Two-part model with measurement error
Modello a due parti con errore di misura

Maria Felice Arezzo, Serena Arima, and Giuseppina Guagnano

Abstract In many applications, there are positive-valued phenomena which show a very high frequency at zero. One major difficulty with this type of data is that the existence of a point-mass at zero makes common distributions unsuited for modeling the data. To cope with these difficulties, some models have been developed. A popular example is the two-part model in which two stochastic models are assumed: the first governs whether the response variable is zero or positive an the second, conditional on its being positive, models the level. We extend the two-part model to cope with measurement error on the dependent variables of both stochastic parts. This situation is common in many applied works.

Abstract *In molte applicazioni la variabile di interesse assume valori positivi con una frequenza molto alta di valori nulli. Una delle principali difficoltà con questo tipo di dati è che l'esistenza di una massa a zero rende le distribuzioni comuni inadatte per la modellazione dei dati. Per far fronte a queste difficoltà, sono stati sviluppati alcuni modelli. Un esempio è il modello in due parti in cui vengono assunti due modelli stocastici: il primo determina se la variabile di risposta è zero o positiva e il secondo, condizionatamente al primo, ne modella il livello. Estendiamo il modello a due parti tenendo conto dell'errore di misura sulle variabili dipendenti di entrambi i modelli stocastici. Questa situazione è comune in molti lavori applicati.*

Key words: Two-part model, Measurement error in the dependent variable

Maria Felice Arezzo
Department MEMOTEF, Sapienza University of Rome e-mail: mariafelice.arezzo@uniroma1.it

Serena Arima
Department DSSSU University of Salento, Lecce e-mail: name@email.address

Giuseppina Guagnano
Department MEMOTEF, Sapienza University of Rome e-mail: giuseppina.guagnano@uniroma1.it

1 Introduction

In many fields, real phenomena with positive values often show a very high frequency at zero. This kind of data can be represented by semi-continuous variables, which are combination of a point-mass at zero and a positive skewed distribution.

One major difficulty with this type of data is that the existence of a point-mass at zero makes common distributions, such as the gamma, unsuited for modeling the data. To cope with these difficulties, the two-part model has been proposed. In it, two stochastic models are assumed: the first, by means of an additional binary variable, governs whether the response variable is zero or positive and the second, conditional on its being positive, models the level.

Also in many applied works, variables are flawed with measurement error. This could easily happen, for example, during an interview if the respondent misunderstands the question.

We extended the two part model to consider two types of measurement errors: the first affects the binary variable that governs whether the response variable is zero or positive, and the second is on the positive part of the response variable.

In the literature, a mismeasurement on a continuous variable is called measurement error while it is called misclassification when it affects a categorical variable. When the fallible variable is continuous, the two dominant error models are the Berkson's [1] and the classical [3]. In the first one, the error-prone observed value is fixed while the true unobservable variable is random and its random structure is specified conditionally on the former. In the classical approach the error-prone variable is specified as a function of the true one with the error component inserted in a multiplicative or additive form and independent from the true variable.

Let us introduce some notation on the measurement error models used in our work. Let Y^O be the fallible/error-prone binary variable and y^O be the observed value. The misclassification model, which specifies the behaviour of Y^O given the true unobserved value $Y^T = y^T$, is characterized by the misclassification probability:

$$P(Y^O = y^O | Y^T = y^T). \tag{1}$$

Following [4], we set the misclassification probabilities $\alpha_1 = P(Y^O = 0|Y^T = 1)$ (the probability of false negative) and $\alpha_0 = P(Y^O = 1|Y^T = 0)$. (the probability of false positive). Since Y^T is random, if we specify the distribution of $Y^O|Y^T$, it follows that:

$$P(Y_i^O = 1) = (1-\alpha_1)\pi_i + \alpha_0(1-\pi_i) = \pi_i(1-\alpha_0-\alpha_1) + \alpha_0 \tag{2}$$

where $\pi_i = P(Y_i^T = 1)$. Such a probability can be estimated as a function of covariates through a generalized linear model.
When we deal with a continuous variables W, the classical error model in the multiplicative and additive form respectively is:

$$W_i^O = \begin{cases} W_i^T \cdot \xi_i, & \xi_i \sim \log N(\mu, \sigma_\xi^2) \\ W_i^T + \xi_i, & \xi_i \sim N(\mu, \sigma_\xi^2) \end{cases} \qquad \text{with } \mu = 0 \tag{3}$$

where μ is usually null and W^T (or its logarithm) can be specified as a linear function of some predictors. In the proposed model, see section (2), we generalize the distribution of ξ_i, admitting $\mu \neq 0$ and specific for each unit.

