma}_1, \ldots, \hat{\sigma}_K\}$, where $\hat{\sigma}_k^2$'s are the diagonal elements of $\widehat{\Sigma}$. Then we can use $M_n = n\widehat{Q}_n^{\mathrm{T}} L^{-1} \widehat{R}^{+} L^{-1} \widehat{Q}_n$, which can be shown to have an asymptotic χ^2_d distribution with $d = \mathrm{rank}(\widehat{R})$.

3·3. *Power of the test statistic*

To study the power of M_n, we use the same alternative to model (1) as in Zhang (2001):

$$\mathrm{pr}(Y = 1 \mid Z, X) = H(\beta_0^* + Z^{\mathrm{T}}\beta_Z + X^{\mathrm{T}}\beta_X + \log\{r^*(Z, X, \eta)\}), \tag{8}$$

where $r^*(Z, X, \eta)$ is a function from $\mathbb{R}^{p_Z+p_X+q} \to \mathbb{R}^{+}, \eta \in \mathbb{R}^q$. We assume that there exists an η_0 such that $r^*(Z, X, \eta_0) = 1$ for all (Z, X) and that the partial derivative $s^*(Z, X, \eta) = \partial r^*(Z, X, \eta)/\partial\eta$ exists in a neighborhood of η_0. This type of alternatives is not directly testable since the covariate X is not observed.

LEMMA 2. *If model (8) is true, there exists an $r(Z, \Delta, \eta) : \mathbb{R}^{p_Z+p_X+q} \to \mathbb{R}^{+}$, such that*

$$g_1(Z, \Delta)/g_0(Z, \Delta) = r(Z, \Delta, \eta)w(Z, \Delta, \Theta) \tag{9}$$

and $r(Z, \Delta, \eta_0) = 1$ for all (Z, Δ) with the partial derivative $s(Z, X, \eta) = \partial r(Z, X, \eta)/\partial\eta$ existing in a neighborhood of η_0, where $\Delta = W + Y\Sigma_U\beta_X$ and $g_0(\cdot)$, $g_1(\cdot)$, $w(\cdot)$ are as in (6).

The proof is given in the supplementary material. By Lemma 2, the null hypothesis of testing the validity of model (1) implies the null hypothesis $H_0 : \eta = \eta_0$ under model (9). Under model (9) with $\eta \neq \eta_0$, the consistency of the proposed test statistic naturally follows from the discussion in Zhang (2001) with some modifications, which states that the proposed test procedure of model (1) based on M_n is consistent against any alternative $\eta \neq \eta_0$ such that $E\{Q_n(\Theta_0)\} \neq 0$.

Define $\eta_n = \eta_0 + n^{-1/2}\tau$ for some fixed point τ in $\mathbb{R}^q$. For $k = 1, \ldots, K$, let $c_k = -(1+\rho)^{-1}\rho \int \{1 + \rho w_0(T)\}^{-2} w_0(T)\{1 - \rho w_0(T)\}\tau^{\mathrm{T}} s(T, \eta_0) f_k(T)\, dG_0(T)$, $\Psi = \int \{1 + \rho w_0(T)\}^{-1} w_0(T)\tau^{\mathrm{T}} s(T, \eta_0) J(T)\, dG_0(T)$, and $\mu_k = c_k + (b_k - b_k^*)^{\mathrm{T}} B^{-1}\Psi$. The next theorem gives the asymptotic local power of M_n under model (8). Its proof is given in the Appendix.

THEOREM 2. *Under model (8) with $(\Theta, \eta) = (\Theta_0, \eta_n)$, with some regularity conditions,*

$$M_n \to \chi^2_K(\delta) \text{ in distribution}, \text{ as } n \to \infty.$$

where $\delta = \mu^{\mathrm{T}}\Sigma^{-1}\mu$ with $\mu = (\mu_1, \ldots, \mu_K)^{\mathrm{T}}$ defined above and Σ as in Theorem 1. Here $\chi^2_K(\delta)$ stands for the non-central chi-square distribution with $df = K$ and noncentrality parameter δ.

3·4. *When Σ_U is unknown*

We have so far assumed that Σ_U is known, which is true in many cases. For example, it is popular in database security management to manually add normal random errors to original data to protect confidential, numerical data from unauthorized queries, see Muralidahar et al. (1999).

