DISCUSSION PAPER SERIES

IZA DP No. 6901

Measuring the Shadow Economy: Endogenous Switching Regression with Unobserved Separation

Tomáš Lichard
Jan Hanousek
Randall K. Filer

October 2012

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Measuring the Shadow Economy: Endogenous Switching Regression with Unobserved Separation

Tomáš Lichard
CERGE-EI, Prague

Jan Hanousek
CERGE-EI, Prague

Randall K. Filer
Hunter College, CUNY Graduate Center,
CERGE-EI, IZA and CESifo

Discussion Paper No. 6901
October 2012

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 6901
October 2012

ABSTRACT

Measuring the Shadow Economy: Endogenous Switching Regression with Unobserved Separation*

We develop an estimator of unreported income, perhaps due to tax evasion, that does not depend on as strict identifying assumptions as previous estimators based on microeconomic data. The standard identifying assumption that the self-employed underreport income whereas wage and salary workers do not is likely to fail in countries where employees are often paid under the table or engage in corrupt activities. Assuming that evading individuals have a higher consumption-income gap than non-evading ones due underreporting both to tax authorities and in surveys, an endogenous switching model with unknown sample separation enables the estimation of consumption-income gaps for both underreporting and truthful households. This avoids the need to identify non-evading and evading groups *ex-ante*. This methodology is applied to data from Czech and Slovak household budget surveys and shows that estimated evasion is substantially higher than found using previous methodologies.

JEL Classification: C34, E01, H26, J39

Keywords: shadow economy, switch regression, income-consumption gap

Corresponding author:

Randall K. Filer
Department of Economics
Hunter College/ City University of New York
695 Park Avenue
New York, NY 10021
USA
E-mail: rfiler@hunter.cuny.edu

* This project was supported by the National Science Foundation of the United States under grant #SES-0752760 to the Research Foundation of the City University of New York. All opinions are those of the authors and should not be attributed to the NSF or CUNY. We wish to express thanks for valuable comments to Orley Ashenfelter, Richard Blundell, Libor Dušek, Štěpán Jurajda, Peter Katuščák, Jan Kmenta, Steven Rivkin, and seminar participants at CERGE-EI. All remaining errors and omissions are entirely ours.

1 Introduction

The measurement of the shadow economy (also known as the grey or underground economy – i.e. income hidden from authorities) is of major interest to both economists and public policy makers. Measures such as Gross Domestic Product (GDP) obviously do not reflect the true productivity of the economy if they omit unofficial production. The standard methods of estimating deadweight loss (Harberger, 1964) understate inefficiencies if they do not reflect the diversion of economic activity into a possibly less efficient hidden sector.[1] Countries that try to offset the income lost in evasion by increasing tax rates can find themselves in a "vicious cycle" (Lyssiotou, Pashardes and Stengos, 2004, p.622) where rising tax rates create incentives for even greater evasion.

Allingham and Sandmo (1972) provided a basic framework for thinking about this problem rigorously. Estimation of the size of the shadow economy, however, is a challenge for numerous reasons, not the least of which is that by definition individuals are attempting to hide such of activities. Schneider and Enste (2002) divide methods of estimation into two main groups: direct and indirect. The first group is composed of surveys and inquiries on tax evasion. It is hard to imagine, however, that individuals who do not report all or part of their income on tax returns would reveal their full income in a survey, even if the survey promises anonymity. If nothing else, memories or records of income reported to the authorities provide an easy reference point when answering survey questions. In another direct method, tax authorities in many countries also attempt to estimate tax evasion from audited tax returns.[2]

In the second group (indirect methods) Schneider and Enste recognize three subgroups:

1. National accounting approaches focusing on the discrepancy between national accounting sources and uses (macroeconomic approach) or the discrepancy between incomes and expenditures of households (microeconomic approach);

2. Monetary approaches focusing on cash velocity, and transaction and cash demand;

3. Physical input methods focusing on electricity consumption.

Frequently several indirect indicators of the size of the shadow economy are combined in a single estimating equation, the so called Multiple Indicators - Multiple Causes (MIMIC) technique. Field and laboratory experiments (see Slemrod, Blumenthal and Christian 2001) should also be added to this set of categories.

[1] Such inefficiencies might be caused by resources being used in evasion effort instead of productive activities. They might also arise because the need to not draw attention from authorities results in inefficiently small enterprise sizes.

[2] One of the most comprehensive examples is probably the US Tax Compliance Measurement Program (TCMP). See Slemrod (2007) for details.

Macroeconomic methods of estimation of the size of the shadow economy have a long tradition dating from Cagan (1958), but have often been criticized for lacking an underlying theory and for flawed econometric techniques (see Hanousek and Palda 2006 or Thomas 1999). Changes in electricity demand inherently confound changes in the size of the shadow economy with changes in the composition of output or production efficiency. We, therefore, focus on the discrepancy between the income and expenditure of households.

A key difficulty with prior work using households' reported income and expenditure is that the researcher has assumed *a priori* that a certain sub-sample of the population does not evade (typically wage and salaried workers), and estimated underreporting for the rest of the sample (self-employed, farmers, etc.). This simplifying assumption is, however, weak both theoretically (see Kolm and Nielsen 2008 for a model with concealment of income by firms and salary workers) and also empirically. For example, the Eurobarometer survey (European Commission 2007) shows that 5 percent of respondents in the EU admitted they carried out undeclared work in the preceding 12 months. National values of this percentage range vary substationally, with the highest share in Denmark (18 percent) and lowest share in Cyprus (1 percent). The Czech and Slovak Republics are at 7 and 6 percent, respectively. In a separate question, 5 percent of respondents in the EU answered they had received at least part of their salary as 'envelope' or 'cash-in-hand' wages (lower bound estimates) in the preceding 12 months. As with the above question, national values differed (being higher in transition countries), with the lowest numbers for UK (1 percent) and the highest for Romania (23 percent). Czech and Slovak employees are somewhere in the middle of the group at 3 and 7 percent, respectively.[3]

Using self-employment as identification for potential income underreporting, Pissarides and Weber (1989) estimate food Engel curves for the employed from the UK 1982 family expenditure survey and then invert these to predict income for the self-employed. The difference between the predicted income and the reported income of the self-employed is interpreted as the size of the "black economy." Lyssiotou, Pashardes and Stengos (2004) criticized this approach, claiming that the use of food expenditures only can cause preference heterogeneity to be interpreted as tax evasion and suggested estimating a complete demand system to account for the heterogeneity in preferences using the generalized method of moments (GMM). Their approach is, however, still limited by the *a priori* assumption that wage income is reported correctly.

