Journal of Data Science **7**(2009), 513-536

A Simple Method for Screening Binary Models with Large Sample Size and Continuous Predictor Variables

Weichung Joe Shih[1,2] and Junfeng Liu[1,2]
[1] *The Cancer Institute of New Jersey and* [2] *University of Medicine and Dentistry of New Jersey*

Abstract: For binary regression model with observed responses (Ys), specified predictor vectors (Xs), assumed model parameter vector (β) and case probability function ($\Pr(Y = 1|X, \beta)$), we propose a simple screening method to test goodness-of-fit when the number of observations (n) is large and Xs are continuous variables. Given any threshold $\tau \in [0, 1]$, we consider classifying each subject with predictor X into Y^*=1 or 0 (a deterministic binary variable other than the observed random binary variable Y) according to whether the calculated case probability ($\Pr(Y = 1|X, \beta)$) under hypothesized true model $\geq$ or $< \tau$. For each τ, we check the difference between the expected marginal classification error rate (false positives [Y^*=1, Y=0] or false negatives [Y^*=0, Y=1]) under hypothesized true model with the observed marginal error rate which is directly observed due to this classification rule. The screening profile is created by plotting τ-specific marginal error rates (expected and observed) versus $\tau \in [0, 1]$. Inconsistency indicates lack-of-fit and consistence indicates good model fit. We note that, the variation of the difference between the expected marginal classification error rate and the observed one is constant ($O(n^{-1/2})$) and free of τ. The smallest homogeneous variation at each τ potentially detects flexible model discrepancies with high power. Simulation study shows that, this profile approach named as CERC (classification-error-rate-calibration) is useful for checking wrong parameter value, incorrect predictor vector component subset and link function misspecification. We also provide some theoretical results as well as numerical examples to show that, ROC (receiver operating characteristics) curve is not suitable for binary model goodness-of-fit test.

Key words: Binary model, classification error rate, goodness-of-fit, receiver operating characteristics (ROC).

1. Introduction

For (generalized) linear regression models with parameter vector (β), k-dimen sional predictor vectors (X_i, $i = 1, \ldots, n$) and responses (Y_i, $i = 1, \ldots, n$), hypothesis test often involves a null (H_0) and an alternative (H_a) assumption. The Neyman-Pearson lemma based test rule usually leads to deriving a distribution of estimated parameters/statistics under H_0 (hypothesized true parameter values and model) and p-value is calculated to determine whether we would be in favor of H_0 or H_a. On the other hand, goodness-of-fit test does not necessarily involve an alternative hypothesis and criteria are mainly proposed to evaluate the consistence between a hypothesized true model and observations without the need for a competing model. Binary regression model goodness-of-fit test deserves special attention due to simple structure of responses. McCullagh and Nelder (1989, Ch.4.4.5) demonstrated that, two commonly applied criteria for generalized linear model goodness-of-fit test, the residual deviance function and Pearson's χ^2 statistic, are not very suitable for binary model goodness-of-fit test. The present work proposes an efficient simple method to evaluate any hypothesized binary model when the sample size (n) is large and predictors (Xs) are continuous variables. The motivation comes from subject classification by case (Y=1) probability $\Pr(Y = 1|X, \beta)$, where Y (subject-specific) is the random binary response for any subject with predictor vector X (subject-specific) and parameter vector β (model-specific), and the probability function Pr (model-specific) for Bernoulli trial could be of any form with range [0,1]. As an example, clinicians may use probability threshold 1/2 to make disease diagnosis based on trained binary regression model from historical data including patient disease (case positive: Y=1 and case negative: Y=0) status as well as multiple medical and/or biological characteristics (X). A new patient will then be diagnosed as disease positive ($Y^* = 1$) if the calculated disease (case) probability $\geq$1/2 or disease negative ($Y^* = 0$) if the disease (case) probability <1/2. Note that, Y^* is not necessarily equal to Y and thus we may expect false positives (Y^*=1, Y=0) and/or false negatives (Y^*=0, Y=1). We consider arbitrary probability threshold and classification error rate after comparing case probability with threshold $\tau \in [0, 1]$. $\Pr(Y = 1) \geq \tau$ results in classifying subject as case positive ($Y^* = 1$) and $\Pr(Y = 1) < \tau$ leads to classifying subject as case negative ($Y^* = 0$). Given observed responses (Ys), predictors (Xs), assumed true model parameter (β) and case probability function (Pr), we will obtain two important values: 1) *The expected marginal classification error rate* (EMCER), i.e., the probability that a randomly selected subject out of the n individuals would be "expected" to be misclassified under the assumed true model; and 2) *The observed marginal classification error rate* (OMCER), i.e., the probability that a randomly selected subject out of n individuals is "ob-

served" to be misclassified based on the classification rule. The goodness-of-fit test is done by comparing EMCER with OMCER across $\tau \in [0, 1]$. Any pair-wise τ-specific error rate difference beyond 95% error bound (significant difference) pinpoints model discrepancy between the assumed true model and the observed binary responses. Among these τs, a large portion of significant difference indicates bad model fit, otherwise a good model fit is very likely obtained. We name the present method as "classification-error-rate-calibration (CERC)". CERC enjoys minimal-variation homogeneity across thresholds (τs) and applies well to binary regression model goodness-of-fit test under large sample size (n) and continuous predictor variables (Xs). When the true parameters (β) are unknown, numerical results demonstrate that, CERC works well after replacing the true parameters (β) by estimated parameters ($\hat{\beta}$) from fitting the true model (case probability function Pr and predictor X). Interestingly, we show that, another profile approach ROC (receiver operating characteristics) curve is not suitable for goodness-of-fit test for binary regression models in this scenario.

The rest of this article is organized as follows: Section 2 introduces subject classification by case probability for binary regression model; Section 3 develops a classification-based criterion for binary model goodness-of-fit test; Section 4 demonstrates the usefulness of this simple approach by simulations; Section 5 develops some theoretical results along with numerical examples to show that ROC curve can not be used for binary regression model goodness-of-fit test; and Section 6 concludes with discussion.

2. Subject Classification by Case Probability

We follow the notations in Section 1. For any binary regression model with response Y, k-dimensional predictor vector X, parameter vector β and a specified case (Y=1) probability, $\Pr(Y = 1|X, \beta)$, the case probability population is created as $p_i = \Pr(Y_i = 1|X_i, \beta) = f(X_i, \beta)$, $i = 1, 2, \ldots, n$, a function of predictor vector (X_i) and parameter vector (β). Note that, this case probability population is fixed once we record the predictor vectors, specify the model (case probability function) and the parameter vector. As described in Section 1, each subject is classified using the following rule

$$\text{If } \; p_i \geq \tau, \text{we classify subject with predictor } \; X_i \; \text{ as } \; Y_i^* = 1, \tag{2.1}$$

$$\text{if } \; p_i < \tau, \text{we classify subject with predictor } \; X_i \; \text{ as } \; Y_i^* = 0. \tag{2.2}$$

In the sequel, we take the hypothesized true model (case probability function Pr) as well as parameter vector (β) as fixed and no model fitting (parameter estimation) is considered, unless parameter (β) estimation is actually applied (details later, Sections 4.1 and 4.2).

2.1 Expected marginal classification error rate (EMCER)

Result 1 If the assumed model (case probability function Pr and parameter vector β) is correct, the sample size is n and the probability threshold is τ, we have the following result for EMCER given τ, which is denoted as $\alpha_T(\tau)$.

$$\alpha_T(\tau) = \frac{1}{n} \sum_{\{p_i<\tau, 1\le i\le n\}} p_i + \frac{1}{n} \sum_{\{p_i\ge\tau, 1\le i\le n\}} (1-p_i) = \text{EFPR} + \text{EFNR},$$

where, EFPR and EFNR are the expected marginal false positive and false negative rates respectively.