With repeated measurements for each X_i, say $W_{i1}, \ldots, W_{ik_i}$, Σ_U can be estimated by $\widehat{\Sigma}_U = \{\sum_{i=1}^n (k_i - 1)\}^{-1} \sum_{i=1}^n \sum_{j=1}^{k_i} (W_{ij} - \bar{W}_{i\cdot})(W_{ij} - \bar{W}_{i\cdot})^{\mathrm{T}}$, where $\bar{W}_{i\cdot} = \sum_{j=1}^{k_i} W_{ij}$, $k_i > 1$ and $i = 1, \ldots, n$. In a more general setting, assume that ϑ_U as the vector of distinct elements in Σ_U can be estimated by a root-n consistent estimator $\hat{\vartheta}_U$, which is independent of the observed case-control data, with asymptotic covariance matrix Ω_U. Using the law of large numbers, one has $n^{-1}\sum_{i=1}^n \partial\phi_i(\Theta_0, \Sigma_U)/\partial\vartheta_U^{\mathrm{T}} = B_U + o_p(1)$ and $\partial q_k(\Theta_0, \Sigma_U)/\partial\vartheta_U^{\mathrm{T}} = b_{U,k} + o_p(1)$, for $k = 1, \ldots, K$. A slight modification of the proof of Theorem 2 yields that both Theorems 1 and 2 still hold except for an extra term in each entry of Σ: $\sigma_{kl} = C_{kl} -$

$F_k b_l - F_l b_k + (1+\rho)^{-1}\rho F_k D F_l^{\mathrm{T}} + \sigma_{kl}^*$, where $\sigma_{kl}^* = \{b_{U,k} + B_U^T B^{-1}(b_k - b_k^*)\}^T \Omega_U \{b_{U,l} + B_U^T B^{-1}(b_l - b_l^*)\}$, for $k, l = 1, \ldots, K$.

When Σ_U is misspecified and estimated by $\tilde{\Sigma}_U$, define $\tilde{\Delta} = W + Y\tilde{\Sigma}_U \beta_X$, $\tilde{g}_i(Z, \tilde{\Delta}) = f(Z, \tilde{\Delta} \mid Y = i)$, $i = 0, 1$. In this case, Theorems 1 and 2 generally do not hold anymore. However, if we further assume that X and Z are conditionally independent given Y and $X \mid Y = 0 \sim N(\mu_X, \Sigma_X)$ as in Example 1, one can show that, under model (1),

$$\tilde{g}_1(Z, \tilde{\Delta})/\tilde{g}_0(Z, \tilde{\Delta}) = \exp(\beta_0' + Z^{\mathrm{T}}\beta_Z + \tilde{\Delta}^{\mathrm{T}}\beta_X'),$$

where β_0' is some scalar and $\beta_X' = \{1 + (\Sigma_U + \Sigma_X)^{-1}(\tilde{\Sigma}_U - \Sigma_U)\}\beta_X$. In this special case one can use our method with a misspecified value of Σ_U and still get consistent testing results. In addition, by using a reasonable $\tilde{\Sigma}_U$, the biases of the estimators of β's are negligible. This implies that in practice it is helpful if the distribution of the covariate is about normal among the controls.

4. SIMULATION STUDIES

In Simulation Examples 1 and 2 below, both Z and X are univariate and the true model is

$$\mathrm{pr}(Y = 1 \mid Z, X) = \{1 + r_n(X, \theta)\exp(2 - 0{\cdot}5Z + X)\}^{-1}, \tag{10}$$

where $r_n(X, \theta) = \exp(-n^{-1/2}\theta X^2)$. Then the null hypothesis becomes H_0: $\theta = 0$. We considered three values of θ, $(\theta_1, \theta_2, \theta_3) = (0, \frac{1}{3}, 1)$, and two combinations of sample sizes. The data were generated as follows. First generate one million observations (Z, X, W, Y) using model (10). Then divide the data into two groups: $\{(Z, W, Y) : Y = i\}$, $i = 0, 1$. For fixed (n_0, n_1), randomly select n_0 and n_1 observations from the control group $(Y = 0)$ and the case group $(Y = 1)$, as one sample. Repeat this for 1000 times. The collection of functions used is $\mathcal{G} = \{f_1, f_2, f_3\}$, where $f_1(T) = 1$, $f_2(T) = Z$ and $f_3(T) = \Delta$. In the classical error examples, Σ_U is assumed to be known. The simulation results are summarized in Table 1 and the supplementary material.