Moreover, the assumption of correctly reported consumption may fail when using a complete demand system. While the argument that certain items are more prone to be purchased from transitory income (assumed to be more closely linked to self-employed income) does not pose problems,[4] the fact that some

[3] These numbers, however, should be taken only as an indication. As the authors put it: "In view of the sensitivity of the subject, the pilot nature of the survey and the low number of respondents who reported having carried out undeclared work or having received 'envelope wages', results should be interpreted with great care (p.3)."

[4] As Lyssiotou, Pashardes and Stengos (2004) put it: "For example, households may decide to use their steady wage income on regular non-luxury goods and then use the self-employment

household consumption items are tax-deductible for the self-employed does. Not only can these items lower net self-employment income, their reporting may not be consistent. Some households may report such a business expense as household spending in the survey, while others may not. This includes a variety of non-durables and even durables. Food expenditure, on the other hand, is very rarely tax-deductible and it is not easy to substitute tax-deductible items for food. Based on this argument, we will not consider a complete demand system and concentrate on food expenditures, which are more stable.[5]

Additional works that identify underreporting based on self-employment status include Hurst, Li and Pugsley (2010) and Tedds (2010). The latter used a non-parametric estimation of Engel curves to avoid the assumption that the ratio of evaded income to total income is constant, although she again relies on the assumption that employed individuals do not evade. Studies that estimate the evasion response to tax changes can provide added insight. Gorodnichenko, Martinez-Vazquez and Sabirianova Peter (2009) used the 2001 flat tax reform in Russia as a natural experiment that produced a "control group" consisting of the part of the population for whom the marginal tax rate did not change whose income underreporting could be compared with a "treatment group" of individuals for whom the marginal tax fell.

It is possible to avoid the problem of arbitrary *a priori* assignment of individuals to evading and non-evading groups econometrically by using endogenous switching regression with unknown sample separation. Such a technique has not here-to-fore been applied to the shadow economy,[6] although they have been used elsewhere. In an early study Dickens and Lang (1985) used such a model to test dual labor market theory. Two more recent papers applied this methodology to family economics. Arunachalam and Logan (2006) incorporated two competing incentives to offer a dowry into one switching regression model, while Kopczuk and Lupton (2007) studied whether having a positive net worth at the time of death implies a bequest motive.

Other examples of the application of switching regressions with an unknown (or partially known) sample separation rule include the estimation of cartel stability by Lee and Porter (1984), and stochastic frontier models by Douglas, Conway and Ferrier (1995), or Caudill (2003). These works showed the feasibility of maximum likelihood and other estimation techniques in this situation.

---

income to buy luxuries. " Furthermore, self-employment income may be more associated with expenditure on certain goods like cars, computers and telephone bills for which it can attract higher deductions for business expenses.

[5] Food Engel curves show relatively stable behavior across different specifications of a full demand system. Beneito (2003) shows that only food and housing exhibit significant and negative price elasticity. Besides, the food Engel curve depicts the highest R^2 from all items. A similar study by Rajapakse (2011) again confirms stable behavior of food-based Engel curves. Moreover, in this study using complete system of EC some other items including housing and clothing expenditures are relative luxuries in the system (demand is more expenditure elastic). These studies, in our opinion, support use of food-based Engel curves for our purposes.

[6] DeCicca, Kenkel and Liu (2010) use an endogenous switching regression to estimate the effect of state differences in cigarette excise taxes on the probability of cross-border cigarette purchases in the US. Their model, however, relies on an observable rather than unobservable separation rule since they know which purchases were made across a border.

The methodology of endogenous switching regression with unobserved separation will thus allow relaxing overly restrictive assumptions including an *ad hoc* specification of under-reporting groups or requiring that evaders underreport income by a constant fraction of their income.

2 Methodology

2.1 Consumption-income gap

Our analysis relies on the consumption-income gap as described by Gorodnichenko, Martinez-Vazquez and Sabirianova Peter (2009) based on three assumptions coming from the permanent income hypothesis (Friedman 1957):

$$\begin{aligned} Y_i^R &= \Gamma_i Y_i^c, \text{ where: } \Gamma_i = \Gamma(\mathbf{S}_i) = \exp(-\mathbf{S}_i\boldsymbol{\gamma} + \text{error}) , \quad (1) \\ Y_i^C &= H_i Y_i^P, \text{ where: } H_i = H(\mathbf{L}_{1,i}) = \exp(\mathbf{L}_{1,i}\boldsymbol{\eta} + \text{error}) , \quad (2) \\ C_i &= \Theta_i Y_i^P, \text{ where: } \Theta_i = \Theta(\mathbf{L}_{2,i}) = \exp(\mathbf{L}_{2,i}\boldsymbol{\theta} + \text{error}) . \quad (3) \end{aligned}$$

where i denotes households. Eq.(1) defines reported income as a fraction Γ of true income, where Γ is a function of household characteristics affecting underreporting ($\mathbf{S}_i$) including age (older people are more risk averse and, therefore, less prone to tax evasion), education, whether the individual is self-employed, works in a large or small firm (small firms are more prone to save labor costs by paying a low "official" wage and paying part of the wage "under the table"), whether the employer of the individual is the public sector or a private firm (government is less likely to pay its employees "under the table"). Eq.(2) is based on the permanent income hypothesis where the current true income is a fraction H_i of the permanent lifelong income. H_i depends on the current stage of the life cycle of the head of household and his or her spouse including their ages, education and work experience (vector $\mathbf{L}_{1,i}$). Eq.(3) tells us that consumption constitutes a fraction Θ_i of the household's permanent income. The characteristics $\mathbf{L}_{2,i}$ affecting a household's consumption patterns (tastes) include the age of the head of household and spouse, number and ages of children, number of other household members, marital status, and education among others. It is clear that in general none of these fractions is constant. Taking logarithms of (1), (2) and (3) and substituting yields a definition of the consumption-income gap:

$$\log C_i - \log Y_i^R = \mathbf{S}_i\boldsymbol{\gamma} + \mathbf{L}_i\boldsymbol{\alpha} + \varepsilon_i , \quad (4)$$

where $\log C_i - \log Y_i^R$ is the consumption-income gap of the household. Note that in our context, where consumption is of food only, C_i will usually be less than Y_i, so that $\log C_i - \log Y_i$ is less than one. All other household characteristics held equal, a lower consumption-income gap of household A compared to household B would imply a higher degree of underreporting on the part of household A.

