Models for limited dependent variables with measurement errors

Maria Felice Arezzo, Serena Arima, Giuseppina Guagnano

XII Giornata della Ricerca MEMOTEF
1 giugno 2022

Introduction

Initial aim of the research:

to estimate the individual propensity to work off-the-books and the earnings deriving from these undeclared activities.

Methodological issues

- earnings may be represented by a semi-continuos variable, which can only be observed for individuals who declared to be working off the books;
- the variable representing the condition of working or not working in the black economy may be misclassified, i.e., one may answer no, while s/he actually worked off-the-book. We expect the other type of misclassification, that is, one may declare to have worked off-the-book while s/he haven't, to be very unlikely;
- earnings may be under-declared;
- data are severely unbalanced, that is the number of events (individuals declaring to work off-the-book) is significantly smaller than the number of nonevents: the percentage of cases in the survey is only 5%.

Solution to the first issue

Two-part model:

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

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

where W is a semi-continuous random variable.

It is useful for phenomena with positive values and a very high frequency at zero.

The two-part model has the following mixture p.d.f.:

$$f\left(W_i^T\right) = (1-\pi_i)g(W_i^T|Y_i^T = 0, \boldsymbol{X}_i) + \pi_i g(W_i^T|Y_i^T = 1, \boldsymbol{X}_i) \tag{3}$$

where $\pi_i = P(Y_i^T = 1)$ may be explained by a set of covariates, so depending on the corresponding vector of parameters, say β, and $g(W_i^T|Y_,^T \boldsymbol{X}_i)$ depends on another vector of parameters, say θ. Note that, when $Y_i^T = 0$, the density of W^T collapses to a unit probability mass.

Measurement error issue

Do we need to worry for measurement error?

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 or the interviewer simply checks the wrong box. If the misclassification is on the variable that governs whether the response variable is zero or positive, some truly positive values are recorded as zeros and some truly zeros are recorded as positive.

Whereas there is a vast literature on measurement error on the right side (independent variables), less attention was given on the left side (dependent variables). Exceptions are the influential works of Berkson (JASA, 1950) and Fuller (Wiley, 1988) for continuous variables and of Hausman et al. (Journal of Econometrics 1998) for binary variables.

Measurement error on the binary dependent variables

Hausman et al.:
Let Y^O be the fallible/error-prone binary variable and y^O be the observed value. The misclassification model, which specifies the behavior of Y^O given the true unobserved value $Y^T = y^T$, is characterized by the misclassification probability:
$\alpha_1 = P(Y^O = 0|Y^T = 1)$ (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:

$$P(Y_i^O = 1) = (1 - \alpha_1)\pi_i + \alpha_0(1 - \pi_i) = \pi_i(1 - \alpha_0 - \alpha_1) + \alpha_0 \quad (4)$$

Measurement error on the positive part of the continuous dependent variables

FULLER:

When we deal with a continuous variables W, the classical error model is:

$$W_i^O = W_i^T \cdot \xi_i, \quad \xi_i \sim \log N(\mu, \sigma_\xi^2) \tag{5}$$

where μ is usually null and W^T (or its logarithm) can be specified as a linear function of some predictors.

Rare events

Rare events data, or imbalanced data, have attracted a lot of attentions in machine learning and other quantitative fields.
A common practice in analyzing rare events data is to over-sample (replicate) the events, which has a higher computational cost.
On the other side, Wang (2020) suggested to keep fixed the ones and to consider a subsample of zeros: he proved that under-sampling a small proportion of the nonevents, the resulting under-sampled estimator may have identical asymptotic distribution to the full data MLE distribution.

Our proposal

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 (3), but we need to consider the observability of W and of Y:

$$P\left(Y_i^O = 1|\boldsymbol{Z}_i\right) = \alpha_0 + (1 - \alpha_0 - \alpha_1)P\left(Y_i^T|\boldsymbol{Z}_i\right) \quad (6)$$

$$\log(W_i^O|W_i^T > 0) = \boldsymbol{X}_i\boldsymbol{\theta} + (u_i + \epsilon_i) \quad (7)$$

where $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are the row vectors containing all information for the i-th individual.

The first part of the model (i.e. equation 6) is consistent with equation (4).
For the second part, coherently with (2), we assume a normal distribution for ϵ, $\epsilon_i \sim N(\mu_i, \sigma^2_\epsilon)$, and consequently for the global error component $u_i + \epsilon_i = v_i \sim N(\mu_i, \sigma^2_v)$.
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.

Referring to the whole sample space, the marginal density of $\log(W_i^O)$ can be so defined:

$$f\left(\log(W_i^O)\right) = \sum_{j=1}^{4} \psi_{ji} \cdot g_j\left(\log(W_i^O | Y_i^O, Y_i^T, \boldsymbol{X}_i)\right) \qquad (8)$$

where the weights corresponds to the following joint probabilities:

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

and the conditional values of $Y_i^O$and Y_i^Ti n g_j are the same than in ψ_j.

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: $\mu_i = h(\boldsymbol{X}_i^* \gamma)$, with $\boldsymbol{X}_i^*$ row vector.
For the sake of simplicity, we just consider a linear function $\mu_i = \boldsymbol{X}_i^* \gamma$.
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 γ, we assume that the sets of covariates X and X^* do not overlap.

Simulation results

Coherently with Hausman et al. (1998), we assumed the following generating model for the error-free dependent variables:

$$\Pr\left(Y^T | Z_1, Z_2, Z_3\right) = \Phi\left(\beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3\right) \tag{10}$$

$$\log W^T = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + u \tag{11}$$

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.

Moreover, we considered different under sampling proportions and, for each of them, we investigated the sensitivity of the estimates.

To generate the observed (i.e. error-prone) binary variable, Y^O, we defined the misclassification matrix based on the probabilities α_0 and α_1 and we sample accordingly.
Finally, the mismeasured continuous part is generated as in equation 7 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: $\theta^T = (10, 0.8, -0.5)$,
$\beta^T = (-1, 0.2, 1.5, -0.6)$, $\sigma_u^2 = 2$, $\sigma_\epsilon^2 = 3$ and $\gamma = -0.2$, while for the misclassification probabilities we considered the following three scenarios:
1) $\alpha_0 = \alpha_1 = 0.05$, 2) $\alpha_0 = 0.05; \alpha_1 = 0.20$; 3) $\alpha_0 = \alpha_1 = 0.20$.

We repeated each simulation scenario 100 times with samples of size $n = 5,000$.

The main results can be summarize as follows:
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 grows.

Conversely, the proposed model provides more accurate estimates, in terms of mean squared error and relative bias, 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.

Finally, as to the undersampling, there is a sort of trade-off between estimator behavior and the cross validation:
reducing the proportion of zeros, and consequently the sample size, relative bias and MSE worsen,
at the same time, however, the percentage of correctly predicted events increases.

Main References

Berkson, J.: Are there Two Regressions? Journal of the American Statistical Association. 45,164-180 (1950)

Fuller, W.: Measurement Error Model. Wiley (1988)

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)

Wang, H. (2020): Logistic regression for massive data with rare events.