Tackling Misclassification in Surveys About Undeclared Work Via the EM Algorithm

Maria Felice Arezzo, Giuseppina Guagnano, Domenico Vitale

27 June 2023

Presentation overview

Shadow economy and undeclared work

Shadow economy (SE) consists of all economic activities hidden to the Government. It can include both legitimate and criminal activities.

Undeclared work (UW) on the other side includes any paid work done in legitimate activities but concealed to government to avoid taxation and/or regulations.

Consequences of SE and UW

The costs associated with SE include distortions in the labor market, forgone revenue due to under-reporting of wages and output, sub-optimal provision of public goods, and lower provision of and access to financing.

The costs of UW include lack of control of official authorities over working conditions with obvious implications concerning the health, safety and fairness of the working conditions themselves. Other negative consequences of UW include unfair competition at the expense of legitimate businesses, which then, inevitably, are pressured to evade regulatory standards, and limited public finances available for social cohesion projects

Main focal points

Over the last 50 years or so, the literature on SE and UW focused on:

- Estimation of the size;
- Identification of the determinants;
- Understanding of the motivation and the characteristics of the firms/individuals involved;
- Definition of a theoretical framework for SE and UW;
- Policy tools (from repression *tout-court* to a path toward emersion)

Estimation of the size of SE and UW

Two main approaches to assess the size of SE and UW:

- Macro approaches include the monetary transaction method, the currency demand approach, the income gap method, the electricity approach, the discrepancy method and the MIMIC estimation procedure;
- Micro approaches include surveys with direct or indirect questioning, estimations with tax audit data as well as laboratory and field experiments.

Macro methods tend to overestimate the size of the shadow economy, whereas micro approaches provide lower-bound estimates. Macro approaches are sensitive to their underlying assumptions and calculations.

Macro methods **cannot differentiate** between the shadow economy and undeclared work.

Direct methods with surveys for UW

Direct methods that use surveys are very attractive because it is possible to:

- estimate the size (prevalence of workers/amount of earnings) of UW;
- assess how each determinant shapes UW;
- study the characteristics of the actors involved

Direct methods with surveys for UW

Direct methods that use surveys are very attractive because it is possible to:

- estimate the size (prevalence of workers/amount of earnings) of UW;
- assess how each determinant shapes UW;
- study the characteristics of the actors involved

But the presence of non-responses and/or of the so called Social Desirability Bias (SDB) makes things complicated.

Social desirability bias

Social desirability bias: the answering behavior can be distorted by self-presentation concerns, making respondents omit or lie about socially undesirable decisions.

SDB results in measurement error in the variables used in the statistical model.

The special Eurobarometer surveys on UW

The special Eurobarometer survey is one of the tools through which the European Commission aims to know the main characteristics of undeclared work in order to develop opportune counteracting policy tools.

There are 3 waves so far: 2007, 2013 and 2019.
The survey is nationally representative of the adults aged 15 years or older living in one of the EU-28 countries. Besides the demographics and other controls, it comprises a rich set of questions on undeclared work.

Relevant items in the questionnaire

There are several variables that can be used as response in a statistical model. Some are:

- Do you personally know any people who work without declaring their income or part of their income to tax or social security institutions?
- **Apart from a regular employment, have you yourself carried out any undeclared paid activities in the last 12 months?**
- APPROXIMATELY how much money have you earned from these undeclared activities in the last 12 months?

Envelope wage

- Sometimes employers prefer to pay all or part of the salary or the remuneration (for extra work, overtime hours or the part above a legal minimum) in cash and without declaring it to tax or social security authorities. Has your employer paid you any of your income in the last 12 months in this way?
- Approximately what percentage of your gross yearly income in your main job did you get this way?

SDB in the Eurobarometer survey

None of the three waves of the Eurobarometer survey suffer from a severe non-response in the binary variable that measure participation in UW (roughly around 3% of non response in each wave).

The problem is the SDB. What evidence do we have to support the statement? In the 2013 survey 35.96% of the respondents said they know someone who work off-the-book, but only 4.62% admitted of being personally involved.

How to treat SDB in binary variables

There are several ways of coping with SDB and reduce the measurement error (misclassification from now on) arising from it.

Attenuating misclassification is crucial not only to properly estimate the prevalence of undeclared workers, but also to estimate the association between the outcome variable and covariates.

How to treat SDB in binary variables

The approaches proposed can be divided in 3 categories:

1. Methods that require supplemental data:
   - Neuhaus (1999)
   - Kuchenhoff et al. (2006)
   - Edwards et al. (2013)
2. Methods that require to be used during the survey administration:
   - randomized response techniques;
   - list experiments;
   - information provision experiments;
   - discrete choice experiments
3. Methods that directly specify false positives and false negatives i the likelihood functions:
   - Hausman et al (1998)
   - Liu and Zhang (2017)
   - Arezzo and Guagnano (2019); Arezzo et al. (2023)

Hausman approach is very appealing because it cancels out the bias from misclassification (leading to consistent estimates of model coefficients) and provides an estimate of SDB.

