

The underreporting of income by self-employed workers. A new interpretation with Spanish data

Diego Martinez-Lopez*

Department of Economics, University Pablo de Olavide

September 26th 2012

Abstract

This paper offers estimates of the underreporting of income by self-employed workers using the Spanish household surveys over the period 2006-2009. We use the well-known model by Pissarides and Weber (1989) but extending its interpretation for admitting also the concealment of income by salary workers. Our results show that the reported income by self-employed has to be increased by about 25 percent to obtain the level of income which would equal the level of underreporting by employees. Although our estimates are robust to changes in specification, endogeneity and non-linearities, the extent of underreporting is indeed sensitive to the types of self-employed and salary workers used in the comparison.

Keywords: underreporting, household surveys, food consumption, tax evasion.

JEL Classification: D12, H26, O17.

*Corresponding author: Diego Martinez, Department of Economics, University Pablo Olavide, Ctra. Utrera Km. 1, 41013 Seville. Spain. Email: dmarlop1@upo.es

1 Introduction

One of the most extended ways of tax evasion is that related to the underreporting of income by self-employed workers. As long as their incomes are not subject to third-party reporting, the probability of being detected by the tax authorities in case of hiding earnings is lower than in the case of salary workers, and this leads to higher levels of tax evasion. This acquires more relevance in a country like Spain, where the high levels of unemployment clearly affect the individual's decisions to work in the underground economy (Ahn and de la Rica, 1997), and the lack of flexible enough rules in the labor markets may encourage the informal economy (Bentolila et al, 1994).

As other phenomena of tax evasion, the first challenge to approach it lies in the difficulty of measuring the extent of such concealment. The standard method is based on the seminal paper by Pissarides and Weber (1989), which uses the Engel curves for food demand. The underlying idea is simple. Both salary and self-employed workers report accurately their food expenditures in household budget surveys. By contrast, when they are asked about their earnings, only the salary workers say their true income. The estimate of underreporting of income by the self-employed workers is then given by the comparison of food expenditures of both groups in function of declared income, given other economic and demographic characteristics. A detailed explanation of this method is provided in the next section.

On this basis, a number of papers have offered estimates of underreporting for different samples. In essence, what is computed is the number by which the reported income of self-employed has to be multiplied to obtain the true income. For the UK economy, Pissarides and Weber (1989) give a central value of 1.55 in 1982. From another point of view, Lyssiotou et al. (2004), using a complete demand system approach and non-parametric estimation methods, suggest that the extent of underreporting by self-employed workers in the UK in 1993 goes from 118 per cent for households with head in blue collar occupation to 64 per cent for white collar jobs.

With data of Canada, Schuetze (2002) finds, for some years between 1969 and 1992, estimates that go from 11 per cent to 23 per cent as average values of lower and upper bound estimates, respectively. For the period 1994-1996, Johansson (2005) gives a range of estimates between 16 and 40 per cent of underreporting in Finland, depending on the definition used for the self-employed household. More recently, Engstrom and Holmlund (2009) conclude that the Swedish households with at least one self-employed member underreport their income by around 30% in early 2000s. And Hurst et al. (2011), using three data samples for the US in the 80s, 90s and early 2000s, estimate the degree of underreporting by between 25 and 35 per cent.

From the very beginning of this literature, most of papers assume that employees do not hide part of their income and underreporting is exclusively concentrated on self-employed workers. But this simplifying assumption is weak from both theoretical (see, for instance, Kolm and Nielsen, 2008, for a

model with concealment of income by firms and salary workers) and empirical points of view. In this sense, the 2007 Eurobarometer shows that 5 per cent of all dependant employees in a representative sample of individuals in the EU admitted having received all or part of their salary as envelope or cash-in-hand wages¹.

Our own sample shows a number of indications leading to think that also the salary workers partly hide their earnings. For instance, about 12% of households declaring that the main source of their income is on their own ("cuenta propia") classify themselves as salary workers. Furthermore, the number of employees that do not inform on their monthly income in the survey is about twice than that of respondents (obviously, this figure is higher in the case of self-employed workers -about three times- but both of them may reveal concealment of income).

This paper applies the methodology by Pissarides and Weber (1989) to get an estimation of the extent of underreporting by the Spanish self-employed over the period 2006-2009. Our data come from the Spanish Household Budget Surveys. The robustness of our results has been checked using alternative specifications, testing for non-linearities in the relationship between income and food expenditure, and dealing with potential problems of endogeneity. Different measures of the key variables have been examined as well.

In this context, we can summarize the main contributions of the paper as follows. Firstly, we replicate the well-known approach of estimating food demand functions for making explicit a measure of concealment of income in a sample that has never been exploited in this regard. Secondly, the interpretation of the standard theoretical model is extended here to consider the possibility that the salary workers also conceal part of their incomes; in fact, this can be seen not only as a realistic assumption but also as a reasonable interpretation of our results. Consequently, our estimates of underreporting by self-employed are always in relation to the employees and we guess this should be also the appropriate interpretation of results found in previous papers.

After the Introduction, we set up the theoretical framework used to measure the extent of underreporting of income. Section 3 explains the main features of data and the criteria followed to build the sample. Section 4 gives details of estimation procedures and shows the results. Finally, section 5 concludes.

2 The model

This section aims to build an analytical framework to estimate the degree of underreporting of income by households with self-employed workers as household heads. The approach used here is based on

¹National values of this percentage range from 23 per cent of Romania to 1 per cent of UK. Spain has the same figure than the EU as a whole; however, when the criterion is the number of hours spent on undeclared work, Spain is clearly above the European average.

the following main assumptions: i) Food expenditures are correctly reported by households in budget surveys; but ii) this not the case of income. Previous studies have qualified this second assumption setting that salary workers are completely honest by reporting their income while self-employed workers hide part of their earnings. However, we really think that a most adjusted picture to the real world involves salary workers (at least some of them) that also conceal partially their income, although in a lower degree than self-employed workers. Consequently, a natural test for measuring the relative extent of such a underreporting by self-employed workers consists of comparing food demand functions -which depend on income- of both groups.

