Revista IECOS, Journal of Economics, Applied Statistics and Social Science
Vol. 24, N° 2, diciembre de 2023, pp. 61-78
DOI: https://doi.org/10.21754/iecos.v24i2.2003

ISSN: 2788-7480 (En línea)
ISSN: 2961-2845 (Impresa)

Binary regression model with misclassification and berkson-type measurement error with Student-t distribution

Modelo de regresión binaria con mala clasificación y error de medición tipo Berkson con Distribución t-Student

Recibido: 09/12/2023
Aceptado: 20/12/2023
Publicado: 30/12/2023

[1] Universidad Federal do Cariri, Céará, Brasil
Correspondencia: marcos.pereira@ufca.edu.br

https://orcid.org/0000-0002-9555-4385

[2] Universidad Federal de Pernambuco, Pernambuco, Brasil
Correspondencia: betsabe@de.ufpe.br

https://orcid.org/0000-0002-1236-0930

Revista de la Facultad de Ingeniería Económica, Ingeniería Estadística y Ciencias Sociales de la Universidad Nacional de Ingeniería

Marco Antonio Alves Pereira[1], Betsabé Grimalda Blas Achic[2]

ABSTRACT

In this article, we introduce a regression model tailored for fitting binary data affected by misclassification in the response variable and Berkson-type measurement error in the covariate. The conventional assumption of a normal distribution for measurement error may inadequately represent atypical observations present in the dataset. To address this limitation, our model incorporates misclassification in the response variable and Berksontype measurement error, employing the Student-t distribution for more robust modeling of these atypical observations. We utilize the cumulative distribution function from the Student-t distribution as the link function, enhancing our ability to capture the dataset's unique characteristics. Model parameters are estimated via the maximum likelihood method. We conduct a comprehensive Monte Carlo simulation study to thoroughly assess the impact of measurement errors and misclassification. Additionally, we apply the proposed model to a real-world dataset of survivors from the atomic bombing in Japan, showcasing its adaptability and suitability in practical scenarios. Our findings highlight the robustness and flexibility of this model in effectively handling complex binary regression scenarios involving measurement errors and misclassification.

Keywords. *Binary regression model; Berkson-type error; misclassification; Student-t*

RESUMEN

En este artículo, presentamos un modelo de regresión diseñado para ajustar datos binarios afectados por error de clasificación en la variable respuesta y error de medición tipo Berkson en la covariable. La suposición convencional de distribución normal para el error de medición puede representar inadecuadamente observaciones atípicas presentes en el conjunto de datos. Para abordar esta limitación, nuestro modelo incorpora error de clasificación en la variable respuesta y error de medición tipo Berkson, empleando la Distribución t de Student para modelar de manera más robusta estas observaciones atípicas. Utilizamos la función de distribución acumulativa de la distribución t de Student como la función de enlace, mejorando la capacidad para capturar las características únicas del conjunto de datos. Estimamos los parámetros del modelo mediante el método de máxima verosimilitud. Realizamos un estudio exhaustivo de simulación de Monte Carlo para evaluar minuciosamente el impacto de los errores de medición y el error de clasificación. Además, aplicamos el modelo propuesto a un conjunto de datos reales de sobrevivientes del bombardeo atómico en Japón, demostrando su adaptabilidad y adecuación en escenarios prácticos. Nuestros resultados resaltan la robustez y flexibilidad de este modelo en el manejo efectivo de escenarios de regresión binaria complejos que involucran errores de medición y error de clasificación.

Palabras clave. *Modelo de regresión binaria, error del tipo Berkson, error de clasificación, Distribución t-Student.*

1. INTRODUCTION

In regression models applied to binary data, it is typical to encounter datasets where certain covariates remain unobserved, leading to biased estimates. Conventional binary regression models operate under the assumption that the observed binary responses are devoid of misclassification, and the independent variables are free from measurement error. However, practical scenarios often involve measurement errors and misclassification, contributing to potential biases and imprecisions in the estimated regression coefficients.

To address these challenges, researchers have proposed various methods to account for measurement error in binary regression models and misclassification in the outcome variable in binary regression models. One approach to tackle measurement error is to assume a known distribution for the measurement errors and estimate the regression parameters using likelihood based methods. This approach, known as the classical measurement error model, has been extensively studied and applied in various fields, including epide-

miology, social sciences, applied social sciences, and environmental sciences. By explicitly modeling the measurement error, it becomes possible to obtain more reliable estimates of the true underlying relationships between variables (Carroll et al., 2006).

Another approach is to model misclassification in the response variable, where the observed binary response may not accurately represent the true underlying response due to misclassification errors. This can be addressed by estimating the probabilities of misclassification and adjusting the model accordingly Ekholm and Palmgren (1982). Carroll et al. (1984) analyzed data from a prospective study on the development of cardiovascular diseases presented in Kannel and Gordon (1968) and demonstrated the impact of measurement errors in binary regression. Burr (1988) investigated measurement errors in Berkson-type covariates in the field of bioassays, employing the probit link function.