Proof. False positive case and false negative case are disjoint events based on aforementioned classification rule. Given the assumed true model with parameter β and predictor population $X = \{X_1, X_2, \ldots, X_n\}$, we have $p_i = \Pr(Y_i = 1|X_i, \beta) = f(X_i, \beta)$ for $i = 1, 2, \ldots, n$. We denote the probability population P as $\{p_1, p_2, \ldots, p_n\}$ and the observed binary response population Y as $\{Y_1, Y_2, \ldots, Y_n\}$ to derive EMCER ($\alpha_T(\tau)$) given τ. Since EMCER (Section 1) simply represents the overall misclassification probability based on foregoing classification rule ((2.1) and (2.2)), we have

$$\begin{aligned}
&\Pr(\text{Misclassification}) \\
&= \Pr(\text{False Positive or False Negative}) \\
&= \Pr(\text{False Positive}) + \Pr(\text{False Negative}) \\
&= \Pr(Y^* = 1, Y = 0) + \Pr(Y^* = 0, Y = 1) \\
&= \Pr(P \ge \tau, Y = 0) + \Pr(P < \tau, Y = 1) \\
&= \Pr(P \ge \tau) \times \Pr(Y = 0|P \ge \tau) + \Pr(P < \tau) \times \Pr(Y = 1|P < \tau) \\
&= \Pr(P \ge \tau) \times \int_{P\ge\tau} (1-P) dF_P + \Pr(P < \tau) \times \int_{P<\tau} P dF_P \\
&= (\int_{P\ge\tau} dF_P) \times \int_{P\ge\tau} (1-P) dF_P + (\int_{P<\tau} dF_P) \times \int_{P<\tau} P dF_P.
\end{aligned}$$

Since case probability population (P) of size n is fixed given predictors (Xs), the hypothesized model (case probability function Pr) and parameter (β), by the definition of EMCER (Section 1) each subject-specific case probability (p_i) has

equal chance of being selected from any specified subspace. So we have

$$\begin{aligned}
&(\int_{P\geq\tau} dF_P) \times \int_{P\geq\tau} (1-P)dF_P \\
&= \frac{\sum_{\{p_i\geq\tau,1\leq i\leq n\}} 1}{n} \times \frac{\sum_{\{p_i\geq\tau,1\leq i\leq n\}}(1-p_i)}{\sum_{\{p_i\geq\tau,1\leq i\leq n\}} 1} \\
&= \frac{\sum_{\{p_i\geq\tau,1\leq i\leq n\}}(1-p_i)}{n}
\end{aligned}$$

and

$$\begin{aligned}
(\int_{P<\tau} dF_P) \times \int_{P<\tau} PdF_P &= \frac{\sum_{\{p_i<\tau,1\leq i\leq n\}} 1}{n} \times \frac{\sum_{\{p_i<\tau,1\leq i\leq n\}} p_i}{\sum_{\{p_i<\tau,1\leq i\leq n\}} 1} \\
&= \frac{\sum_{\{p_i<\tau,1\leq i\leq n\}} p_i}{n}.
\end{aligned}$$

The proof ends.

Remark We consider the clinical example in Section 1. For any true binary regression model, 1/2 threshold always gives minimal EMCER.

Illustration. For any threshold $\tau \neq 1/2$, we assume $\tau > 1/2$ without loss of generality and show that EMCER ($\alpha_T(\tau)$) is no less than that under $\tau = 1/2$.

$$\begin{aligned}
&\alpha_T(\tau) \\
&= \frac{1}{n}[\sum_{\{p_i<\tau,1\leq i\leq n\}} p_i + \sum_{\{p_i\geq\tau,1\leq i\leq n\}} (1-p_i)] \\
&= \frac{1}{n}[\sum_{\{p_i<1/2,1\leq i\leq n\}} p_i + \sum_{\{1/2\leq p_i<\tau,1\leq i\leq n\}} p_i + \sum_{\{p_i\geq\tau,1\leq i\leq n\}} (1-p_i)] \\
&\geq \frac{1}{n}[\sum_{\{p_i<1/2,1\leq i\leq n\}} p_i + \sum_{\{1/2\leq p_i<\tau,1\leq i\leq n\}} (1-p_i) + \sum_{\{p_i\geq\tau,1\leq i\leq n\}} (1-p_i)] \\
&= \frac{1}{n}[\sum_{\{p_i<1/2,1\leq i\leq n\}} p_i + \sum_{\{p_i\geq 1/2,1\leq i\leq n\}} (1-p_i)] \\
&= \alpha_T(1/2). \qquad (2.3)
\end{aligned}$$

Similar procedure applies to $\tau < 1/2$ case.

2.2 Observed marginal classification error rate (OMCER)

OMCER is calculated as follows.

1) If $p_i \geq \tau$, then we classify subject with predictor X_i as $Y_i^* = 1$. The misclassified (false positive) count increases by 1 if the observed Y_i=0.

2) If $p_i < \tau$, then we classify subject with predictor X_i as $Y_i^* = 0$. The misclassified (false negative) count increases by 1 if the observed Y_i=1.

OMCER given τ, denoted by $\alpha_E(\tau)$, is simply the total misclassified count divided by sample size n, i.e.,

$$\alpha_E(\tau) = \frac{1}{n} \sum_{\{p_i<\tau, 1\leq i\leq n\}} 1_{\{y_i=1\}} + \frac{1}{n} \sum_{\{p_i\geq\tau, 1\leq i\leq n\}} 1_{\{y_i=0\}} = \text{OFPR+OFNR}, \quad (2.4)$$

where, $1_{\{Y=1\}}$ and $1_{\{Y=0\}}$ are the indicator functions for case $Y = 1$ and $Y = 0$, OFPR and OFNR are the observed marginal false positive and false negative rates respectively.

Result 2 For any threshold $\tau \in [0, 1]$, the variance of OMCER ($\alpha_E(\tau)$) is

$$V_{\alpha_E(\tau)} = \frac{1}{n^2} \sum_{i=1}^{n} p_i(1 - p_i) = O(n^{-1}),$$

which is free of τ and less than $1/(4n)$.

Proof. From the definition of OMCER ((2.4)), each subject-specific indicator function $1_{\{Y_i=1\}}$ or $1_{\{Y_i=0\}}$ has variance $p_i(1-p_i) \leq 1/4$. All Ys are independently distributed and all subjects are independently classified, the proof ends.

3. Application to Screening Binary Models

We describe why CERC may help to test goodness-of-fit for binary regression models. In Diagram 1, the x-axis represents the subject indices (1 to n) and the y-axis represents the values of ordered case probabilities ($\{p_{(1)},p_{(2)},\ldots,p_{(n)}\}$). Specifically, under the assumed true model, the vertical lines in the lower-left shaded region are case probabilities (subject-specific classification false positive rates) which are less than threshold τ, the vertical lines in the upper-right shaded region are no-case probabilities (subject-specific classification false negative rate) with case probabilities no less than threshold τ. Thus the curved boundary of two shaded regions represents ordered case probabilities under true model. EMCER$\times n$ due to threshold τ is represented by the two shaded regions (Result 1). We are interested in studying the concordance between piecewise true case probabilities and observed case probabilities. In Diagram 1, we use an "abstract" dotted curve to represent the ordered observed case probabilities (like "smoothed" Ys) only for illustration purpose. If the true case probability curve stay close to

observed case probability curve, then the model is likely correct, otherwise the model is to be improved. For the probability curve from true model, we assume probability interval $[\tau,\ \tau+\delta]$ matches subject subset ($S_{\tau,\delta}$) of size $nr_{\tau,\delta}$ (with subject index from m_τ to $m_{\tau+\delta}$, Diagram 1), where $m_{\tau+\delta}-m_\tau \approx nr_{\tau,\delta}$ and $r_{\tau,\delta}$ is a proportion likely associated with τ and δ other than sample size n. We now increase the threshold from τ to $\tau+\delta$ and the change in EMCER is around $r_{\tau,\delta}|(1-(\tau+\delta/2))-(\tau+\delta/2)| = r_{\tau,\delta}|1-2\tau-\delta|$ (Result 1), which is associated with $|$(region I) $-$ region(II+III)$|$, since no change takes place outside $[\tau,\ \tau+\delta]$. Here, we use the middle point $(\tau+\delta/2)$ between τ and $\tau+\delta$ as probability average within $[\tau,\tau+\delta]$ for illustration purpose. If the hypothesized true model (Pr and β) describes the responses (Ys) correctly, within subject subset $S_{\tau,\delta}$, the misclassified responses should also follow this change pattern when threshold increases from τ to $\tau+\delta$: from around $nr_{\tau,\delta}(1-(\tau+\delta/2))$ misclassified subjects $(Y=0)$ to around $nr_{\tau,\delta}(\tau+\delta/2)$ misclassified subjects $(Y=1)$. The change in OMCER is the difference between these two counts divided by sample size n, which is expected to be close to the aforementioned change in EMCER. However, because of possible local discrepancy between true case probabilities and observed case probabilities (curved boundary of shaded areas and "abstract" dotted curve) represented by region III in Diagram 1, the local change in OMCER would be associated with $|$(region I+II)$-$region(III)$|$ in Diagram 1. Consequently, possible discrepancy between EMCER and OMCER changes may occur due to threshold increment from τ to $\tau+\delta$, which can be seen to be associated with region II in Diagram 1. Again, $r_{\tau,\delta}$ is not inherently associated with sample size n, we expect sufficient power to detect minor inconsistency between $\alpha_T(\tau)$ curve and $\alpha_E(\tau)$ curve under large n (Result 2) if the hypothesized true model is actually untrue. On the other hand, cumulative inconsistency over moderate-sized probability interval may lead to appreciable difference between EMCER change and OMCER change. Now we define the misclassification rate difference as