Our starting point is the model by Pissarides and Weber (1989), which we shall hold almost in its totality but introducing the chance of underreporting by salary workers. This innovation not only allows to keep manageable the empirical estimation but also to broaden the interpretation of the results. Particularly, our measure of underreporting by self-employed will be a relative measure which takes as reference a given level (and strictly positive) of hidden income by salary workers.

Let Y_i be the true income of household i . We shall distinguish two types of households, denoted by SW and SE , which refer to salary worker and self-employed worker households, respectively. As usual in the definition of consumption functions, a relation between the observable income Y_i and the permanent income Y_i^p has to be set up:

$$Y_i = p_i Y_i^p, \quad (1)$$

where p_i is a random variable to take into consideration the deviations of observable income from its permanent, long-run value. It is assumed that the mean of p_i is the same for all the households in the economy but the variance of p_i to be higher for self-employed households than for salary workers. This can be seen as a reasonable assumption as long as self-employed workers face more risks and, consequently, a more volatile income is to be expected in their case.

Let Y_i' be the disposable income reported by households in budget expenditure surveys. As said before, previous papers have assumed that salary workers report correctly all their income. In our framework, by contrast, and using a slight modification of the Pissarides and Weber's model, we will allow the phenomenon of underreporting of income also for salary workers. True income Y_i and reported income Y_i' are related as follows:

$$Y_i = k_i Y_i', \quad \text{with } k_i > 1. \quad (2)$$

k_i is a random variable that indicates to what extent household i hides part of her true income Y_i . In other words, k_i is the number by which the reported income Y_i' must be multiplied so as to get the true income Y_i . Both types of workers hide part of their income but in a different proportion: $k_{SE} > k_{SW}$, that is, self-employed households underreport more disposable income than salary households.

Combining equations (1) and (2), and after logarithmical transformation, the log of permanent

income is:

$$\ln Y_i^p = \ln Y_i' - \ln p_i + \ln k_i, \quad (3)$$

which becomes one of the key variables by estimating the following food expenditure function:

$$\ln F_i = \boldsymbol{\alpha} \mathbf{X}' + \beta \ln Y_i^p + \varepsilon_i, \quad (4)$$

where F_i is the food expenditure of household i , $\boldsymbol{\alpha}$ is a vector of parameters common to salary and self-employed worker households, \mathbf{X} is a vector of household characteristics, β is a scalar that can be interpreted as the marginal propensity to consume food, and ε_i is a white noise. In a sense, what expression (4) represents is a log-linear Engel curve for food consumption.

At this point, the main caveat by estimating the above Engel curve is that we have no data on p_i and k_i (in fact, the latter is the measure of underreporting that we are looking for). Thus, we need to make some assumptions on their distribution over the sample. As is usual in literature, we set up:

$$\ln p_i = \mu_i^p + u_i \quad (5)$$

$$\ln k_i = \mu_i^k + v_i, \quad (6)$$

that is, both variables are log-normal distributed, with particular values of μ^p and μ^k for salary and self-employed workers. Disturbances u_i and v_i are assumed to have zero means and constant (but differentiated among both types of workers) variances $\sigma_{u_i}^2$ and $\sigma_{v_i}^2$.

Substituting (5) and (6) into (3), and in turn into (4), we get:

$$\ln F_i = \boldsymbol{\alpha} \mathbf{X}' + \beta \ln Y_i' - \beta(\mu_i^p - \mu_i^k) - \beta(u_i - v_i) + \varepsilon_i. \quad (7)$$

When this Engel curve is adjusted with a dummy variable to reflect the appropriate distribution of k and p across groups $i = SE, SW$, one finds

$$\ln F_i = \boldsymbol{\alpha} \mathbf{X}' + \beta \ln Y_i' - \beta(\mu_{SW}^p + DSE_i(\mu_{SE}^p - \mu_{SW}^p) - (\mu_{SW}^k + DSE_i(\mu_{SE}^k - \mu_{SW}^k))) + \eta_i,$$

where DSE_i is dummy variable that takes the value 1 if the household head of family i is self-employed worker and 0 if salary worker, and η_i is the error of regression that, by construction, includes not only unexplained variations in household food expenditures but also deviations of their actual income from its permanent income and of their reported income from their true income. The estimation of this equation requires further algebra manipulation using the properties of log-normal distributions. Particularly,

$$\ln \bar{p}_i = \mu_i^p + \frac{1}{2} \sigma_{u_i}^2 \quad (8)$$

$$\ln \bar{k}_i = \mu_i^k + \frac{1}{2} \sigma_{v_i}^2, \quad (9)$$

where a bar over a variable denotes its mean. Assuming that the mean of p_i is the same for salary and self-employed workers ($\ln \bar{p}_{SE} = \ln \bar{p}_{SW}$), after substituting for $\mu_{SE}^p - \mu_{SW}^p$, the above Engel curve can be written as

$$\ln F_i = \boldsymbol{\alpha} \mathbf{X}_i' + \beta \ln Y_i' + \gamma DSE_i + \eta_i, \quad (10)$$

where $\gamma = \beta \left[\theta - \frac{1}{2} (\sigma_{v_{SE}}^2 - \sigma_{v_{SW}}^2) + \frac{1}{2} (\sigma_{u_{SE}}^2 - \sigma_{u_{SW}}^2) \right]$ and $\theta = \ln \bar{k}_{SE} - \ln \bar{k}_{SW}$.