In the context of binary response variables, measurement error models become even more challenging due to the presence of misclassification. Several researchers have proposed some models, such as Roy et al. (2005) who developed a measurement error model for misclassified binary responses, where the independent variable is subject to the Berkson-type measurement error which follows the normal distribution. To address departures from normality in measurement errors, Bolfarine and Lachos (2006) considered structural measurement errors following a skew-normal distribution and adopted the probit link function. They employed both classical and Bayesian approaches for parameter estimation, utilizing Markov chain Monte Carlo techniques. Liu and Zhang (2017) conducted a Monte Carlo simulation study with the logistic regression model, employing the logit link function within the classical framework, to demonstrate the presence of non-ignorable biases in parameter estimates when misclassification is disregarded. Bazán et al. (2014) used skew-probit link functions because it deviates from the probit link function in terms of a flexible asymmetry parameter, with Bayesian approach.

In this article, we introduce an innovative regression model designed to tackle the complexities of both measurement error and misclassification in binary data. Berkson-type measurement error occurs when an independent variable isn't directly observed but is derived from a surrogate variable along with measurement error (Roy et al., 2005; Burr, 1988). The conventional assumption of a normal distribution for measurement error often fails to adequately represent unusual observations within the dataset. To address this limitation, we present a flexible modeling framework integrating the Student-t distribution (Lange et al., 1989) to handle the measurement error component.

Moreover, our model incorporates the cumulative distribution function (cdf) from the Student-t distribution as the link function. This link function plays a pivotal role in connecting the linear predictor to the probabilities of the binary response. By employing the cdf from the Student-t distribution as a link function, we enhance our ability to effectively capture the data's unique characteristics and thereby improve the model's overall performance.

For estimating model parameters, we utilize the maximum likelihood method, leveraging the optimx (Nash & Varadhan, 2011) library within the R software (R Core Team, 2021). This method ensures efficient and consistent estimators. Additionally, we conduct a comprehensive Monte Carlo simulation study to evaluate how measurement errors and misclassification impact parameter estimation and prediction accuracy. This study provides insights into the model's robustness under various scenarios involving complex data structures.

To showcase the practical application of our proposed model, we applied it to a genuine dataset featuring survivors of the atomic bombings in Japan. Our analysis of this dataset serves to underscore the model's aptness and effectiveness in capturing the intricacies within the data, offering valuable insights and demonstrating its robustness in handling real-world complexities.

The remainder of the article is organized as follows. In Section 2, we present the proposed regression model that incorporates misclassification and measurement error. We estimate the model parameters numerically using the maximum likelihood method. Section 3 provides a simulation study to evaluate the performance of the maximum likelihood estimators. In Section 4, we apply the proposed model on a real data set of survivors of the atomic bomb attacks in Japan. Finally, Section 5 presents concluding remarks.

2. THE MODEL

The probit model, a binary linear regression using the probit link function, assumes the response variable follows a binary distribution. It models the relationship between predictors and the probability of the outcome. The probit link function is the cdf of the standard normal distribution, transforming the linear predictor into a probability, ensuring a smooth and symmetric relationship between predictors and the likelihood of success.

Using the cdf offers interpretability advantages. Coefficients estimate the change in the odds of success for a one-unit change in the predictor, making the probit model more interpretable than other link functions. It assumes errors

follow a standard normal distribution, typically reasonable for large sample sizes due to the central limit theorem, resulting in reliable estimates and accurate inference.

However, the probit model is just one option among others (e.g., logit, cloglog) for modeling binary responses. The choice of link function relies on factors like the research context, data characteristics, and specific research questions. Researchers often compare different link functions, selecting the one best fitting the data and yielding the most meaningful results.

We present a regression model tailored for binary data handling Berkson-type measurement errors in covariates, where the error follows a Student-t distribution, including it as the link function. Furthermore, we account for misclassification in the response variable.

In linear models with binary responses, average estimates represent proportions. Various link functions are employed to transform the linear predictor, mapping values from the real line to the interval [0, 1]. Consequently, a binary regression model can be defined as

$$(1) \qquad \mathbb{P}(y_i) = F_G(\beta_0 + \beta_1 w_i), \qquad and \qquad F_G^{-1}\big(\mathbb{P}(y_i)\big) = \beta_0 + \beta_1 w_i,$$

Where F_G (.) is the link function, y_i is the binary response variable with Bernoulli distribution and parameter $p_i(\theta) = \mathbb{P}(y_i = 1) = \mathbb{P}(y_i)$ is the model parameter vector, and w_i is the predictor variable, i=1,…,n. We consider the link function to be the cdf of a distribution G_i belonging to the Student-t distribution (Lange et al., 1989) with location parameter 0, scale parameter 1 and v degrees of freedom. Thus, $G_i \sim T(0,1,v)$, which implies that $p_i(\theta) = \mathbb{P}(y_i = 1) = F_G(\beta_0 + \beta_1 w_i, 0,1,v)$.