$$\alpha_{TE}(\tau) = \alpha_E(\tau) - \alpha_T(\tau),\ \tau \in [0,1].$$

Its variance $V_{\alpha_{TE}(\tau)}$ equals $V_{\alpha_E(\tau)}$ (Result 2) and is free of τ since $\alpha_T(\tau)$ is an expected value without random error. For $\alpha_{TE}(\tau)$, we propose a prescribed 95% error bound which is denoted by "Critical α_{TE}" $= 1.96\sqrt{V_{\alpha_{TE}}}$, where the constant $V_{\alpha_{TE}}$ is the variance of $\alpha_{TE}(\tau)$ under hypothesized model and 1.96 is the two-sided 2.5% quantile for standard normal distribution. The normality assumption is reasonable under large sample size n. Note that, a conservative distribution (case probability)-free Critical α_{TE} could be taken as $0.98/\sqrt{n}$. See Result 2 and Table 2.

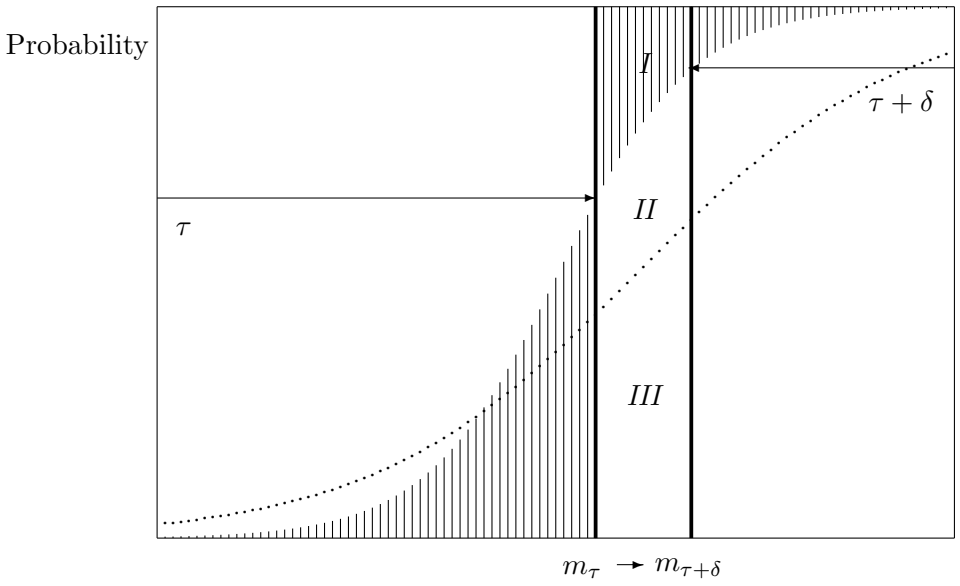

Diagram 1: Shaded area is EMCER$\times n$ given τ under the hypothesized true model, the curved shade boundary represents ordered case probabilities from the true model and the dotted curve represents the corresponding "smoothed" observations (Ys).

One option for model checking is to compare the maximal misclassification rate difference ($\text{Sup}_{\tau\in[0,1]}|\alpha_{TE}(\tau)|$) with Critical α_{TE} and lack-of-fit is to be reported if the former one is greater than the latter one. As a statistical concept, marginal probability simply represents the overall population proportion with certain property, i.e., the probability that a randomly chosen individual from the population has certain property. So far, we applied marginal probability to calculating EMCER (Result 1) and OMCER (Section 2.2) given $\tau \in [0,1]$, where "being misclassified" is our interested property among subject population with size n (Sections 2.1 and 2.2). Now we still use marginal probability idea to construct goodness-of-fit test criterion considering another interested property "$\alpha_{TE}(\tau)$ exceeding 95% error bound" among threshold ($\tau \in [0,1]$) population. Another option for model checking works on the following "exceeding proportion"

$$\frac{\#\{\tau : \tau \in [0,1] \text{ and } |\alpha_{TE}(\tau)| \geq \text{Critical } \alpha_{TE}\}}{\#\{\tau : \tau \in [0,1]\}}, \tag{3.1}$$

where # represents the subset measure on [0,1]. Lack-of-fit is to be reported if this proportion exceeds prescribed 5% (Critical α_{TE} is 95% error bound). Note that, $\alpha_{TE}(\tau)$s across multiple τs are correlated among τs since they are from a

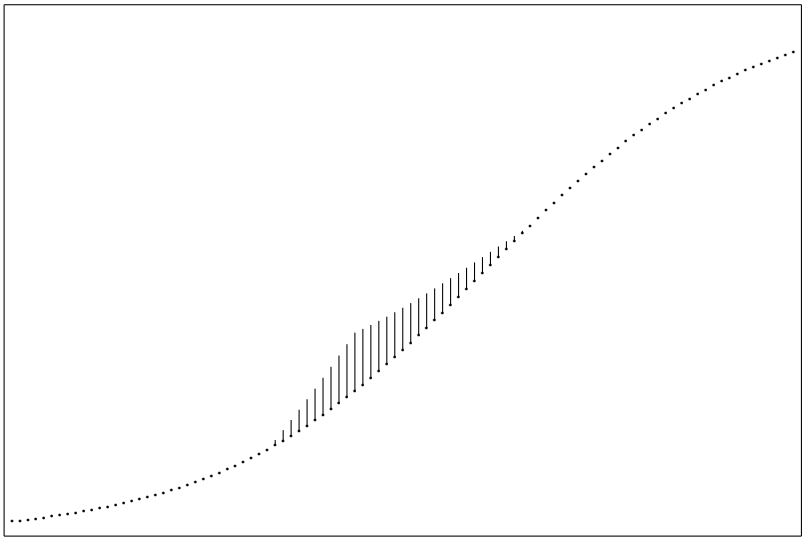

Diagram 2: Case probability curves with a single bump as discrepancy. The x-axis represents subject index and the y-axis represents ordered case probabilities.

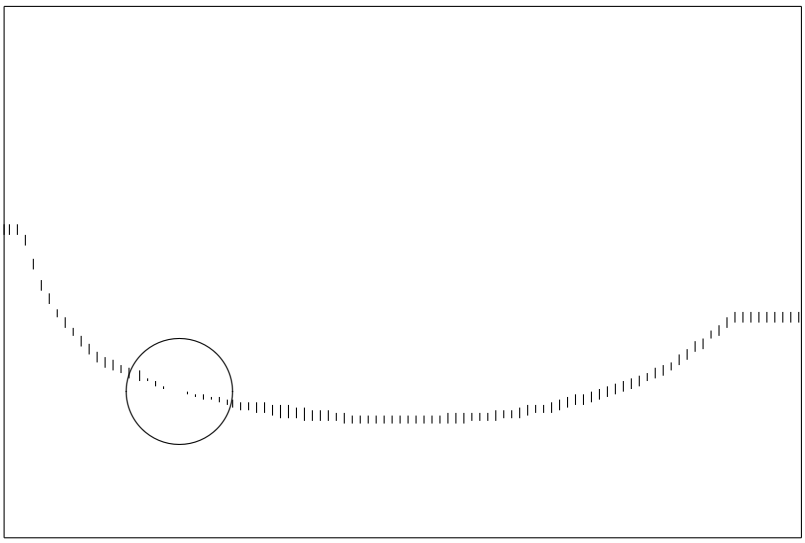

Diagram 3: CERC curves (EMCER and OMCER) due to single bump in probability curves (Diagram 2). The vertical bars represent discrepancy between EMCER and OMCER given $\tau \in [0, 1]$.

single sample of (X,Y). However, marginal probability does not require independence among individuals in the population. Now we illustrate the power of CERC

using a special example. Diagram 2 shows two ordered-case-probability curves with a single bump representing discrepancy between two sets of binary model case probabilities, each subject on the x-axis has two matching probabilities: one is from the true model and the other is from the observed case ($Y = 1$) probability after local smoothing for illustration purpose. The curve with bump is taken as the probability set derived from observed responses. Diagram 3 shows EMCER and OMCER curves across $\tau \in [0, 1]$ for Diagram 2, where the circle identifies minor discrepancy associated with those τs and out-of-circle region shows identical nonignorable discrepancy across other τs.

