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

Tomáš Lichard[1], Jan Hanousek[1], and Randall K. Filer[1,2]

[1]CERGE-EI, Prague, Czech Republic†
[2]Department of Economics, Hunter College and the Graduate Center, CUNY; IZA, Bonn and CESifo, Munich

Abstract

We develop a novel 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 have a secondary source of self-employed income. 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.

Keywords: endogenous switching regression, shadow economy, tax evasion, underreporting

JEL classification: C34, H26

*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. Support from the grant SVV-2013-267 801 is also gratefully acknowledged. We wish to express thanks for valuable comments to Orley Ashenfelter, Richard Blundell, Libor Dušek, Štěpán Jurajda, Peter Katuščák, Jan Kmenta, Steven Rivkin, Karine Torosyan and seminar participants at CERGE-EI. All remaining errors and omissions are entirely ours.

†CERGE-EI is a joint workplace of the Center for Economic Research and Graduate Education, Charles University in Prague, and the Economics Institute of the ASCR, v.v.i.

Abstrakt

Vyvinuli jsme odhad nereportovaného příjmu (pravděpodobně svázaného s daňovými úniky), který využívá mikroekonomických dat a který není založený na tak přísných předpokladech jako předchozí odhady. Standardní předpoklad, že samostatně výdělečně činné osoby nepřiznávají část příjmu, zatím co zamě-stnanci tuto možnost nemají, může selhat v zemích, kde je relativně časté platit část mzdy hotově bez dokladu, nebo kde mají zaměstnanci více zdrojů příjmu. Pokud předpokládáme, že jednotlivci s nepřiznanými příjmy mají vyšší rozdíl mezi spotřebou a příjmem než ti, kteří svůj příjem přiznávají, můžeme odhadovat tento rozdíl pro obě skupiny. Využíváme přitom regresní model s přechodem mezi dvěma stavy (přiznaný a zatajovaný příjem), kde pravidlo přechodu není plně známé a je endogenní (endogenous switching model with unknown sample separation rule). Tím se vyhneme potřebě rozdělit domácnosti do těchto skupin *ex ante*. Tato metodologie aplikovaná na českých a slovenských rodinných účtech vede k vyšším odhadům šedé ekonomiky jako předchozí mikroekonomické metodologie.

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. Estimating 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 activities. Schneider and Enste (2002) divide the methods of estimation into two main groups: direct and indirect. The first group is composed of surveys and other inquiries regarding 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 tax authorities provide an easy reference point when answering survey questions. In another direct method, tax authorities in many countries attempt to estimate tax evasion from audited tax returns.[2]

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

1. national accounting approaches focusing on the discrepancy between national accounting sources and uses data (the so-called "macroeconomic approach") or the discrepancy between reported incomes and expenditures of households ("microeco-

[1] Such inefficiencies might be caused by resources being used in an evasion effort instead of in productive activities. They might also arise because the need not to 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.

nomic approach");

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

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) can also be included as a possible means of measurement.

Macroeconomic methods of estimating 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). The assumption of constant velocity of money implied in many papers using the monetary method is suspect, while 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 the *a priori* identification division of the population into a subset that is assumed not to evade (typically wage and salaried workers) leaving all hidden income to be attributed to the rest of the sample (especially the self-employed or farmers). This simplifying assumption is, however, weak both theoretically (see Kolm and Nielsen, 2008 for a model that includes concealment of income by firms and salary workers) and empirically. For example, the Eurobarometer survey (European Commission, 2007) reports 5 percent of respondents in the EU admitting that they carried out undeclared work in the preceding 12 months. National values of this percentage range substantially, with the highest share in Denmark (18 percent) and the lowest share in Cyprus (1 percent). The Czech and Slovak Republics, which we will analyze below, 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 the 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]

In a pioneering work, Pissarides and Weber (1989) use self-employment to identify households that might under-report income. They 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 et al. (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.

