# Analysis of Mixed Outcomes: Misclassified Binary Responses and Measurement Error in Covariates

**Surupa Roy**
**Tathagata Banerjee**

**W.P. No.2007-01-08**
January 2007



**INDIAN INSTITUTE OF MANAGEMENT**
**AHMEDABAD-380 015**
**INDIA**

# Analysis of Mixed Outcomes: Misclassified Binary Responses and Measurement Error in Covariates

Surupa Roy
St. Xavier's College, Kolkata

Tathagata Banerjee*
Indian Institute of Management, Ahmedabad

**Abstract**

*The focus of this paper is on regression models for mixed binary and continuous outcomes, when the true predictor is measured with error and the binary responses are subject to classification errors. Latent variable is used to model the binary response. The joint distribution is expressed as a product of the marginal distribution of the continuous response and the conditional distribution of the binary response given the continuous response. Models are proposed to incorporate the measurement error and/or classification errors. Likelihood based analysis is performed to estimate the regression parameters of interest. Theoretical studies are made to find the bias of the likelihood estimates of the model parameters. An extensive simulation study is carried out to investigate the effect of ignoring classification errors and/or measurement error on the estimates of the model parameters. The methodology is illustrated with a data set obtained by conducting a small scale survey.*

Keywords: mixed binary-continuous outcomes; classification errors; Berkson model; Maximum likelihood estimate; Misspecified model

*The author is on leave from the Department of Statistics, Calcutta University, India
*Address for correspondence:* Tathagata Banerjee, Wing 14, Indian Institute of Management Ahmedabad, Ahmedabad 380015, India

# Analysis of Mixed Outcomes: Misclassified Binary Responses and Measurement Error in Covariates

## 1. Introduction

Regression models with mixed binary and continuous responses naturally arise in many applied settings. The models find extensive applications in analyzing data arising out in developmental toxicity studies (Catalano and Ryan (1992), Fitzmaurice and Laird (1995), Regan and Catalano (1999, 2000), Geys et al. (2001) and Gueorguieva and Agresti (2001)). The primary impediment to modeling mixed binary continuous outcomes is: no natural choice of a multivariate distribution for modeling such data exists. The model by Olkin and Tate (1961) is the earliest one which considers the factorization of the joint distribution into binary marginal and continuous conditional. Cox (1972), on the other hand, arrives at a model considering the factorization in the reverse sequence viz., continuous marginal and binary conditional. Sammel et al. (1997) subsequently consider a multivariate mixed outcomes model assuming component responses to be independent observations from one parameter exponential families conditional on a common latent variable. Gueorguieva and Agresti (2001) recently consider a correlated probit model that considers an underlying normal latent variable for the binary response. Finally Gueorguieva and Sanacora (2006) extend it for analyzing longitudinal mixed outcome data.

In this paper our primary interest is related to the application of the Cox (1972) model in analyzing data contaminated with measurement error in covariates and/or classification errors in binary responses. In epidemiologic studies, often for some reason, the predictors are not directly observable instead its surrogates are observable though the model is determined by the true predictors. In such cases usually the true predictor is modeled as a linear function of the surrogates plus an error. In measurement error literature such models are usually called the Berkson model (pp.9, Carroll et al. (1995)). On top of it, it may happen that the binary responses recorded may be subject to classification errors. For example, it could be interesting to analyze the data, if available, on the survivors of atomic bomb explosions in Hiroshima and Nagasaki who died after 1945. The continuous response may be the log number of years of survival of a person after his/her exposure to radiations from the explosion and the binary response may be whether he or she died of

cancer or not. One of the important covariates is a measure of exposure to radiation at the time of explosion. The amount of radiation exposure is not observable but one can use the estimated dose using DS86 dosimetry (Roesch 1987, Fujita 1989) as the surrogate. Also the cause of death viz., cancer or not, may be misclassified (Sposto et al.(1992)). The binary regression modeling when the responses (death from cancer or not) are subject to classification errors and covariates (exposure to radiation) are subject to measurement error is considered by Roy et al. (2005).

Surprisingly, however, regression problem with mixed outcomes is not considered in the measurement error literature. And thus the effect on the estimates of the model parameters of misclassification errors in the binary outcome and measurement error in covariates are not known. The problem that we consider here seems to be new and of considerable importance in the epidemiologic studies. To be more specific, the questions that we address here are: in a regression set-up with mixed outcomes how the likelihood estimates of the model parameters would be affected if we consider a naïve model i.e., a model that assumes the surrogates as the true predictors and ignores the presence of classification errors? Which of these errors is more serious? How the proposed models that incorporate these errors would perform compared to the naïve model? We also present some interesting theoretical results that provide strong insight in understanding the effects of these errors on the parameter estimates and also partial answers to the above questions. The proofs of a few others still elude us. We cite them as open problems. Extensive simulation studies that we present at the end support our theoretical findings besides helping us to understand the joint effect of these errors on the estimates of the model parameters.

Regarding the presentation, first we introduce the naïve model (Cox (1972), Catalano and Ryan (1992)) in Section 2. In Sections 3-5, we propose its modifications in the presence of classification errors, measurement errors and lastly, in the presence of both respectively. The Subsections in Sections 2-5 discuss the parameter estimation and some theoretical results regarding the effect of measurement and/or classification errors on the estimates of the model parameters. The results of an extensive simulation study investigating the sensitivity of the estimates of the model parameters to different choices of classification errors and measurement error parameters are presented in Section 6. In

Section 7 analysis of a data set collected by conducting a small scale survey is given. Finally conclusions are drawn in Section 8.

## 2. Naive Model Analysis

*2.1. Model*

Suppose $y_i = (y_{1i}, y_{2i})^T$, i =1,2,…n, denote the bivariate responses where $y_{1i}$ is binary and $y_{2i}$ is continuous. Let $y_{1i}^*$ be the unobserved latent variable such that

$$\begin{aligned} y_{1i} &= 1, \text{ if } y_{1i}^* > 0, \\ &= 0, \text{ if } y_{1i}^* \le 0. \end{aligned} \tag{2.1}$$

Associated with the i[th] observation there is a $p_1 \times 1$ covariate vector $x_{1i}$ thought to predict $y_{1i}^*$ (and hence $y_{1i}$) and a $p_2 \times 1$ covariate vector $x_{2i}$ thought to predict $y_{2i}$. The following bivariate model is considered,

$$\begin{aligned} y_{1i}^* &= \beta_{01} + \beta_1^T x_{1i} + \varepsilon_{1i}, \\ y_{2i} &= \beta_{02} + \beta_2^T x_{2i} + \varepsilon_{2i}, \end{aligned} \tag{2.2}$$

where the joint distribution of $\varepsilon_{1i}$ and $\varepsilon_{2i}$ is normal with zero means, correlation coefficient ρ and variances unity and $\sigma_2^2$ respectively. For the model to be identifiable $V(\varepsilon_{1i}) = 1$ is a standard assumption (Cox, 1972; Catalano and Ryan, 1992). Further $(\varepsilon_{1i}, \varepsilon_{2i})$ 's are independent and are independent of $x_i = (x_{1i}^T, x_{2i}^T)^T$. Thus the joint distribution of $y_{1i}^*$ and $y_{2i}$ is given by,

$$y_{1i}^*, y_{2i} \,|\, x_i \sim N_2(\beta_{01} + \beta_1^T x_{1i}, \beta_{02} + \beta_2^T x_{2i}, 1, \sigma_2^2, \rho). \tag{2.3}$$

Now, the joint distribution of $(y_{1i}, y_{2i})$ given the true predictor can be written as

$$f(y_{1i}, y_{2i} / x_i) = f(y_{1i} / y_{2i}, x_i) f(y_{2i} / x_{2i}) \tag{2.4}$$

where $f(y_{2i} / x_{2i})$ and $f(y_{1i} / y_{2i}, x_i)$ are the marginal distribution of $y_{2i}$ and the conditional distribution of $y_{1i}$ given $y_{2i}$ respectively. From (2.3) it follows that

$$\pi_{1i} = P(y_{1i} = 1 / y_{2i}, x_i) = P(y_{1i}^* \ge 0 / y_{2i}, x_i) = \Phi\left(\frac{\mu_{1i}}{\sqrt{1-\rho^2}}\right) \tag{2.5}$$

where,

$$\mu_{1i} = \beta_{01} + \beta_1^T x_{1i} + \frac{\rho}{\sigma_2}(y_{2i} - \beta_{02} - \beta_2^T x_{2i}). \tag{2.6}$$

It also follows from (2.5) that if ρ=0 then $P(y_{1i}=1|y_{2i},x_i)$ becomes independent of $y_{2i}$. In general the joint distribution of $(y_{1i},y_{2i})$ can be written as

$$f(y_{1i},y_{2i}/x_i)=\pi_{1i}^{\,y_{1i}}(1-\pi_{1i})^{1-y_{1i}}\frac{1}{\sigma_2\sqrt{2\pi}}\exp\left(-\frac{1}{2\sigma_2^2}(y_{2i}-\beta_{02}-\beta_2^T x_{2i})^2\right), \qquad (2.7)$$

where $\pi_{1i}$ is given by (2.5).