4. Simulation Study

This section studies the usefulness of CERC for detecting binary model lack-of-fit due to misspecified parameter value (β), predictor component subset (X) or link function g, where subject-specific linear prediction $x\beta = g$(subject-specific case probability p). We consider the following popular link functions: logit link $g_1(p) = \log(p/(1-p))$, probit link $g_2(p) = \Phi^{-1}(p)$ and complementary log-log link $g_3(p) = \log(-\log(1-p))$.

4.1 Distinction under same link functions

Identical parameter dimension

For $k = 2$ or 3, based on certain distribution, we simulate each component of predictor vector $X_i = (X_{i,0}, X_{i,1}, \ldots, X_{i,k-1})$, $i = 1, 2, \ldots, n$ (=10,000), while the first component ($X_{i,0}$) is 1. No correlation structure within each predictor vector is incorporated for simplicity. We use k-dimensional true parameter ($\beta_0 = (\beta_{0,0}, \beta_{1,0}, \ldots, \beta_{k-1,0})$) to simulate n binary responses under link function $g(\Pr(Y = 1))=X^T\beta_0$. In practice, often we do not know the exact true parameter (β_0) and instead use a hypothesized true model (with working parameter $\beta_* = (\beta_{0,*}, \beta_{1,*}, \ldots, \beta_{k-1,*})$) to describe the binary response mechanism, where β_* has the same dimension as β_0 while different value. Given probability threshold $\tau \in [0, 1]$, we can calculate EMCER (Section 2.1) using hypothesized true parameter β_* and calculate OMCER (Section 2.2) using β_* as well as simulated responses from β_0. Since EMCER is produced from β_* and OMCER is actually produced from both β_0 and β_*, we expect potential discrepancy between these two misclassification rates because the difference between β_0 and β_* is not made by model fitting (parameter estimation). Simulations involving model fitting (parameter estimation) will be discussed in Sections 4.1 and 4.2. We do this cross m(=100) different τs (equal-partition on [0,1]) in order to form a goodness-of-fit evaluation profile. We use logit link function as an example and similar results

Table 1: Simulation results (g_0 = true link function; ($\beta_{0,0}$, $\beta_{1,0}$, $\beta_{2,0}$, $\beta_{3,0}$) is true parameter; g_* = working link function; ($\beta_{0,*}$, $\beta_{1,*}$, $\beta_{2,*}$, $\beta_{3,*}$) is working parameter.)

	true link function					working link function				
No.	g_0	$\beta_{0,0}$	$\beta_{1,0}$	$\beta_{2,0}$	$\beta_{3,0}$	g_*	$\beta_{0,*}$	$\beta_{1,*}$	$\beta_{2,*}$	$\beta_{3,*}$
1	logit	-1.00	2.00			logit	-1.00	2.00		
2	logit	-1.00	2.00			logit	-1.00	2.50		
3	logit	-1.00	2.50			logit	-1.00	2.00		
4	logit	-1.00	2.00			logit	-1.50	2.00		
5	logit	-1.00	2.00			logit	-0.50	2.50		
6	logit	-1.00	2.00	3.00		logit	-1.00	2.00	3.00	
7	logit	-1.00	2.00	3.00		logit	-1.00	2.50	3.50	
8	logit	-1.00	2.00	3.50		logit	-1.00	2.50	3.00	
9	logit	-1.00	2.00	3.00		logit	-1.50	2.50	2.50	
10	logit	-1.00	2.00	3.00		logit	-1.50	2.00	3.00	
11	logit	0.00	4.00	4.00	-12.0	logit	7E-3	3.94	3.82	-11.7
12	logit	0.00	4.00	4.00	-12.0	logit	2.98	-2.01	-2.10	0.00
13	probit	-1.00	2.00			probit	-1.05	2.03		
14	cloglog	-1.00	2.00			cloglog	-1.02	1.97		
15	probit	-1.00	2.00			logit	-1.77	3.57		
16	logit	-1.00	2.00			probit	-0.61	1.15		
17	cloglog	-1.00	2.00			logit	-0.69	2.73		
18	logit	-1.00	2.00			cloglog	-1.21	1.33		

Table 2: Simulation results (Mean $|\alpha_{TE}|$ = average of $|\alpha_{TE}(\tau)|$ across $\tau \in [0,1]$, $0.98/\sqrt{n}$ is the conservative distribution-free 95% error bound for $|\alpha_{TE}|$ under the true model with sample size n.)

No.	Mean $\lvert\alpha_{TE}\rvert$	$\text{Sup}_{1\le\tau\le1}\lvert\alpha_{TE(\tau)}\rvert$	Critical α_{TE}	$0.98/\sqrt{n}$	Exceeding proportion
1	2.36E-3	5.59E-3	7.37E-3	1.00E-2	0.00
2	2.45E-2	4.12E-2	6.91E-3	1.00E-2	0.86
3	1.81E-2	3.27E-2	7.34E-3	1.00E-2	0.86
4	3.90E-2	6.13E-2	7.01E-3	1.00E-2	0.93
5	5.56E-2	9.15E-2	7.04E-3	1.00E-2	0.95
6	1.20E-3	2.76E-3	6.07E-3	1.00E-2	0.00
7	1.78E-2	2.90E-2	5.69E-3	1.00E-2	0.84
8	8.58E-3	1.58E-2	5.88E-3	1.00E-2	0.71
9	2.99E-2	5.18E-2	5.96E-3	1.00E-2	0.90
10	2.02E-2	3.85E-2	5.95E-3	1.00E-2	0.86
11	1.44E-3	6.31E-3	7.85E-3	1.00E-2	0.00
12	2.16E-2	5.44E-2	8.48E-3	1.00E-2	0.58
13	1.55E-3	4.61E-3	5.85E-3	1.00E-2	0.00
14	1.16E-3	4.91E-3	6.73E-3	1.00E-2	0.00
15	2.46E-3	6.16E-3	5.92E-3	1.00E-2	0.06
16	3.42E-3	7.91E-3	7.36E-3	1.00E-2	0.05
17	8.81E-3	1.69E-2	6.79E-3	1.00E-2	0.60
18	1.10E-2	2.72 E-2	7.52E-3	1.00E-2	0.52

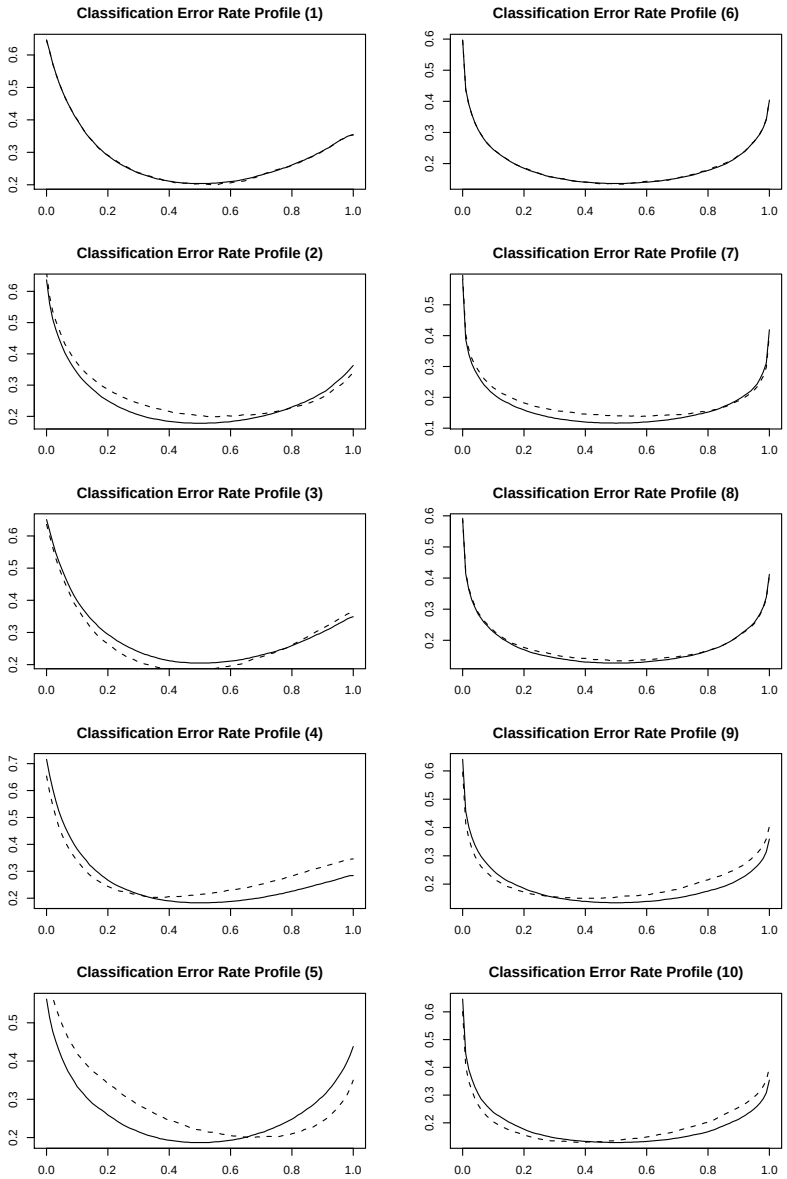

Figure 1: τ-wise comparison between EMCER and OMCER from simulations 1-10. (The solid line is EMCER curve, the dotted line is OMCER curve. τ covers [0,1] on the x-axis.)

occur for other links. Ten simulations are considered in order to take into account diverse scenarios, the configurations as well as model checking results are summarized in Table 1 (panel 1), Table 2 (panel 1) and Figure 1.