Additional work that identifies under-reporting based on self-employment status includes Hurst, Li, and Pugsley (2010) and Tedds (2010). The latter study criticized previous works on three main grounds: (1) that they assumed constant fraction of underreporting on total income, (2) they assumed a specific form of the underreporting function and (3) they relied on monotonicity of the expenditure function w.r.t. income.[4] As a remedy the author used a non-parametric estimation of food Engel curves. This estimation strategy, however, still hinges on the assumption that only self-employed individuals evade.

Studies that estimate the evasion response to tax changes can provide added insight. Gorodnichenko et al. (2009) used the 2001 flat tax reform in Russia as a natural experiment that produced a "control group" consisting of a part of the population for whom the marginal tax rate did not change and thus whose income under-reporting could be

[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]The last criticism applies specifically to Lyssiotou et al. (2004) who used a complete demand system. Goods that were shown to violate this assumption include alcohol and tobacco.

compared with a "treatment group" of individuals for whom the marginal tax fell. As a result, they did not need the *ex ante* assumption about which groups of individuals evade, however they estimated only the change in the shadow economy, not its overall size.

We propose a way to avoid the problem of arbitrary *a priori* assignment of individuals to evading and non-evading groups econometrically by using an endogenous switching regression with an unknown sample separation rule. Such a technique has not here-to-fore been applied to the shadow economy,[5] although it has been used elsewhere. In an early study, Dickens and Lang (1985) used such a model to test the theory of dual labor markets. 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 studies showed the feasibility of maximum likelihood and other estimation techniques in this situation.

The methodology of endogenous switching regression with unobserved separation will thus allow the relaxing of overly restrictive assumptions including an *ad hoc* specification of under-reporting groups or requiring that evaders under-report 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 et al. (2009) based on three assumptions coming from the permanent income hypothesis (Fried-

[5]DeCicca et al. (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.

man, 1957):

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

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 under-reporting ($\mathbf{S}_i$). In estimates presented below this vector includes age (older people are more risk averse and, therefore, less prone to tax evasion), education, whether workers in the household are self-employed, working in a large or small firm (small firms are more prone to save labor costs by paying a low "official" wage combined with a part of the wage paid "under the table"), or employed in the public or private sector (government is usually less likely to pay its employees "under the table", although on the other hand, public employees may be more prone to accepting bribes). 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 the household and his or her spouse including their ages, education and work experience (vector $\mathbf{L}_{1,i}$). Eq.(3) indicates 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 the household and spouse, number and ages of children, number of other household members, marital status, and education among others. Taking the logarithms of (1), (2) and (3) and substituting yields a definition of the consumption-income gap:

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

where $\log C_i - \log Y_i^R$ is the consumption-income gap of the household. Note that if all other household characteristics are held equal, a higher consumption-income gap in household A compared to household B implies a higher degree of under-reporting on the part of household A.

As in Gorodnichenko et al. (2009), our basic definition of consumption is the expenditure on nondurable goods. We chose this measure as reporting on large purchases of durables may be more unreliable than reporting on smaller nondurable consumption. The household may be inclined to hide larger purchases of durables out of caution or fear, especially if it participates in the informal sector. Moreover, purchases of durable goods are more likely than other expenditure to actually be investment, especially if the household derives part of its income from self-employment. By limiting the measure of consumption to nondurables, however, we make an assumption that preferences over nondurable and durable goods are homothetic, implying that the income elasticity of nondurable goods is unitary. This assumption has often been used in macroeconomic literature (see Eichenbaum and Hansen, 1990 or Ogaki and Reinhart, 1998), although Pakoš (2011) criticized it. Even his estimate of income elasticity of nondurable goods is, however, relatively close to one, lying in the interval of [0.882, 0.954].

A possible problem with basing estimates on nondurable consumption is that such items may include tax deductible purchases for self-employed individuals. This is usually not the case with food as used by Pissarides and Weber (1989). Expenditures on food, however, may not meet the homotheticity requirement. We will report estimates based on both food and food plus other nondurables and find these to be gratifyingly consistent, suggesting that neither of these potential problems is critical.