Simulation Example 1. Denote $\mathcal{D}$ as the bivariate normal distribution with mean 0, variance 1 and correlation $0{\cdot}7$. We generated $(Z, \log X) \sim \mathcal{D}$ and the surrogate W using classical error model $W = X + U$ with $U \sim N(0, 0{\cdot}5)$. The D_n and L_n are Zhang's (2001) goodness-of-fit test and Lin et al.'s (2002) test for logistic link function ignoring measurement errors in X, respectively. For L_n, we use b as the median of the range, and $B = 1000$ bootstrap samples. The results show that our method achieves significance levels close to the nominal levels while the naive use of D_n and L_n leads to inflated sizes. In addition, for local alternatives, the power of our test increases as θ moves away from 0 as expected.

Simulation Example 2. We generated $(Z, \log W) \sim \mathcal{D}$ and used the regression calibration model $X = 0{\cdot}1Z + 0{\cdot}9W + U$, where $U \sim N(0, 0{\cdot}5)$. The new parameters defined in (3) are $\beta_Z' = 0{\cdot}4$ and $\beta_X' = -0{\cdot}9$. In this case, as shown in Example 2, D_n is a valid test for model (1). Indeed, seen from the rightmost panel of Table 1, both D_n and L_n achieve significance levels close to the nominal levels when $\theta = 0$ and have increasing power as θ increases. The results in the supplementary material indicate that the estimators of β_Z' and β_X' are essentially unbiased.

Simulation Example 3. To simulate logistic regression data such that $X \mid Y = 0$ has a normal distribution, we generated data in the same way as in Zhang (2001) using model (2): for each (n_0, n_1), generate n_0 values of X from $N(0, 1)$ for the control group and n_1 values of X from $N(\mu, \sigma_n^2)$ for the case group, where $\mu = -1$, and $\sigma_n^2 = (1 - 2n^{-1/2}\theta)^{-1}$. We then generated n independent random numbers from $N(0, 1)$ as the covariate Z. Parameter θ plays the same role as in model (10) and takes values $(\theta_1, \theta_2, \theta_3) = (0, 1{\cdot}5, 3)$. Such a sampling scheme generates data from (10) with $\beta_Z = 0$ and $\beta_X = -1$. Finally, the surrogate W was generated by $W = X + U$ with $U \sim N(0, 0{\cdot}5^2)$. In this case, Table 1 shows that all three methods have appropriate

Table 1. *Empirical levels of the tests in the numerical examples.*

			Classical Error X log-normal Nominal levels (%)			Classical Error $X \| Y=0$ normal Nominal levels (%)			Regression Calibration Nominal levels (%)		
θ	(n_1, n_2)	Method	10·0	5·0	1·0	10·0	5·0	1·0	10·0	5·0	1·0
θ_1	$(100, 200)$	M_n	10·0	5·7	2·0	10·0	5·4	1·5			
		D_n	33·2	21·9	7·2	10·7	5·4	1·7	8·3	4·4	1·7
		L_n	27·4	17·8	5·0	11·0	5·1	0·5	10·5	4·6	0·5
	$(200, 100)$	M_n	9·6	5·2	1·4	10·7	4·7	0·9			
		D_n	27·2	15·2	3·3	10·7	4·8	0·8	8·5	5·0	1·6
		L_n	23·5	14·2	4·1	9·3	5·2	1·2	10·9	4·8	0·7
θ_2	$(100, 200)$	M_n	27·4	23·6	20·4	15·4	7·9	2·9			
		D_n	41·9	32·6	22·5	16·1	9·4	2·9	24·5	19·4	15·1
		L_n	31·3	20·4	8·2	13·5	6·8	1·4	17·7	11·1	4·5
	$(200, 100)$	M_n	13·2	9·7	6·5	13·1	7·1	1·5			
		D_n	25·5	18·4	9·8	15·4	8·1	2·2	16·5	12·5	8·6
		L_n	22·1	13·4	4·8	13·5	7·6	1·0	14·6	7·6	1·7
θ_3	$(100, 200)$	M_n	76·9	68·3	50·5	33·3	23·2	8·8			
		D_n	77·3	68·0	47·6	36·7	24·1	10·3	92·1	88·8	81·9
		L_n	40·7	31·4	16·0	31·8	18·5	6·3	52·4	42·2	26·4
	$(200, 100)$	M_n	89·5	86·9	80·7	31·8	21·0	9·4			
		D_n	86·1	81·1	68·9	36·7	24·6	10·9	87·6	85·4	81·9
		L_n	38·0	25·9	8·8	32·9	18·9	4·7	41·9	28·8	13·5