$\diamond$ (Simulation 1) k=2, $X = (1, X_1)$, $\beta_0 = (-1.00, 2.00)$ and $\beta_* = \beta_0$. We

simulate X_1 from uniform distribution $U[0,1]$. Ys are generated using logit link with linear prediction $-1+2X_1$. This is a control simulation and no discrepancy is detected between model assumption and responses (Ys).

- ◇ (Simulation 2) k=2, $X=(1,X_1)$, $\beta_0=(-1.00,2.00)$ and $\beta_*=(-1.00,2.50)$. We simulate X_1 from uniform distribution $U[0,1]$. Ys are generated using logit link with linear prediction $-1+2X_1$. We apply a working linear term coefficient which is different from the true value and substantial discrepancy is found.

- ◇ (Simulation 3) k=2, $X=(1,X_1)$, $\beta_0=(-1.00,2.50)$ and $\beta_*=(-1.00,2.00)$. We simulate X_1 from uniform distribution $U[0,1]$. Ys are generated using logit link with linear prediction $-1+2.5X_1$. We switch the values between the true parameter and the working parameter in simulation 2 and similar discrepancy is found with opposite directions (left panels in rows 2 and 3, Figure 1.

- ◇ (Simulation 4) k=2, $X=(1,X_1)$, $\beta_0=(-1.00,2.00)$ and $\beta_*=(-1.50,2.00)$. We simulate X_1 from uniform distribution $U[0,1]$. Ys are generated using logit link with linear prediction $-1+2X_1$. We apply a working intercept which is different from the true value and substantial discrepancy is found.

- ◇ (Simulation 5) k=2, $X=(1,X_1)$, $\beta_0=(-1.00,2.00)$ and $\beta_*=(-0.50,2.50)$. We simulate X_1 from uniform distribution $U[0,1]$. Ys are generated using logit link with linear prediction $-1+2X_1$. We apply a working parameter with two components different from the true value and substantial discrepancy is found.

- ◇ (Simulation 6) k=3, $X=(1,X_1,X_2)$, $\beta_0=(-1.00,2.00,3.00)$ and $\beta_*=(-1.00,2.00,3.00)$. We simulate X_1 and X_2 from uniform distribution $U[0,1]$. Ys are generated using logit link with linear prediction $-1+2X_1+3X_2$. For 3-dimensional case, this is a control simulation and no discrepancy is detected between model assumption and responses (Ys).

- ◇ (Simulation 7) k=3, $X=(1,X_1,X_2)$, $\beta_0=(-1.00,2.00,3.00)$ and $\beta_*=(-1.00,2.50,3.50)$. We simulate X_1 and X_2 from uniform distribution $U[0,1]$. Ys are generated using logit link with linear prediction $-1+2X_1+3X_2$. We apply a working parameter with two linear term coefficients different from the true value and substantial discrepancy is found.

- ◇ (Simulation 8) k=3, $X=(1,X_1,X_2)$, $\beta_0=(-1.00,2.00,3.50)$ and $\beta_*=(-1.00,2.50,3.00)$. We simulate X_1 and X_2 from uniform distribution $U[0,1]$. Ys are generated using logit link with linear prediction $-1+2X_1+$

$3.5X_2$. Compared with simulation 7, we still apply a working parameter with two linear term coefficients different from the true value while with identical sum 5.5. For such a case, substantial discrepancy is still found.

- ◇ (Simulation 9) k=3, $X = (1, X_1, X_2)$, $\beta_0 = (-1.00, 2.00, 3.00)$ and $\beta_* = (-1.50, 2.50, 2.50)$. We simulate X_1 and X_2 from uniform distribution $U[0, 1]$. Ys are generated using logit link with linear prediction $-1 + 2X_1 + 3X_2$. We apply a working parameter with three components different from the true value and substantial discrepancy is found.

- ◇ (Simulation 10) k=3, $X = (1, X_1, X_2)$, $\beta_0 = (-1.00, 2.00, 3.00)$ and $\beta_* = (-1.50, 2.00, 3.00)$. We simulate X_1 and X_2 from uniform distribution $U[0, 1]$. Ys are generated using logit link with linear prediction $-1 + 2X_1 + 3X_2$. We apply a working parameter with only intercept different from the true value and substantial discrepancy is found.

To sum up, if the working parameter (β_*) is identical to true parameter (β_0), then the two misclassification rate curves (EMCER and OMCER) overlap exactly; if moderately nonidentical, then two misclassification rate curves divert and a substantial proportion of $\alpha_{TE}(\tau)$s exceed Critical α_{TE}. The models with parameter dimension greater than 3 give similar results.

Before proceeding to the next section, we now introduce a result (Chapter 4.4.2, McCullagh and Nelder 1989) with regard to parameter estimation. If a generalized linear model is correctly specified, then parameter estimation by maximizing likelihood can be implemented by iteration (Fisher scoring) and

$$E(\hat{\beta} - \beta) = O(n^{-1}) \quad \text{and} \quad \text{cov}(\hat{\beta}) = (X^T W X)^{-1}\{1 + O(n^{-1})\}, \tag{4.1}$$

where $\hat{\beta}$ is the estimated parameter, β is the true parameter, X is the design matrix (predictor vector population), W is a component from Fisher information matrix and n is the sample size. This proposition would be applied to the following simulation studies where parameter estimation under the hypothesized true model (predictor vector component subset, link function) is needed. Recall that, model fitting (parameter estimation) is not needed for simulations 1-10.

Missing interaction terms

We use a numerical example used by Landwehr et al (1984).

- ◇ (Simulation 11) We simulate two independent variables (Z_1, Z_2) of size n=10,000 from $U[0, 1]$. Ys are generated using logit link with linear prediction $4Z_1 + 4Z_2 - 12Z_1Z_2$, thus X =(X_0,X_1, X_2,X_3)=(1,Z_1,Z_2,Z_1Z_2), ($\beta_{0,0}$,$\beta_{1,0}$,$\beta_{2,0}$,$\beta_{3,0}$)=(0,4,4,-12). S-Plus glm(family=binomial(logit)) with

interaction term offers estimated working parameter ($\beta_{0,*}$,$\beta_{1,*}$,$\beta_{2,*}$,$\beta_{3,*}$)=(7E-3,3.94,3.82, -11.7). After model fitting (parameter estimation) under the correct model (with interaction) assumption, no $|\alpha_{TE}(\tau)|$ exceeds Critical α_{TE} although the estimated working parameter is slightly different from the true parameter which produces the random responses (Ys). See Table 1 (panel 2), Table 2 (panel 2) and Figure 2 (left panel, row 1).

◇ (Simulation 12) We simulate two independent variables (Z_1, Z_2) of size n=10,000 from $U[0,1]$. Ys are generated using logit link with linear prediction $4Z_1 + 4Z_2 - 12Z_1Z_2$, thus X =(X_0,X_1,X_2, X_3)=(1,Z_1,Z_2,Z_1Z_2), ($\beta_{0,0}$,$\beta_{1,0}$,$\beta_{2,0}$,$\beta_{3,0}$)=(0,4,4,-12). S-Plus glm(family=binomial(logit)) without interaction term offers estimated working parameter ($\beta_{0,*}$,$\beta_{1,*}$,$\beta_{2,*}$,$\beta_{3,*}$) = (2.98,
$-2.01, -2.10, 0.00$). After model fitting (parameter estimation) under the incorrect model (without interaction), many $|\alpha_{TE}(\tau)|$s exceed Critical α_{TE}. See Table 1 (panel 2), Table 2 (panel 2) and Figure 2 (right panel, row 1).