2.2 From consumption-income gap to shadow economy

We now extend the above analysis of the consumption-income gap. Without much loss of generality we that there are two groups of individuals in every economy: those who evade and those who do not. These two groups of agents differ, all other characteristics held constant, by the average size of the gap between their income and consumption. For non-evaders. γ in Eq.(4) is equal to 0 by definition. . Since consumption should be based on true rather than reported 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 et al., 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 agents 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 the costs of evasion including expected fines and costs associated with hiding of the income (including psychic costs) of household i. One can think of Eq.(7) as a reduced form equation of an underlying optimization problem. In this equation, agents compare the maximal net benefits from the optimal level of under-reporting with cheating with those from reporting incomes accurately.

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:

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

For estimating purposes, this system can be expressed as follows:

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

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

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

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

where $\mathbf{X}_i$ is the vector of explanatory variables that affect consumption and income and $\mathbf{Z}_i$ is the vector of variables that affect the tax evasion propensity.

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:

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

If we assume that $(\varepsilon_e, \varepsilon_{ne}, \varepsilon_s) \sim N(0, \Sigma)$, where:

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

the log-likelihood function (13) becomes:

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

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

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

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

Note that, as usual in this type of estimation, $\sigma_{e,ne}$ is unidentifiable, as the two regimes never occur at the same time (see Maddala, 1983). 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 other 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 the consumption-income gap, such as the dummies for 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:

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

which is household i's estimated degree of income under-reporting 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 the 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:

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

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 an economy's officially reported income.

To increase the robustness to the choice of initial values and the presence of outliers, Monte Carlo simulations were used to compute both means and standard errors of the estimators. For each country, 250 random samples with replacement were drawn from the data, with the estimation of Eq.(14) and a computation of the shadow economy from Eqs.(18) and (19) done for each sample.[6] This results in a data series from which the means of these estimates can be computed. Standard errors are then the standard errors of these means.

3 Data

We illustrate the value of our estimator by applying it to recent data from the Czech and Slovak Republics. The choice of these countries was not arbitrary. Rather, they represent modern, EU member economies with the required data collected by Eurostat standards but where the assumption that only self-employed households hide income (as assumed by Pissarides and Weber (1989)) seems particularly questionable. In both countries we use the Household Budget Survey from 2008.

3.1 Czech Republic

The data from the Czech household budget survey for 2008 contain information about the income from various sources and expenditures on different categories of goods and services for 3,271 Czech households. We restrict our analysis to a subsample of households with

[6] See Appendix A for details.

working heads.[7] Summary statistics (weighted means) for this subsample are given in Table 1. The definition of disposable 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 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. We define our main consumption variable as a sum of expenditures on non-durable goods. More specifically, our definition includes expenditure on food both at home and away from home, alcohol and tobacco, clothing and footwear, rents, utilities and other services. 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.

Following the discussion in the Methodology section, X_i contains variables such as number of household members of different categories, education, relationship status, age, and age squared. The last two variables are an indicator for work experience.

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

Variable	**Mean**	**Std. Dev.**
Total no. of households	2,138	N/A
average no. of household members	2.606	1.192
average no. of heads with a spouse or a partner	1,486	N/A
average no. of children	0.817	0.943
average monthly disposable income of households (CZK)	30,979	16,550
average age of head	45.306	11.073
no. of self-employed heads	456	N/A
no. of heads working in public sector	610	N/A
no. of heads working in private sector	1,072	N/A
no. of blue-collar heads	1,170	N/A
no. of heads with secondary education	1,753	N/A
no. of heads with a bachelor's degree or higher	264	N/A

[7]The reduction in sample size is primarily due to the presence of households headed by retirees.