- **Naive model (M1)**

  For the naive model, which assumes an absence of misclassification and measurement error, we consider the parameter vector to be estimated as $\theta 1=(\beta_0,\beta_1,v)^T$. The probability of observing $y_i = 1$, denoted as $p_{1i}(\theta_1) = \mathbb{P}(y_i = 1)$„ for i=1,…,n is given by

  $$(2) \qquad \mathbb{P}(y_i = 1) = F_G(\beta_0 + \beta_1 w_i, 0, 1, v),$$

  and is based on the responses $y = (y_1, \dots, y_n)^T$a nd the predictors

  $w = (w_1, \dots, w_n)^T$, with $G_i \sim T(0,1,v)$,

$$L(\theta_1|y,w) = \prod_{i=1}^{n} p_{1i}\,(\theta_1)^{y_i}\big(1-p_{1i}(\theta)\big)^{1-y_i},$$

So, we will have the log-likelihood function is then obtained as

$$\ell(\theta_1|y,w) = \sum_{i=1}^{n} y_i \log\big(p_{1i}(\theta_1)\big) + \sum_{i=1}^{n}(1-y_i)\log\big(1-p_{1i}(\theta_1)\big), \tag{3}$$

which represents the log-likelihood function for the M1 model, considering (2) and (3), we have

$$\ell(\theta_1|y,w) = \sum_{i=1}^{n} y_i \log\big(F_G(\beta_0+\beta_1 w_i,0,1,v)\big)$$

$$+\sum_{i=1}^{n}(1-y_i)\log\Big(1-F_G\big(\beta_0+\beta_{1w_i},0,1,v\big)\Big). \tag{4}$$

- **Model incorporating misclassification (M2)**

Let y_i, represent the unobserved or true binary response, and y_i, i=1,...,n, denote the observed binary response. We assume that the probabilities ϵ_0 and ϵ_1 of misclassification (Roy et al., 2005) are

$$\mathbb{P}(\tilde{y}_i = 1|y_i = 0) = \epsilon_0, \;\; and \;\; \mathbb{P}(\tilde{y}_i = 0|y_i = 1) = \epsilon_1, \tag{5}$$

Where $\epsilon_0 + \epsilon_1 < 1$.

Considering $\tilde{y}_i, i = 1,\dots,n$, with a Bernoulli distribution parameterized by $p_{2i}(\theta_2) = \mathbb{P}(\tilde{y}_i = 1)$,, we have

$$\mathbb{P}(\tilde{y}_i = 1) = \epsilon_0 + (1-\epsilon_0-\epsilon_1)F_G(\beta_0+\beta_1 w_i,0,1,v). \tag{6}$$

A regression model for binary data with misclassification, where the parameter vector is denoted as $\theta_2 = (\beta_0,\beta_1,\epsilon_0,\epsilon_1,v)^T$, given the predictors $w = (w_1,\dots,w_n)^T$ is represented by the log-likelihood function

$$\ell(\theta_2|\tilde{y},w) = \sum_{i=1}^{n} \tilde{y}_i \log\big(\epsilon_0 + (1-\epsilon_0-\epsilon_1)F_G(\beta_0+\beta_1 w_i,0,1,v)\big) \tag{7}$$

$$+\sum_{i=1}^{n}(1-\tilde{y}_i)\;\log\Big(1-\big(\epsilon_0+(1-\epsilon_0-\epsilon_1)F_G(\beta_0+\beta_1 w_i,0,1,v)\big)\Big),$$

where $\tilde{y} = (\tilde{y}_1, \ldots, \tilde{y}_n)^T$ are the observed responses.

- **Model considering measurement error effects (M3)**

In the regression model with Berkson-type measurements error, the predictor variable X_i is not directly observed. Instead, X_i is obtained as the sum of its surrogate ω_i and a measurement error δ_i,i=1,…,n, with $\delta_i \sim T(0,\sigma^2,v)$ and $X_i \sim T(\omega_i,\sigma^2,v)$. Specifically, we have

$$\mathbb{P}(y_i = 1|X_i) = F_G(\beta_0 + \beta_1 X_i, 0, 1, v), \quad \text{(Probability of } y_i \text{ given } X_i) \tag{8}$$

$$X_i = w_i + \delta_i. \quad \text{(Measurement error model)} \tag{9}$$

Assuming that the random variables X_i and G_i are univariate and independent random variables that constitute the random vector with bivariate Student-t distribution

$$\begin{pmatrix} X_i \\ G_i \end{pmatrix} \sim T_2\left(\begin{pmatrix} w_i \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & 1 \end{pmatrix}, v\right), \tag{10}$$

where $X_i \sim T(\omega_i,\sigma^2,v)$ and $G_i \sim T(0,1,v)$,i=1,…,n, then we can define the random variable $Q_i=G_i-\beta_1 X_i \sim T(-\beta_1 \omega_i, 1+\beta_1^2 \sigma^2, v)$ (Branco & Dey, 2001; Lin, 1972).

