Liu, H., & Zhang, Z. (2017). Logistic Regression with Misclassification in Binary Outcome Variables: A Method and Software. Behaviormetrika, 44(2), 447–476.

Logistic Regression with Misclassification in Binary Outcome Variables: Method and Software

Haiyan Liu and Zhiyong Zhang

University of Notre Dame

Author Note

Haiyan Liu, Department of Psychology, University of Notre Dame; Zhiyong Zhang, Department of Psychology, University of Notre Dame.

This study was supported by a grant from the Institute of Education Sciences (US) (R305D140037). However, the contents of the paper do not necessarily represent the policy of the Department of Education, and you should not assume endorsement by the Federal Government.

Correspondence concerning this article should be addressed to Haiyan Liu, 118 Haggar Hall, Department of Psychology, Univeristy of Notre Dame, Notre Dame, IN 46556. Email: hliu6@nd.edu

Abstract

Misclassification means the observed category is different from the underlying one and it is a form of measurement error in categorical data. The measurement error in continuous, especially normally distributed, data is well known and studied in the literature. But the misclassification in a binary outcome variable has not yet drawn much attention in psychology. In this study, we show through a Monte Carlo simulation study that there are non-ignorable biases in parameter estimates if the misclassification is ignored. To deal with the influence of misclassification, we introduce a model with false positive and false negative misclassification parameters. Such a model can not only estimate the underlying association between the dependent and the independent variables but also provide the information on the extent of misclassification. To estimate the model, the maximum likelihood estimation method based on a Newton-type algorithm is utilized. Simulation studies are conducted to evaluate the performance and a real data example is used to demonstrate the usefulness of the new model. An R package is also developed to aid the application of the model.

Keywords: Binary outcome, Fisher scoring algorithm, Logistic regression, Misclassification, marijuana use

Logistic Regression with Misclassification in Binary Outcome Variables: Method and Software

Introduction

Classical methods for binary data analysis, such as logistic regression and contingency table analysis, assume that there is no measurement error in the variables involved in the model. However, this assumption often does not hold because almost nothing can be measured perfectly in the social and behavioral research. Measurement error, the difference between a measured value of quantity and its true value, is well known to threaten the validity of statistical inference. For example, measurement error can result in diminished correlations or regression coefficients. To capture its categorical attributes, measurement error is often referred to as misclassification in categorical outcome variables, especially dichotomous response variables (e.g., Gustafson, 2003; Kuha et al., 2005). Different from the misclassification due to the prediction error of a model in some other studies, in this study it is purely referred to as measurement error in the data collection process. Misclassification can happen in many settings. For example, it can be due to respondent error such as aberrant responses such as careless errors and lucky guessing. It may also happen in a survey when the participants do not want to provide trustful responses. For instance, in a study of marijuana use, a participant who has used marijuana might choose not to report it due to concerns over potential consequences. In general, misclassification means the recorded value of a discrete response variable is different from its true value.

The essential goal of measurement error analysis is to obtain unbiased parameter estimates and reliable inferences. Measurement error in continuous, especially normally distributed, data is well studied in the literature (e.g., Klepper & Leamer, 1984; Carroll et al., 2006). It is usually assumed to be normally distributed and independent with the underlying variable. There are many techniques/models dealing with continuous measurement error (Bagozzi, 1981; Fuller, 2009; Stefanski, 2000). For example, factor analysis is a multivariate technique that can be used to deal with measurement error in correlated variables (e.g., Cattell, 1952). The association of the observed scores with measurement errors and their underlying true score is modeled by factor loading (e.g., Child, 2006). Nonetheless, relatively fewer studies have investigated the influence of misclassification and proposed methods to handle it. This is partially due to the fact misclassification has very specific forms. For instance, in binary data, it can only be 0 if the true score is 1, and 1 if the true score is 1. As a result, the technique used in continuous measure error analysis is hardly extended to misclassifications.

Misclassification influences the validity of statistical inferences. The marginal misclassification may exist in a two-way contingency table (e.g., Bross, 1954; Goldberg, 1975), and it causes lower power of tests for independence (e.g., Assakul & Proctor, 1967; Chiacchierini & Arnold, 1977). The misclassification in the covariates caused both biases and misleading standard errors of parameter estimates (e.g., Carroll et al., 2006; Copeland et al., 1977; Davidov et al., 2003; Liu et al., 2013). To handle the problems of the misclassified covariates, it has been suggested that external information regarding misclassification rates be incorporated into the model (e.g. Davidov et al., 2003).

Misclassification in binary dependent variables in regression modeling have drawn great attention of researchers. To study the influence of misclassification on the regression coefficients estimates, Neuhaus (1999) derived a consistent estimator for the true association between the covariates and the outcome variable, which was a function of the observed association, the true slope parameter, and misclassification rates. It was shown that the association between the outcome variable and the covariates was attenuated when the outcome variable was subject to misclassification. However, it is hard to apply this method in practice for three reasons. First, this expression is optimal only when the true coefficients are close to 0, because the Taylor expansion technique was used in the derivation. Second, the derived consistent estimator is a function of true slope parameter, which is not available with misclassification in the data. Third, one needs to have prespecified misclassification rates in the data set, which are typically unknown. If the assumed misclassification rates are not consistent with the true misclassification rates, the estimator is still inconsistent. Similarly, to use the simulation and extrapolation (SIMEX) method proposed by Küchenhoff et al. (2006), the misclassification rates are either known or can be estimated from a separate sample available for the analysis.

Some other techniques are also proposed to account for the misclassification in the regression analysis. For instance, Edwards et al. (2013) used a multiple imputation method to reduce the bias, which also required a validation data set with no misclassification to provide information on the misclassification rates. A Bayesian method using data augmentation technique is adopted to do covariate selection when the binary outcome variable is subject to misclassification (Gerlach & Stamey, 2007). In this study, the imperfectly measured sample is treated as missing data and a perfectly measured one is required to augment the missing data. In some practical studies such as in Savoca (2011) and Magder & Hughes (1997), researchers also tried to adjust the influence of misclassification on the parameter estimates with given known misclassification rates or additional information on it. However, we are very often lack of such information and would like to estimate the extents of misclassification using the data at hand.

Hausman et al. (1998) proposed a modified model with two misclassification parameters: *false negative* and *false positive* parameters. The false negative (FN) parameter represents the probability of an observed value 0 having a true value 1 and the false positive (FP) parameter is the probability that an observed 1 is truly 0. Through such a model, one can estimate not only the parameters of the original research questions but also the extent of misclassification. However, the study can still be improved in several ways. First, the simulation study in Hausman et al. assumed that the false positive and false negative parameters were the same. Thus in that model, there was only one misclassification parameter even though there were two types of misclassification in the data. The performance of the model with free false positive and false negative parameters is not known to researchers and deserves further investigation. Second, the focus of the simulation study was on how severe the consequence of ignoring the misclassification, but not on how well the modified model works under different scenarios. Thus more comprehensive simulation studies are needed to understand the performance of the model. Third, in statistical inference, the standard error estimates are important, but it is not clear how reliable the standard error estimates from the modified model are. Fourth, Hausman et al. (1998) did not describe the algorithm they used and there is currently no easy-to-use software that can be used to estimate the models.

Therefore, the purpose of this study is to extend Hausman et al. (1998) with the following aims. First, we introduce the logistic regression model with misclassification parameters proposed by Hausman et al.. Second, we develop a Fisher scoring algorithm to obtain model parameter estimates and standard errors. Third, simulation studies are conducted to demonstrate the consequence of ignoring misclassification and to evaluate the performance of the new models in terms of both parameter estimates and their standard errors. Fourth, we introduce a newly developed R package to facilitate the application of the models.

The rest of the paper is organized in the following way. First, we formulate the model and elucidate the interpretation of the parameters to be estimated. Second, we derive the Fisher scoring algorithm for model estimation as well as the standard errors for parameter estimates, which is lacked in the literature. Third, simulation studies are conducted to address the problems caused by ignoring misclassification and to evaluate the performance of the Fisher scoring algorithm. Fourth, we illustrate how to analyze a set of real data on marijuana use collected by the National Longitudinal Survey of Youth study in year 1997 using the new models. Fifth, we demonstrate the use of our new developed R package “logistic4p” using the same data as in the empirical study. The last section concludes the study with discussion.

Logistic Regression with Misclassification Correction

In this section, we are going to introduce the logistic models with misclassification parameters. Following traditional assumptions on misclassification in binary response variables (e.g., Hausman et al., 1998; Neuhaus, 1999), we assume non-differential misclassification in the binary dependent variable. Non-differential misclassification means that the probability of being misclassified is the same across all subjects (e.g., Jurek et al., 2005). In addition, we consider the model involving at least one covariate and there is no measurement error in covariates as commonly assumed in most statistical models.

In the following, we use $\tilde{Y}$ to represent the true state of the binary response variable. To model the probability of $\tilde{Y}$ being 1, logistic regression model can be fitted to the response variable with a set of predictors $X_1, \ldots, X_p$ (e.g., McCullagh & Nelder, 1989; Nelder & Baker, 1972),

$$\begin{cases} \tilde{Y} & \sim \text{bernoulli}(F) \\ F & = \frac{1}{1+\exp(-\eta)} \\ \eta & = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \end{cases} \tag{1}$$

where $\beta_1, \cdots, \beta_p$ represent the association between covariates and the binary outcome variable $\tilde{Y}$.

Let $\{(y_i, x_i), i = 1, \cdots, n\}$ be a set of data collected from n participants. Without misclassification, the recorded binary data y_i's are the true realization of $\tilde{Y}$. By fitting the above model to the data, we could obtained the estimates of $\beta_1, \cdots, \beta_p$, which are consistent estimates of the population parameters. When some of true status are misclassified, the recorded binary data $\{y_1, \cdots, y_n\}$ are different from the true status $\{\tilde{y}_1, \tilde{y}_2, \cdots, \tilde{y}_n\}$, which are however blind to us. For instance, a participant i smoked marijuana, i.e., $\tilde{y}_i = 1$, but the recorded data indicates he/she did not, i.e, $y_i = 0$. Under the assumption of non-differential misclassification, the chance of misclassification is only related to the true status $\tilde{y}_i$ through the transition probability distribution function as follows,

$$Pr(y_i = 1|\tilde{y}_i = 0) = r_0 \tag{2}$$

$$Pr(y_i = 0|\tilde{y}_i = 0) = 1 - r_0 \tag{3}$$

$$Pr(y_i = 0|\tilde{y}_i = 1) = r_1 \tag{4}$$

$$Pr(y_i = 1|\tilde{y}_i = 1) = 1 - r_1 \tag{5}$$

where r_0 and r_1 are called *false positive* (FP) and *false negative* (FN) rates, respectively, which represent the extent of misclassification (e.g., McCullagh & Nelder, 1989). Subject to misclassification, the observed y_i and the true $\tilde{y}_i$ can be different. If one simply ignores the misclassification and fits a logistic regression model directly to y_i using Equation (1), the estimated logistic regression coefficients will not necessarily represent the true association between $\tilde{Y}$ and its predictors (e.g., Neuhaus, 1999).