Unfortunately, it does come with limitations:

- coefficients estimates have high variance;
- estimates of SDB depends on initial values;

Misclassification in the (binary) variable of interest means that the *observed* variable of interest is different from the *true* one.

It's fairly straightforward to assume that if $Y_{obs} = 1$ then $Y_{true} = 1$, but when $Y_{obs} = 0$ it can be that $Y_{true} = 1$ or $Y_{true} = 0$.

In other words, the only information that we can rely upon is $Y_{obs} = 1$. Our approach is to label as 1 only the data without misclassification and treat the cases where $Y_{obs} = 0$ as if this information were missing.
In 2009, Ward et al proposed an expectation–maximization algorithm to estimate the underlying presence–absence logistic model for presence-only data.

The EM Algorithm for Presence-Only Data

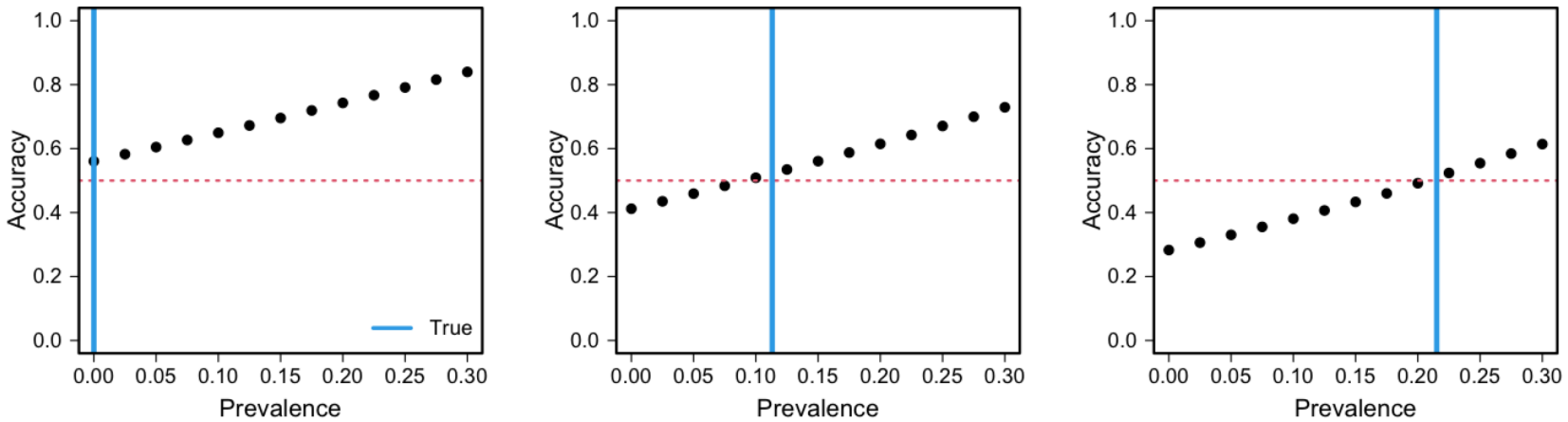
The EM Algorithm for Presence-Only Data is an iterative procedure that impute the unknown y at each iteration and then fits a model using these imputed y. At each iteration we apply the following two steps, until subsequent iterations result in the same model:

1. We replace the unknown y's with our best estimate inline image estimated from the model fit in the previous iteration;
2. We assume that we "know" these y's and thus apply a logit model;