For the regression model with binary response with measurement error Berkson-type (8)-(9), considering $p_{3i}(\theta_3) = \mathbb{P}(y_i = 1)$, where $\theta_3 = (\beta_0, \beta_1, \sigma^2, v)^T$, we have

$$\mathbb{P}(y_i = 1) = \mathbb{E}_X\{F_G(\beta_0 + \beta_1 X_i, 0, 1, v)\} = F_Q(\beta_0, -\beta_1 w_i, 1 + \beta_1^2\sigma^2, v). \tag{11}$$

where G_i is the link function $Q_i=G_i-\beta_1 X_i$ follows a Student-t distribution with location parameter $-\beta_1 \omega_i$, scale parameter $(1+\beta_1^2\sigma^2)^{\frac{1}{2}}$, and v degrees of freedom and F_Q is the cdf of Q_i.

A model for binary data with Berkson-type measurements error, with the parameter vector $\theta_3=(\beta_0,\beta_1,\sigma^2,v)^T$, has the log-likelihood function given by

$$\ell(\theta_3|y,w) = \sum_{i=1}^{n} y_i \log\left(F_Q(\beta_0, -\beta_1 w_i, 1 + \beta_1^2\sigma^2, v)\right) + \sum_{i=1}^{n}(1 - y_i)\log\left(1 - F_Q(\beta_0, -\beta_1 w_i, 1 + \beta_1^2\sigma^2, v)\right). \tag{12}$$

- **Model incorporating both measurement error and misclassification (M4)**
  We delineate a model that integrates both measurement error and misclassification, building upon the foundations laid by the M2 (6) and M3 (11) models. We consider the probabilities ϵ_0 and ϵ_1 of misclassification.

  Given the parameter vector $\theta_4=(\beta_0,\beta_1,\epsilon_0,\epsilon_1,\sigma^2,\nu)^T$ and denoting the unobserved binary response as the true value y_i, the observed binary response as $\tilde{y}_i$, and the observed variable as ω_i,i=1,...,n, the log-likelihood function is expressed as

$$\ell(\theta_4|\tilde{y},w)=\sum_{i=1}^{n}\tilde{y}_i\log\Big(\epsilon_0+(1-\epsilon_0-\epsilon_1)F_Q(\beta_0,-\beta_1 w_i,1+\beta_1^2\sigma^2,\nu)\Big)$$

$$+\sum_{i=1}^{n}(1-\tilde{y}_i)\log\Big(1-\Big(\epsilon_0+(1-\epsilon_0-\epsilon_1)F_Q(\beta_0,-\beta_1 w_i,1+\beta_1^2\sigma^2,\nu)\Big)\Big). \quad (13)$$

3. SIMULATION STUDY

We performed a series of Monte Carlo simulations to examine the effects of misclassification and measurement errors on the coefficient estimates of regression models. In each scenario, we generate 500 Monte Carlo with measurement error following a Student-t distribution and/or with misclassification. Below, we provide a description of the simulation study, mirroring the approach taken by Roy et al. (2005).

1. We generate the variable $w_i, i=1,\dots,n$ with uniform distribution in the interval (-4,4) and these values are kept fixed.

2. We generate the variable x_1, where $x_i=w_i+\delta_i$, with x_i and δ_i following Student-t distributions with $\nu=4$ degrees of freedom, i. e., $\delta_i\sim T(0,\sigma^2,4)$ and $x_i\sim T(w_i,\sigma^2,4), i=1,\dots,10000$.

3. We generate the variable y_i with Bernoulli distribution and probability of success $F_G(\beta_0+\beta_1 x_i,0,1,4), i=1,\dots,10000$, with $G_i\sim T(0,1,4)$, according to the scenario considered and with $\beta_0=0$ and $\beta_1=1$.

4. We generate the variable $\tilde{y}_i, i=1,\dots,10000$, with misclassification with probabilities prefixed as (5).

5. We fit the generated data $(\tilde{y}_i, w_i), i = 1, \dots, 10000$, with the models M1, M2, M3 e M4 presented to estimate the parameters using the maximum likelihood method.

6. Repeat steps 2-5 for 500 replicas and find the estimates $\hat{\theta}_{kj}, k = 1, \dots, 4$ and $j = 1, \dots, 500$, and find the standard errors of $\hat{\theta}$ through Fisher's information matrix.

7. Calculate the average of $\hat{\theta}_{kj}, k = 1, \dots, 4$ and $j = 1, \dots, 500$ and the average of standard errors.

8. Repeat steps 2-7 for different values of ϵ_0, ϵ_1 and σ^2.

Table 1-3 display the outcomes of simulations, featuring mean values and standard errors (SE) statistics for the adjusted model parameters. The results are derived from 500 Monte Carlo samples, each comprising 10, 000 observations, considering the presence of measurement and/or classification errors. Additionally, for the sake of comparison, adjustments were made using the probit link function (Roy et al., 2005) for models M1, M2, M3, and M4. This involved substituting the cdf of the Student-t distribution (T) with that of the normal distribution (N). Throughout all scenarios, σ^2 is assumed to be known.