In order to account for the misclassification, we need to find the true distribution of y_i 's . For an observation $y_i = 1$, there are two possibilities. First, the underlying $\tilde{y}_i = 1$ and the response is not misclassified. Second, the underlying $\tilde{y}_i = 0$ but $y_i = 1$ because of misclassification. Therefore, if π_i is the probability of $y_i = 1$ conditional on the vector of features of subject i, denoted by $\boldsymbol{x}_i = (1, x_{1i}, \cdots, x_{pi})'$, base on the law of total probability, we have,

$$\begin{aligned}
\pi_i &= Pr(y_i = 1|\boldsymbol{x}_i) \\
&= Pr(y_i = 1|\tilde{y}_i = 1, \boldsymbol{x}_i)Pr(\tilde{y}_i = 1|\boldsymbol{x}_i) + Pr(y_i = 1|\tilde{y}_i = 0, \boldsymbol{x}_i)Pr(\tilde{y}_i = 0|\boldsymbol{x}_i) \\
&= (1 - r_1)Pr(\tilde{y}_i = 1|\boldsymbol{x}_i) + r_0[1 - Pr(\tilde{y}_i = 1|\boldsymbol{x}_i)] \\
&= r_0 + (1 - r_0 - r_1)Pr(\tilde{y}_i = 1|\boldsymbol{x}_i) \\
&= r_0 + (1 - r_0 - r_1)F_i. \qquad (6)
\end{aligned}$$

As a consequence, the regular logistic regression model can be extended to include both false positive and false negative misclassification parameters as follows:

$$\begin{cases} y_i & \sim \text{bernoulli}(\pi_i) \\ \pi_i & = r_0 + (1 - r_0 - r_1)F_i \\ F_i & = \frac{1}{1+\exp(-\eta_i)} \\ \eta_i & = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} \end{cases} \tag{7}$$

with r_0 and r_1 defined earlier.

Let $\mathbf{1}$ be a $n-$dimensional column vector of 1, $\boldsymbol{X}_j, j = 1, \ldots, p$ be a vector of observed data for the j'th predictor, and $\mathbf{X} = (\mathbf{1}, \boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_p)$ be a $n \times (p+1)$ design matrix. The model defined in Equation (7) is identifiable if it satisfies two regularity conditions (e.g., Hausman et al., 1998; Newey & McFadden, 1994). One is $r_0 + r_1 < 1$, which is called monotonicity condition. The other is $E(\mathbf{X}'\mathbf{X}) < \infty$ and $\mathbf{X}'\mathbf{X}$ is non-singular. In practice, the misclassification rates r_0 and r_1 are expected to be small, generally less than 0.50. Otherwise, the misclassification would not happen purely due to chance. As a consequence, the monotonicity condition holds automatically in most general cases. The second condition is also required in the regular regression analysis, otherwise the parameter estimates would be extremely unstable from sample to sample. Therefore, the two conditions are usually met in practice.

The proposed model with misclassification parameters defined by Eqn (7) is closely relevant to the four-parameter logistic (4PL) IRT model, in which the predictor is a latent variable(Loken & Rulison, 2010) though. The false positive parameter r_0 corresponds to the guessing parameter in the 4PL IRT model, which is the lower asymptote of the mean curve. While, $1 - r_1$ corresponds to the upper asymptote parameter in the 4PL IRT model. When $r_1 = 0$, the upper asymptote is 1, the model corresponds to the tree-parameter logistic (3PL) IRT model (e.g., van der Linden & Hambleton, 2013). In Figure 1, we plot the probability $Pr(Y = 1)$ with the same regression coefficients $\beta_0 = -1$ and $\beta_1 = 1$ with different false positive and false negative rates. When $r_0 = 0$ and $r_1 = 0$, the lower and upper asymptotes are 0 and 1, which corresponds to the conventional logistic regression model. When $r_0 > 0$, the lower asymptote is larger than 0 and therefore, the probability of $Pr(Y = 1)$ is always at least r_0. When $r_1 > 0$, the upper asymptote can never reach 1.

We denote the model with both misclassification parameters as LG_{FPFN}, where "FP" and "FN" are the short forms of "false positive" and "false negative". When $r_0 = r_1 = 0$, the model reduces to the conventional logistic regression model (LG). In certain situations, one can also constrain the false positive and false negative rates to be the same ($r_0 = r_1 = r$). This model was studied in the simulation of Hausman et al. (1998) and will be referred to as LG_E. Furthermore, if false positive is the primary concern, we do not need to estimate r_1 but only r_0 (LG_{FP}), and if false negative parameter is of interest, we can set $r_0 = 0$ (LG_{FN}). These four models have fewer parameters and are easier to handle than LG_{FPFN}.

Fisher Scoring Algorithm

To estimate the parameters in the logistic models, the maximum likelihood (ML) estimation method is used here because it readily provides standard error estimates. Due to the nonlinear structure and the interaction between the misclassification parameters and the regression coefficients, no direct solution of ML estimates for the logistic regression models with misclassification parameters exists. Therefore we resort to numerical methods. Although the Newton-Raphson method is often used in obtaining ML estimates, we employ the Fisher scoring algorithm because its results are less dependent on the starting values and have better convergence rates (e.g., Schworer & Hovey, 2004; Longford, 1987).

The algorithm is based on the estimating equations from the ML estimation. For any y_i either 0 or 1, and $\boldsymbol{x_i} = (1, x_{i1}, \cdots, x_{ip})$ the conditional probability density function of Y_i

on the features of subject i is

$$Pr(Y_i = y_i|\boldsymbol{x_i}) = \pi_i^{y_i}(1-\pi_i)^{1-y_i} = \exp\{y_i\theta_i - \log(1+\exp(\theta_i))\} \tag{8}$$

with $\theta_i = \log\frac{\pi_i}{1-\pi_i}$, $\pi_i = r_0 + (1-r_0-r_1)F_i$, $F_i = \frac{\exp(\eta_i)}{1+\exp(\eta_i)}$, and $\eta_i = \boldsymbol{x_i'\beta}$. Given n independent observations $(\boldsymbol{x}_i, y_i)_{i=1}^n$, the likelihood function is

$$L = \exp\{\sum_{i=1}^n y_i\theta_i - \sum_{i=1}^n \log(1+\exp(\theta_i))\}$$

with the corresponding log-likelihood,

$$l = \sum_{i=1}^n l_i = \sum_{i=1}^n [y_i\theta_i - \log(1+\exp(\theta_i))]. \tag{9}$$

Recall that the unknown parameters in the model include the misclassification parameters r_0 and r_1 as well as the regression coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)'$. For convenience, we use $\boldsymbol{\gamma} = (r_0, r_1, \boldsymbol{\beta}')'$ to denote the column vector of all parameters.

To obtain the ML estimates of $\boldsymbol{\gamma}$, denoted by $\hat{\boldsymbol{\gamma}} = (\hat{r}_0, \hat{r}_1, \hat{\boldsymbol{\beta}}')'$, we need to get the solutions to the following set of estimating equations:

$$\boldsymbol{g_n} = \begin{cases} \frac{\partial l}{\partial r_0} & = \sum_{i=1}^n \frac{\partial l_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \pi_i}\frac{\partial \pi_i}{\partial r_0} = \sum_{i=1}^n \frac{y_i-\pi_i}{\pi_i(1-\pi_i)}\frac{\partial \pi_i}{\partial r_0} = 0 \\ \frac{\partial l}{\partial r_1} & = \sum_{i=1}^n \frac{\partial l_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \pi_i}\frac{\partial \pi_i}{\partial r_1} = \sum_{i=1}^n \frac{y_i-\pi_i}{\pi_i(1-\pi_i)}\frac{\partial \pi_i}{\partial r_1} = 0 \\ \frac{\partial l}{\partial \beta_0} & = \sum_{i=1}^n \frac{\partial l_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \pi_i}\frac{\partial \pi_i}{\partial \beta_0} = \sum_{i=1}^n \frac{y_i-\pi_i}{\pi_i(1-\pi_i)}\frac{\partial \pi_i}{\partial \beta_0} = 0 \\ \frac{\partial l}{\partial \beta_1} & = \sum_{i=1}^n \frac{\partial l_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \pi_i}\frac{\partial \pi_i}{\partial \beta_1} = \sum_{i=1}^n \frac{y_i-\pi_i}{\pi_i(1-\pi_i)}\frac{\partial \pi_i}{\partial \beta_1} = 0 \\ & \vdots \\ \frac{\partial l}{\partial \beta_p} & = \sum_{i=1}^n \frac{\partial l_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \pi_i}\frac{\partial \pi_i}{\partial \beta_p} = \sum_{i=1}^n \frac{y_i-\pi_i}{\pi_i(1-\pi_i)}\frac{\partial \pi_i}{\partial \beta_p} = 0 \end{cases}. \tag{10}$$