*2.2. Parameter Estimation*

Let the parameter of interest be denoted by $\theta=(\theta_1^T,\theta_2^T)^T$, where $\theta_1=(\beta_{01},\beta_1^T,\rho)^T$ and $\theta_2=(\beta_{02},\beta_2^T,\sigma_2^2)^T$. The log likelihood function is given by

$$L_1(\theta_1,\theta_2)=L_{11}(\theta_1,\theta_2)+L_{12}(\theta_2), \qquad (2.8)$$

where,

$$L_{11}(\theta_1,\theta_2)=\sum_{i=1}^{n}y_{1i}\ln\pi_{1i}+\sum_{i=1}^{n}(1-y_{1i})\ln(1-\pi_{1i}), \qquad (2.9)$$

$$L_{12}(\theta_2)=-\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln\sigma_2-\frac{1}{2\sigma_2^2}\sum_{i=1}^{n}(y_{2i}-\beta_{02}-\beta_2^T x_{2i})^2. \qquad (2.10)$$

The maximum likelihood estimate (mle) of θ is obtained by solving the following likelihood equations iteratively:

$$\frac{\partial L_{11}(\theta_1,\theta_2)}{\partial\theta_1}=0, \qquad (2.11)$$

$$\frac{\partial L_{11}(\theta_1,\theta_2)}{\partial\theta_2}+\frac{\partial L_{12}(\theta_2)}{\partial\theta_2}=0. \qquad (2.12)$$

Starting with an initial value of $\theta_2$ the equations (2.11)-(2.12) can be solved iteratively until convergence is achieved.

## 3. Model with Classification Errors

*3.1. Model and estimation*

Suppose the true binary response $y_{1i}$ is subject to classification errors and instead of $y_{1i}$, an error prone response $\tilde{y}_{1i}$ is observed. We assume a simple probability model linking the manifest response $\tilde{y}_{1i}$ to the true response $y_{1i}$. This is given by,

$$\begin{aligned}P(\tilde{y}_{1i}=1/y_{1i}=0,y_{2i},x_i)&=P(\tilde{y}_{1i}=1/y_{1i}=0)=\varepsilon_0,\\ P(\tilde{y}_{1i}=0/y_{1i}=1,y_{2i},x_i)&=P(\tilde{y}_{1i}=0/y_{1i}=1)=\varepsilon_1,\end{aligned} \qquad (3.1)$$

where $\varepsilon_0$ and $\varepsilon_1$ are unknown probabilities of misclassification. To keep the treatment simple, the misclassification probabilities are assumed to be independent of the covariate $x_i$ and the continuous response $y_{2i}$. Now, straight forward probability calculation gives,

$$\pi_{2i} = P(\tilde{y}_{1i} = 1 / y_{2i}, x_i) = \varepsilon_0 + (1 - \varepsilon_0 - \varepsilon_1)\Phi(\frac{\mu_{1i}}{\sqrt{1-\rho^2}}) \qquad (3.2)$$

where $\mu_{1i}$ is given by (2.6). Note that the above model is no longer a probit model. Note, ρ=0 entails $P(\tilde{y}_{1i} = 1 / y_{2i}, x_i) = \varepsilon_0 + (1 - \varepsilon_0 - \varepsilon_1)\Phi(\mu_{1i}) = P(\tilde{y}_{1i} = 1 / x_i)$. The effect of classification errors on the estimates of regression parameters in this special case of binary regression has been considered by Roy et al. (2005).

Now, the joint probability distribution of $(\tilde{y}_{1i}, y_{2i})$ given the true predictor is factorized as,

$$\begin{aligned} f(\tilde{y}_{1i}, y_{2i} / x_i) &= P(\tilde{y}_{1i} = 1 / y_{2i}, x_i)^{\tilde{y}_{1i}} (1 - P(\tilde{y}_{1i} = 1 / y_{2i}, x_i))^{1-\tilde{y}_{1i}} f(y_{2i} / x_{2i}) \\ &= \pi_{2i}^{\tilde{y}_{1i}} (1 - \pi_{2i})^{1-\tilde{y}_{1i}} \frac{1}{\sigma_2 \sqrt{2\pi}} \exp\left( -\frac{1}{2\sigma_2^2} (y_{2i} - \beta_{02} - \beta_2^T x_{2i})^2 \right), \end{aligned} \qquad (3.3)$$

with $\pi_{2i}$ as defined in (3.2). The resulting log likelihood function is

$$L_2(\theta_1, \theta_2, \varepsilon_0, \varepsilon_1) = L_{21}(\theta_1, \theta_2, \varepsilon_0, \varepsilon_1) + L_{22}(\theta_2), \qquad (3.4)$$

where,

$$L_{21}(\theta_1, \theta_2, \varepsilon_0, \varepsilon_1) = \sum_{i=1}^{n} \tilde{y}_{1i} \log \pi_{2i} + \sum_{i=1}^{n} (1 - \tilde{y}_{1i}) \log(1 - \pi_{2i}) \qquad (3.5)$$

and $L_{22}(\theta_2)$ is identical to $L_{12}(\theta_2)$.

For finding likelihood estimates we need to solve the likelihood equations simultaneously for $\theta_1$, $\theta_2$, $\varepsilon_0$ and $\varepsilon_1$. However, if the observations are such that most of the $\Phi(\frac{\mu_{1i}}{\sqrt{1-\rho^2}})$'s lie in the central part of the probit function, more specifically between, say, 0.1 and 0.9 then $\varepsilon_0$ and $\varepsilon_1$ become almost confounded with $\theta_1$ (Cox and Snell, 1989, pp.22 ) and thus making separate estimation of $\varepsilon_0, \varepsilon_1$ and $\theta_1$ difficult unless the sample size is very large. In such situations estimation of $\theta_1$ is possible if $\varepsilon_0$ and $\varepsilon_1$ are known or its estimates are available from independent validation studies. In epidemiologic studies, separate estimates of $\varepsilon_0$ and $\varepsilon_1$ are often obtained from external validation data (see Holcroft and Spiegelman (1999), Morrissey and Spiegelman (1999) and other references

therein). Maximum likelihood estimate of $\theta_1$ is then obtained by replacing $\varepsilon_0$ and $\varepsilon_1$ in the log likelihood function by their estimates and treating them as if they are known. The asymptotic distribution of such estimates may be obtained as in Roy et al. (2005).

*3.2. Effect of Classification Errors on the Estimates*

In this section we investigate the effect of ignoring classification errors on the likelihood estimates of the parameters assuming that the classification errors are known. The key result that we use follows from the work of White (1982) on misspecified models. It says the likelihood estimate $\hat{\theta}^* = (\hat{\theta}_1^{*T}, \hat{\theta}_2^{*T})^T$ under the false model converges to $\theta^* = (\theta_1^{*T}, \theta_2^{*T})^T$ that minimizes the Kullback-Leibler divergence (See Kullback (1959)) between the true and the false models. In our case it is given by,

$$\mathrm{E}_x \mathrm{E}_{y_2/x} \mathrm{E}_{\tilde{y}_1/y_2,x} \left[ \log \{ f_T(\tilde{y}_1, y_2 / x) / f_F(\tilde{y}_1, y_2 / x) \} \right], \tag{3.6}$$

where, $f_T(\tilde{y}_1, y_2 / x)$ and $f_F(\tilde{y}_1, y_2 / x)$ are given by (3.3) and (2.7) respectively. Also the expectations are taken with respect to the true model.