2.2 From consumption-income gap to shadow economy

The above analysis of the consumption-income gap can be extended in several ways. The first assumption that can be made without much loss of generality is that there are two groups of individuals in every economy: those who evade and those who do not. For the latter, γ in Eq.(4) is equal to 0 by definition. These two groups of agents differ, all other characteristics held constant, by the size of the gap. Since consumption is based on true income, evading households consume a greater share of their reported income. Under the assumption that, unlike income, consumption is measured correctly for both groups (for empirical support of this assumption see Hurst, Li and Pugsley 2010), we can write:

$$\log C_i - \log Y_i^{R,e} = \mathbf{S}_i\boldsymbol{\gamma} + \mathbf{L}_i\boldsymbol{\alpha}_e + \varepsilon_{e,i} \quad \text{if } i \text{ is evading}, \tag{5}$$

$$\log C_i - \log Y_i^{R,ne} = \mathbf{L}_i\boldsymbol{\alpha}_{ne} + \varepsilon_{ne,i} \quad \text{if } i \text{ is not evading}, \tag{6}$$

where $Y_i^{R,e}$ and $Y_i^{R,ne}$ are the reported income if the household i evades and does not evade, respectively. It is reasonable to assume that people evade if their expected gain from evasion exceeds a certain threshold f:

$$\left(\log C_i - \log Y_i^{R,e}\right) - \left(\log C_i - \log Y_i^{R,ne}\right) \geq f_i\,. \tag{7}$$

where f_i represents costs of evasion including expected fines and costs associated with hiding income (including psychic costs) of household i. If we assume that the cost of evasion is equal to a constant average cost k plus an error term $\varepsilon_{f,i}$ (the deviation of household i from this average) we can write the probability of household i being in the evading regime as:

$$P = \Pr\left\{\mathbf{S}_i\boldsymbol{\gamma} + \mathbf{L}_i\left(\boldsymbol{\alpha}_e - \boldsymbol{\alpha}_{ne}\right) - k \geq \varepsilon_{f,i} + \varepsilon_{e,i} - \varepsilon_{ne,i}\right\} = \Pr\left\{\mathbf{Z}_i\boldsymbol{\Gamma} \geq \varepsilon_{s,i}\right\}\,. \tag{8}$$

This system can be expressed as the following econometric model:

$$\left(\log C_i - \log Y_i^R\right)_e = \mathbf{X}_i\boldsymbol{\beta}_e + \varepsilon_{e,i}\,, \tag{9}$$

$$\left(\log C_i - \log Y_i^R\right)_{ne} = \mathbf{X}_i\boldsymbol{\beta}_{ne} + \varepsilon_{ne,i}\,, \tag{10}$$

$$y_i^* = \mathbf{Z}_i\boldsymbol{\Gamma} - \varepsilon_{s,i}\,, \tag{11}$$

$$\log C_i - \log Y_i^R = \begin{cases} \left(\log C_i - \log Y_i^R\right)_e & \text{iff } y_i^* \geq 0\,, \\ \left(\log C_i - \log Y_i^R\right)_{ne} & \text{iff } y_i^* < 0\,, \end{cases} \tag{12}$$

where $\mathbf{X}_i$ and $\mathbf{Z}_i$ are the vectors of explanatory variables that affect consumption and income patterns, and tax evasion propensities, respectively.

The latent variable y_i^* can be interpreted as the propensity to evade. It cannot be observed, but if $y_i^* > 0$, household i's gap is determined by (9), otherwise it is determined by (10). We can express the likelihood contribution of household i as:

$$\begin{aligned} L_i = {} & \Pr\left(\varepsilon_{s,i} \leq \mathbf{Z}_i\boldsymbol{\Gamma} \mid \mathbf{Z}_i, \mathbf{X}_i, \varepsilon_{e,i}\right) \cdot f\left(\varepsilon_{e,i}\right) \\ & + \Pr\left(\varepsilon_{s,i} > \mathbf{Z}_i\boldsymbol{\Gamma} \mid \mathbf{Z}_i, \mathbf{X}_i, \varepsilon_{e,i}\right) \cdot f\left(\varepsilon_{ne,i}\right)\,. \end{aligned} \tag{13}$$

Under the assumption that $(\varepsilon_e, \varepsilon_{ne}, \varepsilon_s) \sim N(0, \Sigma)$, where:

$$\Sigma = \begin{pmatrix} \sigma_e^2 & & \\ 0 & \sigma_{ne}^2 & \\ \sigma_{e,s} & \sigma_{ne,s} & 1 \end{pmatrix},$$

the log-likelihood function (13) becomes:

$$\ln L(\boldsymbol{\beta}_e, \boldsymbol{\beta}_{ne}, \boldsymbol{\Gamma}, \sigma_e, \sigma_{ne}, \sigma_{e,s}, \sigma_{ne,s}) = \sum_{i=1}^{N} \ln \left\{ \frac{1}{\sigma_e} \Phi \left(\frac{\mathbf{Z}_i \boldsymbol{\Gamma} - \frac{\sigma_{e,s}}{\sigma_e^2} \varepsilon_{e,i}}{\left(1 - \frac{\sigma_{e,s}^2}{\sigma_e^2}\right)^{.5}} \right) \cdot \phi \left(\frac{\varepsilon_{e,i}}{\sigma_e} \right) + \frac{1}{\sigma_{ne}} \left[1 - \Phi \left(\frac{\mathbf{Z}_i \boldsymbol{\Gamma} - \frac{\sigma_{ne,s}}{\sigma_{ne}^2} \varepsilon_{ne,i}}{\left(1 - \frac{\sigma_{ne,s}^2}{\sigma_{ne}^2}\right)^{.5}} \right) \right] \cdot \phi \left(\frac{\varepsilon_{ne,i}}{\sigma_{ne}} \right) \right\}, \tag{14}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal density and cumulative distribution function respectively, and:

$$\varepsilon_{e,i} = (\ln C_i - \ln Y_i) - \mathbf{X}_i \boldsymbol{\beta}_e, \tag{15}$$

$$\varepsilon_{ne,i} = (\ln C_i - \ln Y_i) - \mathbf{X}_i \boldsymbol{\beta}_{ne}. \tag{16}$$

Technical details of the maximization of (14) are given in the Appendix. Although identification based solely on functional assumptions is possible, valid exclusion restrictions such that $\mathbf{Z}_i \neq \mathbf{X}_i$ are desirable, ensuring that all the parameters (except σ_s, which is normalized to one) are identifiable. Applied to the case at hand, the switching equation will contain variables that influence activity in the hidden economy rather than consumption-income gap, such public sector or self employment.

2.3 Measure of the shadow economy

Under the initial assumption of correct consumption reporting, the expected value of the difference in the gaps for both regimes of household i is equal to:

$$\mathbb{E}\left[\left(\widehat{\log C_i - \log Y_i^R} \right)_e - \left(\widehat{\log C_i - \log Y_i^R} \right)_{ne} \right] = \mathbb{E}\left[\left(\widehat{\log Y_{i,ne}^R - \log Y_{i,e}^R} \right) \right], \tag{17}$$

which is household i's estimated degree of income underreporting as a fraction of its reported income. The overall size of the shadow economy is therefore defined as the expected value of this difference in gaps, i.e. the sum of the differences between the income-consumption gaps for the respective regimes weighted by probability of each household being in the shadow sector:

$$\widehat{Evasion} = \frac{1}{N} \sum_{i=1}^{N} \left(\mathbf{X}_i \hat{\boldsymbol{\beta}}_e - \mathbf{X}_i \hat{\boldsymbol{\beta}}_{ne} \right) \cdot \hat{P}_{e,i}, \tag{18}$$

The probability of being in the shadow sector $\hat{P}_{e,i}$ can be computed by Bayes' theorem as:

$$\hat{P}_{e,i} = \frac{\frac{1}{\hat{\sigma}_e}\Phi\left(\frac{\mathbf{Z}_i\hat{\mathbf{\Gamma}}-\frac{\hat{\sigma}_{e,s}}{\hat{\sigma}_e^2}e_{e,i}}{\left(1-\frac{\hat{\sigma}_{e,s}^2}{\hat{\sigma}_e^2}\right)^{.5}}\right)\phi\left(\frac{e_{e,i}}{\hat{\sigma}_e}\right)}{\frac{1}{\hat{\sigma}_e}\Phi\left(\frac{\mathbf{Z}_i\hat{\mathbf{\Gamma}}-\frac{\hat{\sigma}_{e,s}}{\hat{\sigma}_e^2}e_{e,i}}{\left(1-\frac{\hat{\sigma}_{e,s}^2}{\hat{\sigma}_e^2}\right)^{.5}}\right)\phi\left(\frac{e_{e,i}}{\hat{\sigma}_e}\right)+\frac{1}{\hat{\sigma}_{ne}}\left[1-\Phi\left(\frac{\mathbf{Z}_i\hat{\mathbf{\Gamma}}-\frac{\hat{\sigma}_{ne,s}}{\sigma_{ne}^2}e_{ne,i}}{\left(1-\frac{\hat{\sigma}_{ne,s}^2}{\hat{\sigma}_{ne}^2}\right)^{.5}}\right)\right]\cdot\phi\left(\frac{e_{ne,i}}{\hat{\sigma}_{ne}}\right)}, \tag{19}$$

where:

$$e_{e,i} = \quad (\ln C_i - \ln Y_i) - \mathbf{X}_i\hat{\boldsymbol{\beta}}_e\,, \tag{20}$$

$$e_{ne,i} = \quad (\ln C_i - \ln Y_i) - \mathbf{X}_i\hat{\boldsymbol{\beta}}_{ne}\,. \tag{21}$$

Eq.18 will thus give us the size of the shadow economy as a fraction of economy's total reported income.

The likelihood ratio test is a natural choice for a test of the assumption that we can divide households into two groups based on their consumption-income gaps. Given that a model consisting of a single gap function is nested in the endogenous switching model, such a test can be used to compare the two models, with the null hypothesis being that both models explain data equally well.[7] Following Dickens and Lang (1985) the degrees of freedom are equal to number of constraints plus the number of unidentified parameters (found only in the switching equation). As argued by Goldfeld and Quandt (1976) this leads to a conservative critical value.

To increase robustness to initial values and outliers, Monte Carlo simulations were used to compute both means and standard errors of the estimators. For each country 500 random samples with replacement were drawn from the data, with estimation of Eq.(14), and computation of the shadow economy from Eqs.(18) and (19) done for each sample.[8] This results in a data series from which means of these estimates can be computed. Standard errors are then the standard errors of these means.

3 Data

3.1 Czech Republic

The data come from the Czech household budget survey for 2008. They contain information about income from various sources and expenditures on different categories of goods and services for 3,271 Czech households. From this dataset a

[7] In that case, one should obviously choose the model with a single gap function, which is more parsimonious.