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 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, 2008

Variable	**Mean**	**Std. Dev.**
Total number of households in the subsample	2,885	N/A
average no. of household members	3.16	1.307
share of married households	2,156	N/A
average no. of children	1.048	1.053
average monthly disposable income of households (SKK)	33365.971	12972.335
average age of head	44.096	9.829
no. of self-employed heads	483	N/A
no. of heads working in public sector	791	N/A
no. of heads working in private sector	1,611	N/A
no. of blue-collar heads	1,216	N/A
no. of heads with a high school degree	1,425	N/A
no. of heads with a bachelor's degree or higher	457	N/A

4 Results

The results of maximum likelihood estimation of the structural endogenous switching model for Czech and Slovak Republics based on non-durable consumption (including food) are shown in Tables B1 and C1 in the Appendix, respectively.[8] These estimates, together with the confidence intervals, were obtained from the Monte Carlo method described above. Note that in both cases the likelihood ratio test rejects the null hypothesis of joint statistical insignificance of estimates at 1% level[9]. Plugging the estimated coefficients in

[8]Those for an equation based on food only are available at

[9]The likelihood ratio test is a natural choice to test the assumption that are divided 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. Following Dickens and Lang (1985), the degrees of freedom are equal to number of constraints plus the number of unidentified

these tables into Eq.(18) yields the estimates of the shadow economy in Tables 3 and 4. Results based on food consumption and total nondurable consumption are quite close and both are very tightly estimated. The key finding is that the shadow economy in the Czech Republic constituted between 20 and 22 percent of reported income in 2008, while in Slovakia this fraction was between 29 and 30 percent. To to arrive at true income in these economies, we have to multiply the officially reported income by 1.2 and 1.3 respectively. These estimates for the Czech Republic are slightly higher than those reported by Schneider et al. (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, under-reporting of income extends to wage and salary workers as well as the self-employed.

Table 3: Shadow economy estimates — Czech Republic (2008)

Consumption	Shadow economy (% of reported income)	SE (bootstrapped)	Shadow economy (% of total income)	SE (bootstrapped)
Nondurables	21.99%	0.99%	17.16%	0.36%
Food	20%	1.78%	16.7%	0.29%

Table 4: Shadow economy estimates — Slovak Republic (2008)

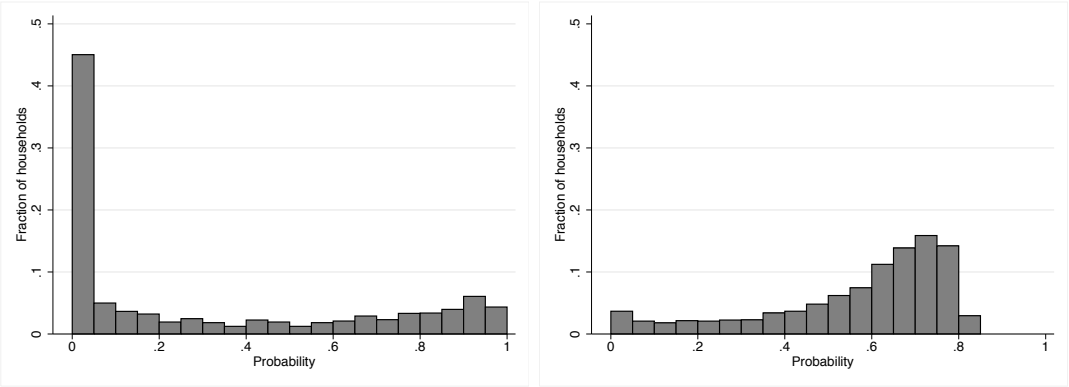
Consumption	Shadow economy (% of reported income)	SE (bootstrapped)	Shadow economy (% of total income)	SE (bootstrapped)
Nondurables	30.44%	1.12%	22.29%	0.29%
Food	28.6%	1.63%	22.24%	0.44%

Equation(19) enables calculation of the predicted probability of hiding income for every household in the sample defined on the interval [0, 1]. As might be expected from Tables 3 and 4, the mean of this estimated probability is substantially higher in Slovakia, where

parameters (found only in the switching equation). As argued by Goldfeld and Quandt (1976), this leads to a conservative critical value.

the average household has an estimated 54 percent probability of hiding at least some income than it is in the Czech Republic where the corresponding estimated probability is 34 percent. As can be seen in Fig.1, which plots the distribution of probabilities across the samples, there is a bimodal pattern with mass concentrated at or near zero in both countries and then a second concentration at higher probabilities, with the main mass at a substantially higher probability in Slovakia.