Let the parameters under the true and the false models be denoted by $\theta = (\theta_1^T, \theta_2^T)^T$ and $\theta^* = (\theta_1^{*T}, \theta_2^{*T})^T$ respectively. Taking the derivatives of (3.6), we find that $\theta^* = (\theta_1^{*T}, \theta_2^{*T})^T = (\beta_{01}^*, \beta_1^{*T}, \beta_{02}, \beta_2^{*T}, \sigma_2^*, \rho^*)^T$ solve the system of equations given below:

$$\mathrm{E}_x \mathrm{E}_{y_2/x} \left\{ \lambda\left(1|\mathrm{x},\mathrm{y}_2\right) P_F'\left(1|x, y_2\right) + \lambda\left(0|x, y_2\right) P_F'\left(0|x, y_2\right) \right\} = 0 \tag{3.7}$$

$$\mathrm{E}_x \mathrm{E}_{y_2/x} \left[ x_1 \left\{ \lambda\left(1|x, y_2\right) P_F'\left(1|x, y_2\right) + \lambda\left(0|x, y_2\right) P_F'\left(0|x, y_2\right) \right\} \right] = 0 \tag{3.8}$$

$$\mathrm{E}_x \mathrm{E}_{y_2/x} \left[ \left\{ \lambda\left(1|\mathrm{x},\mathrm{y}_2\right) P_F'\left(1|x, y_2\right) + \lambda\left(0|x, y_2\right) P_F'\left(0|x, y_2\right) \right\} \left( \mathrm{y}_2 - \beta_{02}^* - \beta_2^{*T} x_2 \right) \right] = 0 \tag{3.9}$$

$$\begin{aligned} &\mathrm{E}_x \mathrm{E}_{y_2/x} \left[ \left\{ \lambda\left(1|\mathrm{x},\mathrm{y}_2\right) P_F'\left(1|x, y_2\right) + \lambda\left(0|x, y_2\right) P_F'\left(0|x, y_2\right) \right\} \rho^* \sigma_2^{*-1} \right] \\ &+ \mathrm{E}_x \left[ \sigma_2^{*-2} \left\{ \left(\beta_{02} - \beta_{02}^*\right) + \left(\beta_2 - \beta_2^*\right)^T x_2 \right\} \right] = 0 \end{aligned} \tag{3.10}$$

$$\begin{aligned} &\mathrm{E}_x \mathrm{E}_{y_2/x} \left[ x_2 \left\{ \lambda\left(1|x, y_2\right) P_F'\left(1|x, y_2\right) + \lambda\left(0|x, y_2\right) P_F'\left(0|x, y_2\right) \right\} \rho^* \sigma_2^{*-1} \right] + \\ &\mathrm{E}_x \left[ x_2 \sigma_2^{*-2} \left\{ \left(\beta_{02} - \beta_{02}^*\right) + \left(\beta_2 - \beta_2^*\right)^T x_2 \right\} \right] = 0 \end{aligned} \tag{3.11}$$

$$\mathrm{E_x E_{y_2/x}}\left[\left\{\lambda\left(1|x,y_2\right)P_F'\left(1|x,y_2\right)+\lambda\left(0|x,y_2\right)P_F'\left(0|x,y_2\right)\right\}\rho^*\sigma_2^{-*2}(y_2-\beta_{02}^*-\beta_2^{*T}x_2)\right]$$
$$+\mathrm{E_x E_{y_2/x}}\left\{\sigma_2^{*-3}(y_2-\beta_{02}^*-\beta_2^{*T}x_2)^2\right\}-(1/\sigma_2^*)=0 \qquad (3.12)$$

where $\lambda\left(u|x,y_2\right)=\dfrac{P_T(\tilde{y}_1=u/y_2,x)}{P_F(\tilde{y}_1=u/y_2,x)}$ and $P_F'\left(u|x,y_2\right)=\dfrac{\partial}{\partial\beta_{01}^*}P_F\left(\tilde{y}_1=u/x,y_2\right)$; u = 0,1.

For a given distribution of x, one can solve the system (3.7) - (3.12) and compare $\theta^*$ with the true value of θ. In general $\theta^*\neq\theta$, i.e. ignoring the classification errors produces biased estimates. It does not seem possible to find an explicit general solution to the above system of equations and hence a general result on the effects of classification errors on the estimates of the model parameters. However, in the following we find some interesting results in specific cases and then discuss its implications.

To be specific, we consider solving the system of equations (3.7)-(3.12) assuming that $\theta_2^*=\theta_2$. It is interesting to note that this assumption is valid in case the covariates for the mixed outcome responses are same. This is generally the case in all teratological applications (Catalano and Ryan (1992), Fizmaurice and Laird (1995), Regan and Catalano (1999, 2000), Geys et al. (2001), Gueorguieva and Agresti (2001)) as well as in most of the applications related to the epidemiologic studies. For example in the cohort study of the effects of radiation exposures among the survivors of atom bomb explosions in Hiorshima and Nagasaki in Japan the covariates are radiation exposure level besides other demographic characteristics.

In this specific case since $\theta_2^*=\theta_2$ we consider equations (3.7)-(3.9) with $(\beta_{02}^*,\beta_2^*)$ in (3.9) replaced by $(\beta_{02},\beta_2)$. Still a general solution $\theta_1^*=\left(\beta_{01}^*,\beta_1^*,\rho^*\right)^T$ to (3.7)-(3.9) eludes us. However, an approximate relationship can be established between $\theta_1^*$ and $\theta_1$. Note that the values of $(\beta_{01}^*,\beta_1^*,\rho^*)$ that yield $\lambda(1/x,y_2)=\lambda(0/x,y_2)=1$ for all x and $y_2$ solve (3.7)-(3.9). Equations (2.5) and (2.6) entail

$$\rho^*\sigma_2^{-1}\ (1-\beta_2)(1-\rho^{*2})^{-1/2}=\Phi^{-1}\{P_F(\tilde{y}_1=1/y_2+1,x)\}-\Phi^{-1}\{P_F(\tilde{y}_1=1/y_2,x)\}$$

where, x is the common covariate. Replacing the false probabilities above by the true ones we get

$$\rho^{*}\sigma_2^{-1}\ (1-\rho^{*2})^{-1/2} = \Phi^{-1}\{P_T(\tilde{y}_1 = 1/\, y_2 + 1, x)\} - \Phi^{-1}\{P_T(\tilde{y}_1 = 1/\, y_2, x)\}$$

$$= \mathrm{H}_1\left(\frac{\rho}{\sqrt{1-\rho^2}}\right)$$

where,

$$H_1(\frac{\rho}{\sqrt{1-\rho^2}}) = \Phi^{-1}(\varepsilon_0 + (1-\varepsilon_0-\varepsilon_1)\,\Phi(\frac{\beta_0+\beta_1 x_1 + \rho\sigma_2^{-1}(y_2+1-\beta_{02}-\beta_2 x_2)}{\sqrt{1-\rho^2}}))$$

$$-\Phi^{-1}(\varepsilon_0 + (1-\varepsilon_0-\varepsilon_1)\,\Phi(\frac{\beta_0+\beta_1 x_1 + \rho\sigma_2^{-1}(y_2-\beta_{02}-\beta_2 x_2)}{\sqrt{1-\rho^2}}).$$

This entails

$$\frac{\rho^{*}}{\sqrt{1-\rho^{*2}}} = \sigma_2 H_1(\frac{\rho}{\sqrt{1-\rho^2}}).$$

Taking $\mathrm{G}_1(\frac{\rho}{\sqrt{1-\rho^2}}) = \sigma_2 H_1(\frac{\rho}{\sqrt{1-\rho^2}})$ and expanding it about 0 we obtain,

$$\mathrm{G}_1\left(\frac{\rho}{\sqrt{1-\rho^2}}\right) \cong \frac{\rho}{\sqrt{1-\rho^2}}\,\mathrm{G}_1'(0) = \frac{\rho}{\sqrt{1-\rho^2}}\,\frac{(1-\varepsilon_0-\varepsilon_1)\phi(\beta_{01}+\beta_1 x_1)}{\phi[\Phi^{-1}\{\varepsilon_0+(1-\varepsilon_0-\varepsilon_1)\Phi(\beta_{01}+\beta_1 x_1)\}]}$$

for small values of ρ. It is now easy to check that $0 \le \mathrm{G}_1'(0) \le 1$ whatever $x_1$ may be. For proof of this we refer to Neuhaus (1999). For $x_1$=0, the attenuation factor reduces to that of Neuhaus (1999) and Li and Duan (1989). Ignoring classification errors thus results in the attenuation of the numerical value of the estimate of $\frac{\rho}{\sqrt{1-\rho^2}}$. Making use of the fact that $\frac{\rho}{\sqrt{1-\rho^2}}$ is an increasing function of ρ it is easy to see that under the naïve model the numerical value of ρ is attenuated.

To investigate the effect on $\beta_1$, we follow the above logic to arrive at

$$\beta_1^{*} \cong \beta_1 \frac{\sqrt{1-\rho^{*2}}}{\sqrt{1-\rho^2}} \frac{(1-\varepsilon_0-\varepsilon_1)\phi\left(\dfrac{\beta_{01}+\rho\sigma_2^{-1}(y_2-\beta_{02}-\beta_2 x_2}{\sqrt{1-\rho^2}}\right)}{\phi\left[\Phi^{-1}\left\{\varepsilon_0+(1-\varepsilon_0-\varepsilon_1)\Phi\left(\dfrac{\beta_{01}+\rho\sigma_2^{-1}(y_2-\beta_{02}-\beta_2 x_2}{\sqrt{1-\rho^2}}\right)\right\}\right]} \qquad (3.13)$$

In case ρ=0, (3.13) reduces to the result obtained by Neuhaus (1999). Notice here that the first factor on the right hand side of (3.13) is more than unity while the second is less than unity. Thus the two attenuation effects are confounded. Interestingly, however, comparing

the results of the simulation studies presented in Section 8, Tables 4-6 with that given in Table 1 of Roy et al. (2005) we observe that the attenuation in the estimate of $\beta_1$ is more in case $\rho$=0.6 than in $\rho$=0. This shows that for $\rho$=0.6 the effect of attenuation of the second factor is more than compensate for the inflation of the first factor compared to $\rho$=0.