If a probability density function is from the exponential family, the following relationship holds (e.g., Agresti, 2013),

$$E[\frac{\partial l_i^2}{\partial \gamma_1 \gamma_2}] = -E[\frac{\partial l_i}{\partial \gamma_1}\frac{\partial l_i}{\partial \gamma_2}]$$

for a pair of parameters γ_1, γ_2. According to Equation (8), the density function of Y_i is from the exponential family even with the misclassification parameters. Therefore, for the logistic model with misclassification, we have for $j, k = 0, 1, \cdots, p$,

$$\begin{aligned}
E(\frac{\partial^2 l_i}{\partial r_0^2}) &= -E(\frac{\partial l_i}{\partial r_0})^2 = -\frac{1}{\pi_i(1-\pi_i)}(\frac{\partial \pi_i}{\partial r_0})^2 \\
E(\frac{\partial^2 l_i}{\partial r_1^2}) &= -E(\frac{\partial l_i}{\partial r_1})^2 = -\frac{1}{\pi_i(1-\pi_i)}(\frac{\partial \pi_i}{\partial r_1})^2 \\
E(\frac{\partial^2 l_i}{\partial \beta_j \partial \beta_k}) &= -E[(\frac{\partial l_i}{\partial \beta_j})(\frac{\partial l_i}{\partial \beta_k})] = -\frac{1}{\pi_i(1-\pi_i)}(\frac{\partial \pi_i}{\partial \beta_j})(\frac{\partial \pi_i}{\partial \beta_k}) \\
E(\frac{\partial^2 l_i}{\partial r_0 \partial r_1}) &= -E(\frac{\partial l_i}{\partial r_0}\frac{\partial l_i}{\partial r_1}) = -\frac{1}{\pi_i(1-\pi_i)}(\frac{\partial \pi_i}{\partial r_0})(\frac{\partial \pi_i}{\partial r_1}) \\
E(\frac{\partial^2 l_i}{\partial r_0 \partial \beta_j}) &= -E(\frac{\partial l_i}{\partial r_0}\frac{\partial l_i}{\partial \beta_j}) = -\frac{1}{\pi_i(1-\pi_i)}(\frac{\partial \pi_i}{\partial r_0})(\frac{\partial \pi_i}{\partial \beta_j}) \\
E(\frac{\partial^2 l_i}{\partial r_1 \partial \beta_j}) &= -E(\frac{\partial l_i}{\partial r_1}\frac{\partial l_i}{\partial \beta_j}) = -\frac{1}{\pi_i(1-\pi_i)}(\frac{\partial \pi_i}{\partial r_1})(\frac{\partial \pi_i}{\partial \beta_j})
\end{aligned}$$

with $\frac{\partial \pi_i}{\partial r_0} = 1 - F_i$, $\frac{\partial \pi_i}{\partial r_1} = -F_i$, and $\frac{\partial \pi_i}{\partial \beta_j} = (1 - r_0 - r_1)F_i(1 - F_i)x_{ij}$ with $x_{i0} = 1$ The Fisher information is

$$\mathcal{I}(\boldsymbol{\gamma}) \;=\; -E\left(\sum_{i=1}^{n} \frac{\partial^2 l_i}{\partial \gamma_k \partial \gamma_s}\right)_{k,s} = \sum_{i=1}^{n} E\left(\frac{\partial l_i}{\partial \gamma_k}\frac{\partial l_i}{\partial \gamma_s}\right)_{k,s}.$$

Because we have $p+3$ parameters in the model, thus $\mathcal{I}(\boldsymbol{\gamma})$ is a $(p+3) \times (p+3)$ matrix. Let $\mathbf{D}$ be a n by $p+3$ matrix with i'th row being the gradient of π_i with respect to the parameters, i.e., $(\frac{\partial \pi_i}{\partial r_0}, \frac{\partial \pi_i}{\partial r_1}, \frac{\partial \pi_i}{\partial \beta_0}, \frac{\partial \pi_i}{\partial \beta_1}, \cdots, \frac{\partial \pi_i}{\partial \beta_p})$ and $\mathbf{W}$ be a diagonal matrix with the diagonal elements $\frac{1}{\pi_i(1-\pi_i)}$. As a result,

$$\mathcal{I}(\boldsymbol{\gamma})_{(p+3)\times(p+3)} = \mathbf{D}'_{(p+3)\times n}\mathbf{W}_{n\times n}\mathbf{D}_{n\times(p+3)}. \tag{11}$$

Let $\mathbf{u} = [\frac{\partial l}{\partial r_0}, \frac{\partial l}{\partial r_1}, \frac{\partial l}{\partial \beta_0}, \cdots, \frac{\partial l}{\partial \beta_p}]'$, the gradient of the likelihood function in Equation (9) with respect to the parameters. Using the same notation, we have

$$\begin{aligned}\mathbf{u} &= [\sum_{i=1}^{n}\frac{y_i-\pi_i}{\pi_i(1-\pi_i)}\frac{\partial\pi_i}{\partial r_0},\sum_{i=1}^{n}\frac{y_i-\pi_i}{\pi_i(1-\pi_i)}\frac{\partial\pi_i}{\partial r_1},\sum_{i=1}^{n}\frac{y_i-\pi_i}{\pi_i(1-\pi_i)}\frac{\partial\pi_i}{\partial\beta_0},\cdots,\sum_{i=1}^{n}\frac{y_i-\pi_i}{\pi_i(1-\pi_i)}\frac{\partial\pi_i}{\partial\beta_p}]' \\ &= \mathbf{D}'\mathbf{W}(\boldsymbol{y}-\boldsymbol{\pi}) \end{aligned} \tag{12}$$

where $\boldsymbol{y}=(y_1,\ldots,y_n)'$ and $\boldsymbol{\pi}=(\pi_1,\ldots,\pi_n)'$.

With the Fisher information matrix, the parameter estimates can be obtained using the Fisher scoring algorithm. Given a set of starting values, we update the parameters at step $t+1$ using

$$\boldsymbol{\gamma}^{(t+1)}=\boldsymbol{\gamma}^{(t)}+(\mathcal{I}^{(t)})^{-1}\mathbf{u}^{(t)}=[(\mathbf{D}^{(t)'}\mathbf{W}^{(t)}\mathbf{D}^{(t)})^{-1}\mathbf{D}^{(t)'}\mathbf{W}^{(t)}][\boldsymbol{y}-\boldsymbol{\pi}^{(t)}+\mathbf{D}^{(t)}\boldsymbol{\gamma}^{(t)}]. \tag{13}$$

where $\boldsymbol{\gamma}^{(t)}$ are the parameter estimates at step t. Note that $\mathbf{D}^{(t)}$, $\mathbf{W}^{(t)}$, and $\boldsymbol{\pi}^{(t)}$ are evaluated with $\boldsymbol{\gamma}^{(t)}$ at step t. The iterative procedure stops when it satisfies certain stopping criterion. In the study, we stop the algorithm if $\max(|\boldsymbol{\gamma}^{(t+1)}-\boldsymbol{\gamma}^{(t)}|)<10^{-6}$, which means that in two consecutive steps, the maximum absolute difference for all parameters is smaller than 10^{-6}. The parameter estimates obtained in the last step is an approximation of the ML estimates for the model, denoted by $\hat{\boldsymbol{\gamma}}$. A good starting value can improve the speed of convergence. In our current algoritm, the default starting values are based on the parameter estimates from the conventional logistic regression (LG), which is best guess of parameter values without considering misclassifications.

Under some regularity conditions (e.g., Newey & McFadden, 1994), $\hat{\boldsymbol{\gamma}}$ is asymptotically unbiased and follows a normal distribution with the covariance matrix as the inverse of the Fisher information matrix,

$$\sqrt{n}(\hat{\gamma}-\gamma)\to N(\mathbf{0},\mathcal{I}^{-1}), \quad \text{asymptotically.}$$

where $\mathcal{I}$ is the population Fisher information matrix. Therefore, the asymptotic covariance matrix for $\hat{\boldsymbol{\gamma}}$ can be estimated by the inverse of estimated Fisher information matrix evaluated at the parameter estimates $\hat{\boldsymbol{\gamma}}$,

$$\widehat{cov}(\hat{\boldsymbol{\gamma}}) = \hat{\mathcal{I}}^{-1}(\hat{\boldsymbol{\gamma}}) = \left(\hat{\mathbf{D}}'\hat{\mathbf{W}}\hat{\mathbf{D}}\right)^{-1}.$$

The standard errors of the parameter estimates are readily available as the square roots of the corresponding diagonal elements of the covariance matrix.

Although the above Fisher scoring algorithm is derived for the model with both false positive and false negative misclassification parameters, its extension to other models is the same and thus is not repeated here.

In practice, a critical question is how to select a model that fits the data best. Because of the use of the ML estimation method, we can conduct a likelihood ratio test for two nested models. Let M_0 be a null model, e.g., a logistic regression model, which is nested in the model M_1, e.g., the logistic model with false positive and/or false negative parameters. Because M_1 contains more parameters than M_0, it fits the data at least as well as M_0. Whether M_1 fits the data significantly better than M_0 can be evaluated through hypothesis testing. The test statistics

$$D = -2[\log L(M_0) - \log L(M_1)].$$

asymptotically follows a Chi-squared distribution with degrees of freedom being the difference between the numbers of parameters in the two models.

For the non-nested models, Akaike information criterion (AIC) and Bayesian information criterion (BIC) can be used to compare the relative fit of models,

$$\begin{aligned} AIC &= -2\log L(M) + 2k \\ BIC &= -2\log L(M) + k\log n \end{aligned}$$

where k is the number of parameters and n is the sample size. A model with smaller AIC and/or BIC is preferred.

Simulation Study

In the previous section, we derived an iterative procedure to obtain parameter estimates, whose performance is still not clear. Thus, the goal of the simulation study is twofold. First, we would like to demonstrate the influence of misclassification on covariate parameter estimates. Second, we will evaluate the performance of the algorithm that we developed.

Study design

The data are generated according to the population model with four predictors in Equation (7). The population regression coefficients are set to be $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)' = (-3.5, -0.5, 3, 0.6, -1)'$, which are similar to those in the empirical study introduced in the next section. In addition, we consider three potentially influential factors in the simulation study: the sample size, the population distributions of predictors, and the misclassification rates.

Sample size. In practice, the misclassification rates are usually very small and thus hard to be detected. A relatively larger sample size is required to detect such small effects. For the 4PL IRT model, Loken & Rulison (2010) used the Bayesian estimation method and a sample size at least 600 is used. Hausman et al. (1998) used the sample size $n = 5000$ in the simulation study to estimate the model with only one misclassification parameter. In our model, we consider two misclassification parameters, thus a larger sample size is needed. In addition, we are interested in how the sample size influences the performance of the estimation procedure. Hence, we consider three different sample sizes $n = 1,000$, $2,000$, and $5,000$, which are smaller than both the one used by Hausman et al. (1998) and the one in the empirical study. For sample size less than $1,000$, we still could fit the the model with misclassification parameters, but the convergence rates might be low.