In Table 1, showcasing simulation results where data exclusively incorporated misclassification, we note smaller biases in adjustments employing the Student-t distribution in contrast to adjustments with the normal distribution. This pattern holds true when comparing models M2 and M1. The superiority of the M2 model with Student-t becomes more pronounced with escalating probabilities of misclassification, as expected. Notably, the Student-t degrees of freedom estimated with the M1 model are significantly smaller than those estimated with M2, emphasizing the need for a distribution with heavy tails, particularly since M1 does not factor in misclassification. Additionally, it's worth mentioning that the SEs of M2 are larger due to the incorporation of additional parameters into the model.

Table 1.
The mean and SE of model parameters for M1 and M2 are derived from 500 Monte Carlo samples, each comprising 10, 000 observations. The data were generated with misclassification and without measurement error in three distinct scenarios.

θ	N	T	N	T
	M1		M2	
$(\epsilon_1, \epsilon_2, \sigma^2) = (0.01, 0.01, 0)$				
β_0	-0.003 (0.017)	0.000 (0.026)	0.000 (0.026)	0.000 (0.030)
β_1	0.687 (0.011)	1.044 (0.055)	0.882 (0.028)	1.004 (0.071)
ϵ_0	—— (——)	—— (——)	0.024 (0.004)	0.008 (0.011)
ϵ_1	—— (——)	—— (——)	0.024 (0.004)	0.008 (0.011)
ν	—— (——)	2.741 (0.349)	—— (——)	4.662 (3.451)
$(\epsilon_1, \epsilon_2, \sigma^2) = (0.05, 0.05, 0)$				
β_0	-0.004 (0.015)	-0.005 (0.032)	-0.003 (0.031)	0.000 (0.036)
β_1	0.535 (0.008)	1.116 (0.077)	0.894 (0.036)	1.008 (0.085)
ϵ_0	—— (——)	—— (——)	0.065 (0.006)	0.048 (0.017)
ϵ_1	—— (——)	—— (——)	0.061 (0.006)	0.045 (0.017)
ν	—— (——)	1.316 (0.124)	—— (——)	4.762 (4.983)
$(\epsilon_1, \epsilon_2, \sigma^2) = (0.1, 0.1, 0)$				
β_0	0.005 (0.015)	-0.005 (0.039)	0.001 (0.038)	0.000 (0.044)
β_1	0.427 (0.007)	1.166 (0.109)	0.907 (0.048)	1.022 (0.104)
ϵ_0	—— (——)	—— (——)	0.112 (0.007)	0.095 (0.021)
ϵ_1	—— (——)	—— (——)	0.114 (0.007)	0.098 (0.021)
ν	—— (——)	0.781 (0.072)	—— (——)	4.695 (6.622)

In Table 2, we present simulation results based on data generated to incorporate measurement error. The models utilizing the Student-t distribution stand out, demonstrating superior performance with the smallest biases. Notably, as we increase σ^2, all models exhibit a noticeable rise in bias in estimating β_1. This trend is similarly observed concerning ν when employing the Student-t distribution.

Table 2.
The mean and SE of model parameters for M1 and M3 are calculated from 500 Monte Carlo samples, each comprising 10, 000 observations. The datasets were generated to include measurement error and exclude misclassification in three distinct scenarios.

θ	N	T	N	T
	M1		M3	
$(\epsilon_1, \epsilon_2, \sigma^2) = (0, 0, 0.01)$				
β_0	-0.002 (0.018)	-0.001 (0.024)	-0.001 (0.018)	-0.001 (0.024)
β_1	0.747 (0.012)	0.987 (0.047)	0.749 (0.013)	0.994 (0.048)
ϵ_0	— (—)	— (—)	— (—)	— (—)
ϵ_1	— (—)	— (—)	— (—)	— (—)
ν	— (—)	4.252 (0.730)	— (—)	4.228 (0.726)
$(\epsilon_1, \epsilon_2, \sigma^2) = (0, 0, 0.1)$				
β_0	0.000 (0.017)	0.002 (0.023)	0.000 (0.018)	0.000 (0.024)
β_1	0.717 (0.012)	0.916 (0.043)	0.733 (0.013)	0.956 (0.049)
ϵ_0	— (—)	— (—)	— (—)	— (—)
ϵ_1	— (—)	— (—)	— (—)	— (—)
ν	— (—)	4.651 (0.897)	— (—)	4.631 (0.895)
$(\epsilon_1, \epsilon_2, \sigma^2) = (0, 0, 0.5)$				
β_0	-0.002 (0.017)	0.000 (0.020)	0.000 (0.019)	0.000 (0.024)
β_1	0.619 (0.010)	0.741 (0.035)	0.687 (0.014)	0.871 (0.057)
ϵ_0	— (—)	— (—)	— (—)	— (—)
ϵ_1	— (—)	— (—)	— (—)	— (—)
ν	— (—)	5.697 (1.636)	— (—)	5.735 (1.640)

Table 3.
The mean and SE of parameters for models M1, M2, M3, and M4 are computed from 500 Monte Carlo samples, each with a size of 10, 000, generated to incorporate both measurement error and misclassification.