**Note:**

In case the covariates for the two responses are different theoretical results on the effect of the classification errors on the likelihood estimates of the parameters seem to be intractable. We leave this as an open problem.

## 4. Model with Measurement Error

*4.1. Model and estimation*

Without loss of generality we assume that the measurements on $x_i = (x_{1i}^T, x_{2i}^T)^T$ are available only through the recording of an imperfect surrogate $z_i = (z_{1i}^T, z_{2i}^T)^T$, where $z_{1i}$ and $z_{2i}$ are the surrogates for $x_{1i}$ and $x_{2i}$ respectively. We also assume a non-differential measurement error model, i.e., given the true predictors, the surrogates add nothing to the prediction of the response. Mathematically, it means

$$P(y_{1i}=1, y_{2i} \le t / x_i, z_i) = P(y_{1i}=1, y_{2i} \le t / x_i). \quad (4.1)$$

Since we consider Berkson model, the measurement error distribution is modeled by the conditional distribution of the true predictors given its surrogates. In particular, we assume that

$$(x_{1i}, x_{2i}) \big| (z_{1i}, z_{2i}) \sim N_p \left\{ (z_{1i}, z_{2i}), \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{pmatrix} \right\} \quad (4.2)$$

where, $p=p_1+p_2$ and $\Sigma$ is completely known from external validation studies (Carroll et al., 1995). With the strength of the assumptions (2.3) and (4.2) we have the following latent variable formulation.

$$(y_{1i}^*, y_{2i}) / z_i \sim N_2(\beta_{01} + \beta_1^T z_{1i}, \beta_{02} + \beta_2^T z_{2i}, 1 + \beta_1^T \Sigma_{11} \beta_1, \sigma_2^2 + \beta_2^T \Sigma_{22} \beta_2, \rho^*)$$

$$\text{where } \rho^* = \frac{\rho\sigma_2 + \beta_1^T \Sigma_{12} \beta_2}{\sqrt{1 + \beta_1^T \Sigma_{11} \beta_1}\sqrt{\sigma_2^2 + \beta_2^T \Sigma_{22} \beta_2}}. \quad (4.3)$$

As an implication of the above result it follows that the marginal distribution of the continuous variable $y_{2i}$ given $z_{2i}$ is normal with mean $\beta_{02} + \beta_2^T z_{2i}$ and variance

$$\sigma_2^{*2} = \sigma_2^2 + \beta_2^T \Sigma_{22} \beta_2 . \tag{4.4}$$

Moreover the conditional distribution of $y_{1i}^*$ given $y_{2i}$ and $z_i$ is again normal with mean

$$\mu_{1i}^* = \beta_{01} + \beta_1^T z_{1i} + \frac{\rho^* \sqrt{1 + \beta_1^T \Sigma_{11} \beta_1}}{\sqrt{\sigma_2^2 + \beta_2^T \Sigma_{22} \beta_2}} (y_{2i} - \beta_{02} - \beta_2^T z_{2i}) \tag{4.5}$$

and variance $(1 + \beta_1^T \Sigma_{11} \beta_1)(1 - \rho^{*2})$ where $\rho^*$ is given by (4.3). Thus we obtain,

$$\begin{aligned} \pi_{3i} &= P(y_{1i} = 1 / y_{2i}, z_i) = P(y_{1i}^* \geq 0 / y_{2i}, z_i) = \Phi\left(\frac{\mu_{1i}^*}{\sqrt{1 + \beta_1^T \Sigma_{11} \beta_1}\sqrt{1 - \rho^{*2}}}\right) \\ &= \Phi\left(\left(\frac{\gamma_{01} + \gamma_1^T z_{1i}}{\sqrt{1 - \rho^{*2}}}\right) + \frac{\rho^*}{\sigma_2^* \sqrt{1 - \rho^{*2}}} (y_{2i} - \beta_{02} - \beta_2^T z_{2i})\right) \end{aligned} \tag{4.6}$$

Where,

$$\gamma_{01} = \frac{\beta_{01}}{\sqrt{1 + \beta_1^T \Sigma_{11} \beta_1}}, \quad \gamma_1^T = \frac{\beta_1^T}{\sqrt{1 + \beta_1^T \Sigma_{11} \beta_1}} . \tag{4.7}$$

Now, using factorization representation of the joint distribution of the mixed outcome we obtain,

$$f(y_{1i}, y_{2i} / z_i) = \pi_{3i}^{\,y_{1i}} (1 - \pi_{3i})^{1 - y_{1i}} \frac{1}{\sigma_2^* \sqrt{2\pi}} \exp\left( -\frac{1}{2} \frac{(y_{2i} - \beta_{02} - \beta_2^T z_{2i})^2}{\sigma_2^{*2}} \right), \tag{4.8}$$

where, $\sigma_2^{*2}$ and $\pi_{3i}$ are given by (4.4) and (4.6) respectively.

**Note:**

It is interesting to note from the above that even when ρ=0 the conditional distribution of $y_{1i}$ given $y_{2i}$ still depends on $y_{2i}$ which should not be the case if the covariates were directly observable. This happens because unlike the presence of classification errors ρ=0 does not entail $\rho^*$ to be zero in this case. Hence unlike the previous two models discussed in Sections 2 and 3, the conditional probit model in this case does not reduce to the unconditional probit model.

Maximum likelihood estimating equations are identical to (2.11) - (2.12) with $\beta_{01}, \beta_1, \sigma_2$ and $\rho$ replaced by $\gamma_{01}, \gamma_1, \sigma_2^*$ and $\rho^*$ respectively. Once the estimates of $(\gamma_{01}, \gamma_1^T, \beta_{02}, \beta_2^T, \sigma_2^*, \rho^*)^T$ are found we can retrieve the estimates of $(\beta_{01}, \beta_1^T, \beta_{02}, \beta_2^T, \sigma_2, \rho)$ using the relationships given by the equations (4.3), (4.4) and (4.7). Sometimes, it may happen that the estimates of some of the parameters do not exist. Of course, the probabilties of such events go to zero as the sample size increases.

*4.2. Effect of measurement error*

Let us define $\xi_1 = (\gamma_{01}, \gamma_1^T, \rho^*)^T$ and $\xi_2 = (\beta_{02}, \beta_2^T, \sigma_2^{*2})^T$, where $\rho^*, (\gamma_{01}, \gamma_1^T)^T, \sigma_2^{*2}$ are given by (4.3), (4.7) and (4.4) respectively. Note that, the estimates of $\beta_{02}$ and $\beta_2$ remain unaffected by the presence of measurement error which is similar to the case observed in normal linear model set-up for Berkson model.

From equation (4.4) it is to be noted that the maximum likelihood estimate of $\sigma_2^2$, say, $\hat{\sigma}_2^2$ under the measurement error model is related to the naïve estimate, say, $\hat{\sigma}_2^{*2}$ by $\hat{\sigma}_2^2 = \hat{\sigma}_2^{*2} - \hat{\beta}_2^T \Sigma_{22} \hat{\beta}_2$, provided $\hat{\sigma}_2^{*2} > \hat{\beta}_2^T \Sigma_{22} \hat{\beta}_2$, otherwise the maximum likelihood estimate does not exist. It shows that the naïve estimate of $\sigma_2^2$ gets inflated.

Equation (4.7) shows that the estimates of $\beta_{01}$ and $\beta_1$ under the measurement error model are given by $\hat{\beta}_{01} = \dfrac{\hat{\gamma}_{01}}{\sqrt{1 - \hat{\gamma}_1^T \Sigma_{11} \hat{\gamma}_1}}$ and $\hat{\beta}_1 = \dfrac{\hat{\gamma}_1}{\sqrt{1 - \hat{\gamma}_1^T \Sigma_{11} \hat{\gamma}_1}}$, provided $1 > \hat{\gamma}_1^T \Sigma_{11} \hat{\gamma}_1$, otherwise the maximum likelihood estimates do not exist. The equations clearly show that the naïve estimates of $\beta_{01}$ and $\beta_1$ are attenuated.

From equation (4.3) we observe that the estimate of $\rho$, say, $\hat{\rho}$ under the measurement error model and $\hat{\rho}^*$ under the naïve model are related by $\hat{\rho} = \left( \hat{\rho}^* \sqrt{1 + \hat{\beta}_1^T \Sigma_{11} \hat{\beta}_1} \sqrt{\hat{\sigma}_2^2 + \hat{\beta}_2^T \Sigma_{22} \hat{\beta}_2} - \hat{\beta}_1^T \Sigma_{12} \hat{\beta}_2 \right) \hat{\sigma}_2^{-1}$.