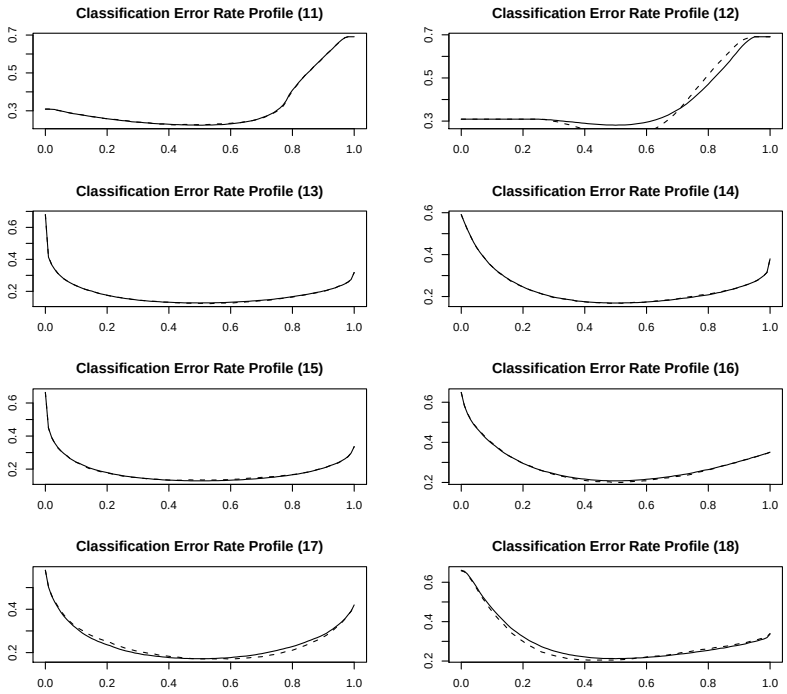

Figure 2: τ-wise comparison between EMCER and OMCER from simulations 11-18. (The solid line is EMCER curve, the dotted line is OMCER curve. τ covers [0,1] on the x-axis.)

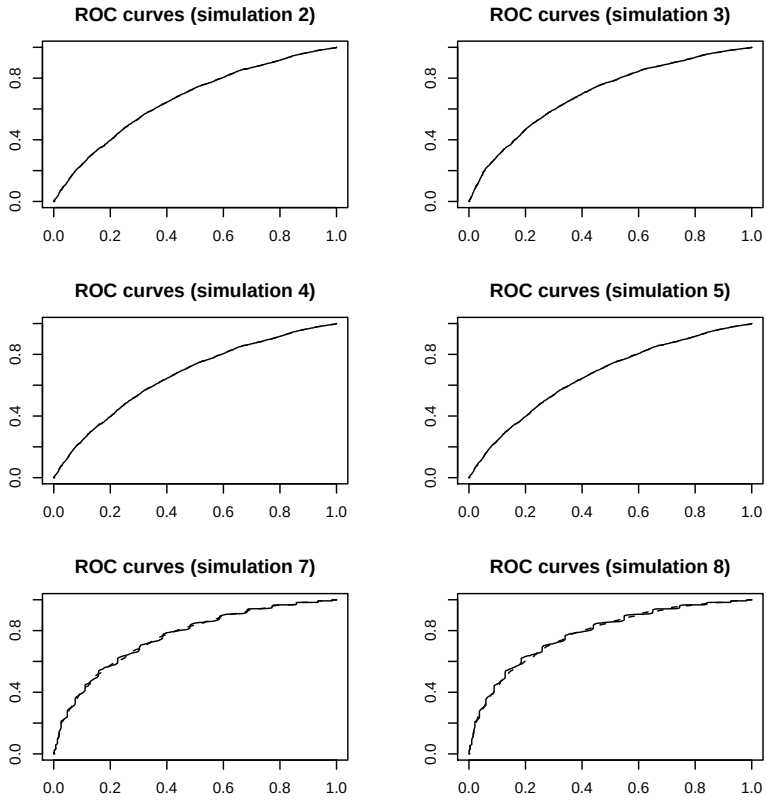

Figure 3: ROC comparison (I). (The solid line is ROC curve under correct model, the dash line is ROC under misspecified model. Both use same simulated Ys from correct model.)

4.2 Distinction among different link functions

Control simulations

- ◇ (Simulation 13) We simulate predictor population X_1 of size n=10,000 from $N(0,1)$ and Ys are generated by using probit link with linear prediction $-1+2X_1$, thus X =(1,X_1) and β_0=($\beta_{0,0}$, $\beta_{1,0}$)=(-1,2). glm(family=binomial(probit)) in S-Plus offers estimated working parameter for probit link: β_*=($\beta_{0,*}$, $\beta_{1,*}$)=(-1.05, 2.03). After model fitting (parameter estimation) under correct links, no $|\alpha_{TE}(\tau)|$ exceeds Critical α_{TE}. See Table 1 (panel 3), Table 2 (panel 3) and Figure 2 (left panel, row 2).

- ◇ (Simulation 14) We simulate predictor population X_1 of size n=10,000 from $N(0,1)$ and Ys are generated by using complementary log-log link with

linear prediction $-1 + 2X_1$, thus X =(1,X_1) and β_0=($\beta_{0,0}$, $\beta_{1,0}$)=(-1,2). glm(family=binomial(cloglog)) in S-Plus offers estimated working parameter for complementary log-log link: β_*=($\beta_{0,*}$, $\beta_{1,*}$)=(-1.02, 1.97). For simulation 14, we observe similar result from simulation 13. See Table 1 (panel 3), Table 2 (panel 3) and Figure 2 (right panel, row 2).

Probit and Logit

- ◇ (Simulation 15) We simulate predictor population X_1 of size n=10,000 from N(0,1) and Ys are generated by using probit link with linear prediction $-1+2X_1$, thus X =(1,X_1) and β_0=($\beta_{0,0}$, $\beta_{1,0}$)=(-1,2). glm(family=binomial(logit)) in S-Plus offers estimated working parameter for logit link: β_*=($\beta_{0,*}$, $\beta_{1,*}$)=(-1.77, 3.57). After model fitting (parameter estimation) under incorrect links, EMCER and OMCER are not very easy to distinguish. Around 5% of $|\alpha_{TE}(\tau)|$s exceed Critical α_{TE}, especially in the middle range of $\tau \in [0, 1]$. See Table 1 (panel 4), Table 2 (panel 4) and Figure 2 (left panel, row 3).

- ◇ (Simulation 16) We simulate predictor population X_1 of size n=10,000 from N(0,1) and Ys are generated by using logit link with linear prediction $-1+2X_1$, thus X =(1,X_1) and β_0=($\beta_{0,0}$, $\beta_{1,0}$)=(-1,2). glm(family=binomial(probit)) in S-Plus offers estimated working parameter for probit link: β_*=($\beta_{0,*}$, $\beta_{1,*}$)=(-0.61, 1.15). For simulation 16, we observe similar result from simulation 15. See Table 1 (panel 4), Table 2 (panel 4) and Figure 2 (right panel, row 3).

Complementary log-log and logit

- ◇ (Simulation 17) We simulate predictor population X_1 of size n=10,000 from N(0,1) and Ys are generated by using complementary log-log link with linear prediction $-1 + 2X_1$, thus X =(1,X_1) and β_0=($\beta_{0,0}$, $\beta_{1,0}$)=(-1,2). glm(family=binomial(logit)) in S-plus offers estimated working parameter for logit link: β_*=($\beta_{0,*}$, $\beta_{1,*}$)=(-0.69, 2.73). After model fitting (parameter estimation), EMCER and OMCER are clearly distinguished, a substantial proportion of $|\alpha_{TE}(\tau)|$s exceed Critical α_{TE}. See Table 1 (panel 4), Table 2 (panel 4) and Figure 2 (left panel, row 4).

- ◇ (Simulation 18) We simulate predictor population X_1 of size n=10,000 from N(0,1) and Ys are generated by using logit link with linear prediction $-1+2X_1$, thus X =(1,X_1) and β_0=($\beta_{0,0}$, $\beta_{1,0}$)=(-1,2). glm(family=binomial(cloglog)) in S-plus offers estimated working parameter for complementary log-log link: β_*=($\beta_{0,*}$, $\beta_{1,*}$)=(-1.21, 1.33). For simulation 18, we observe similar result from simulation 17. See Table 1 (panel 4), Table 2 (panel 4) and Figure 2 (right panel, row 4).