θ	N	T	N	T
$(\epsilon_1, \epsilon_2, \sigma^2) = (0.05, 0.05, 0.1)$				
	M1		M2	
β_0	0.005 (0.016)	-0.003 (0.030)	0.002 (0.032)	-0.002 (0.036)
β_1	0.523 (0.008)	1.002 (0.068)	0.833 (0.034)	0.929 (0.077)
ϵ_0	— (—)	— (—)	0.062 (0.006)	0.043 (0.020)
ϵ_1	— (—)	— (—)	0.065 (0.006)	0.044 (0.021)
ν	— (—)	1.422 (0.148)	— (—)	4.781 (5.556)
	M3		M4	
β_0	-0.005 (0.016)	0.002 (0.031)	0.000 (0.032)	-0.001 (0.038)
β_1	0.537 (0.009)	1.057 (0.080)	0.867 (0.038)	0.979 (0.090)
ϵ_0	— (—)	— (—)	0.063 (0.006)	0.041 (0.022)
ϵ_1	— (—)	— (—)	0.064 (0.006)	0.042 (0.022)
ν	— (—)	1.421 (0.146)	— (—)	4.738 (5.709)

Table 3 highlights that models utilizing the Student-t distribution generally yield superior results, exhibiting smaller biases across most scenarios. Notably, estimates of the parameter β_1 in models M1 and M3, under the normal distribution, displayed the most significant biases. Additionally, under the Student-t distribution, the estimated values of ν are consistently below 2, emphasizing the requirement for a distribution with heavy tails and rendering adjustments with the normal distribution inappropriate. The M4 model under the Student-t distribution, on the whole, delivered accurate estimates for all parameters.

4. APPLICATION

The dataset under examination in this analysis pertains to survivors of the atomic bombings conducted by the United States on the cities of Hiroshima and Nagasaki in Japan. Those who survived or resided in nearby areas experienced the effects of radiation exposure, leading to health issues, including cancer. The data utilized in this study, as sourced from Sposto et al. (1992), originates from a research initiative commenced 5 years after the atomic bombings. The primary objective of this study was to assess the impact of radiation exposure on cancer-related deaths. The cohort consisted of 86,520 survivors of the attacks, categorized into exposed and non-exposed groups based on their proximity to the bomb blast (< 2km, 2 to 10 km). These survivors were monitored from 1950 to 1985.

Table 4 presents information on radiation exposure dose, mean radiation exposure dose, number of cancer deaths, number of deaths from other causes, and the proportion of cancer deaths among the 31, 037 individuals studied. Measurement errors in radiation doses depend on location and biological reasons, as individuals can absorb different amounts of radiation despite having the same exposure conditions. Radiation exposure dose is measured using dosimetry, which quantifies the radiation doses to which an individual (or living being) may be exposed. Data were collected on various types of cancer, including lung, mouth, intestine, breast, prostate, among others. However, the radiation doses absorbed by the intestine at the time of exposure were selected as the reference dose.

Table 4.
Number of cancer and non-cancer deaths among the atomic bomb survivors in Hiroshima and Nagasaki corresponding to 10 dose categories.

Dose	Mean dose	Cancer deaths	Non-cancer deaths	Proportion
0.00	0.000	2784	10201	0.2144
0.01 - 0.05	0.018	2105	7451	0.2203
0.06 - 0.09	0.072	439	1509	0.2253
0.10 - 0.19	0.137	523	1701	0.2352
0.20 - 0.49	0.324	586	1785	0.2471
0.50 - 0.99	0.693	339	826	0.2910
1.00 - 1.99	1.350	204	369	0.3560
2.00 - 2.99	2.350	57	86	0.3986
3.00 - 3.99	3.520	21	51	0.2917
$\geq$ 4.00	4.430	13	26	0.3611

We applied the four studied models to fit the dataset, considering the cdf of both the Student-t (T) and normal (N) distributions as link functions in each model. For models involving measurement error, the substitute variable wi represents the average dose observed for each category, while the variable xi represents the true dose. Thus, we make the assumption that $X_i \sim T(\omega_i, \sigma_i^2, v)$, where $\sigma_i^2 = c\omega_i^2$ and c=0.5 (Roy et al., 2005), i=1,…,31,037. The focal point of this application is to assess the adequacy of the M4 model with Student-t in describing the data. Table 5 provides the estimated values of the parameters for models M1, M2, M3, and, M4 along with the corresponding SEs and the p-values obtained through Wald statistics.

Table 5.
Model comparison: parameter estimates, SEs, p-values, AIC and BIC criteria for M1, M2, M3, and M4 fitted to data from atomic bomb survivor in Hiroshima and Nagasaki, Japan.