2 The proposed model

Let us consider a semi-continuous random variable W, whose observability depends on a binary variable Y: when $Y_i = 1$, we observe a positive value for W_i; otherwise, when $Y_i = 0$, we have $W_i = 0$. Referring to the two-part model: in the first part, we have to specify a binary choice model for the probability of observing a positive-versus-zero outcome and then, in the second part, a regression model is fit for the positive outcome *conditional* on a positive outcome. Under this framework, let us initially assume that there is no measurement error in any of the response variables (the true W and Y coincide with the observable ones). Let us denote themwith W^T and Y^T, the probability of a positive response as $P(W_i^T > 0|\boldsymbol{Z}_i) = P(Y_i^T = 1|\boldsymbol{Z}_i) = \pi_i$, and the conditional distribution of the positive responses as $g(W_i^T|W_i^T > 0, \boldsymbol{X}_i)$, where Z and X are two sets of possibly overlapping explanatory variables.

The two-part model has the following mixture p.d.f. [2] and likelihood:

$$f\left(W_i^T\right) = (1-\pi_i)I(Y_i^T = 0) + \pi_i g(W_i^T|Y_i^T = 1, \boldsymbol{X}_i) \tag{4}$$

$$L(\beta,\theta) = \underbrace{\left[\prod_{Y^T=0}(1-\pi_i)\cdot\prod_{Y^T=1}\pi_i\right]}_{L_1(\boldsymbol{\beta})}\cdot\underbrace{\left[\prod_{Y^T=1}g(W_i^T|Y_i^T=1,\boldsymbol{X}_i)\right]}_{L_2(\boldsymbol{\theta})} \tag{5}$$

where β and θ are vectors of parameters that govern the binary and the continuous part respectively, and $I(Y_i^T = 0)$ is an indicator function such that it equals 1 if $Y_i^T = 0$ and 0 otherwise; it is motivated by the fact that when $Y_i^T = 0$, the density of W^T collapses to a unit probability mass.

More precisely, $L_1(\beta)$ is the likelihood of a standard binary regression model and the corresponding link function is usually specified as probit or logit. $L_2(\theta)$ refers to the regression model for the continuous variable W^T, usually involving gamma or log-Normal distributions. We model L_1 with a probit link and L_2 as a log-Normal regression dealing with the following two-part model:

$$\text{Part one:} \qquad P(W_i^T > 0|\boldsymbol{Z}_i) = P(Y_i^T = 1|\boldsymbol{Z}_i) = \pi_i = \Phi(\boldsymbol{Z}_i\boldsymbol{\beta}) \tag{6}$$

$$\text{Part two:} \qquad \log(W_i^T) = \boldsymbol{X}_i\boldsymbol{\theta} + u_i, \quad u_i \sim N(0;\sigma_u^2) \tag{7}$$

where $\Phi(\cdot)$ is the standard normal c.d.f.

Suppose, now, that both variables W and Y may be affected by measurement errors, so we have the observable Y_i^O and W_i^O and the true Y_i^T and W_i^T. We assume that the measurement error that affects the observable W^O acts in a multiplicative way as in the first line of equation (3), so that $\log(W^O) = \log(W^T) + \varepsilon$, with $\varepsilon = \log(\xi)$. When we admit the possibility of measurement error for W and Y, we can no longer refer only to the p.d.f. of the true W as in (4), but we need to consider the observability of W^O and define its p.d.f.:

$$f\left(\log(W_i^O)\right) = \sum_{j=1}^{4} \psi_{ji} \cdot g_j\left(\log(W_i^O)\right) \tag{8}$$

where the weights are defined as:

$$\begin{aligned}
\psi_{1i} &= P(Y_i^O = 0, Y_i^T = 0) = (1-\alpha_0)\cdot(1-\pi_i) \\
\psi_{2i} &= P(Y_i^O = 1, Y_i^T = 1) = (1-\alpha_1)\cdot\pi_i \\
\psi_{3i} &= P(Y_i^O = 1, Y_i^T = 0) = \alpha_0\cdot(1-\pi_i) \\
\psi_{4i} &= P(Y_i^O = 0, Y_i^T = 1) = \alpha_1\cdot\pi_i
\end{aligned} \tag{9}$$

and the conditional densities in equation (8) are:

$$\begin{aligned}
g_1\left(\log(W_i^O)\right) &= f\left(\log(W_i^O)|Y_i^O = 0, Y_i^T = 0\right) = 1 \\
g_2\left(\log(W_i^O)\right) &= f\left(\log(W_i^O)|Y_i^O = 1, Y_i^T = 1, \boldsymbol{X}_i\right) \\
g_3\left(\log(W_i^O)\right) &= f\left(\log(W_i^O)|Y_i^O = 1, Y_i^T = 0\right) \\
g_4\left(\log(W_i^O)\right) &= f\left(\log(W_i^O)|Y_i^O = 0, Y_i^T = 1\right) = 1
\end{aligned} \tag{10}$$