[8] See Appendix A for details.

sub-sample of households with working heads was chosen. The summary statistics (weighted means) of this sub-sample are given in Table 1. The definition of income used for the computation of the gap is the monthly average of the total gross income of the household from all sources minus all taxes and obligatory payments (such as obligatory health insurance, which is technically a tax in the Czech Republic). To account for possible consumption smoothing and precautionary saving (which may be greater for certain types of households) net dissavings were included in income. When defining consumption, we used only expenditures on food. Given that a great many items in the nondurable category are tax deductible for the self-employed in both the Czech and Slovak Republics, and can easily be used not only for business purposes but also for private consumption, using food consumption is more robust. As discussed above, Z_i contains dummies for public sector or self-employment status of the head of household or spouse, blue collar head or spouse, age, square of age (previous research shows that risk aversion increases with age up to certain point, but then it decreases again) and education of head. Explicit marital status cannot be determined from the Czech data, which only reports whether the household head has a life partner, not the exact legal status of the relationship. We also control for potential shocks to social norms (reporting propensities) produced by the fall of communism in the former Czechoslovakia by including whether individuals were raised before or after 1989, i.e. those who are older than 42 in 2008 (see e.g. Večerník, 1996).

As discussed above, X_i contains variables such as number of household members of different categories, education, relationship status, and age and age squared. The last two variables are an indicator for work experience. Given that manual jobs usually have higher calorie requirements, dummies for having a blue collar head or blue collar spouse are also included in the outcome equations.

Table 1: Summary statistics of the subsample in the Czech HBS, 2008

Total no. of households	2,138
average no. of household members	2.26
average no. of heads with a spouse or a partner	1,338
average no. of children	0.56
average annual disposable income of households (CZK)	309,318
average age of head	53
no. of self-employed heads	286
no. of heads working in public sector	383
no. of heads working in private sector	1,469
no. of blue-collar heads	733
no. of heads with a high school degree	1,783
no. of heads with a bachelor's degree (or higher)	220

The Czech HBS uses stratified sampling, so each household's log-likelihood contribution (Eq.(14)) was multiplied by the probability of the household being

in the random sample. Similarly, observations in the estimations of the shadow economy (Eq.(18)) are weighted both by the estimated probability of the given household being in the shadow sector and its sampling weights.

3.2 Slovak Republic

Similar to the Czech case, the HBS for 2008 collected by the Slovak Statistical Office was used. Overall, the sample contains 4,718 households. Estimation was done on a subsample of 2,991 households whose head was working (either employed or self-employed) during 2008. Summary statistics for Slovak households included in the subsample can be seen in Table 2. The Slovak Statistical office chooses sample households through simple random sampling without sample weights. Therefore, as opposed to the Czech case, log-likelihood contributions are automatically weighted by the probability of the household appearing in the sample. However, definitions of variables are almost an exact copy of those of their Czech counterparts, except for marital status, which is explicitly observed in the Slovak data.

Table 2: Summary statistics of the subsample in the Slovak HBS

Total number of households	2,992
average no. of household members	3.19
average no. of married households	2,259
average no. of children	1.055
average annual disposable income of households (SKK)	421,104
average age of head	44.17
no. of self-employed heads	506
no. of heads working in public sector	793
no. of heads working in private sector	1,657
no. of blue-collar heads	1,221
no. of heads with a high school degree	1,497
no. of heads with a bachelor's degree (or higher)	494

4 Results

4.1 Main results

The results of maximum likelihood estimation of the structural endogeneous switching model for Czech and Slovak Republics are shown in Tables B1 and C1, respectively. These estimates, together with the confidence intervals, were obtained from the Monte Carlo method described above. Note that the likelihood ratio test rejects the null hypothesis, implying that this model is significantly better in explaining the data than an OLS model describing behavior of all households with a single consumption-income gap equation. Plugging the estimated coefficients in these tables into Eq.(18) yields the estimates of the

shadow economy in Table 3. The main result is that the shadow economy in the Czech Republic constituted approximately 20 percent of reported income in 2008, while in Slovakia this fraction was approximately 28.6 percent. Therefore, if we want to arrive at true income in these economies, we have to multiply the officially reported income by 1.2 and 1.286 respectively. These estimates for the Czech Republic are slightly higher than those reported by Schneider, Buehn and Montenegro (2010) for 2007 (17.0 percent) and substantially higher than those derived using self-employment status as an *ex ante* mechanism for defining evaders as in Pissarides and Weber (1989) where the share of unreported income was estimated by Lichard (2012) to be 4 percent of GDP. For Slovakia, our estimates of the share of the shadow economy in GDP are substantially higher than reported by Schneider et al. (16.8 percent for 2007) or Lichard (6.8 percent). From these results it is obvious that in post-communist countries at least, underreporting of income extends to wage and salary workers as well as the self-employed.

Table 3: Shadow economy estimates

Country	Year	Share of shadow economy on total income	95% confidence interval (bootstrapped)
Czech Republic	2008	20%	±3.5%
Slovak Republic	2008	28.6%	±3.2%

4.2 Discussion of marginal effects

Marginal effects for the Czech Republic and Slovakia are shown in Tables B2 and C2 respectively.The effects of changes in variables on the probability of being in the shadow sector can be seen in the sixth column of the respective tables.[9] The Czech results suggest that households with a white collar head have a higher propensity of evading, while households headed by both blue and white collar self-employed workers are more likely to be in the shadow sector, as intuition predicts (by 15.7 and 9.6 percentage points, respectively). By way of contrast, the estimates suggest that there is no systematic difference between blue collar and white collar employees in Slovakia, although households headed by white collar self-employed workers are 8.7 percentage points more likely to be in the shadow sector than those headed by white collar employees. If the head of a household working in the public sector, the probability of its under-reporting income decreases in both countries, although in the Slovak case by only 3 percentage points compared to almost 19 percentage points in the Czech Republic. Having children decreases the probability of being in the shadow sector for Czech households slightly (by 2.5 percentage points), and more substantially for Slovak households (around 14 percentage points). Education effects also differ between

[9]Note that the reference category is a household with an unmarried, white collar head employed in a private company.

countries. In the Czech Republic the probability of being in the evading sector decreases slightly with education, while Slovak households are more likely to underreport when their head is more educated. Age decreases the propensity to evade in both countries. While age itself has a negative coefficient, the square of age has a positive one. This convex relationship is, however, very weak – due to the small coefficient on square of age, the convex effect is negligible for the range of applicable ages. For both countries every spousal employment characteristic (being self-employed, blue collar, or white collar) decreases the probability of the household being in the shadow economy relative to the spouse not working (or is insignificant). One plausible explanation is that, given that in both the Czech Republic and Slovakia heads are predominantly male (implying that spouses are female), the higher risk aversion of females[10] means that women are less prone to underreport income when they work. This is corroborated by the marginal effect of the head of the household being married, which decreases the probability of a household being in the shadow sector in both countries. Concerning the possible effect of the shock to social norms brought about by the fall of the communism, heads raised during communism seem to be more likely to withhold income in both countries, although the effect decreases with education.