Predictors. In the simulation, we manipulate four predictors, among which the first three follow the Bernoulli distribution with parameter values $p = 0.5, 0.4$, and 0.75 respectively. The fourth predictor follows the standard normal distribution. This design covers both continuous and categorical predictors, which is the same as in the empirical example.

Misclassification rates. In the study, both r_0 and r_1 take one of the 4 values: 0, 0.05, 0.10, and 0.20. Therefore, there are 16 different combinations for (r_0, r_1) in total.

Data generating and model fitted

Combing the sample sizes and misclassification rates, we evaluate 48 different conditions in total. Under each condition, we simulate $1,000$ data sets. For each generated data set, we estimate the conventional logistic regression model (LG), the model with both misclassification parameters (LG_{FPFN}), and the model used to generate the data set. However, when the data generating model is the LG or LG_{FPFN} model, the true model is the same as LG or LG_{FPFN}, hence only two models are actually estimated. The data generating model and model fitted are summarized in Table 1.

Evaluation criteria

The performance of the models are evaluated according to the relative bias, standard errors estimates, coverage rates of confidence intervals, and convergence rates. Each of these are described below.

Let γ represent a parameter. And let R be the number of converged solutions among T replications. The convergence rate is

$$CV = \frac{R}{T} \times 100\%.$$

With R sets of parameter estimates $\hat{\gamma}_r, r = 1, \ldots, R$, the average parameter estimate is ,

$$\bar{\hat{\gamma}} = \sum_{r=1}^{R} \hat{\gamma}_r / R.$$

The relative bias is the relative discrepancy of the parameter estimate from its true value,

$$\text{bias} = \begin{cases} 100 \times \bar{\hat{\gamma}} & \gamma = 0 \\ 100 \times \frac{\bar{\hat{\gamma}} - \gamma}{|\gamma|} & \gamma \neq 0 \end{cases}, \tag{14}$$

which evaluates the accuracy of the parameter estimates. Typically, a bias less than 5% is *ignorable*, a bias between 5% and 10% is *moderate*, and a bias above 10% is *significant* (Muthén & Muthén, 2002). For each replicate $\hat{\gamma}_r$, its estimated standard error is denoted by $se(\hat{\gamma}_r)$. The average of estimated standard errors (a.se) of the parameter estimate is

$$\text{a.se} = \frac{1}{R} \sum_{r=1}^{R} se(\hat{\gamma}_r)$$

and the empirical standard error $(e.se)$ is the standard deviation of R converged replicates:

$$\text{e.se} = \sqrt{\frac{1}{R-1} \sum_{r=1}^{R} (\hat{\gamma}_r - \bar{\hat{\gamma}})^2}.$$

If the standard error is estimated well, we expect the average of estimated standard errors (a.se) is close to the empirical standard error (e.se). We construct the 95% confidence interval of γ in the r'th replication as $[\gamma_L^r, \gamma_U^r]$ with $\gamma_L^r = \hat{\gamma}_r - 1.96 \cdot se(\hat{\gamma}_r)$ and $\gamma_U^r = \hat{\gamma}_r + 1.96 \cdot se(\hat{\gamma}_r)$. The coverage rate of the 95% confidence interval is

$$CR = \frac{1}{R} \sum_{r=1}^{R} I_r,$$

where $I_r = 1$ if $\gamma_L^r \leq \gamma \leq \gamma_U^r$, otherwise, 0. With R independent replications, according to the Central Limit Theorem, the CR converges to a normal distribution with mean 0.95 and standard error $\sqrt{\frac{0.95\times0.05}{R}}$ asymptotically. Hence, a CR that falls in the range $[0.95 - 1.96\sqrt{0.95 \times 0.05/R}, 0.95 + 1.96\sqrt{0.95 \times 0.05/R}]$ is considered to be acceptable. In the case $R = 1000$, the range should be about $[0.935, 0.965]$.

Results

For the sake of space, only parts of the results are included in the manuscript. Complete results are available on request and on our website. In reporting the results, we focus on (1) whether the model with misclassification parameters can fit the data generated from a logistic model without misclassification, and (2) how much better the model with misclassification parameters performs compared to the regular logistic regression model (LG) if there is misclassification in the data.

Data without misclassification. We first investigate the performance of the logistic model with misclassification parameters when analyzing data without misclassification. Under this scenario, we first generated data from a logistic regression model with the regression coefficients specified in the simulation design. Then, we fit both the logistic regression model (LG) and and the model with both false positive and false negative misclassification parameters (LG_{FPFN}) to each generated data set. Results under this scenario are provided in Table 2.

When the logistic regression model was fitted to the data, our estimation algorithm never failed to converge. The biases of parameter estimates were ignorable ($< 5\%$) even when the sample size was as small as $1,000$. The coverage rates of 95% confidence intervals were generally close to the nominal level. In addition, the average of estimated standard errors ($a.se$) were close to the empirical standard errors ($e.se$), indicating the standard errors were also estimated accurately.

When the logistic model with both false positive and false negative parameters was fitted to the data, the convergence rate was low, although it increased along the sample size. When the sample size was $1,000$, the convergence rate was 38.4% and when the sample size was $5,000$, it was 67.8%. The bias was ignorable when $n = 5,000$, moderate when $n = 2,000$ and significant when $n = 1,000$. Although the biases for the misclassification parameters were generally small, they were overestimated consistently. The coverage rates of 95% confidence intervals were underestimated for the misclassification parameters but reasonable for the regression coefficients.

To summarize, when data were generated from a logistic model without misclassification, the logistic model performed very well. When the sample size was large, the model with misclassification parameters can also recover the regression parameters reasonably well.

Data with equal false positive and false negative parameters ($r_0 = r_1 = r$). With $r_0 = r_1 = r$, the true model is thus LG_E, the logistic model with equal false positive and false negative parameters. For each generated data set from the LG_E model, we fitted the logistic model (LG), the LG_{FPFN} model assuming unequal false positive and false negative parameters, and the LG_E models to it. Note that the logistic model was misspecified and the LG_{FPFN} model overfitted the data. The simulation results with $r_0 = r_1 = 0.05$ are presented in Table 3.

When the true model LG_E was fitted to the data, our algorithm converged well and the biases in parameter estimates were ignorable for $n = 1,000$ and they were smaller when the sample size increased. The coverage rate of confidence intervals were also generally acceptable. The average of estimated standard errors ($a.se$) were close to the empirical standard errors ($e.se$). Thus the algorithm provided reliable standard error estimates. Although it is not clear to us which algorithm was used by Hausman et al. (1998), our parameter estimates are very close to those reported by them and the discrepancy of relative biases are within 1%, which is purely due to random seeds of data generating process.

When we ignored the misclassification and fitted the LG model to the generated data, the parameter estimates were all biased, around $25 - 30\%$. The coverage rates of the

95% confidence intervals were lower than the nominal level, especially for β_0, β_2 and β_4.

When the LG$_{FPFN}$ model was fitted to the simulated data, the convergence rate was low, 70.3%, with $n = 1000$ but increased to 94.9% with $n = 5,000$. The biases in parameter estimates decreased as the sample size increased. The biases for all parameter estimates were ignorable with the sample size $n = 5,000$. The coverage rates and standard error estimates generally performed well.

When the misspecification was more severe such as $r_0 = r_1 = r = 0.10, 0.20$, the performance of LG$_E$ model was still very well, but the problems of fitting the LG model became even worse. The LG$_{FPFN}$ model still offered acceptable results especially when the sample size was large.

Therefore, when the data was generated from the model with equal false positive and false negative rates, the LG$_E$ model worked well even with the sample size not larger than $1,000$. The LG$_{FPFN}$ performed well too but required a large sample size to converge due to extra parameters to be estimated. The LG model caused severely biased parameter estimates and extremely low coverage rates. The problems of fitting the LG model did not disappear even when the sample size was large.

Data with misclassification, unequal false positive and false negative parameters ($r_0 = 0.05, r_1 = 0.1$ and $r_0 = 0.1, r_1 = 0.05$). The results for data with unequal false positive and false negative parameters are presented in Table 4 when $r_0 = 0.05$ and $r_1 = 0.1$, and in Table 5 when $r_0 = 0.1$ and $r_1 = 0.05$.

When the LG model was fitted to the data, the biases in the regression coefficients were all significant, about 30% when $r_0 = 0.05, r_1 = 0.1$ and 40% when $r_0 = 0.1, r_1 = 0.05$, and the coverage rates were very problematic. The results from the LG$_{FPFN}$ model seemed to be related to the sample size. When the sample size was $1,000$, the he convergence rates were low and the biases in both regression coefficients and the false negative parameter were substantial. When the sample size was $2,000$, both convergence rates and parameter estimates were improved. Finally, when the sample size was $5,000$, everything seemed to perform reasonably well.

Data with either false positive ($r_0 = 0.1, r_1 = 0$) or false negative ($r_0 = 0, r_1 = 0.1$). With false positive misclassification only, the LG$_{FP}$ is the true model and with false negative misclassification only, the LG$_{FN}$ is the correct model. The LG model under-fits the data while the LG$_{FPFN}$ over-fits the data. The simulation results are summarized in Table 6 and Table 7.

First, when the true model, either the LG$_{FP}$ or LG$_{FN}$, was fitted to the data, the results were generally good with ignorable biases in parameter estimates and reasonably good coverage rates of confidence intervals except for data with the false negative misclassification and small size ($n = 1,000$). When the misclassification was ignored by fitting the LG model to the generated data sets, the parameter estimates had severe biases and the coverage rates of the 95% the confidence intervals were low. For data with only false negative misclassification, the LG model provided reasonable parameter estimates but still bad coverage rates. Especially, the results from the LG model did not improve with the increase of sample size. When the LG$_{FPFN}$ was fitted to the simulated data, the convergence can be a problem but improved with the increase of the sample size. The biases of parameters became ignorable in general when the sample size was as large as $5,000$.

Summary of simulation findings

When ignoring misclassification in data, the use of ordinary logistic regression led to severe biases in parameter estimates. The estimated regression coefficients were biased towards 0, thus the association between the predictors and outcome variables were underestimated. The logistic regression with both false positive and false negative parameters was able to correctly recover both regression coefficients and misclassification parameters but required a large sample size. For example, with a sample size $2,000$, the results were acceptable and with a sample size $5,000$, the results were generally accurate. It was also worth noting that for the model with either false positive or false negative parameter, the results can be very good even with a smaller sample size $1,000$.