The density g_2 represents the main contribution in explaining $\log(W^O)$, but its weight ψ_2 tends to zero as α_1 goes to 1.

The density g_3 does not depend on the covariates $\boldsymbol{X}$ because $W_i^T = 0$ when $Y_i^T = 0$; hence it only refers to the erratic component ε. Its weight ψ_3 increases as α_0 gets higher. As a last step, we model Y_i^O and $\log(W_i^O)$ as:

$$P\left(Y_i^O = 1|Z_i\right) = \alpha_0 + (1-\alpha_0-\alpha_1)P\left(Y_i^T|Z_i\right) \tag{11}$$

$$\log(W_i^O) = \boldsymbol{X}_i\boldsymbol{\theta} + (u_i + \varepsilon_i) \tag{12}$$

where $\boldsymbol{X}_i$ is the row vector containing all information for the i-th individual. The first part of the model (i.e. equation 11) is consistent with equation (2). For the second part, coherently with (7), we assume a normal distribution for ε, $\varepsilon_i \sim N(\mu_i, \sigma_\varepsilon^2)$, and consequently for the global error component $u_i + \varepsilon_i = v_i \sim N(\mu_i, \sigma_v^2)$. Furthermore, we assume that u_i and ε_i are uncorrelated. It's important to stress that the above specification extends the classical measurement error model, allowing each unit to

have a different expected value μ_i. In other words, the measurement error may act with a different intensity for each population unit. We model the expected value as a function of individual characteristics: $\mu_i = h(\boldsymbol{X}_i^* \boldsymbol{\gamma})$, with $\boldsymbol{X}_i^*$ row vector. For the sake of simplicity, we just consider a linear function $\mu_i = \boldsymbol{X}_i^* \boldsymbol{\gamma}$. Admitting a varying μ_i implies that the conditional densities g_2 and g_3 must be conditioned to $\boldsymbol{X}_i^*$. Since $\log(W_i^O)$ and μ_i are both specified as linear functions of the predictors, to avoid any problem of identifiability of the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$, we assume that the sets of covariates X and X^* do not overlap.

The contribution of the $i-th$ unit to the likelihood is:

$$L_i(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \alpha_0, \alpha_1) = \{(1-\alpha_0) \cdot (1-\pi_i) + \alpha_1 \cdot \pi_i\}^{(1-y_i^O)} \cdot \left\{(1-\alpha_1) \cdot \pi_i \cdot N(\boldsymbol{X}_i \boldsymbol{\theta} + \boldsymbol{X}_i^* \boldsymbol{\gamma}; \sigma_v^2) + \alpha_0 \cdot (1-\pi_i) \cdot N(\boldsymbol{X}_i^* \boldsymbol{\gamma}; \sigma_\varepsilon^2)\right\}^{y_i^O} \tag{13}$$

where $N(\boldsymbol{X}_i \boldsymbol{\theta} + \boldsymbol{X}_i^* \boldsymbol{\gamma}; \sigma_v^2)$ and $N(\boldsymbol{X}_i^* \boldsymbol{\gamma}; \sigma_\varepsilon^2)$ are the densities g_2 and g_3 in equation (10) respectively. In the following, for the sake of brevity and to highlight the dependency on the parameters, we denote them as $g_2(\boldsymbol{\theta}, \boldsymbol{\gamma})$ and $g_3(\boldsymbol{\gamma})$.

3 Simulation study

We present the finite sample performances of the proposed model and compare them to the classical probit/ols two part model via Monte Carlo simulations. We assumed the following generating model for the error-free dependent variables:

$$\Pr\left(Y^T | Z_1, Z_2, Z_3\right) = \Phi\left(\beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3\right) \tag{14}$$

$$\log W^T = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + u \tag{15}$$

where $\Phi(\cdot)$ is the c.d.f of a standard normal. The covariates are generated as follows: Z_1 is log-normal with zero mean and unit variance, X_2 and Z_2 are binomial with $p = 1/3$, X_1 and Z_3 are uniformly distributed over the unit interval. To generate the observed (i.e. error-prone) binary variable, Y^O, we define the misclassification matrix based on equation (1) and we sample accordingly. Finally, the mismeasured continuous part is generated as in equation 12 allowing $\mu_i = \gamma X_4$, with $X_4 \sim Mult_4(p_1 = 0.01; p_2 = 0.06; p_3 = 0.33, p_4 = 0.60)$.