As can be seen from the expression which relates γ , β and θ , the extent of underreporting of income estimated is an interval whose limits depend upon the extreme values for variances of u and v in each type of household. The usual approach to get estimates of such as variances involves the computation of residual variances in the following regression for income:

$$\ln Y_i' = \boldsymbol{\Omega} \mathbf{X}_i' + \boldsymbol{\Gamma} \mathbf{Z}_i' + \xi_i, \quad (11)$$

where \mathbf{Z}_i is a vector of variables used as instruments in IV-2SLS estimates of expression (10), given the potential endogeneity of Y_i' . Again, the error term ξ_i has three components: unexplained variations in household permanent income, deviations of their actual income from its permanent income and deviations of their reported income from their true income. If the first component is assumed to be the same in both the salary and self-employed workers -which seems to be a reasonable assumption given that the risks of omitting variables related to the distinction between self-employed vs salary workers are null when a dummy is included or a separate estimation by type of household is considered-, we can write

$$\sigma_{\xi_{SE}}^2 - \sigma_{\xi_{SW}}^2 = \sigma_{u_{SE}}^2 + \sigma_{v_{SE}}^2 - 2cov(uv)_{SE} - \sigma_{u_{SW}}^2 - \sigma_{v_{SW}}^2 + 2cov(uv)_{SW}. \quad (12)$$

On the other hand, given the value of γ above, the relative underreporting of income by self-employed households is given then by

$$\theta = \frac{\gamma}{\beta} + \frac{1}{2} (\sigma_{v_{SE}}^2 - \sigma_{u_{SE}}^2 + \sigma_{u_{SW}}^2 - \sigma_{v_{SW}}^2). \quad (13)$$

Note that (13) is quite similar to the expression (18) of Pissarides and Weber (1989), where the level of underreporting of income by salary workers is fixed at zero, and consequently the term $\sigma_{v_{SW}}^2$ does not appear. If we set up that the covariance between u and v are null for both types of households, lower and upper bounds for the relative underreporting of income by self-employed households are obtained². Taken the variances for salary workers as parameters, we see that the minimum value for θ is obtained when $\sigma_{v_{SE}}^2$ reaches its lowest value, that is, when it is equal to $\sigma_{v_{SW}}^2$. Under such a case,

$$\theta = \frac{\gamma}{\beta} - \frac{1}{2} (\sigma_{\xi_{SE}}^2 - \sigma_{\xi_{SW}}^2), \quad (14)$$

²As Pissarides and Weber (1989) show, alternative assumptions on partial correlation coefficient between u and v have not a significant impact on estimates of underreporting.

where (12) has been used. By contrast, it is easy to see that (13) reaches its maximum value when $\sigma_{u_{SE}}^2$ is at its minimum feasible value, which in our model is like in Pissarides and Weber (1989): $\sigma_{u_{SE}}^2 = \sigma_{u_{SW}}^2$.³ This gives an upper bound for the extent of underreporting of income by self-employed households:

$$\theta = \frac{\gamma}{\beta} + \frac{1}{2}(\sigma_{\xi_{SE}}^2 - \sigma_{\xi_{SW}}^2) \quad (15)$$

Given the fact that salary workers also partially hide their income, the computation of these lower and upper bounds of the degree of underreporting shows a caveat in Pissarides and Weber (1989)'s approach. Particularly, it can be seen that the value of θ they estimate is not the same θ' (to distinguish from the previous one) which is derived from their theoretical framework. In the case of the lower bound, they initially set up $\sigma_{v_{SE}}^2$ equal to 0; this leads to the following expression of θ' :

$$\theta' = \frac{\gamma}{\beta} - \frac{1}{2}(\sigma_{u_{SE}}^2 - \sigma_{u_{SW}}^2). \quad (16)$$

After that, on the basis of this setting, Pissarides and Weber (1989) substitute the terms between parenthesis of (16) by $(\sigma_{\xi_{SE}}^2 - \sigma_{\xi_{SW}}^2)$ and that is what they really estimate. But comparing the parenthesis of (16) with (12), we see that both of them are not equal with employees concealing income ($\sigma_{v_{SW}}^2 \neq 0$). Consequently, the expression of lower bound of underreporting that they derived from their theoretical framework has to be adjusted in order to have full correspondence with the empirical estimation of θ , namely, $\theta = \theta' + \frac{1}{2}\sigma_{v_{SW}}^2$.

A similar argument can be managed for the case of the upper bound. The model by Pissarides and Weber (1989) sets $\sigma_{u_{SE}}^2 = \sigma_{u_{SW}}^2$ (as in this paper), that applied on (13) gives

$$\theta' = \frac{\gamma}{\beta} + \frac{1}{2}\sigma_{v_{SE}}^2, \quad (17)$$

where again their assumption that there is no concealment of income by salary workers is held. However, this is not what they estimate. By contrast, they consider $(\sigma_{\xi_{SE}}^2 - \sigma_{\xi_{SW}}^2)$ in their estimation, that is, $\sigma_{v_{SE}}^2 - \sigma_{v_{SW}}^2$ instead of $\sigma_{v_{SE}}^2$ in (17). Therefore, their definition of θ' must be corrected in order to have what is obtained from their empirical estimation: $\theta = \theta' - \frac{1}{2}\sigma_{v_{SW}}^2$.

In sum, our approach closely follows that of Pissarides and Weber (1989) but admitting the chance that salary workers also may conceal part of their income. Both the manageability and main equations of the original model keep unchanged and only a slight modification in the interpretation of the results must be taken into account: our measure of underreporting of income by self-employed households is *in relation to* a given (and in principle lower) degree of underreporting of income by salary workers. Section 4 shall offer empirical evidence reinforcing this point as long as the extent of concealment by

³There is an erratum at this point in p. 26 of Pissarides and Weber (1989); using their notation, they write $\sigma_{u_{SE}}^2 = \sigma_{u_{SE}}^2$ when it should be $\sigma_{u_{SE}}^2 = \sigma_{u_{EE}}^2$.

self-employed will dramatically hinge upon the particular group of salary worker households used in the comparison.