Thus, the maximum likelihood estimate of $\rho$ under the measurement error model exists provided

$$\frac{-\hat{\sigma}_2 + \hat{\beta}_1^T \Sigma_{12} \hat{\beta}_2}{\sqrt{1 + \hat{\beta}_1^T \Sigma_{11} \hat{\beta}_1} \sqrt{\hat{\sigma}_2^2 + \hat{\beta}_2^T \Sigma_{22} \hat{\beta}_2}} < \hat{\rho}^* < \frac{\hat{\sigma}_2 + \hat{\beta}_1^T \Sigma_{12} \hat{\beta}_2}{\sqrt{1 + \hat{\beta}_1^T \Sigma_{11} \hat{\beta}_1} \sqrt{\hat{\sigma}_2^2 + \hat{\beta}_2^T \Sigma_{22} \hat{\beta}_2}}.$$

It is also clear from above that the estimate of ρ is affected by the presence of measurement error. However, the effect of measurement error in this case does not show any clear cut pattern. Note $\rho^*$ given by (4.3) can be written as $\rho^* = \frac{\rho\sigma_2 + Cov(\beta_1^T x_{i1}, \beta_2^T x_{i2})}{\sqrt{(1+V(\beta_1^T x_{i1}))(\sigma_2^2 + V(\beta_2^T x_{i2}))}}$. Here, we observe that if either one of $V(\beta_1^T x_{i1})$ and $V(\beta_2^T x_{i2})$ is big enough to offset the contribution of $\rho\sigma_2$ in the numerator, then the value of $\rho^*$ tends to $Corr(\beta_1^T x_{i1}, \beta_2^T x_{i2})$. In case of a scalar and common covariate, say, $x_{1i} = x_{2i} = x_i$ and measurement error variance $\sigma^2$ we have $\rho^* = \frac{\rho\sigma_2 + \beta_1\beta_2\sigma^2}{\sqrt{(1+\beta_1^2\sigma^2)}\sqrt{(\sigma_2^2 + \beta_2^2\sigma^2)}}$. Note that $\sigma^2$=0 implies $\rho^* = \rho$. At the other end $\sigma^2 = \infty$ entails $\rho^* = 1$ or -1 depending on whether $\beta_1\beta_2 > 0$ or $< 0$. However it is easy to see that $\rho^*$ is not necessarily a monotonic function of $\sigma^2$. Thus effect of measurement error on the naive estimate of $\rho$ can be in both directions. This is an interesting observation per se.

## 5. Model with Measurement Error and Classification Errors

Finally the model with the binary responses subject to classification errors and the true covariates subject to measurement errors are considered. In this case $\tilde{y}_{1i}$ acts as the manifest response and $z_i$'s are considered to be the surrogate for the true predictors $x_i$. Now the conditional probability of the event $\tilde{y}_{1i}$=1 given the continuous response $y_{2i}$ and the surrogates $z_i$, is given by

$$\begin{aligned}\pi_{4i} = P(\tilde{y}_{1i} = 1/ y_{2i}, z_i) &= P(\tilde{y}_{1i} = 1/ y_{1i} = 1, y_{2i}, z_i)P(y_{1i} = 1/ y_{2i}, z_i) + \\ &\quad P(\tilde{y}_{1i} = 1/ y_{1i} = 0, y_{2i}, z_i)P(y_{1i} = 0/ y_{2i}, z_i) \qquad (5.1)\\ &= \varepsilon_0 + (1 - \varepsilon_0 - \varepsilon_1)\Phi\left(\frac{\mu_{1i}^*}{\sqrt{1+\beta_1^T\Sigma_{11}\beta_1}\sqrt{1-\rho^{*2}}}\right)\end{aligned}$$

where $\rho^*$ and $\mu_{1i}^*$ are given by (4.3) and (4.5) respectively. Thus the joint distribution of the manifest binary response $\tilde{y}_{1i}$ and the continuous response $y_{2i}$ given the surrogate $z_i$ is given by

$$\begin{aligned}&f(\tilde{y}_{1i}, y_{2i} | z_i) = f(\tilde{y}_{1i} | y_{2i}, z_i)\, f(y_{2i} | z_{2i}) \\ &= \pi_{4i}^{\tilde{y}_{1i}}(1-\pi_{4i})^{1-\tilde{y}_{1i}} \frac{1}{\sigma_2^*\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(y_{2i} - \beta_{02} - \beta_2^T z_{2i})^2}{\sigma_2^{*2}}\right) \qquad (5.2)\end{aligned}$$

where $\pi_{4i}$ and $\sigma_2^{*2}$ are given by (5.1) and (4.4) respectively. The estimates of the parameters in this case will be affected in a similar way as discussed in Sections 3 and 4. This model is particularly useful in understanding the joint effect of the errors on the estimates of the model parameters. Interestingly the above two errors may work in opposite directions to cancel out each other's effect on the estimate of ρ. We investigate it further by taking up a simulation study in the next section.

## 6. Simulation Study

An extensive simulation study is carried out to investigate the marginal and the joint effects of measurement error and classification errors on the estimates of the parameters. We consider a common covariate for both the binary and the continuous outcomes. The naïve model obtained from equation (2.7) by replacing $y_{1i}$ by $\tilde{y}_{1i}$ and $x_i$ by $z_i$ is denoted by $M_1$.

The classification error model obtained from equation (3.3) by replacing $x_i$ by $z_i$ is denoted by $M_2$. The measurement error model obtained from equation (5.8) by replacing $y_{1i}$ by $\tilde{y}_{1i}$ is denoted by $M_3$ and finally the model incorporating both the classification errors and measurement error as given by equation (5.2) is denoted by $M_4$. The details of the study are given below.

Step 1: The surrogate $z_i$, $i=1,2,...n$ are generated from uniform(-4, 4), and are kept fixed.

Step2: $x_i$'s are generated from univariate $N(z_i, \sigma^2)$, $i$=1,2,….n for a prefixed value of the measurement error variance $\sigma^2$.

Step3: $(y_{1i}^*, y_{2i})$ i=1,2,…n are generated from $N_2(\beta_{01}+\beta_1 x_i, \beta_{02}+\beta_2 x_i, 1, \sigma_2^2, \rho)$. Here $\beta_{01}=0$, $\beta_1=1.0$, $\beta_{02}=0$, $\beta_2=1.0$ $\sigma_2^2=1.0$ and $\rho=0.6$.

Step4: $y_{1i}$ (i =1,2,…n), are generated as follows:

$$y_{1i} = 1, \text{ if } y_{1i}^* \geq 0$$
$$= 0, \text{ otherwise.}$$

Step5: $\tilde{y}_{1i}$'s (i=1,2,…,n) are generated from $y_{1i}$ using $P(\tilde{y}_{1i}=1 | y_{1i}=0)=\varepsilon_0$ and $P(\tilde{y}_{1i}=0 | y_{1i}=1)=\varepsilon_1$ where ($\varepsilon_0$ , $\varepsilon_1$) are prefixed numbers.

Step 6: Given the data ($\tilde{y}_{1i}$, $y_{2i}$, $z_i$, i=1, 2,..,n) the likelihood estimates are obtained under models $M_1$-$M_4$, by solving the likelihood equations.

Step 7: Steps 2-6 are repeated a large number of times and the estimates $\hat{\theta}_{(l)} = (\hat{\theta}_{1(l)}, \hat{\theta}_{2(l)}) = (\hat{\beta}_{01(l)}, \hat{\beta}_{1(l)}, \hat{\beta}_{02(l)}, \hat{\beta}_{2(l)}, \hat{\sigma}^2_{2(l)}, \hat{\rho}_{(l)})$ and $\hat{\varepsilon}_{0(l)}, \hat{\varepsilon}_{1(l)}$ are obtained (l=1,2,…R). Also the standard errors of $\hat{\theta}_{1(l)}, \hat{\theta}_{2(l)}, \hat{\varepsilon}_{0(l)}$ and $\hat{\varepsilon}_{1(l)}$ are obtained from the inverse of Fisher Information matrix.

Step 8: The average of $\hat{\theta}_{(l)}$'s $\hat{\varepsilon}_{0(l)}$'s and $\hat{\varepsilon}_{1(l)}$'s (l=1,2,…R) i.e. $\bar{\theta}$ , $\bar{\varepsilon}_0$ and $\bar{\varepsilon}_1$ are computed and reported in the tables below along with the average of the standard errors (given in parenthesis) obtained from the repeated calculation of Fisher Information matrix . The simulated standard errors given by $\sqrt{\frac{1}{R}\sum_{l=1}^{R}(\hat{\theta}_{(l)} - \bar{\theta})^2}$ , $\sqrt{\frac{1}{R}\sum_{l=1}^{R}(\hat{\varepsilon}_{0(l)} - \bar{\varepsilon}_0)^2}$ and $\sqrt{\frac{1}{R}\sum_{l=1}^{R}(\hat{\varepsilon}_{1(l)} - \bar{\varepsilon}_1)^2}$ are also obtained. However, their values being very close to those obtained from Fisher information matrix they are not reported in the Tables furnished below.