Table 2. *Empirical levels of M_n using misspecified Σ_U when $\theta = 0$.*

	Classical error: X log-normal							Classical error: $X \| Y=0$ normal						
	Nominal levels			$\hat\beta_Z$		$\hat\beta_X$		Nominal levels			$\hat\beta_Z$		$\hat\beta_X$	
δ	10·0	5·0	1·0	Bias	RMSE	Bias	RMSE	10·0	5·0	1·0	Bias	RMSE	Bias	RMSE
0·0	21·5	11·9	2·5	-0·24	0·30	0·47	0·49	10·7	5·4	1·7	0·00	0·14	0·18	0·22
0·5	12·5	6·9	1·3	-0·16	0·25	0·33	0·38	9·9	5·2	0·8	0·00	0·14	0·10	0·20
1·0	8·9	4·7	1·5	0·06	0·32	-0·12	0·57	10·1	5·1	1·1	0·00	0·16	-0·02	0·18
1·1	10·0	6·1	2·3	0·14	0·39	-0·31	0·76	10·1	5·2	1·1	0·00	0·14	-0·05	0·19
1·5	17·4	13·4	6·5	1·37	2·25	-3·92	5·61	10·3	5·1	1·1	0·00	0·15	-0·18	0·29

RMSE stands for root mean square error.

sizes and the results in the supplementary material also indicate that only M_n produces unbiased estimates for β_X. The L_n appears to have generally somewhat lower power than the other two methods.

Sensitivity analysis. To study the effects of mis-specification of Σ_U, we generated data as in Simulation Examples 1 and 3 with $(n_0, n_1) = (100, 200)$ and ran a series of sensitivity tests by plugging $\tilde\Sigma_U = \delta\Sigma_U$ in M_n instead of the true Σ_U with $\delta = 0, 0{\cdot}5, 1, 1{\cdot}1$ and $1{\cdot}5$. The results are summarized in Table 2. When X is log-normal and δ moves away from 1, the biases in $\hat\beta_Z$ and $\hat\beta_X$ increase and the empirical levels of M_n remain reasonable for δ close to 1. When $X \mid Y = 0$ is normal, whatever value δ takes, the empirical levels of M_n are all close to the nominal levels but better estimates of Σ_U results in smaller bias in $\hat\beta_X$, which confirms the conclusion in § 3·4.

One important issue is the choice of $\mathcal{G} = \{f_k : \mathbb{R}^{p_Z+p_X} \to \mathbb{R}^1, k = 1, \ldots, K\}$. What forms of f_k's should be taken and how many functions are enough? In Zhang (2001), if we have p covariates and let $X = (1, x_1, \ldots, x_p)^{\mathrm{T}}$, then $d = (p+1)(p+2)/2$ functions are used with $f_k(X)$'s as the upper triangular elements of matrix of XX^{T}. This choice may not be ideal especially when p is large. Our limited empirical experience suggests that it often leads to a singular estimated covariance matrix $\widehat\Sigma$ of $\widehat Q_n$. The singularity of $\widehat\Sigma$ implies that some compo-

Table 3. *Empirical power (%) at significance level* $\alpha = 0.05$.