3 The data

The data used are drawn from the Spanish Household Budget Surveys (EPF in Spanish) from 2006 to 2009 elaborated by the Spanish National Institute of Statistics (INE in Spanish). The sample size is approximately 24,000 households per year, with half of the sample renewed each year⁴. The food consumption expenditures registered in the EPF refer to both the monetary flow on the payment of certain goods and the value of the consumption made by the households in terms of self-consumption and self-supply as well. In this paper, we work with the sum of both of them not only because the econometric estimates become worse if self-consumption and self-supply are not taken into account but also due to the differences between salary and self-employed workers in these items⁵.

In a number of cases (about 25% of households), the INE makes imputations in food expenditures to correct missing values, errors, absence of answer, etc. Our estimates distinguish these two different situations. Anyway, the differences between self-employed workers and salary workers in the percentage of imputation over the total food expenditures are practically null⁶. We also show results with and without meals away home included in the household food expenditures.

There are two variables of interest regarding the household income in the EPF. The first is the net (after taxes) income of household as a whole and the second is the net income of the household head. Both of them are measured in nominal terms. Since both food expenditures and household incomes as nominal variables could be subject to the effect of price changes, we have deflated the former using the food CPI and the latter using the GDP deflator. Estimates only change insignificantly, thus we have decided to report here only the regressions with nominal data.

In a high number of cases (around 70% for salary workers and almost 80% of self-employed workers), the INE makes imputations of the monthly net total income received by the households. This is because a huge number of households do not inform about how much they earn. All these observations based on imputed values have been removed in our sample. This is not the case of net income of the household head, where all the data available here come from the answers of participants.

Salary worker household is defined as that in which the household head is self-reported as salary worker and the corresponding for self-employed worker household. Other criteria have been considered in this key distinction (such as the main source of income for the households) in order to avoid a number

⁴Details on the methodology followed by the INE can be seen at http://www.ine.es/en/daco/daco42/daco4213/resmeto06_en.pdf

⁵The difference between the broad concept of food expenditures and the narrower monetary flow of expenditures is double for self-employed workers when comparing to salary workers.

⁶For the sake of simplicity, we only report here estimates with no imputations in food expenditures.

of inconsistencies⁷. As is usual in this type of papers, families in which the household head works in agriculture, cattle farming or fishing have been removed from the sample; this way, we aim at avoiding that the relationship between food consumption and income to be affected by the particular consumption pattern of these households.

It is not straightforward to set up a clear correspondence between our data and those from other statistical sources. Unfortunately, the Spanish National Accounts do not distinguish between self-employed and salary workers in terms of the primary generation and allocation of income. However, we can see how the composition of our sample is closely similar to that of Economically Active Population Survey (EAPS; EPA in Spanish). While the shares of self-employed and salary workers over total of non-agricultural jobs are respectively about 15 and 84 per cent in the EAPS, the corresponding weights in our sample are 17 and 82 per cent.

Table 1: Differences between SE and SW households in food expenditure and income

	SE	SW	t test on equality of means
Ln (food w/o meals out)	8.17 (0.70)	8.15 (0.74)	0.12 (0.56)
Ln (food)	8.73 (0.68)	8.73 (0.65)	-1.20 (0.89)
Ln (net household income)	9.64 (0.68)	9.80 (0.61)	9.39 (0.00)
Ln (net household head income)	9.30 (0.58)	9.52 (0.51)	14.56 (0.00)

Note: Standard deviations between parentheses except in column (3) where p-value of H_0 : equality of means

Table 1 shows the main intuition behind this paper. Households whose head is a self-employed worker declare to spend in food practically the same than the households headed by a salary worker. But households with a self-employed head systematically always report less income in the EPF than the corresponding salary worker households. These differences are statistical significant (column (3)). In line with previous research, this is a clear indication that self-employed households underreport part of their income⁸. On the other hand, standard deviations of income (whatever the definition used) is always higher in the case of self-employed households than in the case of salary worker households, reflecting a positive sign for the difference $\sigma_{\xi_{SE}}^2 - \sigma_{\xi_{SW}}^2$; this is compatible with a more volatile pattern for self-employed income, as we set up in the theoretical framework⁹.

⁷For instance, a high number of household heads auto-classified as salary workers declare receive no income from firms or government as payment of their labor.

⁸In a different context but also with data from EPF, the informal economy becomes evident as there exists an income-expenditure gap. Ratio between total expenditures per household and household income is 1.30 for salary workers and 1.36 for self-employed.

⁹A classical test has been carried out to check whether the difference between the standard deviations of income for

As we are interested in isolating the effect of the self-employed condition on the extent of underreporting, we need to control for the factors which are involved in determining the food demand function of both groups. Table 2 gives information about some economic and demographic variables with some expected impact on household food expenditures. On this basis we can characterize the average self-employed household in relation to the salary worker family.

Table 2: Differences in economic and demographic variables between SE and SW households

	Variable	SE	SW
Demographic characteristics	# of members	2.43 (1.20)	2.55 (1.23)
	# of dependent children	0.45 (0.84)	0.58 (0.90)
	age of household head	59.43 (16.03)	52.66 (15.78)
	# of labor active members	0.93 (0.98)	1.15 (0.94)
	Dummy for Spanish nac.	0.95 (0.20)	0.93 (0.23)
	Compliance	0.015 (0.16)	0.007 (0.11)
	Dummy for male household head	0.72 (0.44)	0.72 (0.44)
	Dummy for married household head	0.11 (0.32)	0.18 (0.38)
Schooling	# of members recipient of income	1.60 (0.73)	1.56 (0.74)
	Dummy for primary school or less	0.44 (0.49)	0.31 (0.46)
	Dummy for secondary school I (16 y.o.)	0.29 (0.45)	0.27 (0.44)
	Dummy for secondary school II (18 y.o.)	0.11 (0.31)	0.15 (0.36)
Housing	Dummy for University	0.15 (0.36)	0.24 (0.43)
	Dummy for housing-owner with mortgage	0.21 (0.41)	0.31 (0.46)
	Dummy for towns <10,000 inhabitants	0.26 (0.44)	0.20 (0.40)
	# of others housing owned by household	0.18 (0.43)	0.15 (0.40)
Other consumptions	Neighbourhood	3.69 (1.59)	3.56 (1.42)
	Log of food expenditures away home	7.21 (1.43)	7.33 (1.34)
	Log of alcoholic drink expenditures	5.69 (1.47)	5.84 (1.46)
	Log of durables for housing expenditures	6.07 (1.33)	6.11 (1.37)
	Log of durables for leisure expenditures	5.99 (1.54)	6.02 (1.49)
	Log of car expenditures	5.51 (2.71)	5.53 (2.76)