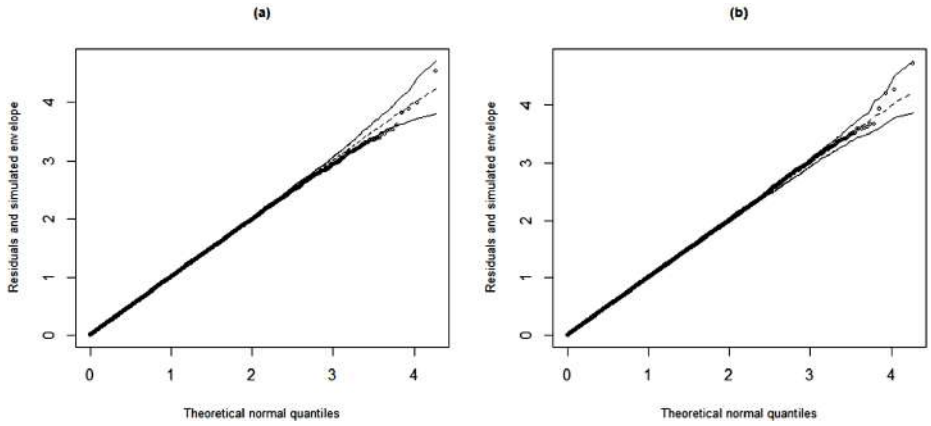
	N		T		N		T	
	M1				M2			
θ	Estim. (S.E.)	p-value	Estim. (S.E.)	p-value	Estim. (S.E.)	p-value	Estim. (S.E.)	p-value
β_0	-0.775 (0.008)	<0.001	-0.775 (0.008)	<0.001	-0.054 (0.229)	0.008	-0.594 (0.209)	0.002
β_1	0.225 (0.023)	<0.001	0.226 (0.022)	<0.001	0.282 (0.034)	<0.001	0.309 (0.006)	<0.001
ϵ_0	—(—)	—	—(—)	—	0.000 (0.093)	0.500	0.000 (0.091)	0.500
ϵ_1	—(—)	—	—(—)	—	0.250 (0.159)	0.058	0.250 (0.115)	0.015
ν	—(—)	—	>100 (356)	0.390	—(—)	—	4.110 (4.552)	0.183
AIC	33188		33190		33189		33192	
BIC	33205		33215		33223		33234	
	M3				M4			
β_0	-0.775 (0.008)	<0.001	-0.779 (0.017)	<0.001	-0.607 (0.289)	0.018	-0.727 (0.296)	0.007
β_1	0.229 (0.024)	<0.001	0.230 (0.025)	<0.001	0.284 (0.046)	<0.001	0.357 (0.021)	<0.001
ϵ_0	—(—)	—	—(—)	—	0.001 (0.096)	0.495	0.001 (0.116)	0.496
ϵ_1	—(—)	—	—(—)	—	0.021 (0.144)	0.083	0.200 (0.089)	0.012
ν	—(—)	—	>100 (489)	0.419	—(—)	—	1.095 (2.101)	0.164
AIC	33187		33189		33187		33190	
BIC	33204		33214		33221		33232	

In the results presented in Table 5, we observed that the estimated values of parameters for models M1 and M3 are notably similar. The Wald statistics indicate that the estimated degrees of freedom under the Student-t distribution are not significant; they are, in fact, zero, suggesting a distribution with heavy tails for a more appropriate fit. In the case of the M2 model, the estimated value of ϵ_1 is not significant when considering the normal distribution at a 0.05 significance level, but it becomes significant with the Student-t distribution. Similarly, for the M4 model with the Student-t distribution, ϵ_1 is significant, but the degree of freedom, as per the Wald test, is not, indicating the necessity for an adjustment with a distribution featuring heavy tails.

According to the AIC criterion, the most suitable models are M3 and M4 under the normal distribution. Notably, the M3 model with the normal distribution also presents the lowest BIC. Given the observed existence of classification and measurement errors alongside a distribution with heavy tails, we propose the use of the M4 model for this application.

In Figure 1, we present graphs featuring randomized quantile residuals, as proposed by Dunn and Smyth (1996). These residuals tend to converge to the standard normal distribution when the model parameters are estimated consistently (Pereira & Russo, 2019). Additionally, we include their simulated confidence bands, constructed at a 0.95 confidence level using the hnp (de Andrade et al., 2017) library, for both the fit M1 (N) and the more appropriate fit M4 (T). Notably, the simulated envelope graph under the M1 model and the normal distribution appears unsuitable for the dataset. In such cases, the preferred option is adjusting with the M4 model and the Student-t distribution, characterized by heavy tails and accounting for both measurement error and misclassification.