Real Data Analysis

We now illustrate how to apply the proposed model by analyzing a set of empirical data. The data were from the National Longitudinal Survey of Youth 1997 (NLSY). All the data used in the current analysis were collected in 1997. The outcome variable of interest is whether a participant has ever used marijuana and the predictors include gender, residence area, smoking cigarettes, and peer's life style reported by participants. The primary interest of the analysis is to estimate the true proportion of marijuana use and evaluate the relationship between marijuana use and the four predictors.

The sample size of the data is 5399. About half (49.2%) of the participants were identified as female and 74.8% of the participants lived in urban areas. Around 40% of the participants reported ever tried cigarettes. In the data, 20.3% of the participants reported that they had used marijuana ever before the survey. Peer's life style was measured by self-reported scores on six items. The higher score, the healthier their peers lived.

Because we did not know which model would fit the data best, we fitted and compared five models: the ordinary logistic regression model (LG), and four models with misclassification parameters (LG_{FPFN}, LG_{FP}, LG_{FN}, LG_E). Among the five models, the LG_{FP} and LG_E model did not converge. If they were the true model, they should converge almost surely according to our simulation results in Table 3 and Table 6. Thus, the nonconvergence of the two models was owing to the lack of fit of the models to the data. The results for other three models were provided in Table 8.

To determine which model fitted the data best, we compared the three converged models based on AIC, BIC and likelihood ratio tests. The AIC and BIC indices for the three models were offered in Table 8a. The LG_{FN} model had the smallest AIC and BIC, indicating that it fitted the data best among the three converged models. The results for the likelihood ratio tests were provided in Table 8b. First, comparing the LG against the

LG_{FPFN} and LG_{FN} models, the χ^2 statistics were 14.29 and 13.84 with p-values 0.0008 and 0.0002, respectively. Therefore, the LG model fitted the data significantly worse than both the LG_{FPFN} and LG_{FN} models. However, the LG_{FN} and LG_{FPFN} models appeared to fit the data equally well with the estimated χ^2 statistic 0.46 and p-value 0.4986. Since the LG_{FN} model had one parameter less, we accepted it as the best fit model for the NLSY data based on the parsimony principle. Thus, we used the LG_{FN} model as our final model for further analysis and interpretation.

In the LG_{FN} model, the estimated false negative rate (FN) was 0.1947 with p-value less than 0.001, which indicated among the people who had used marijuana indeed, 19.47% of them reported they did not. As a consequence, the observed proportion of marijuana use was smaller than the true proportion. According to Equation (6), the proportions of the true marijuana use (F) and the observed marijuana use (π) satisfy the following relationship

$$\pi = r_0 + (1 - r_0 - r_1)F$$

or equivalently $F = (\pi - r_0)/(1 - r_0 - r_1)$. For the NLSY data, the observed proportion of marijuana use was 20.3%. Therefore, the estimated proportion of true marijuana use after the correction of misclassification should be

$$\text{true proportion} = \frac{\text{observed proportion} - r_0}{1 - r_1 - r_0} = \frac{20.3\% - 0}{1 - 19.47\% - 0} \approx 25.21\%,$$

which was about 5% larger than the observed proportion on average.

In terms of the association between the predictors and marijuana use, we observed the following. First, girls were less likely to use marijuana than boys as indicated by the coefficients for gender (-0.6139) given other covariates the same. Second, if a participant smoked cigarettes, it is more likely for him/her to use marijuana. Third, participants who lived in urban areas were more likely to use marijuana than those who lived in rural areas when other predictors were controlled at the same level. Finally, for a participant whose peers lived healthier lives, he or she was less likely to use marijuana.

R Package

The R package "`logistic4p`" is developed to facilitate the use of logistic models with misclassification parameters. The package computes the misclassification rates, regression coefficients, and their standard errors based on the model and iterative procedure introduced in Section 2 and 3. In addition, it also offers the p-values, log-likelihood, and model fit indices such as AIC and BIC. The codes will run in any system that can run R for they are created within R. The NLSY data set is included as an example in the R package. In the remainder of this section, we illustrate how to use the R package using the NLSY data set.

In order to use the R package, one needs to install it on your computer first with `install.packages("logistic4p", repos="http://r-forge.r-project.org")` and then load it using the command `library(logistic4p)`. To estimate a model, users can use the R function `logistic4p(x, y, initial, model = c("lg", "fp.fn", "fp", "fn", "equal"), max.iter = 1000, epsilon = 1e-06, detail = FALSE)`, in which `x` is the matrix or data frame including the predictors and `y` is the vector of the binary dependent variable. The users may provide initial values for the parameters to be estimated, otherwise the default one, which is based on the estimates of the conventional LG model, will be used. Through this function, users can fit the five models discussed in the study to the data. The default model is the logistic model without misclassification parameter (lg) but can be changed by the `model` argument. The default maximum number of iterations and tolerance are $1,000$ and $1e-06$, which are subject to change by users.

The R input and output of analyzing the nested data is provided in Figure 2. First load the data using `data(nlsy)`. The dependent variable is the marijuana use, which is the first variable in the data set. The other four variables are the predictors. For illustration, we ran the logistic model with both false positive and false negative misclassification parameters with command `logistic4p(x, y, model="fp.fn")` using the default initial values. The output is provided in Figure 2. The algorithm converged after 299 iterations. The log-likelihood, AIC, and BIC are -1725.302, 3464.605, and 3510.763, respectively. In addition, the parameter estimates, standard errors, z.values, and two-sided p-values are also provided.

Discussion and Conclusion

Binary data are often collected in the social and behavioral research, such as in cognitive testing (e.g., right or wrong) and in diagnostic analysis (e.g., cancer or not). To analyze the binary outcome data, logistic regression models are typically used. In the conventional logistic regression (LG) analysis, it is assumed that there is no response error or misclassification on the outcome variable. However, in practice, this assumption hardly holds. As a consequence, the parameter estimates and statistical inference based on the conventional logistic regression may not be trustworthy.

In this study, we investigated the consequences of ignoring misclassification in binary outcome variables and presented several alternative models that can handle misclassification. The alternative models included the logistic model with only the false positive parameter (LG_{FP}), the logistic model with only the false negative parameter (LG_{FN}), the logistic model with equal false positive and false negative parameters (LG_{E}) and the logistic model with free false positive and false negative parameters (LG_{FPFN}). To estimate the models, we employed a Fisher scoring algorithm that provided both parameter estimates and standard error estimates.

Through simulation studies, we showed that the parameters in the models with misclassification parameters can be estimated well with correctly specified models and sufficient large sample size. Blindly fitting a logistic regression model to the data with misclassification resulted in severely biased parameter estimates. However, overfitting the data without misclassification with a model with misclassification parameters can still provide reasonable results. In the real data analysis, we showed that different models can be compared using AIC and BIC, and a model with smaller AIC and BIC is usually suggested. For nested models, the likelihood ratio test can also be used. The alternative model is preferred over the null model when it is significantly better; otherwise the null model is recommended.

Our simulation results showed that both the parameter estimates and coverage rates suffered a lot if the misclassification in the data was ignored. The algorithm we developed offers accurate parameter and standard error estimates when the population model was fitted to the data. Although the LG$_{FPFN}$ model contains extra parameters when fitted to the data set with no or only one type of misclassification, it still works well especially when the sample size is large. Compared to the true model, the LG$_{FPFN}$ requires relatively larger sample sizes to perform well. In general, a sample size at least $5,000$ can ensure the parameters are well recovered. And to estimate the model with just one misclassification parameter, a sample size of $2,000$ is a safe bet although a smaller sample size, e.g., $1,000$, can also achieve reasonable results.

If a model is badly misspecified, our software and algorithm may not provide converged results, although intermediate results are still available for diagnostic purposes. For example, if the data are truly from a model with the false positive parameter but the model with the false negative parameter is used, it almost never converges. Therefore, when getting non-convergent results, one may consider fitting a different model. There are situations that even with the correct model, our algorithm might not converge. To deal with the problem, our R package allows a user to provide customized starting values to improve convergence. The default starting values are based on the parameter estimates from the conventional logistic regression (LG).

As in other regression analysis, we assume that there are no measurement errors in predictors. However, it is possible to extend the model to account for the measurement error in them. In addition, although this study has focused on the binary outcome variable, the idea of introducing misclassification parameters in the model can be extended to ordinal data or nominal data analysis.

For the misclassification rates are generally small and hard to detect, a relatively large sample is required for the estimation of a logistic model with misclassification parameters. Bayesian estimation method can be useful taking its advantages of incooperating relevant prior information on the misclassification parameters (McInturff et al., 2004), if such kind of prior information is available. A systematic evaluation is lacked in the literature. In addition, it has some potential problems such as the boundary issues. Bayesian estimation of misclassification parameters are always positive, regardless the fact that the misclassification rates could be exactly 0 in the population. Model selection among the five different forms of models is subtle and further investigation is still demanded.

To summarize, if one suspects that a binary outcome variable is not reliably measured, a logistic regression model with misclassification parameters can be applied. The comparison between the new model and a logistic regression model can provide insight on whether it is necessary to estimate the misclassification parameters.

References

Agresti, A. (2013). *Categorical data analysis* (3rd ed., Vol. 996). New York: John Wiley & Sons.

Assakul, K., & Proctor, C. H. (1967). Testing independence in two-way contingency tables with data subject to misclassification. *Psychometrika*, *32*(1), 67–76.

Bagozzi, R. P. (1981). Evaluating structural equation models with unobservable variables and measurement error: a comment. *Journal of Marketing Research*, 375–381.

Bross, I. (1954). Misclassification in 2 x 2 tables. *Biometrics*, *10*(4), 478–486. doi: 10.2307/3001619

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective.* CRC press.

Cattell, R. B. (1952). *Factor analysis: an introduction and manual for the psychologist and social scientist.* Harper.

Chiacchierini, R. P., & Arnold, J. C. (1977). A two-sample test for independence in 2× 2 contingency tables with both margins subject to misclassification. *Journal of the American Statistical Association*, *72*(357), 170–174. doi: 10.1080/01621459.1977.10479933

Child, D. (2006). *The essentials of factor analysis.* A&C Black.