Notes: Variable "compliance" is defined as the difference between the theoretical records and the actual records effectively collected in a household. Variable neighbourhood ranges between 1 (luxury urban) and 7 (agrarian rural). All data are with no imputations in food expenditure answers. Standard deviations between parentheses.

salary and self-employed workers are statistically significant or not. The results have been clear: they are statistically significant. The test has been performed for the household total income and the income of household head as well.

Although the self-employed households consist of less members, dependent children and labour active members than the salary worker households, the former have a slightly higher number of income recipients than the latter. Self-employed households also are headed by an older person than the corresponding salary worker family, whose nationality is mainly Spanish and male sex (with very small differences with respect to the salary worker households). Human capital accumulation is bigger in the case of employee households¹⁰.

Regarding housing characteristics, the average self-employed household lives more in towns below 10,000 inhabitants, has a less recourse to mortgages, and owns slightly more houses (other than the main one) when comparing to the average salary worker household. If other types of expenditures are analysed, the self-employed households spend less money in alcoholic drinks, meals out of home, cars and durables goods for housing or leisure than the salary worker households. Finally, the interpretation of variable "compliance" says that the higher its value, the less the implication of the household in providing the information required in the survey; in this sense, self-employed households are less collaborative than employee households.

4 Estimations and results

The model of section 2 suggests an equation which allows us to obtain an estimate of underreporting of income by self-employed worker households in relation to salary workers households. In essence, expression (10) states that the food consumption of both types of households depends on reported income, on a dummy distinguishing whether the household head of the family is a self-employed worker or not, and a number of variables controlling for different socio-economic and demographic characteristics.

On the basis of equation (10), we have run a number of regressions under several specifications and methods. Particularly, we report 2SLS estimates in which both the reported household income and the dummy variable for SE have been instrumented. It is clear, by assumption, that the first one is measured with error given the existence of transitory variations around its permanent value and the own concealment phenomenon. The dummy for self-employed workers is also treated as endogenous as long as we have evidence of potential misclassification of self-employed households as families with a salary worker head¹¹; otherwise, this would likely lead to a downward bias in the estimated coefficient for self-employed dummy (Schuetze, 2002). Not surprisingly, the Hausman specification test supports the idea that OLS estimates are inconsistent and the IV approach is required. Selection of instruments has been done using the Sargan test; a complete list of those that have been used

¹⁰This is an interesting fact that also occurs in the Swedish case (Holmlund and Engstrom, 2009) but not in the US sample (Hurst et al, 2011).

¹¹Recall the data we provided in the Introduction.

can be seen in the Appendix A. Moreover, the lack of correlations between the disturbances and the unobserved individual effects, as also Hausman-type tests show for all the specifications -not reported here-, leads to a random effects model.

We have used different definitions for the dependent variable: the log of total food expenditures (food purchases plus meals away home) per household or the log of food expenditures (only food purchases) per household. Similarly, two measures of income have been considered: the log of net total income (called in tables *total income*) or the log of net income earned by the household head (called *HH income*).

In all specifications the control variables are the age of household head and its square, the number of members in the households (or the number of dependant children), a dummy for marital status (1 if married, 0 otherwise), a time dummy for 2009, and a constant. Other specifications were estimated but those reported here are the best ones in terms of econometric guarantees and economic sense. Particularly, regional dummies, the log of expenditures in clothes, cars, health, and other household spending items, dummies controlling for the size of the city, and time dummies for others years were included but they were not statistically significant. Additionally, among the rejected variables as potential instruments, we have dummies for housing ownership (if financed with a mortgage, if rented), the number of labour active members, a dummy for sex of household head, the log of durable goods for housing expenditures and the log of durable goods for leisure expenditures.

Table 3 reports IV estimates of expression (10) and shows that the degree of underreporting by the Spanish self-employed workers ranges between 20 and 30 per cent. Table 3 also displays the lower and upper values of degree of underreporting, according to the expressions (14) and (15), respectively, and the variances of Table B1 in the Appendix B¹². Recall that on the basis of standard assumptions, the values for $\sigma_{\xi_{SE}}^2$ and $\sigma_{\xi_{SW}}^2$ can be obtained as the residual variances of (11) when a separate estimation for each type of household is done.

When the dependent variable is the log of food purchases, the concealment of earnings is higher than in the case of taking into consideration the meals away home as well; it makes sense to find less underreporting when explicit expenditures in bars and restaurants is regarded, specially as receipts and invoices may be involved. In the two first columns, we also see higher levels of underreporting when only the income earned by the self-employment head of household is considered. By contrast, the concealment of earnings by the households with self-employed head is lower if we take all the familiar income regardless its sources. In other words, the higher the share of self-employment income over the total family income, the higher the level of underreporting.