θ	$\mathcal{G}_1$	$\mathcal{G}_2$	$\mathcal{G}_3$	$\mathcal{G}_4$	$\mathcal{G}_5$	$\mathcal{G}_6$
0	4·1	5·2	6·0	5·7	5·9	6·2
$\frac{1}{3}$	15·2	14·2	15·0	15·9	16·5	16·4
1	73·2	73·3	79·5	77·9	76·3	77·2

nents of $\widehat{Q}_n$ are redundant. In general, each f_k can be viewed as a source of information and including additional functions might increase the power of our test. However, including redundant functions in the collection would not help increase the power. We ran a simulation study to show the effects of the choices of $\mathcal{G}$. The data were generated as in Simulation Example 1 with a sample size $(n_0, n_1) = (100, 100)$. Define functions $f_1(T) = 1$, $f_2(T) = Z$, $f_3(T) = \Delta$, $f_4(T) = \Delta Z$, $f_5(T) = Z^2$ and $f_6(T) = \Delta^2$. Then for $k = 1, \ldots, 6$, the collections are defined as $\mathcal{G}_k = \{f_1, \ldots, f_k\}$. One thousand case-control samples were generated. For each data set we conducted our hypothesis test using $\mathcal{G}_1, \ldots, \mathcal{G}_6$. The power is calculated at a significance level $\alpha = 0.05$ and the results are summarized in Table 3. We observe that when $\theta \neq 0$, the power of the test increases from $\mathcal{G}_1$ to $\mathcal{G}_3$ or $\mathcal{G}_4$ and then stabilized. Another observation is that when using $\mathcal{G}_6$, most of the times the estimated $\widehat{\Sigma}$ is singular and the degrees of freedom is 5, not 6. We suggest including as many functions as possible until the estimated $\widehat{\Sigma}$ becomes singular.

5. ILLUSTRATION OF AN APPLICATION

Carroll et al. (2006) used a data set from the Framingham heart study to illustrate the conditional score method for linear logistic regression. Here we use the same data set to test the validity of using linear logistic regression. The response variable Y is the occurrence of coronary heart disease within eight years following the Exam 3 with 128 cases and 1487 controls. We use the same covariates as in Carroll et al. (2006), with two error free covariates: Z_1 is the patient's age at Exam 2 and Z_2 is the smoking status at Exam 1, and two covariates measured with errors $X_1 = \log(\text{SC})$ and $X_2 = \log(\text{SBP} - 50)$, where SC is the serum cholesterol level at Exam 3 and SBP is the systolic blood pressure level at Exam 3. We assume a classical error model for the data set: $(W_1, W_2) = (X_1, X_2) + (U_1, U_2)$ with $(U_1, U_2) \sim N(0, \Sigma_U)$ and unknown Σ_U. There are also two measurements of X_2 and one measurement of X_1 at Exam 2. Making use of the repeated measures as described in Carroll et al. (2006, p.118), one has $\widehat{\Sigma}_U = ((0{\cdot}0085, 0{\cdot}0007)^{\mathrm{T}}, (0{\cdot}0007, 0{\cdot}0127)^{\mathrm{T}})$ with an estimated correlation of 0·065.

This was originally a prospective study. To make it into a case-control study setting, we randomly took a 100 cases and 200 control as our data set. To test the validity of logistic linear model, let $T = (Z_1, Z_2, \Delta_1, \Delta_2)$ and use the function collection $\mathcal{G} = \{f_1(T) = 1, f_2(T) = Z_1, f_3(T) = Z_2, f_4(T) = \Delta_1, f_5(T) = \Delta_2\}$. Simple computations yield $M_n = 3{\cdot}62$ with $df = 5$ and a p-value 0·604 and $D_n = 5{\cdot}31$ with $df = 5$ and a p-value 0·379. While the two p-values are quite different, both tests fail to reject the null hypothesis. The facts that W_1 appears to be normally distributed and that W_2 is reasonably normal but slightly right skewed might argue for the use of D_n as in Example 1 in §2.

The estimated coefficients and their standard errors are given in the supplementary material. As discussed in §2, even when it is valid to use Zhang (2001), the resulting estimators may be biased. We can see that the estimates are quite different for the coefficients of Log-Chol and Log-SBP using two methods and our method would correct the bias of the estimators caused by the measurement error in Log-Chol and Log-SBP if the error structure assumptions are met.