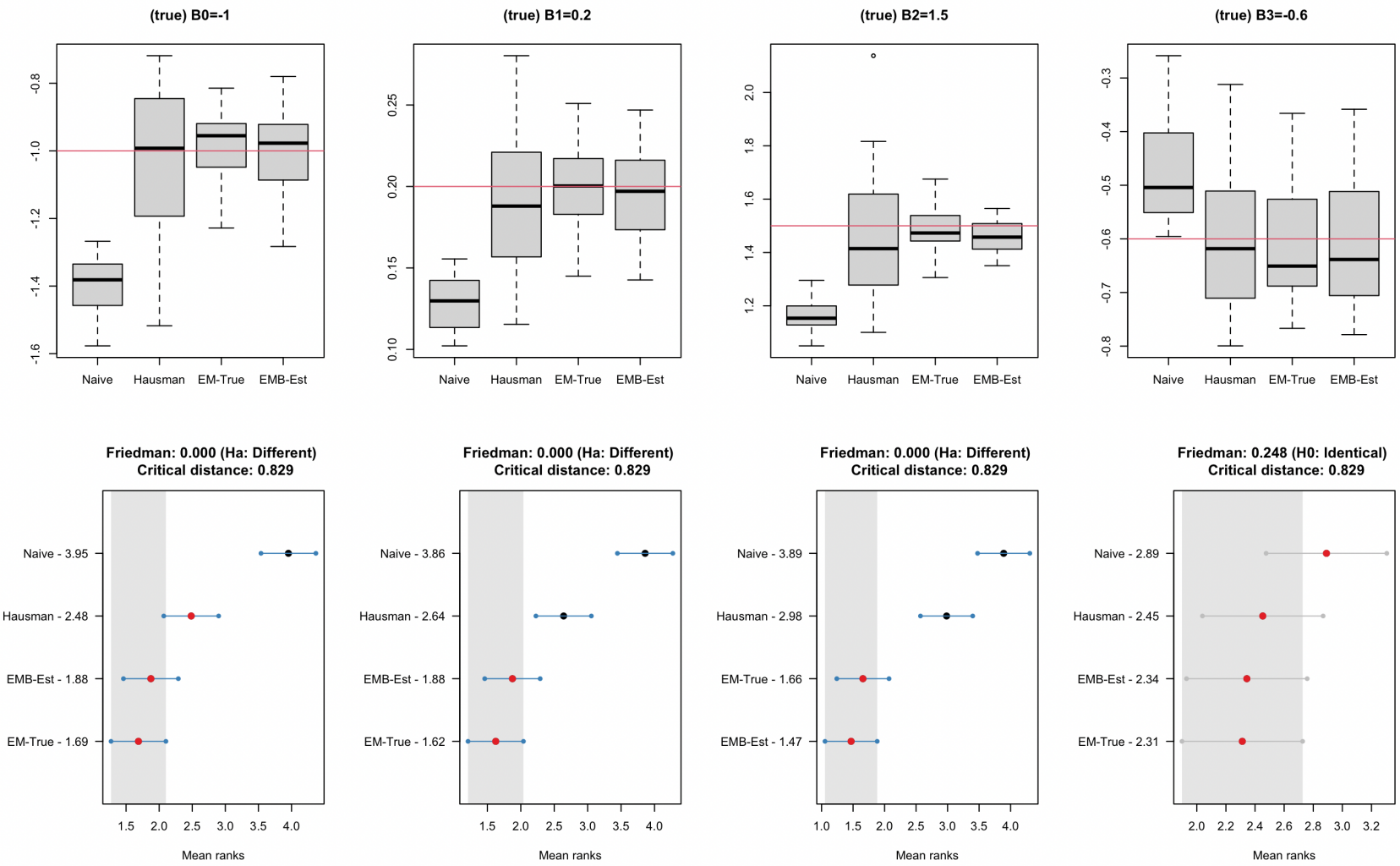
A major problem in the above algorithm is that it requires the knowledge of the prevalence of ones. Using an out-of-sample approach, we provide an estimate of the prevalence of ones as the first value at which model accuracy exceeds 0.5.

Simulation results: prevalence

We conducted a wide simulation study, considering data with different amount of misclassification and different prevalence of ones.



Simulation results: coefficients



Application to the Eurobarometer survey in 2013

Table 1 Summary fits of the Logit model estimated with standard procedure (naive) and through the EM algorithm with $\pi = 0.2$

	EM			Naive		
Regressors	Estimate	Std error	$\Pr(>\|z\|)$	Estimate	Std error	$\Pr(>\|z\|)$
(Intercept)	+2.371	0.153	0.000	-0.680	0.236	0.004
Female	-0.870	0.039	0.000	-0.623	0.064	0.000
Age	-0.035	0.002	0.000	-0.021	0.003	0.000
Tax morale	+0.736	0.014	0.000	+0.369	0.018	0.000
Urban	-0.068	0.040	0.091	-0.086	0.065	0.190
Occupation (Ref. Cat.: Unemployed)						
Self employed	-0.550	0.086	0.000	-0.036	0.126	0.774
Employed	-1.159	0.063	0.000	-0.612	0.092	0.000
Inactive	-0.968	0.074	0.000	-0.532	0.111	0.000
Retired	-1.475	0.084	0.000	-1.034	0.141	0.000
Financial problems (Ref. Cat.: Most of the time)						
Occasional	-0.834	0.058	0.000	-0.503	0.087	0.000
None	-1.441	0.060	0.000	-0.940	0.090	0.000
Detection risk (Ref. Cat,: Very small)						
Very high	-1.185	0.076	0.000	-0.890	0.130	0.000
Fairly high	-1.256	0.058	0.000	-0.877	0.092	0.000

Conclusions

Based on the results obtained in the simulations, we can state that the proposed approach:

- always outperform the naive model (i.e. ignore misclassification);
- it is comparable to Hausman's in term of bias reduction;
- outperforms Hausman's in term of variability of the estimates most of the time;

Based on the results on real data, we can state that the proposed approach:

- is in strong agreement with similar estimates founded in the literature on the topic;
- it shows that some covariates (i.e. tax morale) plays a more important role in involvement in undeclared activities;