Across simulations we fixed: $\theta^T = (10, 0.8, -0.5)$, $\beta^T = (-1, 0.2, 1.5, -0.6)$, $\sigma_u^2 = 2$ and set the remaining parameters according three scenarios: 1) $\alpha_0 = \alpha_1 = 0.05; \sigma_\varepsilon^2 = 3; \gamma = -0.2$, 2) $\alpha_0 = 0.05; \alpha_1 = 0.20; \sigma_\varepsilon^2 = 3; \gamma = -0.2$ and 3) $\alpha_0 = \alpha_1 = 0.20; \sigma_\varepsilon^2 = 3; \gamma = -0.2$. We repeat each simulation scenario 100 times with samples of size $n = 5{,}000$. In table 1 we report the results.

For the binary part of the model, even in the case of a small amount of misclassification ($\alpha_0 = \alpha_1 = 0.05$), ordinary probit produces estimates that are biased by 14-22%. As expected, the problem worsens as the amount of misclassification

Table 1 Empirical mean and standard errors, over 100 simulated data sets, of the parameter estimates based on the proposed model and the two-part model.

	Scenario 1			Scenario 2			Scenario 3		
	True value	Proposed model	Two-part model	True value	Proposed model	Two-part model	True value	Proposed model	Two-part model
θ_0	10	9.837	0.881	10	9.854	0.738	10	9.937	0.643
		0.020	*0.004*		*0.021*	*0.004*		*0.014*	*0.004*
θ_1	0.8	0.805	11.138	0.8	0.795	9.394	0.8	0.781	9.521
		0.019	*0.028*		*0.021*	*0.030*		*0.021*	*0.031*
θ_2	-0.5	-0.493	0.155	-0.5	-0.494	0.128	-0.5	-0.486	0.136
		0.003	*0.011*		*0.003*	*0.010*		*0.003*	*0.010*
γ	-0.2	-0.151		-0.2	-0.153		-0.2	-0.172	
		0.005			*0.005*			*0.002*	
σ_v^2	5	4.731	8.371	5	4.708	9.707	5	4.468	10.138
		0.017	*0.032*		*0.019*	*0.030*		*0.018*	*0.030*
β_0	-1	-1.023	-0.843	-1	-1.028	-0.914	-1	-1.003	-0.515
		0.014	*0.005*		*0.018*	*0.005*		*0.035*	*0.004*
β_1	0.2	0.201	0.157	0.2	0.201	0.119	0.2	0.196	0.086
		0.003	*0.002*		*0.004*	*0.001*		*0.006*	*0.001*
β_2	1.5	1.529	1.293	1.5	1.516	1.093	1.5	1.464	0.795
		0.016	*0.004*		*0.023*	*0.004*		*0.035*	*0.004*
β_3	-0.6	-0.626	-0.495	-0.6	-0.619	-0.411	-0.6	-0.578	-0.272
		0.012	*0.007*		*0.015*	*0.007*		*0.019*	*0.006*
α_0	0.05	0.054		0.05	0.050		0.2	0.178	
		0.003			*0.004*			*0.007*	
α_1	0.05	0.046		0.2	0.185		0.2	0.173	
		0.003			*0.006*			*0.006*	

Note: The standard error of the simulation results are reported in italic.

grows. Conversely, the proposed model provides more accurate estimates, in terms of mean squared errors, for all levels of misclassification.

For the continuous part, the results of the proposed model are very encouraging since the estimates of all parameters are trustworthy. These results hold for all simulations scenarios (all tables are available upon request). Although satisfactory, the estimates of γ and σ_v^2 showed some variability in the accuracy when $\sigma_\varepsilon^2 = 3$.

References

1. Berkson, J.:Are there Two Regressions? Journal of the American Statistical Association. **45**,164–180 (1950).
2. Cragg, J. G.: Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods. Econometrica. **39**, 829–844 (1971).
3. Fuller, W.: Measurement Error Model. Wiley (1988).
4. Hausman, J., Abrevaya, J., Scott-Morton, F. M.: Misclassification of the dependent variable in a discrete-response setting. Journal of Econometrics. **87**, 239–269 (1998).