Copeland, K. T., Checkoway, H., McMichael, A. J., & Holbrook, R. H. (1977). Bias due to misclassification in the estimation of relative risk. *American Journal of Epidemiology*, *105*(5), 488–495.

Davidov, O., Faraggi, D., & Reiser, B. (2003). Misclassification in logistic regression with discrete covariates. *Biometrical Journal*, *45*(5), 541–553. doi: 10.1002/bimj.200390031

Edwards, J. K., Cole, S. R., Troester, M. A., & Richardson, D. B. (2013). Accounting for misclassified outcomes in binary regression models using multiple imputation with internal validation data. *American journal of epidemiology*, *177*(9), 904–912.

Fuller, W. A. (2009). *Measurement error models* (Vol. 305). John Wiley & Sons.

Gerlach, R., & Stamey, J. (2007). Bayesian model selection for logistic regression with misclassified outcomes. *Statistical Modelling*, *7*(3), 255–273.

Goldberg, J. D. (1975). The effects of misclassification on the bias in the difference between two proportions and the relative odds in the fourfold table. *Journal of the American Statistical Association*, *70*(351a), 561–567. doi: 10.1080/01621459.1975.10482472

Gustafson, P. (2003). *Measurement error and misclassification in statistics and epidemiology: impacts and bayesian adjustments*. CRC Press.

Hausman, J. A., Abrevaya, J., & Scott-Morton, F. M. (1998). Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics*, *87*(2), 239–269. doi: 10.1016/S0304-4076(98)00015-3

Jurek, A. M., Greenland, S., Maldonado, G., & Church, T. R. (2005). Proper interpretation of non-differential misclassification effects: expectations vs observations. *International journal of epidemiology*, *34*(3), 680–687. doi: 10.1093/ije/dyi060

Klepper, S., & Leamer, E. E. (1984). Consistent sets of estimates for regressions with errors in all variables. *Econometrica: Journal of the Econometric Society*, 163–183.

Küchenhoff, H., Mwalili, S. M., & Lesaffre, E. (2006). A general method for dealing with misclassification in regression: the misclassification simex. *Biometrics*, *62*(1), 85–96.

Kuha, J., Skinner, C., & Palmgren, J. (2005). Misclassification error. *Encyclopedia of Biostatistics*. doi: 10.1002/0470011815.b2a03084

Liu, Y., Liu, J., & Zhang, F. (2013). Bias analysis for misclassification in a multicategorical exposure in a logistic regression model. *Statistics & Probability Letters*, *83*(12), 2621–2626. doi: 10.1016/j.spl.2013.08.014

Loken, E., & Rulison, K. L. (2010). Estimation of a four-parameter item response theory model. *British Journal of Mathematical and Statistical Psychology*, *63*(3), 509-525. doi: 10.1348/000711009X474502

Longford, N. T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika*, *74*(4), 817–827. doi: 10.1002/j.2330-8516.1987.tb00217.x

Magder, L. S., & Hughes, J. P. (1997). Logistic regression when the outcome is measured with uncertainty. *American Journal of Epidemiology*, *146*(2), 195–203.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models.* CRC press.

McInturff, P., Johnson, W. O., Cowling, D., & Gardner, I. A. (2004). Modelling risk when binary outcomes are subject to error. *Statistics in medicine*, *23*(7), 1095–1109.

Muthén, L. K., & Muthén, B. O. (2002). How to use a monte carlo study to decide on sample size and determine power. *Structural Equation Modeling*, *9*(4), 599–620.

Nelder, J. A., & Baker, R. J. (1972). *Generalized linear models.* Wiley Online Library.

Neuhaus, J. M. (1999). Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika*, *86*(4), 843–855. doi: 10.1093/biomet/86.4.843

Newey, W. K., & McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics*, *4*, 2111–2245.

Savoca, E. (2011). Accounting for misclassification bias in binary outcome measures of illness: The case of post-traumatic stress disorder in male veterans. *Sociological Methodology*, *41*(1), 49–76.

Schworer, A., & Hovey, P. (2004). Newton-raphson versus fisher scoring algorithms in calculating maximum likelihood estimates. *Electronic Proceedings of Undergraduate Mathematics Day*, *1*, 1–11. Retrieved from http://ecommons.udayton.edu/cgi/viewcontent.cgi?article=1006&context=mth_epumd

Stefanski, L. (2000). Measurement error models. *Journal of the American Statistical Association*, *95*(452), 1353–1358.

van der Linden, W. J., & Hambleton, R. K. (2013). *Handbook of modern item response theory*. Springer Science & Business Media.

Data generating model		Model fitted
LG	$r_0 = 0, r_1 = 0$	LG, LG_{FPFN}
LG_E	$r_0 = r_1 \in \{0.05, 0.10, 0.20\}$	LG, LG_{FPFN}, LG_E
LG_{FPFN}	$r_0 \neq r_1 \in \{0.05, 0.10, .20\}$	LG, LG_{FPFN}
LG_{FP}	$r_0 \in \{0.05, 0.10, 0.20\}, r_1 = 0$	LG, LG_{FP}, LG_{FPFN}
LG_{FN}	$r_0 = 0, r_1 \in \{0.05, 0.10, 0.20\}$	LG, LG_{PN}, LG_{FPFN}

Table 1
Data generating model and fitted models

Table 2

Analysis of data from the model without misclassification ($r_0 = 0, r_1 = 0$).

		LG				LG_{FPFN}			
Par	True	bias(%)	a.se	e.se	CR(%)	bias(%)	a.se	e.se	CR(%)
		$n = 1,000$							
r_0	0	-	-	-	-	0.73	0.0099	0.0091	92.2
r_1	0	-	-	-	-	6.51	0.1128	0.1083	**85.4**
β_0	-3.5	-1.38	0.2975	0.3008	94.8	-9.36	0.4901	0.4738	97.1
β_1	-0.5	-1.25	0.1947	0.1940	95.3	**-20.66**	0.2478	0.2474	95.1
β_2	3	1.20	0.2349	0.2395	95.2	**14.90**	0.5593	0.5443	94.5
β_3	0.6	2.67	0.2357	0.2426	94.7	**19.64**	0.2955	0.2932	96.4
β_4	-1	-1.05	0.1122	0.1146	94.8	**-17.54**	0.2362	0.234	95.1
CV(%)		100				38.4			
		$n = 2,000$							
r_0	0	-	-	-	-	0.87	0.0072	0.0701	**86.1**
r_1	0	-	-	-	-	3.86	0.0857	0.1058	**84.1**
β_0	-3.5	-0.65	0.2087	0.2123	95.0	-4.30	0.3132	0.5583	94.0
β_1	-0.5	-0.13	0.137	0.1369	95.3	-5.80	0.1606	0.1652	96.4
β_2	3	0.55	0.1647	0.1676	94.0	6.85	0.3575	0.5083	95.6
β_3	0.6	0.98	0.1657	0.1676	94.9	8.25	0.1937	0.2099	95.4
β_4	-1	-0.57	0.079	0.0775	96.6	-8.34	0.1538	0.1955	95.2
CV(%)		100				49.6			
		$n = 5,000$							
r_0	0	-	-	-	-	0.27	0.0042	0.0406	**87.9**
r_1	0	-	-	-	-	0.92	0.0562	0.0732	**86.1**
β_0	-3.5	-0.16	0.1313	0.1356	95.0	-1.13	0.1791	0.3105	94.0
β_1	-0.5	-0.79	0.0864	0.0873	94.3	-2.74	0.0957	0.106	95.0
β_2	3	0.33	0.1037	0.1075	94.7	1.98	0.2048	0.3031	94.7
β_3	0.6	-0.55	0.1043	0.1058	95.3	2.18	0.1148	0.1222	94.5
β_4	-1	-0.29	0.0498	0.049	95.0	-2.33	0.0894	0.1137	95.0
CV(%)		100				67.8			

Note. A bold number is either a significant bias (bias>10%) or a bad coverage rate (CR< 90%). LG represents the logistic regression with no misclassification parameter and LG_{FPFN} is the logistic regression model with both false positive and false negative parameters. The CR and CV denote the coverage rates and convergence rates respectively.

Table 3

Analysis of data from the model with equal false positive and false negative parameters ($r_0 = r_1 = r = 0.05$)

		LG				LG_{FPFN}				LG_E			
Par	True	bias(%)	a.se	e.se	CR(%)	bias(%)	a.se	e.se	CR(%)	bias(%)	a.se	e.se	CR(%)
		$n = 1,000$											
r_0	0.05	-	-	-	-	4.37	0.0199	0.0203	93.2	-1.85	0.0194	0.0208	92.7
r_1	0.05	-	-	-	-	**49.96**	0.1524	0.1586	**81.9**				
β_0	-3.5	**26.32**	0.2356	0.2345	**4.1**	-9.59	0.7091	0.7475	97.0	-4.12	0.6085	0.6334	95.5
β_1	-0.5	**27.93**	0.1719	0.1754	**86.7**	**-17.65**	0.2973	0.3181	95.4	-3.62	0.2424	0.2468	95.7
β_2	3	**-27.08**	0.1833	0.1879	**0.9**	**14.31**	0.8033	0.8702	97.7	4.09	0.523	0.5423	94.9
β_3	0.6	**-28.39**	0.206	0.2065	**85.1**	**16.46**	0.3556	0.3976	95.3	4.88	0.2937	0.2964	95.2
β_4	-1	**28.64**	0.0931	0.097	**15.3**	**-15.32**	0.3108	0.3266	97.3	-3.13	0.184	0.1928	93.9
CV(%)		100				70.3				99.1			
		$n = 2,000$											
r_0	0.05	-	-	-	-	-1.34	0.0149	0.0153	94.9	-3.20	0.014	0.0146	94.7
r_1	0.05	-	-	-	-	2.52	0.1139	0.1193	**86.0**				
β_0	-3.5	**26.82**	0.1657	0.1752	**0.1**	-3.15	0.4508	0.4798	93.5	-0.89	0.4111	0.4287	93.9
β_1	-0.5	**27.62**	0.1209	0.1202	**78.2**	-6.28	0.1916	0.2019	94.7	-2.07	0.1678	0.1722	94.5
β_2	3	**-27.44**	0.1289	0.1301	**0.0**	4.84	0.5172	0.5432	94.6	1.10	0.3525	0.3717	94.3
β_3	0.6	**-29.01**	0.1450	0.1494	**74.7**	5.62	0.2299	0.2352	94.8	0.72	0.203	0.2067	94.3
β_4	-1	**28.63**	0.0655	0.069	**1.6**	-5.56	0.2032	0.2152	94.5	-0.81	0.1272	0.1338	93.0
CV(%)		100				82.9				100			
		$n = 5,000$											
r_0	0.05	-	-	-	-	1.54	0.0094	0.0095	94.4	1.06	0.0089	0.0087	94.9
r_1	0.05	-	-	-	-	3.79	0.0713	0.0805	91.0				
β_0	-3.5	**27.02**	0.1044	0.1061	**0.0**	-2.09	0.2792	0.2848	95.9	-1.25	0.2610	0.2641	95.5
β_1	-0.5	**30.21**	0.0762	0.0771	**47.7**	-1.59	0.1181	0.1201	94.4	0.33	0.1063	0.1091	95.1
β_2	3	**-27.80**	0.0812	0.0835	**0.0**	2.64	0.3197	0.3282	93.5	1.17	0.2232	0.2287	95.5
β_3	0.6	**-29.71**	0.0914	0.0924	**49.7**	3.41	0.1426	0.1467	94.2	1.31	0.1290	0.1314	94.9
β_4	-1	**29.30**	0.0413	0.0427	**0.0**	-2.97	0.1261	0.128	94.4	-0.95	0.0809	0.0847	93.9
CV(%)		100				94.9				100			