Remark In simulations from Section 4.1, we neither fit the model nor estimate the parameter while only simply took working parameter as the hypothesized true parameter for downstream goodness-of-fit test. However, we point out that, in simulations from Sections 4.1 and 4.2, we did fit the model and estimate the working parameter under the hypothesized true model assumption other than simply assigning certain value to working parameter (Section 4.1) for downstream goodness-of-fit test. In this case, we must say that, Results 1 and 2 would be approximate since the parameter vector β is estimated from random responses (Ys) and is random other than fixed (simulations 1-10, Section 4..1). However, proposition (4.1) shows that the random error for parameter estimation is at the order of $O(n^{-1/2})$, which is small under large sample size (n). On the other hand, in terms of consistence between EMCER and OMCER, large sample size (n) always makes the parameter estimation sufficiently good even if the estimated parameter vector components may be slightly different from the original components used for simulation (simulations 11, 13 and 14). So our simulations demonstrate that, under large sample size (n), the error evolved from working parameter (β_*) estimation is ignorable and Results 1 and 2 still sufficiently hold with the unknown true model parameter (β) replaced by the estimated parameter ($\hat{\beta} = \beta_*$). In other words, the lack-of-fit is likely due to model misspecification other than model fitting (parameter estimation). In the above simulations, we did not try diverse combinations of wrong parameter value (Section 4.1), incorrect predictor vector component subset (Section 4.1 and/or link function misspecification (Section 4.2), while apply only one of them to each simulation study. The combination case will lead to similar results.

5. A Note on Receiver Operating Characteristics (ROC)

Given assumed binary regression model and observed responses, the receiver operating characteristics (ROC) plots pairs of (1-specificity, sensitivity) based on threshold (τ)-specific classification rules ((2.1) and (2.2)) for all $\tau \in [0, 1]$ to form a model-specific ROC curve (profile). That is,

$$\text{sensitivity} = \Pr(Y^* = 1|Y = 1) \quad \text{and} \quad \text{specificity} = \Pr(Y^* = 0|Y = 0),$$

which are simply the proportions of correct classifications given observed $Y = 1$ or $Y = 0$ subsets. Note that, τ-specific (1-specificity, sensitivity) is in fact τ-specific (false positive rate, true positive rate). The higher area under the curve (AUC) represents better prediction (Chapter 6.2.6, Agresti 2002). However, we point out that, ROC can not be used for testing goodness-of-fit. We follow the notations in Section 4. Given continuous predictor variables Xs, we assume the true binary regression model has case probability function $\Pr_0(Y = 1|X, \beta_0) = g_0^{-1}(X^T\beta_0)$,

where $\beta_0 = (\beta_{0,0}, \beta_{1,0}, \beta_{2,0}, \ldots, \beta_{k-1,0})$ and $g_0^{-1}(\cdot)$ is a monotonically increasing function with range [0,1]. The responses (Ys) are produced from the true model. We also assume that, the incorrect model has case probability function $\Pr_*(Y = 1|X, \beta_*) = g_*^{-1}(X^T\beta_*)$, where $\beta_* = (\beta_{0,*}, \beta_{1,*}, \beta_{2,*}, \ldots, \beta_{k-1,*})$ and $g_*^{-1}(\cdot)$ is another monotonically increasing function with range [0,1]. The responses (Ys) are the same for both the true model and the incorrect model. We show that, under certain circumstances, identical (1-specificity, sensitivity) may exist for the true and the incorrect models with different thresholds (τs) and ROC curves may overlap under two different model assumptions. We use subscript "0" for the true model and "$*$" for the incorrect model. We consider some special situations where two intercepts ($\beta_{0,0}$ and $\beta_{0,*}$) and/or ($\beta_{1,0}$ and $\beta_{1,*}$) may have different values for two models (true and incorrect). The j-th component of predictor vector X is denoted by X_j for $0 \le j \le k-1$ and $X_0 = 1$. We assume parameter components $\beta_{j,0} = \beta_{j,*}$ for $2 \le j \le k-1$ and use notations β_j, $2 \le j \le k-1$ for both the true and incorrect models for notational simplicity. For generality, we first consider the case where β_0 and β_* have dimension $k > 2$.
Under the true model,

$$\begin{aligned} \text{1-specificity}(\tau_0) &= \Pr(g_0^{-1}(\beta_{0,0} + X_1\beta_{1,0} + \sum_{2\le j\le k-1} X_j\beta_j) \ge \tau_0 | Y = 0) \\ \text{sensitivity}(\tau_0) &= \Pr(g_0^{-1}(\beta_{0,0} + X_1\beta_{1,0} + \sum_{2\le j\le k-1} X_j\beta_j) \ge \tau_0 | Y = 1) \end{aligned} \quad (5.1)$$

Under the incorrect model,

$$\begin{aligned} \text{1-specificity}(\tau_*) &= \Pr(g_*^{-1}(\beta_{0,*} + X_1\beta_{1,*} + \sum_{2\le j\le k-1} X_j\beta_j) \ge \tau_* | Y = 0) \\ \text{sensitivity}(\tau_*) &= \Pr(g_*^{-1}(\beta_{0,*} + X_1\beta_{1,*} + \sum_{2\le j\le k-1} X_j\beta_j) \ge \tau_* | Y = 1) \end{aligned} \quad (5.2)$$

Note that, predictor population $(X_0, \ldots, X_{k-1})$, true parameter β_0 and incorrect parameter β_* are all known. For any τ_0, if we can find a corresponding τ_* which induces the same true (false) positive rate under the incorrect model given $Y = 1$ ($Y = 0$) as τ_0 does for the true model, then we will have two exactly overlapping ROC curves although the identical (1-specificity, sensitivity) pairs may be induced from different thresholds (τs). In order to find a pair of identical (1-specificity, sensitivity) from (5.1) and (5.2), a sufficient condition is that, the predictor vector subspaces induced from the following paired inequalities conditional on $Y = 1$

and $Y = 0$ response subspaces are all identical, i.e.,

$$\begin{aligned}&\{\mathbf{X}_{1,k-1} : g_0^{-1}(\beta_{0,0} + X_1\beta_{1,0} + \sum_{2\le j\le k-1} X_j\beta_j) \ge \tau_0, Y = 1\}\\ &= \{\mathbf{X}_{1,k-1} : g_*^{-1}(\beta_{0,*} + X_1\beta_{1,*} + \sum_{2\le j\le k-1} X_j\beta_j) \ge \tau_*, Y = 1\} \qquad (5.3)\end{aligned}$$

where $\mathbf{X}_{i,j}$ is a short notation for the vector $(X_i, \ldots, X_j)$ and

$$\begin{aligned}&\{\mathbf{X}_{1,k-1} : g_0^{-1}(\beta_{0,0} + X_1\beta_{1,0} + \sum_{2\le j\le k-1} X_j\beta_j) \ge \tau_0, Y = 0\}\\ &= \{\mathbf{X}_{1,k-1} : g_*^{-1}(\beta_{0,*} + X_1\beta_{1,*} + \sum_{2\le j\le k-1} X_j\beta_j) \ge \tau_*, Y = 0\} \qquad (5.4)\end{aligned}$$

Note that, the restricted predictor subspaces (on $Y = 1$ or $Y = 0$) are portions of arbitrarily collected predictor population X other than product space $\prod_{j=1}^{k-1}[-\infty, +\infty]$ for $(X_1, X_2, \ldots, X_{k-1})$ random predictor variable. Moreover, the distributions of X on subspaces restricted by $Y = 1$ or $Y = 0$ is not easy to study. Thus it is desirable that, τ_0 induces τ_* which is free of X. From (5.3) and (5.4), we have

$$\begin{aligned}&\{\mathbf{X}_{1,k-1} : \sum_{2\le j\le k-1} X_j\beta_j \ge g_0(\tau_0) - (\beta_{0,0} + X_1\beta_{1,0}), Y = 1\}\\ &= \{\mathbf{X}_{1,k-1} : \sum_{2\le j\le k-1} X_j\beta_j \ge g_*(\tau_*) - (\beta_{0,*} + X_1\beta_{1,*}), Y = 1\} \qquad (5.5)\end{aligned}$$

and

$$\begin{aligned}&\{\mathbf{X}_{1,k-1} : \sum_{2\le j\le k-1} X_j\beta_j \ge g_0(\tau_0) - (\beta_{0,0} + X_1\beta_{1,0}), Y = 0\}\\ &= \{\mathbf{X}_{1,k-1} : \sum_{2\le j\le k-1} X_j\beta_j \ge g_*(\tau_*) - (\beta_{0,*} + X_1\beta_{1,*}), Y = 0\} \qquad (5.6)\end{aligned}$$

We have two tracks to follow.