5 Conclusion

The size of the shadow economy was estimated based on microeconomic data without assumptions that hampered previous estimators thereby possibly underestimating of the size of the shadow economy by excluding under-reporting among the group assumed to fully report. The application of the methodology to Czech and Slovak data and its comparison to the standard exclusion restriction adopted by Pissarides and Weber (1989) and others corroborates this hypothesis. We find that, in these economies at least, employees being paid under the table or having a secondary, undeclared, source of income constitutes a major source of unreported income.

[10] Previous studies offer some support for the proposition that women are more risk averse than men. For an overview of experimental results see Eckel and Grossman (2008).

References

Allingham, Michael G., and Agnar Sandmo. 1972. "Income tax evasion: a theoretical analysis." *Journal of Public Economics*, 1(3-4): 323–338.

Arunachalam, Raj, and Trevon D. Logan. 2006. "On the heterogeneity of dowry motives." National Bureau of Economic Research Working Paper 12630.

Beneito, Pilar. 2003. "A complete system of Engel curves in the Spanish economy." *Applied Economics*, 35(7): 803 – 816.

Cagan, Philip. 1958. *The demand for currency relative to total money supply.* UMI.

Caudill, Steven B. 2003. "Estimating a mixture of stochastic frontier regression models via the EM algorithm: A multiproduct cost function application." *Empirical Economics*, 28(3): 581–598.

DeCicca, Philip, Donald S. Kenkel, and Feng Liu. 2010. "Excise tax avoidance: The case of state cigarette taxes." National Bureau of Economic Research Working Paper 15941.

Dickens, William T., and Kevin Lang. 1985. "A test of dual labor market theory." *The American Economic Review*, 75(4): 792–805.

Douglas, Stratford M., Karen Smith Conway, and Gary D. Ferrier. 1995. "A switching frontier model for imperfect sample separation information: With an application to constrained labor supply." *International Economic Review*, 36(2): 503–526.

Dutoit, Laure C. 2007. "Heckman's selection model, endogenous and exogenous switching models: A survey." *Available at* `http://works.bepress.com/laure_dutoit/3`.

Eckel, Catherine C., and Philip J. Grossman. 2008. "Men, women and risk aversion: Experimental evidence." In *Handbook of Experimental Economics Results.* , ed. Charles R. Plott and Vernon L. Smith, Chapter 113, 1061 – 1073. Elsevier.

European Commission. 2007. "Undeclared work in the European Union." *Special Eurobarometer*, 284.

Friedman, Milton. 1957. "The relation between the permanent income and relative income hypotheses." *A Theory of the Consumption Function*, 157–182.

Goldfeld, Steven M., and Richard E. Quandt. 1976. "Techniques for estimating switching regressions." *Studies in Nonlinear Estimation*, 3–35. Cambridge, MA: Ballinger.

Gorodnichenko, Yuriy, Jorge Martinez-Vazquez, and Klara Sabirianova Peter. 2009. "Myth and reality of flat tax Reform: Micro estimates of tax evasion response and welfare effects in Russia." *Journal of Political Economy*, 117(3): 504–554.

Hanousek, Jan, and Filip Palda. 2006. "Problems measuring the underground economy in transition." *Economics of Transition*, 14(4): 707–718.

Harberger, Arnold. 1964. "Taxation, resource allocation, and welfare." *The Role of Direct and Indirect Taxes in the Federal Reserve System*, 25–80. Princeton University Press.

Hurst, Erik, Geng Li, and Benjamin Pugsley. 2010. "Are household surveys like tax forms: Evidence from income underreporting of the self employed." National Bureau of Economic Research Working Paper 16527.

Kolm, Ann-Sofie, and Søren Bo Nielsen. 2008. "Under-reporting of income and labor market performance." *Journal of Public Economic Theory*, 10(2): 195–217.

Kopczuk, Wojciech, and Joseph P. Lupton. 2007. "To leave or not to leave: The distribution of bequest motives." *The Review of Economic Studies*, 74(1): 207–235.

Lee, Lung-Fei, and Robert H. Porter. 1984. "Switching regression models with imperfect sample separation information–With an application on cartel stability." *Econometrica*, 52(2): 391–418.

Lichard, Tomas. 2012. "Shadow economy in the Czech Republic, Russia, Slovakia and Ukraine: Food Engel curve approach." Unpublished manuscript.

Lyssiotou, Panayiota, Panos Pashardes, and Thanasis Stengos. 2004. "Estimates of the black economy based on consumer demand approaches." *Economic Journal*, 114(497): 622–640.

Pissarides, Christopher A., and Guglielmo Weber. 1989. "An expenditure-based estimate of Britain's black economy." *Journal of Public Economics*, 39(1): 17–32.

Rajapakse, Suri. 2011. "Estimation of a complete system of nonlinear Engel curves: further evidence from Box-Cox Engel curves for Sri Lanka." *Applied Economics*, 43(3): 371–385.

Schneider, Friedrich, and Dominik H. Enste. 2002. *The Shadow Economy: An International Survey.* Cambridge (UK): Cambridge University Press.

Schneider, Friedrich, Andreas Buehn, and Claudio Montenegro. 2010. "New estimates for the shadow economies all over the world." *International Economic Journal*, 24(4): 443–461.

Slemrod, Joel. 2007. "Cheating ourselves: The economics of tax evasion." *The Journal of Economic Perspectives*, 21(1): 25–48.

Slemrod, Joel, Marsha Blumenthal, and Charles Christian. 2001. "Taxpayer response to an increased probability of audit: evidence from a controlled experiment in Minnesota." *Journal of Public Economics*, 79(3): 455–483.