Figure 1.
Simulated envelopes for the randomized quantile residuals: (a) M1 normal (b) M4 Student-t



5. CONCLUDING REMARKS

In this study, we adopt the assumption that the measurement error in the covariate adheres to a Student-t distribution, and the binary response is subject to misclassification. Utilizing the cdf of the Student-t distribution as a link function, the M4 model presented in this article proves valuable for modeling the mean of a binary response with both classification and measurement errors in the covariate. This model is particularly suited for adjusting data where the measurement error does not follow a normal distribution.

Parameter estimation was performed using the maximum likelihood method with the R software (R Core Team, 2021) and the optimx (Nash & Varadhan, 2011) library.

Simulations demonstrate the superiority of models considering some form of error and employing the Student-t distribution compared to the M1 model and models with the normal distribution, especially when ϵ_0,ϵ_2, and σ^2 are involved. Generally, models M2 and M4 exhibit higher SEs due to their greater number of parameters.

In the final application, we employed data concerning the health effects on survivors of the atomic bombings in Hiroshima and Nagasaki in 1945, revealing a Berkson-type measurement error with a distribution featuring heavy tails. Among the models studied, it was observed that M1 and M3 provide similar estimates of coefficients β_0 and β_1. However, the M4 model with the Student-t distribution yields notably different parameter estimates from the other models. Importantly, considering measurement error and misclassification observed in the data, the M4 model with the Student-t distribution emerges as the most suitable choice, supported by Wald statistics and simulated envelope graphs.

6. REFERENCES

Bazán, J. L., Romeo, J. S., & Rodrigues, J. (2014). Bayesian skew-probit regression for binary response data. *Brazilian Journal of Probability and Statistics, 28*(4), 467-482. https://doi.org/10.1214/13-BJPS218

Bolfarine, H., & Lachos, V. H. (2006). Skew binary regression with measurement errors. *Statistics, 40*(6), 485-494. https://doi.org/10.1080/02331880600589270

Branco, M. D., & Dey, D. K. (2001). A general class of multivariate skew-elliptical distributions. *Journal of Multivariate Analysis, 79*(1), 99-113. https://doi.org/10.1006/jmva.2000.1960

Burr, D. (1988). On errors-in-variables in binary regression—Berkson case. *Journal of the American Statistical Association, 83*(403), 739-743. https://doi.org/10.1080/01621459.1988.10478656

Carroll, R. J., Spiegelman, C. H., Lan, K. G., Bailey, K. T., & Abbott, R. D. (1984). On errors-in-variables for binary regression models. *Biometrika, 71*(1), 19-25. https://doi.org/10.1093/biomet/71.1.19

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC. https://doi.org/10.1201/9781420010138

de Andrade Moral, R., Hinde, J., & García Borges Demétrio, C. (2017). Half-normal plots and overdispersed models in R: the hnp package. *Journal of Statistical Software, 81*(10). https://doi.org/10.18637/jss.v081.i10

Dunn, P. K., & Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and graphical statistics, 5*(3), 236-244. https://doi.org/10.1080/10618600.1996.10474708

Ekholm, A., & Palmgren, J. (1982). A model for a binary response with misclassifications. In *GLIM 82: Proceedings of the international conference on generalised linear models* (pp. 128-143). Springer New York. https://doi.org/10.1007/978-1-4612-5771-4_13

Kannel, W. B., & Gordon, T. (1968). *The Framingham Study: an epidemiological investigation of cardiovascular disease*. United States. Department of Health, Education, and Welfare, National Institutes of Health.

Lange, K. L., Little, R. J., & Taylor, J. M. (1989). Robust statistical modeling using the t distribution. *Journal of the American Statistical Association, 84*(408), 881-896. https://doi.org/10.1080/01621459.1989.10478852

Lin, P. E. (1972). Some characterizations of the multivariate t distribution. *Journal of Multivariate Analysis, 2*(3), 339-344. https://doi.org/10.1016/0047-259X(72)90021-8

Liu, H., & Zhang, Z. (2017). Logistic regression with misclassification in binary outcome variables: a method and software. *Behaviormetrika, 44*(2), 447-476. https://doi.org/10.1007/s41237-017-0031-y

Nash, J. C., & Varadhan, R. (2011). Unifying optimization algorithms to aid software system users: optimx for R. *Journal of Statistical Software, 43*, 1-14. https://doi.org/10.18637/jss.v043.i09

Pereira, M. A. A., & Russo, C. M. (2019). Nonlinear mixed-effects models with scale mixture of skew-normal distributions. *Journal of Applied Statistics, 46*(9), 1602-1620. https://doi.org/10.1080/02664763.2018.1557122

R Core Team, R. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Roy, S., Banerjee, T., & Maiti, T. (2005). Measurement error model for misclassified binary responses. *Statistics in medicine, 24*(2), 269-283. ttps://doi.org/10.1002/sim.1886

Sposto, R., Preston, D. L., Shimizu, Y., & Mabuchi, K. (1992). The effect of diagnostic misclassification on non-cancer and cancer mortality dose response in A-bomb survivors. *Biometrics, 48*(2), 605-617. https://www.jstor.org/stable/2532315