Step 9: Steps 2-8 are repeated for different choices of prefixed $\sigma^2$, $\varepsilon_0$ and $\varepsilon_1$.

Here we have taken R= 500, n=10000. Selection of large sample size is not unjustified in view of the fact that the applications of such models mostly arise in the analysis of epidemiological data where such sample size is common enough. We investigate through simulation studies three different aspects viz., (i) The effect of measurement error and its recovery via model $M_3$; (ii) The effect of classification errors and its recovery via model $M_2$ and (iii) The joint effects of measurement error and classification errors and its recovery via model $M_4$.

**Measurement error:** Tables 1, 2 and 3 describe the effect of measurement error on the estimates of $\theta_1$ and $\theta_2$. Here the misclassification probabilities $\varepsilon_0$ and $\varepsilon_1$ are chosen to be zero. As discussed in Section 4 the estimates of $\beta_{02}$ and $\beta_2$ remain unaffected. The results in Table 1 reveal that for small measurement error variance the effect on the estimates of the parameters $\theta_1$ and $\sigma_2^2$ are negligible. However for large measurement error variance say $\sigma^2 = 1.0$ (see Table 3) the estimate of $\beta_1$ shows appreciable attenuation while those of $\rho$ and $\sigma_2^2$ are overestimated under model $M_1$. Model $M_3$ recovers the point estimates of the affected parameters at the expense of increased standard errors.

Table 1 $(\sigma^2, \varepsilon_0, \varepsilon_1) = (.01, 0, 0)$

| Estimates | $M_1$ | $M_3$ |
|---|---|---|
| $\hat{\beta}_{01}$ | .0004(.0200) | -.0004(.0202) |
| $\hat{\beta}_1$ | .9961(.0187) | 1.0061(.0193) |
| $\hat{\beta}_{02}$ | -.0002(.0100) | -.0002(.0100) |
| $\hat{\beta}_2$ | .9999(.0041) | .9999(.0041) |
| $\hat{\rho}$ | .6040(.0143) | .6000(.0145) |
| $\hat{\sigma}_2^2$ | 1.0101(.0143) | 1.0100 (.0143) |

Table 2 $(\sigma^2, \varepsilon_0, \varepsilon_1) = (.5, 0, 0)$

| Estimates | $M_1$ | $M_3$ |
|---|---|---|
| $\hat{\beta}_{01}$ | .0008(.0173) | .0013(.0260) |
| $\hat{\beta}_1$ | .8165(.0136) | 1.0006(.0409) |
| $\hat{\beta}_{02}$ | -.0001(.0123) | -.0001(.0123) |
| $\hat{\beta}_2$ | .9999(.0050) | .9999(.0050) |
| $\hat{\rho}$ | .7335(.0164) | .6004(.0184) |
| $\hat{\sigma}_2^2$ | 1.5005(.0222) | 1.0005(.0223) |

Table 3 $(\sigma^2, \varepsilon_0, \varepsilon_1) = (1.0, 0, 0)$

| Estimates | $M_1$ | $M_3$ |
|---|---|---|
| $\hat{\beta}_{01}$ | .0007(.0155) | .0014(.0311) |
| $\hat{\beta}_1$ | .7076(.0107) | 1.0001(.0649) |
| $\hat{\beta}_{02}$ | -.0002(.0142) | -.0001(.0142) |
| $\hat{\beta}_2$ | .9999(.0058) | .9999(.0058) |
| $\hat{\rho}$ | .8002(.0082) | .6001(.0198) |
| $\hat{\sigma}_2^2$ | 2.0069(.0294) | 1.0008(.0304) |

**Classification errors:**

Tables 4, 5 and 6 summarize the effect of classification errors on the estimates of the parameters and its recovery via model $M_2$. As noted in Section 4, use of common covariate for both the binary and the continuous responses leaves $\theta_2$ unchanged under models $M_1$ and $M_2$. This is just the case here. Results reveal that unlike the measurement error, in ignoring small classification error the attenuation effect on the estimate of $\theta_1$ is perceptible. The attenuation effect becomes more prominent with increase in the

magnitudes of $\varepsilon_0$ and $\varepsilon_1$ (See Tables 5 and 6). Model $M_2$ clearly recovers the point estimates of $\theta_1$ at the expense of increased standard errors.

Table 4 $(\sigma^2, \varepsilon_0, \varepsilon_1) = (0,.01,.01)$

| Estimates | $M_1$ | $M_2$ |
|---|---|---|
| $\hat{\beta}_{01}$ | .0009(.0188) | .0010(.0220) |
| $\hat{\beta}_1$ | .8360(.0183) | 1.0011(.0220) |
| $\hat{\beta}_{02}$ | .0000(.0099) | .0000(.0099) |
| $\hat{\beta}_2$ | .9998(.0042) | .9998(.0042) |
| $\hat{\rho}$ | .4790(.0180) | .6033(.0188) |
| $\hat{\sigma}_2^2$ | .9996(.0067) | .9996(.0067) |
| $\hat{\varepsilon}_0$ | - | .0100(.0020) |
| $\hat{\varepsilon}_1$ | - | .0100(.0019) |

Table 5 $(\sigma^2, \varepsilon_0, \varepsilon_1) = (0,.05,.05)$

| Estimates | $M_1$ | $M_2$ |
|---|---|---|
| $\hat{\beta}_{01}$ | .0018(.0161) | -.0010(.0257) |
| $\hat{\beta}_1$ | .5933(.0110) | 1.0022(.0288) |
| $\hat{\beta}_{02}$ | .0000(.0099) | .0000(.0099) |
| $\hat{\beta}_2$ | .9998(.0043) | .9998(.0043) |
| $\hat{\rho}$ | .3067(.0166) | .6008(.0230) |
| $\hat{\sigma}_2^2$ | .9991(.0133) | .9991(.0133) |
| $\hat{\varepsilon}_0$ | - | .0500(.0043) |
| $\hat{\varepsilon}_1$ | - | .0500(.0040) |

Table 6 $(\sigma^2, \varepsilon_0, \varepsilon_1) = (0,.10,.10)$

| Estimates | $M_1$ | $M_2$ |
|---|---|---|
| $\hat{\beta}_{01}$ | .0022(.0147) | -.0006(.0315) |
| $\hat{\beta}_1$ | .4580(.0084) | 1.0024(.0346) |
| $\hat{\beta}_{02}$ | .0000(.0099) | .0000(.0099) |
| $\hat{\beta}_2$ | .9998(.0042) | .9998(.0042) |
| $\hat{\rho}$ | .2213(.0148) | .6004(.0279) |
| $\hat{\sigma}_2^2$ | .9996(.0067) | .9996(.0067) |
| $\hat{\varepsilon}_0$ | - | .1000(.0059) |
| $\hat{\varepsilon}_1$ | - | .1000(.0055) |

**Measurement error and Classification errors:**

Tables 7-15 describe the joint effects of classification errors and measurement error. Tables 7-9 show the joint effect when the measurement error variance is held fixed at 0.01 while the misclassification rates gradually increase from (.01,.01) to (.10,.10). Tables 10-12 show the joint effect of both the errors when measurement error variance is held fixed at .5 and Tables 13-15 show the same when the measurement error variance is 1.0

Comparisons of the estimates of the parameters show that $\beta_{02}$ and $\beta_2$ are same for all the four models. The estimate of $\sigma_2^2$ is same under models $M_1$ and $M_2$. This common estimated value shows inflation when compared with the estimates of $\sigma_2^2$ under $M_3$ and $M_4$. The estimate of the regression parameter $\beta_1$ is attenuated under the naïve model $M_1$. The attenuation effect becomes pronounced with increase in the value of the misclassification rates as well as measurement error variance. Ignoring classification errors causes attenuation of the estimate of ρ while ignoring measurement error causes inflation. When measurement error is small the effect of classification errors dominate and the estimate of ρ under naïve model shows attenuation compared to that under the correct model $M_4$ (See Tables 8 and 9). When measurement error is pronounced and classification errors are very small the estimate of ρ shows inflation when compared with the same under the correct model $M_4$ since in this case measurement error dominates (See Table 13). For moderate measurement error and classification errors the effects of ignoring these errors work in opposite directions; one results in attenuation and the other results in inflation of the estimate of ρ. As a result we might chance upon a situation when the estimate of ρ is close to the true value under model $M_1$ (See Table 10). On the whole the model $M_4$ works well in all situations.