Tedds, Lindsay M. 2010. "Estimating the income reporting function for the self-employed." *Empirical Economics*, 38(3): 669–687.

Thomas, Jim. 1999. "Quantifying the black economy: 'measurement without theory' yet again?" *The Economic Journal*, 109(456): 381–389.

Večerník, Jiří. 1996. *Markets and people: The Czech reform experience in a comparative perspective.* Avebury Aldershot.

A Technical Appendix

The estimation was done in TSP 5.1 (64-bit) via the command 'ml'. This command maximizes the log-likelihood function numerically[11] and, therefore, choosing appropriate initial values is essential for convergence. The initial values were set by a procedure described in Dutoit (2007). We initially separate the sample through a dummy $I_i = 1$ if the household i's gap is above a certain threshold (initial evading group) or $I_i = 0$ if it is below that threshold (initial non-evading group). To obtain initial values of $\mathbf{\Gamma}$, a probit regression of I_i on $\mathbf{Z}_i$ is run. After that we use these values $\left(\hat{\Gamma}\right)$ to estimate initial values of the $\boldsymbol{\beta}$'s by running the following OLS regressions:

$$\ln C_i - \ln Y_i = \mathbf{X}_i\boldsymbol{\beta}_e - \sigma_{e,s}\frac{\phi\left(\mathbf{Z}_i\hat{\mathbf{\Gamma}}\right)}{\Phi\left(\mathbf{Z}_i\hat{\mathbf{\Gamma}}\right)} + \varepsilon_{i,e} \text{ if } I_i = 1\,, \tag{22}$$

and

$$\ln C_i - \ln Y_i = \mathbf{X}_i\boldsymbol{\beta}_{ne} + \sigma_{ne,s}\frac{\phi\left(\mathbf{Z}_i\hat{\mathbf{\Gamma}}\right)}{1-\Phi\left(\mathbf{Z}_i\hat{\mathbf{\Gamma}}\right)} + \varepsilon_{i,ne} \text{ if } I_i = 0\,. \tag{23}$$

Then we get initial values of σ_e and $\sigma_{e,s}$ by running the following OLS estimation:

$$\hat{u}_{e,i}^2 = \sigma_e^2 - \sigma_{e,s}\frac{\phi\left(\mathbf{Z}_i\hat{\mathbf{\Gamma}}\right)}{\Phi\left(\mathbf{Z}_i\hat{\mathbf{\Gamma}}\right)}\,,$$

where $\hat{u}_{e,i} = (\ln C_i - \ln Y_i) - X_i\hat{\beta}_e$, where $\hat{\beta}_e$ is the estimate of β_e coming from Eq.(22). The initial values of σ_{ne} and $\sigma_{ne,s}$ are obtained analogously by running:

$$\hat{u}_{ne,i}^2 = \sigma_{ne}^2 - \sigma_{ne,s}\frac{\phi\left(\mathbf{Z}_i\hat{\mathbf{\Gamma}}\right)}{1-\Phi\left(\mathbf{Z}_i\hat{\mathbf{\Gamma}}\right)}\,.$$

These initial values of Γ, $\boldsymbol{\beta}$'s and σ's are then used as starting values for the numerical optimization procedure.

To make the results robust, for each random sample within the Monte Carlo simulation the initial sample separation is in turn set to the first, second and third quartiles, and the mean of the consumption-income gap. After applying the above procedure to each of these initial splits, we choose the results of the one that yields the highest log-likelihood as final results for the given Monte Carlo sample. This results in the data series from which the statistics (such as shadow economy size and standard errors) are computed.

[11] For more detailed information on this command including stopping rules, see the TSP manual at http://www.tspintl.com/products/manuals.htm.

B Estimation Results - Czech Republic

Table B1: Structural model coefficients - Czech Republic (2008)

	Shadow sector		Official sector		Switching equation	
VARIABLES	$\ln C - \ln Y$		$\ln C - \ln Y$		Latent variable	
constant	-1.792***	(0.0752)	-3.6550***	(0.0923)	2.2509***	(0.4040)
# of children	0.0635***	(0.0026)	-0.0944***	(0.0032)		
# of employed	-0.0848***	(0.0047)	0.0008	(0.0066)		
# of unemployed	-0.1112***	(0.0047)	0.0195***	(0.0067)		
is married	-0.3483***	(0.0420)	0.7484***	(0.0116)	-2.3025***	(0.1380)
high school degree	-0.0125	(0.0087)	0.0583***	(0.0091)	0.2567***	(0.0764)
bachelor's degree<	-0.0015	(0.0144)	0.0816**	(0.0354)	-0.4453	(0.2890)
age	0.0134***	(0.0025)	0.0465***	(0.0037)	0.0039	(0.0171)
age^2	-0.0002***	(2.7e-05)	-0.0005***	(3.9e-05)	-0.0002	(0.0002)
blue collar	0.1782***	(0.0098)	0.1178***	(0.0097)	-0.4468***	(0.0771)
blue collar spouse	-0.0179	(0.0123)	0.0606	(0.0692)	-0.3241	(0.4284)
σ_1	0.3144***	(0.0051)				
σ_2			0.2622***	(0.0035)		
has children					0.1534***	(0.0315)
works in public sector					-0.1898***	(0.0197)
self-employed × white collar					0.9722***	(0.0974)
self-employed × blue collar					0.5947***	(0.0285)
spouse in public sector					-0.4318***	(0.0310)
white collar spouse					0.7071***	(0.0591)
self-employed spouse					0.0575	(0.5013)
age>42					0.7211***	(0.0554)
high school × age>42					0.1557**	(0.0722)
university × age>42					-0.0233	(0.6704)
σ_{13}					0.2723***	(0.0090)
σ_{23}					-0.0546***	(0.0154)
Observations			2,079			
Log likelihood			-164.456			
LR test			589.0139			
Prob>χ^2(32)			0.0000			

Bootstrapped standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table B2: Czech Republic (2008) - Marginal effects