These results are line with previous papers, although in the low range. Engstrom and Holmlund (2009), by focussing the difference between underreporting in self-employed households and underre-

¹²Model (1) in Table B1 refers to specification of column (1) in Table (3), and so on.

porting of self-employed income in self-employed households, see how their estimates of such a measure goes from 30 per cent to around 35 per cent in Sweden. Kleven et al (2011), using experimental methods, find that evasion rate for total positive self-employment income is 17.7 per cent in Denmark while the corresponding value for third-party reported income (among other things, salary worker incomes) is below 1 per cent. Hurst et al (2011), with a very close methodology to this one, find that the self-employed workers underreport their income by between 25 and 35 per cent in the US.

Regarding the impact of control variables on the dependent variable, the results exhibit reasonable patterns and again similar to previous studies. Food expenditures are positively affected by the age of household head but negatively by its square; dummies for marital status of household head and for year 2009, when the economic crisis was specially hard, have also a negative impact on food expenditure. By contrast, the number of members or dependant children in the household have a positive effect on food spending.

Table 3: IV estimates of Engel curve for food

	$\ln(\text{food})$	$\ln(\text{food})$	$\ln(\text{total food})$	$\ln(\text{total food})$
$\ln(\text{total income})$	0.278(0.029)		0.423(0.031)	
$\ln(\text{HH income})$		0.325(0.029)		0.497(0.030)
<i>DSE</i>	0.062(0.018)	0.085(0.018)	0.080(0.019)	0.093(0.018)
age	0.051(0.002)	0.039(0.002)	0.034(0.002)	0.024(0.002)
age ²	-0.0004(0.00002)	-0.0002(0.00002)	-0.0003(0.00002)	-0.0002(0.00002)
# children	0.172(0.006)		0.093(0.005)	
# members		0.270(0.004)		0.207(0.004)
marital status	-0.340(0.159)	-0.147(0.014)	-0.251(0.015)	-0.094(0.014)
time dummy 2009	-0.037(0.011))	-0.041(0.010)	-0.065(0.010)	-0.071(0.010))
constant	3.998(.279)	3.221(0.273)	3.686(0.291)	2.762(0.278)
$\hat{\theta}$	1.253	1.300	1.210	1.205
$\hat{\theta}_{\text{Lower}}$	1.230	1.274	1.190	1.183
$\hat{\theta}_{\text{Upper}}$	1.277	1.327	1.230	1.228
\bar{R}^2	0.24	0.30	0.31	0.32
Hausman	16.40(0.01))	13.74(0.03)	18.48(0.005)	17.50(0.007)
Sargan	7.74(0.17)	10.96(0.05)	4.81(0.18)	4.10(0.39)
<i>Obs.</i>	16451	16095	13475	13158

Notes: Standard errors in parenthesis. *Total food* is the sum of food and meals away home expenditures.

HH income is the income of household head and *total income* the income of household.

One of the contributions of this paper lies in the more precise interpretation of the estimates of underreporting by taking into consideration the chance of hiding income by salary workers too. To what extent changes in the set of salary workers used in the estimations affect the degree of underreporting by self-employed workers? How sensitive is this measure with respect to different subsets of salary workers, which in a sense can be seen as control group? Table 4 gives a suggestive answer by reporting meaningful differences in the degree of underreporting of income by self-employed households depending on the group of salary workers with which they are compared to. When families with salary workers as head receive no pensions, the degree of concealment found for the self-employed workers is about 45% more than in the standard approach (1.809 vs 1.253); by contrast, if the estimates are computed regarding households with salary workers and earnings from pensions as well, the level of underreporting by self-employed decreases by 11% in comparison with the standard estimate (1.111 vs 1.253). Similarly, the extent of hidden earnings by self-employed notably goes up when salary households with no income from unemployment benefits are regarded.

These results underline an issue that has been ignored since the very beginning of this literature, namely, that all the measures of degree of underreporting by self-employed workers are in relation to a given level of underreporting by salary workers, which is difficult to be assumed that is null. On the basis that the employees also hide part of their income, our estimates of concealment by self-employed workers are relative to these hidden earnings by salary workers. Furthermore, these estimates are very sensitive to whether the households are entitled to receive social benefits or not.

Table 4: IV estimates of Engel curve for food and different sets of SW

	All SW	SW w/o pensions	SW & pensions	SW w/o unem. ben.
$\ln(\text{total income})$	0.278(0.029)	0.197(0.045)	0.312(0.040)	0.266(0.032)
<i>DSE</i>	0.062(0.018)	0.117(0.027)	0.033(0.023)	0.099(0.021)
age	0.051(0.002)	0.043(0.004)	0.058(0.004)	0.053(0.002)
age ²	-0.0004(0.00002)	-0.0003(0.00004)	-0.0005(0.00004)	-0.0004(0.00002)
# children	0.172(0.006)	0.185(0.007)	0.171(0.015)	0.173(0.007)
marital status	-0.340(0.159)	-0.385(0.020)	-0.292(0.023)	-0.358(0.016)
time dummy 2009	-0.037(0.011)	-0.050(0.014)	-0.027(0.016)	-0.038(0.011)
constant	3.998(.279)	5.007(0.408)	3.629(0.413)	4.082(0.309)
$\hat{\theta}$	1.253	1.809	1.111	1.455
\bar{R}^2	0.24	0.26	0.23	0.24
<i>Obs.</i>	16451	9852	7971	15171

Notes: Dependent variable: Log of food expenditures by household. Standard errors in parenthesis.

Using individuals' own assessment of their employment status could be problematic since there could be workers whose status may not be clear. Although the potential misclassification of self-employed households as families with a salary worker head has been dealt with the above IV approach, we include now an additional check. Column (1) in Table 5 shows estimates of food expenditure equation but requiring that those households stating to be self-employed also declare that the income coming from self-employment is the main one. The degree of underreporting as well as the coefficients of control variables are practically the same than in the canonical specification.