Figure 1: Histograms of evasion probabilities



(a) Czech Republic (b) Slovakia

The impact of various factors on the probability of a household under-reporting income (computed for each observation and then averaged) corresponds with intuition as can be seen in (B2) and (C2). Households headed by women are substantially less likely to underreport income (by 12 percentage points in each country). This is likely to be due to higher risk aversion on the part of women.[10] The same is true for married households in Slovakia (in other words, households headed by single males are the most likely to underreport.)Job characteristics (blue collar employment, self-employment and working in the public sector) of household heads are uniformly more predictive than that of their spouses, again probably due to greater variation in males' behavior with respect to underreporting. In both countries households working in the public sector are less likely to hide income although the effect is higher when the head is so employed than when the spouse. Results with respect to self-employment are somewhat puzzling. Such status, as expect, has a substantial effect for both household heads and their spouses in Slovakia

[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).

while in the Czech Republic both effects are actually negative. In both countries households headed by blue collar workers (or containing spouses with blue collar jobs) are less likely to underreport. Workers with high school degrees are less likely to underreport than those with either more or less education.

Overall, these results suggest that, in addition to being more extensive overall in Slovakia, the propensity to under-report income in more generalized there than in the Czech Republic. The findings with respect to both extent and composition of under-reporting are consistent with the lower overall level of economic development in Slovakia.[11]

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 and that excluding the possibility of such hidden income will seriously underestimate the size of the shadow economy.

[11] In 2008 when our data was collected, the GDP per capita was 75 percent greater in the Czech Republic than in Slovakia (at $23,833 as opposed to $13,603). Schneider (2012) reports that among OECD countries the lower GDP per capita in a country, the higher is the incentive to work in the shadow economy.

References

Allingham, M. G., Sandmo, A., 1972. Income tax evasion: A theoretical analysis. Journal of Public Economics 1 (3-4), 323–338.

Arunachalam, R., Logan, T. D., October 2006. On the heterogeneity of dowry motives. Working Paper 12630, National Bureau of Economic Research.

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

Caudill, S. 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, P., Kenkel, D. S., Liu, F., April 2010. Excise tax avoidance: The case of state cigarette taxes. Working Paper 15941, National Bureau of Economic Research.

Dickens, W. T., Lang, K., 1985. A test of dual labor market theory. The American Economic Review 75 (4), 792–805.

Douglas, S. M., Conway, K. S., Ferrier, G. D., 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, L. C., 2007. Heckman's selection model, endogenous and exogenous switching models: A survey.

Eckel, C. C., Grossman, P. J., 2008. Men, women and risk aversion: Experimental evidence. In: Plott, C. R., Smith, V. L. (Eds.), Handbook of Experimental Economics Results. Elsevier, Ch. 113, pp. 1061 – 1073.

Eichenbaum, M., Hansen, L. P., 1990. Substitution using time with intertemporal aggregate series data. Journal of Business & Economic Statistics 8 (1), 53–69.

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

Friedman, M., 1957. The relation between the permanent income and relative income hypotheses. pp. 157–182.

Goldfeld, S. M., Quandt, R. E., 1976. Techniques for estimating switching regressions. Cambridge, MA: Ballinger, pp. 3–35.

Gorodnichenko, Y., Martinez-Vazquez, J., Sabirianova Peter, K., 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, J., Palda, F., 2006. Problems measuring the underground economy in transition. Economics of Transition 14 (4), 707–718.

Harberger, A., 1964. Taxation, resource allocation, and welfare. Princeton University Press, pp. 25–80.

Hurst, E., Li, G., Pugsley, B., November 2010. Are household surveys like tax forms: Evidence from income underreporting of the self employed. Working Paper 16527, National Bureau of Economic Research.

Kolm, A.-S., Nielsen, S. B., 2008. Under-reporting of income and labor market performance. Journal of Public Economic Theory 10 (2), 195–217.