Note. A bold number is either a significant bias (bias>10%) or a bad coverage rate (CR< 90%). LG represents the logistic regression with no misclassification parameter, LG_{FPFN} is the logistic regression model with both false positive and false negative parameters, and LG_E is the model with equal false positive and false negative parameters. The CR and CV denote the coverage rates and convergence rates respectively.

Table 4

Analysis of data from the model with $r_0 = 0.05, r_1 = 0.10$.

		LG				LG_{FPFN}			
Par	True	bias(%)	a.se	e.se	CR(%)	bias(%)	a.se	e.se	CR(%)
		$n = 1,000$							
r_0	0.05	-	-	-	-	4.44	0.0202	0.0209	93.04
r_1	0.1	-	-	-	-	**13.59**	0.17	0.1794	**82.20**
β_0	-3.5	**26.44**	0.2361	0.2437	**5.1**	**-11.20**	0.7808	0.8709	96.92
β_1	-0.5	**31.32**	0.172	0.1748	**84.2**	**-19.72**	0.3203	0.358	95.45
β_2	3	**-30.32**	0.1829	0.1908	**1.0**	**16.13**	0.8945	1.0136	95.45
β_3	0.6	**-31.32**	0.2068	0.2102	**83.8**	**19.42**	0.3818	0.4052	96.52
β_4	-1	**32.13**	0.0926	0.0952	**9.8**	**-18.36**	0.3406	0.3709	97.05
CV(%)		100				74.7			
		$n = 2,000$							
r_0	0.05	-	-	-	-	0.49	0.015	0.0156	93.59
r_1	0.1	-	-	-	-	**-12.91**	0.1273	0.1392	**87.18**
β_0	-3.5	**26.54**	0.1663	0.1747	**0.1**	-4.26	0.4809	0.4999	95.39
β_1	-0.5	**32.68**	0.1213	0.1246	**71.4**	-4.44	0.2035	0.2125	94.49
β_2	3	**-30.46**	0.1288	0.1303	**0.0**	4.91	0.5603	0.5820	94.26
β_3	0.6	**-31.50**	0.1458	0.1419	**74.6**	7.00	0.2452	0.2443	95.73
β_4	-1	**32.69**	0.0651	0.0665	**0.3**	-5.28	0.2186	0.2294	94.15
CV(%)		100				88.9			
		$n = 5,000$							
r_0	0.05	-	-	-	-	-0.06	0.0095	0.0095	94.32
r_1	0.1	-	-	-	-	-8.31	0.0805	0.0827	91.43
β_0	-3.5	**26.78**	0.1048	0.1054	**0.0**	-1.59	0.2909	0.2903	95.56
β_1	-0.5	**31.43**	0.0766	0.0763	**44.7**	-3.47	0.1254	0.1273	95.05
β_2	3	**-30.48**	0.0813	0.0806	**0.0**	1.68	0.3438	0.3507	93.70
β_3	0.6	**-32.47**	0.0918	0.0895	**43.7**	2.62	0.1501	0.1481	96.28
β_4	-1	**32.40**	0.0411	0.0423	**0.0**	-1.95	0.1353	0.1381	94.01
CV(%)		100				96.9			

Note. A bold number is either a significant bias (bias>10%) or a bad coverage rate (CR< 90%). LG represents the logistic regression with no misclassification parameter and LG_{FPFN} is the logistic regression model with both false positive and false negative parameters. The CR and CV denote the coverage rates and convergence rates respectively.

Table 5

Analysis of data from the model with $r_0 = 0.10, r_1 = 0.05$

		LG				LG_{FPFN}			
Par	True	bias(%)	a.se	e.se	CR(%)	bias(%)	a.se	e.se	CR(%)
		$n = 1000$							
r_0	0.1	-	-	-	-	1.69	0.025	0.026	92.14
r_1	0.05	-	-	-	-	**72.07**	0.1539	0.1573	**78.78**
β_0	-3.5	**43.13**	0.2050	0.2138	**0.0**	**-13.73**	0.9133	1.1235	96.59
β_1	-0.5	**42.57**	0.1574	0.1559	**72.2**	**-22.04**	0.3461	0.4406	94.51
β_2	3	**-40.89**	0.1612	0.1695	**0.0**	**20.69**	1.0082	1.2644	96.59
β_3	0.6	**-44.76**	0.1873	0.1991	**68.4**	**20.80**	0.4130	0.4874	94.66
β_4	-1	**43.34**	0.0832	0.0872	**0.2**	**-23.12**	0.3783	0.4938	95.55
CV(%)			100				67.4		
		$n = 2,000$							
r_0	0.1	-	-	-	-	1.74	0.0186	0.0439	93.24
r_1	0.05	-	-	-	-	**13.63**	0.1140	0.1238	**85.28**
β_0	-3.5	**43.16**	0.1443	0.1453	**0.0**	-4.65	0.5519	0.6522	94.21
β_1	-0.5	**43.19**	0.1109	0.1097	**51.2**	-9.12	0.2165	0.2316	95.42
β_2	3	**-41.26**	0.1136	0.1121	**0.0**	6.63	0.6135	0.6969	94.45
β_3	0.6	**-43.76**	0.1319	0.1293	**48.4**	8.85	0.2597	0.2701	95.17
β_4	-1	**43.11**	0.0587	0.0564	**0.0**	-7.76	0.2344	0.2543	95.17
CV(%)			100				82.9		
		$n = 5,000$							
r_0	0.1	-	-	-	-	-0.50	0.0117	0.0119	94.54
r_1	0.05	-	-	-	-	-8.24	0.0735	0.0765	90.90
β_0	-3.5	**43.09**	0.0911	0.093	**0.0**	-1.51	0.3283	0.3369	94.75
β_1	-0.5	**42.82**	0.0701	0.0666	**12.2**	-2.32	0.1292	0.1273	95.29
β_2	3	**-41.07**	0.0718	0.0718	**0.0**	1.87	0.367	0.3683	94.43
β_3	0.6	**-43.69**	0.0832	0.0827	**11.5**	2.47	0.1556	0.1601	94.33
β_4	-1	**43.18**	0.0371	0.0385	**0.0**	-2.20	0.1409	0.1421	93.58
CV(%)			100				93.4		

Note. A bold number is either a significant bias (bias>10%) or a bad coverage rate (CR< 90%). LG represents the logistic regression with no misclassification parameter and LG_{FPFN} is the logistic regression model with both false positive and false negative parameters. The CR and CV denote the coverage rates and convergence rates respectively.

Table 6

Analysis of data from the model with only false positive misclassification: $r_0 = 0.10, r_1 = 0$

		LG				LG_{FPFN}				LG_{FP}			
Par	True	bias(%)	a.se	e.se	CR(%)	bias(%)	a.se	e.se	CR(%)	bias(%)	a.se	e.se	CR(%)
		$n = 1,000$											
r_0	0.10	-	-	-	-	1.90	0.0248	0.0249	94.5	-2.54	0.0249	0.0247	95.6
r_1	0	-	-	-	-	4.74	0.1323	0.1238	**86.9**	-	-	-	-
β_0	-3.5	**42.75**	0.2051	0.2112	**0.0**	**-11.12**	0.8181	0.8761	95.5	-3.49	0.7532	0.7118	96.2
β_1	-0.5	**39.25**	0.1576	0.1576	**76.3**	**-19.40**	0.3129	0.3348	96.0	-2.03	0.2658	0.2437	97.2
β_2	3	**-37.75**	0.1621	0.162	**0.0**	**16.82**	0.8814	0.9323	96.8	3.87	0.6171	0.5901	96.3
β_3	0.6	**-40.87**	0.1872	0.187	**73.8**	**21.42**	0.3759	0.4000	95.0	1.30	0.3244	0.3023	96.1
β_4	-1	**39.96**	0.0838	0.0869	**0.6**	**-18.12**	0.3309	0.3524	96.5	-1.46	0.1848	0.1785	95.9
CV(%)		100				60.1				99.4			
		$n = 2,000$											
r_0	0.10	-	-	-	-	1.71	0.0176	0.0178	94.2	-1.33	0.0175	0.0170	95.6
r_1	0	-	-	-	-	3.07	0.0922	0.0886	**85.8**	-	-	-	-
β_0	-3.5	**43.06**	0.1445	0.1466	**0.0**	-5.21	0.5252	0.5490	96.0	-1.44	0.4809	0.4642	96.6
β_1	-0.5	**39.41**	0.1111	0.1099	**55.1**	**-10.92**	0.2040	0.1997	97.0	-1.86	0.1867	0.1692	97.5
β_2	3	**-37.94**	0.1142	0.1155	**0.0**	8.88	0.5676	0.5608	96.2	1.94	0.3852	0.378	96.7
β_3	0.6	**-41.28**	0.132	0.1317	**53.5**	7.84	0.2444	0.2489	95.3	-0.57	0.2275	0.2090	96.7
β_4	-1	39.79	0.0591	0.0629	**0.0**	-9.39	0.2156	0.2026	96.3	-1.00	0.1301	0.1186	97.0
CV(%)		100				70.4				100			
		$n = 5,000$											
r_0	0.10	-	-	-	-	-0.04	0.0115	0.012	93.8	-0.95	0.0111	0.0107	96.4
r_1	0	-	-	-	-	0.41	0.0611	0.0701	**83.9**	-	-	-	-
β_0	-3.5	**43.26**	0.091	0.0929	**0.0**	-1.43	0.3089	0.3157	96.2	-0.19	0.2971	0.2819	96.2
β_1	-0.5	**39.78**	0.0701	0.069	**19.2**	-2.66	0.1204	0.1218	95.2	0.04	0.1173	0.1092	97.2
β_2	3	**-38.26**	0.0720	0.0740	**0.0**	2.11	0.3309	0.3373	94.9	0.21	0.2358	0.2235	96.8
β_3	0.6	**-41.24**	0.0832	0.0826	**15.6**	2.65	0.1455	0.1474	95.0	0.10	0.1430	0.1323	96.6
β_4	-1	**39.89**	0.0373	0.0382	**0.0**	-2.68	0.1272	0.1315	93.8	-0.16	0.0818	0.0775	96.0
CV(%)		100				81.6				100			