Table 7 $(\sigma^2, \varepsilon_0, \varepsilon_1) = (.01, .01, .01)$

| Estimates | $M_1$ | $M_2$ | $M_3$ | $M_4$ |
|---|---|---|---|---|
| $\hat{\beta}_{01}$ | .0010(.0191) | -.0006(.0222) | .0010(.0193) | -.0007(.0224) |
| $\hat{\beta}_1$ | .8334(.0181) | .9963(.0220) | .8392(.0185) | 1.0063(.0227) |
| $\hat{\beta}_{02}$ | -.0000(.0100) | -.0000(.0100) | -.0000(.0100) | -.0000(.0100) |
| $\hat{\beta}_2$ | .9998(.0043) | .9998(.0043) | .9998(.0043) | .9998(.0043) |
| $\hat{\rho}$ | .4828(.0189) | .6039(.0190) | .4785(.0190) | .5999(.0191) |
| $\hat{\sigma}_2^2$ | 1.0099(.0137) | 1.0099(.0137) | .9999(.0137) | .9999(.0137) |
| $\hat{\varepsilon}_0$ | - | .0100(.0198) | - | .0100(.0198) |
| $\hat{\varepsilon}_1$ | - | .0100(.0194) | - | .0100(.0194) |

Table 8 $(\sigma^2, \varepsilon_0, \varepsilon_1) = (.01, .05, .05)$

| Estimates | $M_1$ | $M_2$ | $M_3$ | $M_4$ |
|---|---|---|---|---|
| $\hat{\beta}_{01}$ | .0018(.0160) | -.0008(.0270) | .0018(.0160) | -.0007(.0273) |
| $\hat{\beta}_1$ | .5923(.0109) | .9972(.0284) | .5944(.0111) | 1.0073(.0292) |
| $\hat{\beta}_{02}$ | -.0000(.0100) | -.0000(.0100) | -.0000(.0100) | -.0000(.0100) |
| $\hat{\beta}_2$ | .9998(.0043) | .9998(.0043) | .9998(.0043) | .9998(.0043) |
| $\hat{\rho}$ | .3097(.0165) | .6044(.0224) | .3058(.0166) | .6004(.0226) |
| $\hat{\sigma}_2^2$ | 1.0099(.0137) | 1.0099(.0137) | .9999(.0137) | .9999(.0137) |
| $\hat{\varepsilon}_0$ | - | .0500(.0043) | - | .0500(.0043) |
| $\hat{\varepsilon}_1$ | - | .0500(.0040) | - | .0500(.0043) |

Table 9 $(\sigma^2, \varepsilon_0, \varepsilon_1) = (.01, .10, .10)$

| Estimates | $M_1$ | $M_2$ | $M_3$ | $M_4$ |
|---|---|---|---|---|
| $\hat{\beta}_{01}$ | .0021(.0145) | -.0009(.0325) | .0021(.0146) | -.0009(.0324) |
| $\hat{\beta}_1$ | .4574(.0084) | .9977(.0337) | .4584(.0085) | 1.0077(.0348) |
| $\hat{\beta}_{02}$ | -.0000(.0100) | -.0000(.0100) | -.0000(.0100) | -.0000(.0100) |
| $\hat{\beta}_2$ | .9998(.0043) | .9998(.0043) | .9998(.0043) | .9998(.0043) |
| $\hat{\rho}$ | .2236(.0148) | .6043(.0271) | .2204(.0149) | .6003(.0273) |
| $\hat{\sigma}_2^2$ | 1.0099(.0137) | 1.0099(.0137) | .9999(.0137) | .9999(.0137) |
| $\hat{\varepsilon}_0$ | - | .1000(.0059) | | .1000(.0059) |
| $\hat{\varepsilon}_1$ | - | .1000(.0055) | | .1000(.0055) |

Table 10 $(\sigma^2, \varepsilon_0, \varepsilon_1) = (.5, .01, .01)$

| Estimates | $M_1$ | $M_2$ | $M_3$ | $M_4$ |
|---|---|---|---|---|
| $\hat{\beta}_{01}$ | .0002(.0169) | -.0012(.0193) | .0003(.0229) | -.0018(.0292) |
| $\hat{\beta}_1$ | .7211(.0133) | .8169(.0160) | .9750(.0307) | 1.2270(.0484) |
| $\hat{\beta}_{02}$ | -.0002(.0124) | -.0002(.0124) | -.0002(.0124) | -.0002(.0124) |
| $\hat{\beta}_2$ | .9997(.0054) | .9997(.0054) | .9997(.0054) | .9997(.0054) |
| $\hat{\rho}$ | .6222(.0124) | .7340(.0164) | .4383(.0247) | .5770(.0214) |
| $\hat{\sigma}_2^2$ | 1.5013(.0226) | 1.5013(.0226) | 1.0016(.0230) | 1.0016(.0230) |
| $\hat{\varepsilon}_0$ | - | .0100(.0020) | - | .0100(.0020) |
| $\hat{\varepsilon}_1$ | - | .0100(.0019) | - | .0010(.0019) |

Table 11 $(\sigma^2, \varepsilon_0, \varepsilon_1) = (.5, .05, .05)$

| Estimates | $M_1$ | $M_2$ | $M_3$ | $M_4$ |
|---|---|---|---|---|
| $\hat{\beta}_{01}$ | .0009(.0158) | -.0007(.0234) | .0010(.0186) | -.0010(.0354) |
| $\hat{\beta}_1$ | .5450(.0097) | .8179(.0190) | .6401(.0154) | .2307(.0579) |
| $\hat{\beta}_{02}$ | -.0002(.0124) | -.0002(.0124) | -.0002(.0124) | -.0002(.0124) |
| $\hat{\beta}_2$ | .9997(.0054) | .9997(.0054) | .9997(.0054) | .9997(.0054) |
| $\hat{\rho}$ | .4295(.0156) | .7346(.0159) | .2574(.0211) | .5778(.0264) |
| $\hat{\sigma}_2^2$ | 1.5013(.0226) | 1.5013(.0226) | 1.0016(.0230) | 1.0016(.0230) |
| $\hat{\varepsilon}_0$ | - | .0500(.0044) | - | .0500(.0044) |
| $\hat{\varepsilon}_1$ | - | .0500(.0041) | - | .0500(.0041) |

Table 12 $(\sigma^2, \varepsilon_0, \varepsilon_1) = (.5, .10, .10)$

| Estimates | $M_1$ | $M_2$ | $M_3$ | $M_4$ |
|---|---|---|---|---|
| $\hat{\beta}_{01}$ | .0014(.0147) | -.0013(.0289) | .0015(.0162) | -.0019(.0438) |
| $\hat{\beta}_1$ | .4305(.0081) | .8187(.0226) | .4744(.0107) | 1.2336(.0692) |
| $\hat{\beta}_{02}$ | -.0002(.0124) | -.0002(.0124) | -.0002(.0124) | -.0002(.0124) |
| $\hat{\beta}_2$ | .9997(.0054) | .9997(.0054) | .9997(.0054) | .9997(.0054) |
| $\hat{\rho}$ | .3191(.0147) | .7347(.0194) | .1751(.0194) | .5781(.0312) |
| $\hat{\sigma}_2^2$ | 1.5013(.0226) | 1.5013(.0226) | 1.0016(.0230) | 1.0016(.0230) |
| $\hat{\varepsilon}_0$ | - | .1000(.0061) | - | .1000(.0061) |
| $\hat{\varepsilon}_1$ | - | .1000(.0055) | - | .1000(.0055) |

Table 13 $(\sigma^2, \varepsilon_0, \varepsilon_1) = (1.0, .01, .01)$

| Estimates | $M_1$ | $M_2$ | $M_3$ | $M_4$ |
|---|---|---|---|---|
| $\hat{\beta}_{01}$ | .0006(.0158) | -.0004(.0173) | .0010(.0267) | -.0008(.0349) |
| $\hat{\beta}_1$ | .6416(.0113) | .7069(.0124) | 1.0920(.0463) | 1.0003(.0752) |
| $\hat{\beta}_{02}$ | -.0003(.0143) | -.0003(.0143) | -.0003(.0143) | -.0003(.0143) |
| $\hat{\beta}_2$ | .9996(.0063) | .9996(.0063) | .9996(.0063) | .9996(.0063) |
| $\hat{\rho}$ | .7013(.0100) | .8010(.0138) | .3773(.0328) | .5486(.0292) |
| $\hat{\sigma}_2^2$ | 2.0022(.0305) | 2.0022(.0305) | 1.0028(.0326) | 1.0028(.0326) |
| $\hat{\varepsilon}_0$ | - | .0100(.0020) | - | .0100(.0020) |
| $\hat{\varepsilon}_1$ | - | .0100(.0020) | - | .0100(.0020) |

Table 14 $(\sigma^2, \varepsilon_0, \varepsilon_1) = (1.0, .05, .05)$

| Estimates | $M_1$ | $M_2$ | $M_3$ | $M_4$ |
|---|---|---|---|---|
| $\hat{\beta}_{01}$ | .0008(.0153) | .0003(.0206) | .0010(.0206) | .0007(.0415) |
| $\hat{\beta}_1$ | .5042(.0089) | .7072(.0148) | .6763(.0202) | .9798(.0894) |
| $\hat{\beta}_{02}$ | -.0003(.0143) | -.0003(.0143) | -.0003(.0143) | -.0003(.0143) |
| $\hat{\beta}_2$ | .9996(.0063) | .9996(.0063) | .9996(.0063) | .9996(.0063) |
| $\hat{\rho}$ | .5089(.0130) | .8015(.0148) | .1929(.0283) | .5501(.0352) |
| $\hat{\sigma}_2^2$ | 2.0022(.0305) | 2.0022(.0305) | 1.0028(.0327) | 1.0028(.0327) |
| $\hat{\varepsilon}_0$ | - | .0500(.0044) | - | .0500(.0044) |
| $\hat{\varepsilon}_1$ | - | .0500(.0041) | - | .0500(.0041) |