Kopczuk, W., Lupton, J. P., 2007. To leave or not to leave: The distribution of bequest motives. The Review of Economic Studies 74 (1), 207–235.

Lee, L.-F., Porter, R. H., 1984. Switching regression models with imperfect sample separation information–With an application on cartel stability. Econometrica 52 (2), 391–418.

Lichard, T., 2012. Shadow economy in the Czech republic, Russia, Slovakia and Ukraine: Food Engel curve approach, unpublished manuscript.

Lyssiotou, P., Pashardes, P., Stengos, T., 2004. Estimates of the black economy based on consumer demand approaches. Economic Journal 114 (497), 622–640.

Maddala, G. S., 1983. Limited-dependent and qualitative variables in econometrics. No. 3. Cambridge University Press, Cambridge.

Ogaki, M., Reinhart, C. M., 1998. Measuring intertemporal substitution: The role of durable goods. Journal of Political Economy 106, 1078–1098.

Pakoš, M., Jul. 2011. Estimating intertemporal and intratemporal substitutions when both income and substitution effects are present: The role of durable goods. Journal of Business & Economic Statistics 29 (3), 439–454.

Pissarides, C. A., Weber, G., 1989. An expenditure-based estimate of Britain's black economy. Journal of Public Economics 39 (1), 17–32.

Schneider, F., Mar. 2012. The shadow economy and work in the shadow: What do we (not) know? Discussion Paper 6423, IZA.

Schneider, F., Buehn, A., Montenegro, C., 2010. New estimates for the shadow economies all over the world. International Economic Journal 24 (4), 443–461.

Schneider, F., Enste, D. H., 2002. The Shadow Economy: An International Survey. Cambridge (UK): Cambridge University Press.

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

Slemrod, J., Blumenthal, M., Christian, C., 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, L. M., 2010. Estimating the income reporting function for the self-employed. Empirical Economics 38 (3), 669–687.

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

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[12] 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 $\boldsymbol{\delta}$, a probit regression of I_i on $\mathbf{Z}_i$ is run. After that we use these values $\left(\hat{\delta}\right)$ to estimate initial values of the $\boldsymbol{\beta}$'s by running the following OLS regressions:

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

and

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

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

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

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:

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

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

These initial values of δ, $\boldsymbol{\beta}$'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 the shadow economy size and standard errors) are computed.

B Estimation Results - Czech Republic

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

VARIABLES	Shadow sector $\ln C - \ln Y$		Official sector $\ln C - \ln Y$		Switching equation Latent variable	
constant	-0.452***	(0.012)	0.264***	(0.637)	3.061***	(0.459)
# of children	-0.000	(0.001)	0.007	(0.007)		
# of employed	0.000	(0.002)	-0.034***	(0.012)		
# of unemployed	0.004**	(0.002)	-0.054***	(0.015)		
is married	0.003	(0.004)	0.097	(0.608)	-1.045**	(0.434)
high school degree	-0.000	(0.003)	-0.018	(0.017)	-0.010	(0.029)
bachelor's degree or higher	-0.007***	(0.002)	-0.047*	(0.027)	0.108***	(0.039)
high school degree (spouse)	0.003	(0.003)	0.046	(0.029)	-0.071*	(0.041)
bachelor's degree or higher (spouse)	1.242	(1.419)	-0.080	(0.151)	0.867	(1.044)
age	0.001	(0.000)	0.010*	(0.005)	-0.029***	(0.007)
age^2	-0.000	(0.000)	-0.001*	(0.001)	-0.000***	(0.000)
hoh is female	0.001	(0.004)	0.072	(0.610)	-1.021**	(0.432)
has children					-0.000	(0.022)
blue collar					-0.012	(0.016)
works in public sector					0.028	(0.017)
self-employed					0.026	(0.018)
spouse in public sector					0.010	(0.022)
white collar spouse					0.043	(0.028)
blue collar spouse					0.040	(0.034)
self-employed spouse					0.945	(0.675)
σ_1	0.286***	(0.001)				
σ_2			0.847***	(0.017)		
σ_{13}					0.254***	(0.004)
σ_{23}					-0.721***	(0.030)
Observations			2,138			
Log likelihood			-342320			
LR test			59814			
Prob>$\chi^2(40)$			0.0000			

Bootstrapped standard errors in parentheses: *** p<0.01, ** p<0.05, * p<0.1. The structural coefficients for consumption-income gap equations express also the marginal effects given variables have on consumption-income gap. The structural coefficients for switching equation do not have a straightforward interpretation. The marginal effects on probability are shown in Table B2.