Substantial differences do arise when the group of self-employed workers is filtered to take into account some particular, relevant features. Column (2) in Table 5 reports the level of concealment of income when the self-employed has hired at least one worker as employee. The central value of θ is 1.428, notably higher than for the whole sample of self-employed; remaining regressors hardly differ from the previous ones. This fact is specially intense if the self-employed workers with employees are technicians or professionals such as lawyers, doctors, architects and so on (column 3 in Table 5); in this case their reported earnings should be multiplied by 2.4 to obtain the true income, though this result should be interpreted with caution due to the weak statistical significance of the coefficient for the self-employed dummy. By contrast, households with skilled self-employed head (and no distinction is now made whether they have employees or not) are found to underreport less than in the standard case (1.147 vs 1.253).

Table 5: IV estimates of Engel curve for food with different definitions of SE

	(1)	(2)	(3)	(4)
$\ln(\text{total income})$	0.282(0.030)	0.284(0.030)	0.205(0.092)	0.480(0.078)
DSE	0.064(0.028)	0.101(0.028)	0.182(0.112)	0.065(0.028)
age	0.050(0.002)	0.051(0.002)	0.056(0.007)	0.054(0.004)
age ²	-0.0004(0.00002)	-0.0004(0.00002)	-0.0004(0.00007)	-0.0004(0.00004)
# children	0.171(0.006)	0.171(0.006)	0.159(0.014)	0.178(0.011)
marital status	-0.342(0.016)	-0.342(0.016)	-0.443(0.043)	-0.280(0.031)
time dummy 2009	-0.038(0.011)	-0.035(0.011)	0.006(0.026)	-0.049(0.017)
constant	3.991(0.286)	3.959(0.284)	4.557(0.848)	1.959(0.750)
$\hat{\theta}$	1.254	1.428	2.430	1.147
\bar{R}^2	0.24	0.24	0.28	0.23
<i>Obs.</i>	15621	15669	3039	6447

Notes: Dependent variable: Log of food expenditure. Columns: (1): self-employed declaring that the income from self-employment is the main one; (2): self-employed with employees;(3): professional and technician self-employed with employees; (4): skilled self-employed. Standard errors in parenthesis.

Additional robustness analyses have been carried out in order to verify whether the specification chosen is the most appropriate. Firstly, we have checked the assumption of equal propensity to consume for self-employed and salary workers. Columns (1) and (2) of Table 6 report the estimates by regressing the log of food expenditures on both the income of self-employed and of salary worker households, in terms of household income and household head income as well. It is straightforward to see that the differences between the two relevant coefficients are negligible, keeping the remaining coefficients practically unchanged. This result is in line with previous findings by other authors, for instance Pissarides and Weber (1989).

Regarding the potential existence of non-linear relationships between food consumption and income (Lyssiotou et al, 2004 and Tedds, 2010; both of them using non-parametric techniques), we have run regressions where the log of income is assumed to have both first and second order effects on consumption, in a log quadratic version of the Engel curve (4). In the columns (3) and (4) of Table 6, it is clearly seen that the statistical insignificance of the quadratic coefficients reject the presence of non-linear relationships, in line with other papers (Pissarides and Weber, 1989; Hurst et al, 2011).

Table 6: IV estimates of Engel curve for food with changes in specification

	(1)	(2)	(3)	(4)
<i>DSW</i>	0.277(0.029)	0.324(0.029)		
<i>DSE</i>	0.284(0.030)	0.333(0.030)		
ln(total income)			-2.250(4.719)	
ln(total income) ²			0.129(0.241)	
ln(HH income)				-3.731(5.480)
ln(HH income) ²				0.213(0.288)
<i>DSE</i>			0.052(0.027)	0.063(0.035)
age	0.051(0.002)	0.039(0.002)	0.049(0.005)	0.036(0.005)
age ²	-0.0004(0.00002)	-0.002(0.00002)	-0.0004(0.00004)	-0.0002(0.00004)
# children	0.172(0.006)		0.166(0.012)	
# members		0.270(0.004)		0.269(0.005)
marital status	-0.340(0.015)	-0.147(0.014)	-0.345(0.018)	-0.154(0.017)
time dummy 2009	-0.037(0.011)	-0.041(0.010)	-0.036(0.011)	-0.042(0.011)
constant	4.005(0.279)	3.233(0.272)	16.360(23.066)	22.516(26.065)
\bar{R}^2	0.24	0.30	0.22	0.28
<i>Obs.</i>	16451	16095	16451	16095

Notes: Dependent variable: Log of food expenditures. Standard errors in parenthesis.

An interesting extension to make the estimates more reliable consists of applying quantile regression to see whether the underreporting is sensitive to different quantiles. Indeed, it may be the case that the extent of concealment to be dependant on the conditional distribution of food expenditures. Table 7 shows how the value of θ changes with respect to the 0.25th, 0.50th and 0.75th quantile. While the degree of underreporting substantially increases in the comparison between the 0.25th and the 0.50th quantiles, there is an stabilization of the estimate for the last quantile. In other words, it is found an increasing relationship between the level of underreporting and the food expenditures: the higher the households expenditures in food, the higher their underreporting of income. But this is only true for the two first quantiles. Beyond the median household, this relationship disappears, following the standard pattern of decreasing marginal propensity to consume food¹³.

¹³This result is also found when other specifications (total food, income of household head) are estimated.

Table 7: Quantile regression of Engel curve for food

	q25	q50	q75
$\ln(\text{total income})$	0.256(0.010)	0.237(0.006)	0.202(0.008)
<i>DSE</i>	0.031(0.023)	0.051(0.010)	0.040(0.016)
age	0.047(0.003)	0.043(0.002)	0.038(0.002)
age ²	-0.0003(0.00003)	-0.0003(0.00002)	-0.0002(0.00002)
# members	0.209(0.005)	0.199(0.003)	0.183(0.003)
marital status (3)	0.272(0.021)	0.1999(0.009)	0.152(0.012)
time dummy 2009	-0.036(0.017)	-0.038(0.010)	-0.041(0.010)
constant	3.280(0.150)	4.003(0.109)	4.890(0.100)
$\hat{\theta}$	1.132	1.242	1.223
Pseudo R^2	0.21	0.20	0.18
<i>Obs.</i>	16451	16095	16451

Notes: Dependent variable: Log of food expenditures. Standard errors in parenthesis.