Note. *A bold number is either a significant bias (bias>10%) or a bad coverage rate (CR< 90%). LG represents the logistic regression with no misclassification parameter, LG_{FPFN} is the logistic regression model with both false positive and false negative parameters, and LG_{FP} is the logistic model with the false positive parameter. The CR and CV denote the coverage rates and convergence rates respectively.*

Table 7

Analysis of data from the model with only false negative misclassification: $r_0 = 0.00, r_1 = 0.10$.

Par	True	LG bias(%)	LG a.se	LG e.se	LG CR(%)	LG_{FPFN} bias(%)	LG_{FPFN} a.se	LG_{FPFN} e.se	LG_{FPFN} CR(%)	LG_{FN} bias(%)	LG_{FN} a.se	LG_{FN} e.se	LG_{FN} CR(%)
		$n = 1,000$											
r_1	0	-	-	-	-	1.04	0.0103	0.0625	92.81	-	-	-	-
r_1	0.1	-	-	-	-	**39.31**	0.1365	0.145	**83.21**	-9.98	0.1389	0.1467	**85.8**
β_0	-3.5	-2.01	0.3021	0.2997	95.9	**-10.25**	0.5403	0.6827	95.68	-1.87	0.3363	0.3372	96.2
β_1	-0.5	8.37	0.1962	0.1994	93.4	**-15.37**	0.2711	0.278	96.16	-2.89	0.2267	0.2339	94.0
β_2	3	-4.50	0.2367	0.2371	**90.0**	**15.24**	0.6376	0.7216	96.16	2.99	0.3671	0.3906	94.9
β_3	0.6	-5.92	0.2386	0.2376	94.5	**18.81**	0.3255	0.3681	95.44	4.88	0.2724	0.2784	95.8
β_4	-1	6.88	0.1110	0.1083	**89.3**	**-18.57**	0.2684	0.2932	94.48	-4.23	0.1760	0.1836	94.8
CV(%)		100				41.7				88.7			
		$n = 2,000$											
r_0	0	-	-	-	-	0.60	0.0069	0.0477	**88.61**	-	-	-	-
r_1	0.1	-	-	-	-	**20.11**	0.1066	0.1125	**87.72**	-6.86	0.0999	0.108	**89.7**
β_0	-3.5	-1.86	0.2127	0.2097	95.8	-5.20	0.3417	0.466	94.48	-1.34	0.2329	0.2368	95.6
β_1	-0.5	9.26	0.1383	0.1426	92.8	-5.85	0.1779	0.1908	95.37	0.46	0.1577	0.1669	94.0
β_2	3	-4.68	0.1666	0.1613	**86.0**	7.64	0.414	0.4869	94.84	1.67	0.2559	0.2651	94.0
β_3	0.6	-6.44	0.1681	0.1676	94.8	7.92	0.2135	0.2254	95.73	2.14	0.1896	0.1931	95.2
β_4	-1	7.81	0.0782	0.0785	**82.9**	-7.95	0.1754	0.1986	94.31	-1.61	0.123	0.1300	92.9
CV(%)		100				56.2				94.9			
		$n = 5,000$											
r_0	0	-	-	-	-	0.37	0.0042	0.0489	**82.88**	-	-	-	-
r_1	0.1	-	-	-	-	7.91	0.0695	0.0878	90.16	-5.34	0.0638	0.0694	92.8
β_0	-3.5	-1.31	0.1337	0.1298	94.6	-1.28	0.1947	0.4094	93.80	-0.53	0.1449	0.142	96.2
β_1	-0.5	8.17	0.0873	0.0859	93.1	-1.74	0.1062	0.1171	94.74	-0.04	0.0987	0.0973	94.7
β_2	3	-5.09	0.1048	0.103	**65.8**	1.99	0.2382	0.3904	94.47	0.51	0.1599	0.1629	94.7
β_3	0.6	-7.37	0.1059	0.1066	**92.7**	3.20	0.1277	0.1441	94.07	0.60	0.1184	0.1196	94.7
β_4	-1	7.47	0.0493	0.0514	**65.6**	-2.83	0.1035	0.1488	94.61	-0.99	0.0772	0.0812	94.9
CV(%)		100				74.2				99.2			

Note. *A bold number is either a significant bias (bias>10%) or a bad coverage rate (CR< 90%). LG represents the logistic regression model with no misclassification parameter, LG_{FPFN} is the logistic regression model with both false positive and false negative parameters, and LG_{FN} is the logistic model with the false negative parameter. The CR and CV denote the coverage rates and convergence rates respectively.*

Table 8
Analysis of the NLSY1997 data

(a) *Parameter estimates. Gender: 0, boy; 1, girl. Smoke: 0, not smoking cigarettes; 1, smoking cigarettes. Residence: 0, urban; 1, rural. Peer: the higher score, the healthier their peers lived. AIC represents the Akaike information criterion and BIC is the short form of Bayesian information criterion. The model with smaller AIC and BIC is preferred in general. FP and FN mean the false positive and false negative rates. LG, LG_{FPFN}, and LG_{FN} are the models with no misclassification parameter, with both false positive and false negative misclassificaiton parameters, and with only false negative parameter, respectively.*

	LG			LG_{FPFN}			LG_{FN}		
Par	est	s.e	p(>|z|)	est	s.e	p(>|z|)	est	s.e	p(>|z|)
FP	-	-	-	-0.0017	0.0031	0.5826	-	-	-
FN	-	-	-	0.1816	0.0478	< 0.001	0.1947	0.0392	< 0.001
intercept	-3.5914	0.1370	< 0.001	-3.500	0.2045	< 0.001	-3.5753	0.1613	< 0.001
gender	-0.4582	0.0870	< 0.001	-0.5837	0.1181	< 0.001	-0.6139	0.1140	< 0.001
smoke	2.8980	0.1097	< 0.001	3.1976	0.2395	< 0.001	3.2975	0.1667	< 0.001
residence	0.4270	0.1021	< 0.001	0.5549	0.1315	< 0.001	0.5822	0.1300	< 0.001
peer	-0.9384	0.0471	< 0.001	-1.1462	0.1136	< 0.001	-1.1888	0.0862	< 0.001
-2log$L$	3464.896			3450.604			3451.062		
#pars	5			7			6		
AIC	3474.896			3464.605			3463.063		
BIC	3507.866			3510.763			3502.626		

(b) *Model comparison*

Comparison		Test summary		
M_0	M_1	χ^2 statistic	df	p.value
LG	LG_{FPFN}	14.29	2	0.0008
LG	LG_{FN}	13.83	1	0.0002
LG_{FN}	LG_{FPFN}	0.458	1	0.4986

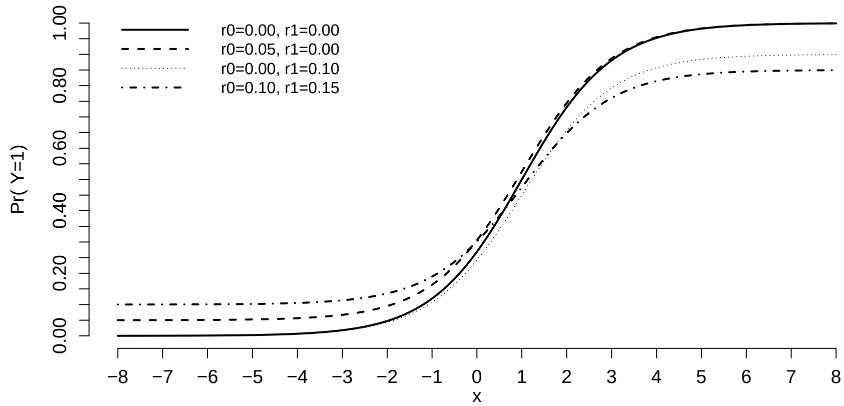

Figure 1. Plot of the conditional probability with one predictor:
$Pr(Y = 1|X = x) = r_0 + (1 - r_0 - r_1)/(1 + exp(-x + 1))$

```
#-----Input-----#
data(nlsy)
head(nlsy)
y=nlsy[, 1]
x=nlsy[, -1]
logistic4p(x,y, model="fp.fn")


#-----Output-----#


The algorithm converged in  299  iterations.
LogLikelihood =  -1725.302
AIC =  3464.605    BIC=  3510.763
Parameter estimates:
            Estimates   Std.Error     z.value     Pr(>|z|)
FP        -0.001698437 0.003090713  -0.5495293 5.826423e-01
FN         0.181610754 0.047847397   3.7956246 1.472722e-04
Intercept -3.499980460 0.204547275 -17.1108633 0.000000e+00
gender    -0.583676758 0.118093479  -4.9424978 7.712798e-07
smoke      3.197646524 0.239526685  13.3498551 0.000000e+00
residence  0.554866523 0.131470417   4.2204667 2.437970e-05
peer      -1.146240383 0.113631343 -10.0873610 0.000000e+00
```

Figure 2. R input and output for the logistic regression model with both false positive and false negative parameters