Table B2: Marginal effects - Czech Republic

VARIABLES	Probability of being in the shadow sector
is married	0.009
age	0.039
age^2	-0.000
female	-0.124
has children	0.045
high school degree	-0.081
bachelor's degree or higher	-0.047
high school degree (spouse)	0.032
bachelor's degree or higher (spouse)	0.061
blue collar	-0.009
self-employed	-0.059
works in public sector	-0.015
blue collar spouse	-0.004
white collar spouse	0.058
self-employed spouse	-0.006
spouse in public sector	-0.006

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)

VARIABLES	Evading regime $\ln C - \ln Y$		Non-evading regime $\ln C - \ln Y$		Switching equation N/A (latent)	
constant	-0.167***	(0.069)	-0.132	(0.139)	-2.751***	(0.828)
# of children	-0.008***	(0.002)	-0.019***	(0.001)		
# of employed	-0.113***	(0.002)	-0.080***	(0.002)		
# of unemployed	-0.038***	(0.002)	-0.046***	(0.002)		
is married	0.025***	(0.0048)	-0.050***	(0.0120)	0.021***	-0.004
high school degree	0.038***	(0.013)	0.042***	(0.005)	-0.180***	(0.037)
bachelor's degree or higher	0.012	(0.013)	-0.050***	(0.011)	-0.209***	(0.039)
high school degree (spouse)	-0.007	(0.010)	-0.041	(0.126)	0.058	(0.125)
bachelor's degree or higher (spouse)	-0.116***	(0.017)	-0.016	(1.355)	-0.007	(0.933)
age	0.005*	(0.003)	-0.021***	(0.002)	0.128***	(0.011)
age^2	0.000	(0.000)	0.000	(0.000)	-0.001***	(0.000)
female	0.001	(0.011)	0.050***	(0.007)	0.037	(0.743)
has children					0.165***	(0.017)
blue collar					-0.035***	(0.013)
works in public sector					-0.056***	(0.010)
self-employed					-0.218***	(0.026)
blue collar spouse					-0.013	(0.760)
white collar spouse					0.212	(1.229)
spouse in public sector					-0.021	(0.019)
self-employed spouse					-0.021	(0.932)
σ_1	0.250***	(0.001)				
σ_2			0.547***	(0.009)		
σ_{13}					0.184***	(0.022)
σ_{23}					0.487***	(0.023)
Observations			2,885			
Log likelihood			-510636			
LR test			434086			
Prob>$\chi^2(40)$			0.0000			

Bootstrapped standard errors in parentheses: *** p<0.01, ** p<0.05, * p<0.1. The structural coefficients for consumption-income gap equations express also the marginal effects given variables have on consumption-income gap. The structural coefficients for switching equation do not have a straightforward interpretation. The marginal effects on probability are shown in Table C2.

Table C2: Marginal effects - Slovak Republic (2008)

VARIABLES	Probability of being in the shadow sector
is married	-0.125
age	0.044
age^2	-0.000
female	-0.124
has children	-0.017
high school degree	-0.073
bachelor's degree or higher	-0.015
high school degree (spouse)	0.050
bachelor's degree or higher (spouse)	0.089
blue collar	-0.075
self-employed	0.152
works in public sector	-0.050
blue collar spouse	-0.018
white collar spouse	-0.010
self-employed spouse	0.087
spouse in public sector	-0.008

The average marginal effects were computed as the average of marginal effects predicted for every observation in the subsample.