Table 15 $(\sigma^2, \varepsilon_0, \varepsilon_1) = (1.0, .10, .10)$

| Estimates | $M_1$ | $M_2$ | $M_3$ | $M_4$ |
|---|---|---|---|---|
| $\hat{\beta}_{01}$ | .0013(.0147) | -.0000(.0250) | .0015(.0176) | -.0000(.4320) |
| $\hat{\beta}_1$ | .4054(.0077) | .7080(.0176) | .4853(.0129) | .9983(.0465) |
| $\hat{\beta}_{02}$ | -.0003(.0143) | -.0003(.0143) | -.0003(.0143) | -.0003(.0143) |
| $\hat{\beta}_2$ | .9996(.0063) | .9996(.0063) | .9996(.0063) | .9996(.0063) |
| $\hat{\rho}$ | .3875(.0145) | .8020(.0158) | .1240(.0253) | .5678(.0345) |
| $\hat{\sigma}_2^2$ | 2.0022(.0305) | 2.0022 (.0305) | 1.0028(.0327) | 1.0028(.0327) |
| $\hat{\varepsilon}_0$ | - | .1000(.0061) | - | .1000(.0061) |
| $\hat{\varepsilon}_1$ | - | .1000(.0055) | - | .1000(.0055) |

## 7. Example

A survey was conducted among 121 male undergraduate students studying Statistics as a subsidiary subject in St Xavier's college, Kolkata, India. On a particular day, in the class, the students were requested to provide information on the following items keeping their anonymity:

1. Total family income per month (z)
2. Pocket money available per month ($y_2$)
3. Whether the student takes alcohol or not ($y_1$).

Regarding the alcohol intake there are 2 categories: (i) Never (ii) At least once a week
In our society consuming alcohol is still considered to be a taboo especially among the students coming from the middle class. Thus, students do not feel free to speak out the truth even if they consume alcohol. On the other hand, there are a few teetotalers who might be tempted to provide wrong information just for fun. Thus $y_1$ is subject to classification errors. It is expected that the binary outcome ($y_1$) and the pocket money available ($y_2$) are correlated. Moreover, $y_1$ and $y_2$ depend upon the family income. However, true income (x) of a family or a person is usually subject to measurement error and the total family income (z) reported by the students can be taken to be a surrogate for the true income. Thus the binary responses in the above data are subject to classification errors and the true covariate (family income) is subject to measurement error.

While carrying out the analysis we expressed $y_2$ and z in the unit of thousand rupees. The analysis was done for all the four models described in Section 8. In the absence of validation data the measurement error variance $\sigma^2$ was assigned a prefixed value 1. The results are reported in Table 16. The results show that the chance of a student reporting that he consumes alcohol when he, in fact, doesn't is small (.0997) whereas the chance of reporting that he doesn't consume when he, in fact, does is high (.5379). The results support our contention made above.

The results show that the measurement error does not affect the estimates of the regression parameters $\beta_{02}$ and $\beta_2$. However the naïve estimate of $\sigma_2^2$ shows slight inflation. The estimates of $\beta_{01}$ and $\beta_1$ under model $M_3$ clearly indicate that ignoring measurement error results in attenuation of the estimates. The estimate of ρ under model

$M_3$ shows attenuation compared to the naïve estimate. Also it is observed that the effect of classification error dominates measurement error.

Table 16

| Estimates | $M_1$ | $M_2$ | $M_3$ | $M_4$ |
|---|---|---|---|---|
| $\hat{\beta}_{01}$ | -.8072(.1151) | -.5394(.1642) | -.8750(.1512) | -.6294(.2245) |
| $\hat{\beta}_1$ | .3860(.0570) | .9997(.0125) | .4184(.0980) | 1.0013(.1015) |
| $\hat{\beta}_{02}$ | .4549(.0819) | .4549(.0819) | .4549(.0819) | .4549(.0819) |
| $\hat{\beta}_2$ | .0624(.0043) | .0624(.0043) | .0624(.0043) | .0624(.0043) |
| $\hat{\rho}$ | .1736(.1328) | .4118(.2007) | .1278(.1330) | .3125(.2089) |
| $\hat{\sigma}_2^2$ | .1783(.0227) | .1783(.0227) | .1744(.0226) | .1744(.0226) |
| $\hat{\varepsilon}_0$ | - | .0997(.1245) | - | .0997(.1245) |
| $\hat{\varepsilon}_1$ | - | .5379(.1322) | - | .5379(.1322) |

## 8. Concluding Remarks

In this paper we consider modeling mixed binary and continuous outcomes when binary outcomes may be subject to classification errors and/or some of the covariates are not observable in the main study but its surrogates are observed. We model the joint distribution of the binary and continuous responses by using a model proposed by Cox (1972). The advantage of using this model is, we are able to find analytical results that throw interesting lights about the behaviour of likelihood estimates of the model parameters in the presence of the above errors. There are still unanswered questions eluding theoretical justification that we left as an open problem. Developing similar methodologies for multivariate mixed outcomes possibly with ordered categorical variables would be worth studying.

## References

Catalano, J. & Ryan, L. M. (1992) Bivariate latent variable models for clustered discrete and continuous outcomes. *J. Am. Statist. Ass.*, 87, 651-658.

Cox, D. R. (1972) The analysis of multivariate binary data. *Appl. Statist.*, 21, 113-120.

Cox, D.R. and Snell, E.J. (1989) *Analysis of Binary Data.* 2nd edn., London: Chapman and Hall.

Fitzmaurice, G. M. & Laird, N. M. (1995) Regression models for a bivariate discrete and continuous outcome with clustering. *J. Am. Statist. Ass.*, 90, 845-852.

Fujita, S. (1989) *Version of DS86. Radiation Effects Research Foundation Update 1, 3.* Hiroshima, Japan.

Geys, H., Regan, M.M., Catalano, P.J. & Molenberghs, G. (2001) Two latent variable risk assessment approaches for mixed continuous and discrete outcomes from developmental toxicity data. *J. Agric. Bio. Envir. Statist*., 6, 340-355.

Gueorguieva, R.V. and Agresti, A. (2001) A correlated probit model for joint modeling of clustered binary and continuous responses. *J. Am. Statist. Ass.* **96**, 1102-1112.

Gueorguieva, R.V. & Sanacora, G. (2006) Joint analysis of repeatedly observed continuous and ordinal measures of disease severity. *Statist. Medi.*, 25, 1037-1322.

Holcroft, C.A. and Spiegelman, D. (1999) Design validation studies for estimating the odds ratio of exposure-disease relationships when exposure in misclassified. *Biometrics*, **55**, 1193-1201.

Kullback, S. (1959) *Information Theory and Statistics*. New York: John Wiley.

Morrissey, M.J. and Spigelman, D. (1999) Matrix methods for estimating odds ratios with misclassified exposure data: extensions and comparison. *Biometrics*, **55**, 338-344.

Neuhaus, J.M. (1999) Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika*, **86**, 843-855.

Olkin, I. & Tate, R. F. (1961) Multivariate correlation models with mixed discrete and continuous variables. *Ann. Math. Statist.*, 32, 448-465.

Regan, M.M. & Catalano, P.J. (1999) Bivariate dose-response modeling and risk estimation in developmental toxicology. *J. Agric. Bio. Envir. Statist.,* 4, 217-237.

Regan, M.M. & Catalano, P.J. (2000) Regression models for mixed discrete and continuous outcomes with clustering. *Risk Analysis.* 20, 363-376.

Roesch, W. C. (1987) *U.S.-Japan joint reassessment of atomic bomb radiation dosimetry in Hiroshima and Nagasaki: final report.* Hiroshima, Japan. Radiation Effects Research Foundation.

Roy, S., Banerjee, T. & Maiti, T. (2005) Measurement error model for misclassified binary responses. *Statist. Med.* 24, 269-283.

Sammel, M. D., Ryan, L. M. & Legler, J. M. (1997) Latent variable models for mixed discrete and continuous outcomes. *J. R. Statist. Soc. B*, 3, 667-678.

Sposto, R., Preston, D. L., Shimizu, Y. and Mabuchi, K. (1992) The effect of diagnostic misclassification on non cancer and cancer mortality dose response in A-bomb survivors. *Biometrics*., 48, 605-617.

White, H. (1982) Maximum likelihood estimation in misspecified models. *Econometrica,* **50,** 1-25.