APPENDIX

Proofs

Proof of Theorems 1 and 2. For any given f_k, under (Θ_0, η_n), the results of Lemma 1 hold and thus $\widehat{\Theta} - \Theta_0 = O_p(n^{-1/2})$. The Taylor expansion and weak law of large numbers yield:

$$0 = \frac{1}{n}\sum_{i=1}^{n} \phi_i(\widehat{\Theta}) = \frac{1}{n}\sum_{i=1}^{n} \phi_i(\Theta_0) + \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \phi_i(\Theta_0)}{\partial \Theta^{\mathrm{T}}}(\hat{\Theta} - \Theta_0) + o_p(n^{-1/2}),$$

$$q_k(\widehat{\Theta}) = q_k(\Theta_0) + (b_k - b_k^*)^{\mathrm{T}} B^{-1}\{n^{-1}\sum_{i=1}^{n} \phi_i(\Theta_0)\} + o_p(n^{-1/2}).$$

Define $\mu = (\mu_1, \ldots, \mu_k)^{\mathrm{T}}$. By some extensive algebra similar to that in Zhang (2001), one can show that, under (Θ_0, η_n), $n^{1/2}E\{q_k(\widehat{\Theta})\} = \mu_k + o(1)$ and

$$n\mathrm{cov}\{q_k(\widehat{\Theta}), q_l(\widehat{\Theta})\} = C_{kl} - F_k b_l - F_l b_k + (1+\rho)^{-1}\rho F_k D F_l^{\mathrm{T}} + o(1) = \sigma_{kl} + o(1)$$

for $k, l = 1, \ldots, K$. By the multivariate Central Limit Theorem and Slutsky's Theorem, one has

$$n^{1/2}(\widehat{Q}_n - \mu) \to N(0, \Sigma) \text{ in distribution},$$

where $\mu = 0_k$ under model (1). Thus Theorem 1 has been shown. Since $\widehat{\Sigma}$ is a consistent estimator of Σ under the local alternative model with parameters (Θ, η_n), Theorem 2 follows. □

REFERENCES

AMSTRONG, B.G., WHITTEMORE, A.S. & HOWE, G.R. (1989). Analysis of case-control data with covariate measurement error: application to diet and colon cancer. *Statist. in Medicine* **8**, 1151-63.

AZZALINI, A., BOWMAN, A. & HARDLE, W. (1989). On the use of nonparametric regression for model checking. *Biometrika* **76**, 1-11.

BONDELL, H.D. (2007). Testing goodness-of-fit in logistic case-control studies. *Biometrika* **94**, 487-95.

CARROLL, R.J. GAIL, M.H. & LUBIN, J.H. (1993). Case control studies with errors in covariates. *J. Am. Statist. Assoc.* **88**, 185–99.

CARROLL, R.J., RUPPERT, D., STEFANSKI, L.A., CRAINICEANU, C.M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective.* Second Edition. London: Chapman and Hall.

HUBER, P.J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the 5th Berkeley Symposium.* **1**, 221–33.

KIM, M.Y.& JACQUOTTE, A.Z. (2002). Correcting for measurement error in the analysis of case-control data with repeated measurements of exposure. *Am. J. Epidemiol.* **145**, 1003–10.

LIN, D.Y., WEI, L.J. & YING, Z. (2002). Model-checking techniques based on cumulative residuals. *Biometrics* **58**, 1–12.

MURALIDAHAR, K., PARSA, R., & SARATHY, R. (1999). A general additive data pertubation method for database security. *Management Sci.* **45**, 1399–451.

PRENTICE, R. L.& PYKE, R. (1979). Logistic disease incidence models and case control studies. *Biometrika* **66**, 403-11.

QIN, J. & ZHANG, B. (1997). A goodness-of-fit test for logistic regression models based on case-control data. *Biometrika* **84**, 609-18.

STEFANSKI, L.A. & CARROLL, R.J. (1987). Conditional scores and optimal scores for generalized linear measurement-error models. *Biometrika* **74**, 703-16.

SU, J.Q. & WEI, L.J. (1987). A lack-of-fit test for the mean function in a generalized linear model. *J. Am. Statist. Assoc.* **86**, 420–26.

WANG, C.Y., WANG, S., & CARROLL, R.J. (1997). Estimation in choice-based sampling with measurement error and bootstrap analysis. *J. Econ.* **77**, 65–86.

ZHANG, B. (2001). An information matrix test for logistic regression models based on case-control data. *Biometrika* **88**, 921-32.