VARIABLES	Shadow sector $\ln C - \ln Y$		Official sector $\ln C - \ln Y$		Probability of being in the shadow sector	
# of children	-0.0502***	(0.0021)	0.0512***	(0.0020)		
# of employed	0.0671***	(0.0036)	-0.0004	(0.0042)		
# of unemployed	0.0880***	(0.0036)	0.0106***	(0.0042)		
is married	-0.2757***	(0.0100)	0.4058***	(0.0079)	-0.5793***	(0.0185)
high school degree	0.0099	(0.0071)	0.0316***	(0.0055)	-0.0423***	(0.0152)
bachelor's degree<	0.0012	(0.0097)	0.0442***	(0.0071)	-0.0792***	(0.0201)
age	-0.0106***	(0.0020)	0.0252***	(0.0018)	-0.0103***	(0.0037)
age^2	0.0001***	(0.0000)	-0.0003***	(0.0000)	0.0002***	(0.0000)
blue collar	-0.1410***	(0.0076)	0.0639***	(0.0055)	-0.1526***	(0.0159)
blue collar spouse	-0.0142	(0.1103)	0.0329***	(0.0085)	-0.0652***	(0.0182)
has children					0.0248***	(0.0053)
works in public sector					-0.0306***	(0.0036)
self-employed × white collar					0.1570***	(0.0061)
self-employed × blue collar					0.0961***	(0.0053)
spouse in public sector					-0.0697***	(0.0056)
white collar spouse					-0.1142***	(0.0093)
self-employed spouse					0.0093	(0.0094)
age>42					0.1165***	(0.0071)
high school × age>42					0.0252***	(0.0081)
university × age>42					-0.0038	(0.0128)

Bootstrapped standard errors in parentheses (*** p<0.01, ** p<0.05, * p<0.1). The basic category is a household with unmarried white collar head employed in a private company. The average marginal effects were computed as the average of marginal effects predicted for every observation in the subsample.

C Estimation Results - Slovak Republic

Table C1: Structural model coefficients - Slovak Republic (2008)

	Evading regime		Non-evading regime		Switching equation	
VARIABLES	$\ln C - \ln Y$		$\ln C - \ln Y$		N/A (latent)	
constant	-1.8365***	(0.0482)	-2.666***	(0.0928)	0.1766	(0.1650)
# of children	-0.0685***	(0.0014)	0.0874***	(0.0034)		
# of employed	-0.1389***	(0.0021)	-0.1040***	(0.0036)		
# of unemployed	-0.0024*	(0.0013)	-0.0169***	(0.0026)		
is married	-0.0505***	(0.0048)	0.0229*	(0.0120)	-0.0279	(0.0297)
high school degree	-0.0672***	(0.0042)	-0.2492***	(0.0135)	-0.7332***	(0.0404)
bachelor's degree<	-0.1686***	(0.0079)	-0.5306***	(0.0160)	-1.5277***	(0.1190)
age	0.0131***	(0.0018)	0.0697***	(0.0039)	0.0823***	(0.0067)
age^2	0.0001***	(2.0e-05)	-0.0006***	(4.5e-05)	0.0004***	(7.6e-05)
blue collar	0.0298***	(0.0062)	0.1134***	(0.0116)	-0.0848***	(0.0259)
blue collar spouse	0.0526***	(0.0072)	0.2446***	(0.0129)	0.0064	(0.0278)
σ_1	0.3571***	(0.0099)				
σ_2			0.7231***	(0.0094)		
has children					-0.6455***	(0.0140)
works in public sector					-0.1753***	(0.0118)
self-employed × white collar					-0.4059***	(0.0583)
self-employed × blue collar					0.4022***	(0.0598)
spouse in public sector					-0.0033	(0.0115)
white collar spouse					0.8959***	(0.0238)
self-employed spouse					0.3797***	(0.0301)
age>42					-1.9746***	(0.0347)
high school × age>42					0.1895***	(0.0351)
university × age>42					0.6860***	(0.1150)
σ_{13}					0.0029	(0.0168)
σ_{23}					-0.6754***	(0.0220)
Observations			2,922			
Log likelihood			-1498.2785			
LR test			193.9166			
Prob>χ^2(32)			0.0000			

Bootstrapped standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table C2: Marginal effects - Slovak Republic (2008)

	Shadow sector		Official sector		Probability of being	
VARIABLES	$\ln C - \ln Y$		$\ln C - \ln Y$		in the shadow sector	
# of children	-0.0274***	(0.0006)	-0.0029***	(0.0007)		
# of employed	0.0326***	(0.0008)	0.0059***	(0.0008)		
# of unemployed	0.0053***	(0.0006)	-0.0001	(0.0006)		
is married	-0.0072***	(0.0019)	0.0022	(0.0032)	-0.0145**	(0.0068)
high school degree	0.0781***	(0.0015)	0.0029	(0.0029)	0.1920***	(0.0075)
bachelor's degree<	0.1662***	(0.0026)	0.0072**	(0.0035)	0.3992***	(0.0102)
age	-0.0218***	(0.0006)	-0.0006	(0.0011)	-0.0279***	(0.0019)
age^2	0.0002***	(0.0000)	0.0000	(0.0000)	0.0002***	(0.0000)
blue collar	-0.0355***	(0.0019)	-0.0013	(0.0030)	0.0024	(0.0068)
blue collar spouse	-0.0766***	(0.0020)	-0.0023	(0.0027)	-0.0339***	(0.0063)
has children					-0.1386***	(0.0027)
works in public sector					-0.0064**	(0.0026)
self-employed × white collar					0.0872***	(0.0073)
self-employed × blue collar					-0.0864***	(0.0077)
spouse in public sector					0.0007	(0.0025)
white collar spouse					0.1923***	(0.0049)
self-employed spouse					-0.0815***	(0.0053)
age>42					0.4239***	(0.0054)
high school × age>42					-0.0407***	(0.0059)
university × age>42					-0.1473***	(0.0080)

Bootstrapped standard errors in parentheses (*** p<0.01, ** p<0.05, * p<0.1). The basic category is a household with unmarried white collar head employed in a private company. The average marginal effects were computed as the average of marginal effects predicted for every observation in the subsample.