1) First, to make (5.5) and (5.6) free of $(X_2, X_3, \ldots, X_{k-1})$, we simply have

$$g_0(\tau_0) - (\beta_{0,0} + X_1\beta_{1,0}) = g_*(\tau_*) - (\beta_{0,*} + X_1\beta_{1,*})$$

i.e.,

$$X_1(\beta_{1,0} - \beta_{1,*}) = g_0(\tau_0) - g_*(\tau_*) - (\beta_{0,0} - \beta_{0,*}).$$

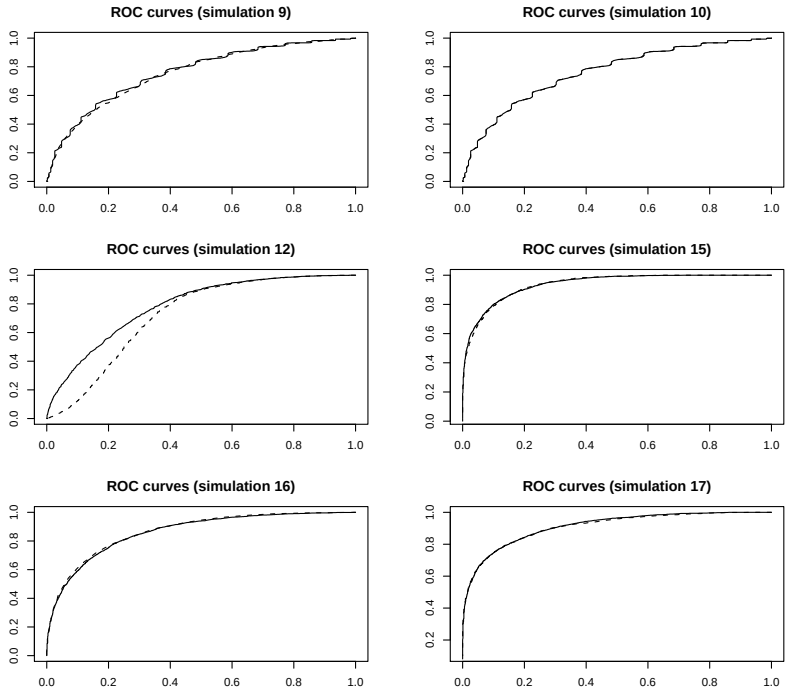

Figure 4: ROC comparison (II). (The solid line is ROC curve under correct model, the dash line is ROC under misspecified model. Both use same simulated Ys from correct model.)

Second, to make it free of X_1, we only require $\beta_{1,0} = \beta_{1,*}$ to obtain the following one-to-one correspondence between τ_0 and τ_*

$$\tau_* = g_*^{-1}(g_0(\tau_0) - (\beta_{0,0} - \beta_{0,*})). \tag{5.7}$$

2) First, to make (5.5) and (5.6) free of X_1, we simply have

$$\frac{\sum_{2\le j\le k-1} X_j\beta_j - g_0(\tau_0) + \beta_{0,0}}{\beta_{1,0}} = \frac{\sum_{2\le j\le k-1} X_j\beta_j - g_*(\tau_*) + \beta_{0,*}}{\beta_{1,*}}.$$

i.e.,

$$(\sum_{2\le j\le k-1} X_j\beta_j)(\beta_{1,0} - \beta_{1,*}) = g_0(\tau_0)\beta_{1,*} - g_*(\tau_*)\beta_{1,0} + \beta_{0,0}\beta_{1,*} - \beta_{1,0}\beta_{0,*}.$$

Second, to make it free of $(X_2, X_3, \ldots, X_{k-1})$, we only require $\beta_{1,0} = \beta_{1,*}$. Either track leads to the same correspondence between τ_0 and τ_* (5.7).

Result 3 The predictor and parameter are k-dimensional ($k > 2$) for the true and incorrect models, no matter the case probability functions are identical or not, if only the intercept parameter in the linear prediction may be different between the true model and the incorrect model, i.e., $\beta_{0,0} \neq \beta_{0,*}$, then ROC curves from two models overlap exactly.

The numerical example is specified in simulation 10, see Table 1 and Figure 4.

When parameter is 2-dimensional, from (5.3) and (5.4), we may require

$$\begin{aligned} & \{(X_0 = 1, X_1) : \beta_{0,0} + X_1\beta_{1,0} \geq g_0(\tau_0)\} \\ = \quad & \{(X_0 = 1, X_1) : \beta_{0,*} + X_1\beta_{1,*} \geq g_*(\tau_*)\}. \end{aligned}$$

To make it free of X_1, we may require $\beta_{1,0} \times \beta_{1,*} > 0$ (identical signs) and

$$\frac{g_0(\tau_0) - \beta_{0,0}}{\beta_{1,0}} = \frac{g_*(\tau_*) - \beta_{0,*}}{\beta_{1,*}}.$$

We have one-to-one correspondence

$$\tau^* = g_*^{-1}(\frac{\beta_{1,*}}{\beta_{1,0}}(g_0(\tau_0) - \beta_{0,0}) + \beta_{0,*}). \tag{5.8}$$

Result 4 If the predictor and parameter are 2-dimensional and the linear effects have equal signs between the true model and the incorrect model, then no matter the parameters and/or case probability functions are identical or not, ROC curves from two models overlap exactly.

The numerical examples are specified in simulations 2, 3, 4, 5, 15, 16 and 17 (Table 1, Figures 3 and 4. We also observe that, ROC may be useful for checking mistakenly dropping necessary effects from true model (simulation 12). See Table 1 and Figure 4.

6. Discussion

In this paper, a simple classification-error-rate-calibration (CERC) method is proposed for binary model screening (goodness-of-fit test) under large sample size and continuous predictor variables. The uniform convergence rate ($O(n^{-1/2})$) of the deviation test statistic tends to identify minor departure from the true model with appreciable power. On the other hand, when the predictor population realistically follows a well-shaped distribution, the discrepancy pattern between EMCER and OMCER is not inherently associated with sample size and CERC

screens threshold (τ)-specific pattern difference with homogeneous minimal variation (maximal power). As an alternative, for p_is less than τ, we may calculate the expected marginal case probability $\sum_{\{p_i<\tau,1\le i\le n\}} p_i/\sum_{\{p_i<\tau,1\le i\le n\}} 1$ and the observed marginal case probability $\sum_{\{p_i<\tau,1\le i\le n\}} 1_{\{Y=1\}}/\sum_{\{p_i<\tau,1\le i\le n\}} 1$ for τ-profiling comparison. However, the variance for the latter is $\sum_{\{p_i<\tau,1\le i\le n\}} p_i(1-p_i)/(\sum_{\{p_i<\tau,1\le i\le n\}} 1)^2$, which may suffer from testing power heterogeneity across different thresholds (τs). Note that this can be taken as τ-based classification-false-positive-rate profile (the lower-left shaded area in Diagram 1), while CERC considers both false positive and negative rates. Finally, we show that, ROC curve is not capable of goodness-of-fit test for binary regression models in the scenario discussed in the present work.

Acknowledgements

The authors are very grateful to the anonymous referee for careful reading of our manuscript and insightful comments which greatly improved our discovery and presentation. We also thank Min-Te Chao for his guidance during the manuscript reviewing process, and Narayanaswamy Balakrishnan and Peter Westfall for their constructive suggestions on an earlier draft.

References

Agresti, A. (2002). *Categorical Data Analysis*, 2nd edition. John-Wiley & Sons.

Landwehr, J.M., Pregibon, D. and Shoemaker, A.C. (1984). Graphical methods for assessing logistic regression models. *Journal of the American Statistical Association* **79**, 61-71.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd edition. Cambridge University Press.

Received December 3, 2007; accepted March 7, 2008.

Weichung Joe Shih
Biometrics Division
The Cancer Institute of New Jersey
New Brunswick, NJ 08904 USA
Shihwj@umdnj.edu

Also:
Department of Biostatistics
School of Public Health
University of Medicine and Dentistry of New Jersey
Piscataway, NJ 08854, USA

Junfeng Liu
Biometrics Division
The Cancer Institute of New Jersey
New Brunswick, NJ 08904 USA

Also:
Department of Biostatistics
School of Public Health
University of Medicine and Dentistry of New Jersey
Piscataway, NJ 08854, USA
Liu16@umdnj.edu