Finally, it is reasonable to think that business cycle may impact on tax evasion and more generally on the extent of informal economy. Although this is not the central point of this paper, Table 8 gives some insights on this potential link for several, different specifications. Contrary to what previous intuition may conjecture, the degree of underreporting in our sample is higher in the years of expansion (2006-2007) than when the economy shrinks (2008-2009).

Table 8: IV estimates of Engel curve for food by subperiods

	$\ln(\text{food})$		$\ln(\text{total food})$	
	2006-2007	2008-2009	2006-2007	2008-2009
$\ln(\text{total income})$			0.447(0.047)	0.399(0.041)
$\ln(\text{HH income})$	0.351(0.047)	0.287(0.038)		
DSE	0.101(0.028)	0.070(0.024)	0.100(0.029)	0.058(0.026)
age	0.036(0.003)	0.041(0.003)	0.032(0.004)	0.037(0.003)
age ²	-0.0002(0.00003)	-0.0003(0.00003)	-0.0003(0.00004)	-0.0003(0.00003)
# children			0.086(0.009)	0.099(0.007)
# members	0.276(0.008)	0.267(0.006)		
marital status	-0.151(0.022)	-0.143(0.018)	-0.263(0.023)	-0.239(0.019)
constant	3.030(0.435)	3.499(0.353)	3.546	3.807(0.391)
$\hat{\theta}$	1.335	1.27	1.252	1.159
\bar{R}^2	0.29	0.31	0.31	0.31
<i>Obs.</i>	7132	8963	5969	7506

Notes: Standard errors in parenthesis.

5 Concluding remarks

At first sight, one could say that the extent of underreporting of income by the Spanish self-employed workers would be above the estimates found for USA, Sweden or UK. This view would be supported firstly by the fact that tax morale in Spain is not so strong as in other OECD countries (Alm and Torgler, 2006). And secondly, as Mediterranean country, the Spanish self-employment rate is higher than in north European countries (Torrini, 2005), and this makes more difficult and costlier the control of such income source by tax authorities.

This paper shows evidence on the extent of underreporting by self-employed in a sample that has never been used with this purpose. Our estimates range this magnitude by around 25 per cent of the reported income recognized by the households headed by self-employed workers. These figures are very close to those corresponding to other countries such as Sweden or USA. Our result has been obtained using data drawn from the Spanish Household Budget Surveys over the period 2006-2009 and after running a number of regressions to control for changes in specification, non-linearities and endogeneity.

Having said that, we have proposed here a broader interpretation of the standard Pissarides and Weber's (1989) model. Instead of assuming that salary workers honestly report all their incomes, we have also admitted the chance of hiding earnings by employees. In this context, our measure of income underreported by self-employed workers must be interpreted as a relative extent of such concealment, taking as reference a given level of underreporting of income by the salary workers.

In other words, our estimate of 25 per cent of underreporting by self-employed households is in relation to the income of self-employed worker that equals the degree of underreporting of income by salary workers, which is strictly positive in our approach. Indeed, we see how the extent of concealment is greatly sensitive to the type of salary worker household taken into consideration in the sample. Recall that the range of underreporting goes from 1.111 when households with employee head also receive pension incomes to 1.809 in the opposite case: no pension income is obtained.

Consequently, our estimates must be seen as lower bounds in the absolute extent of underreporting of income, beyond the standard maximum and minimum thresholds derived from the canonical approach. In this context, the extent of black economy stemmed from the underreporting of self-employed is around 2.5 per cent of GDP¹⁴. Previous estimates of the informal economy in Spain are substantially higher than those reported here. Arrazola et al. (2011) finds that the size of Spanish underground economy in the 2000s is around 20 per cent of GDP, using both the currency demand and electricity models. Scheneider (2012) also reaches practically the same figure over the period 2003-2010. Anyway, the comparison between both of them and this paper only aims for inserting our estimates within the general framework of previous estimates; indeed, the methodologies are quite different and the focus is completely distinct (the whole economy versus the underreporting by self-employed with respect to employees).

A line for further research could be motivated by the consequences of this concealment of income on tax revenues and progressivity. While the effect of progressivity on tax evasion has been examined by some authors, the inverse effect (the impact of the concealment of income by self-employed workers on progressivity) has hardly studied. Although there are some theoretical papers dealing with this issue (see, for instance, the recent paper by Freire-Seren and Panades, 2008), the scope for empirical papers is wide. Precisely on the basis of this new research avenue, it is clear that basic principles of vertical and horizontal equity are damaged in the presence of underreporting.

Additionally, as the salary workers have to pay more taxes compared to self-employed workers, other things equal, an inefficient incentive to allocate more resources (than socially optimal) in the self-

¹⁴As National Accounts in Spain do not provide information on the share of income over the GDP generated by the self-employed workers, we have computed an approximation to this variable. We have used the share they represent over the total employment, and assumed that their productivity is about 30-40 per cent (depending on the years) higher than that of salary workers; this assumption lies in the fact that the self-employed work more than employees (about 30-40 per cent).

employment activities arises. As result of this, individuals see how their employment choice between paid employment and self-employment is distorted in favour of the latter.

ACKNOWLEDGEMENTS

The author would like to thank the hospitality of Department of Economics at Uppsala University and the comments by Bertil Holmlund, Per Engstrom, Ben Pugsley, a referee of IVIE WP Series, two anonymous referees and the editor. This paper was awarded with the Alexandre Pedros Best Paper Prize at the XIX Encuentro de Economía Pública. All errors are my sole responsibility. I also acknowledge financial support from Junta de Andalucía (Proyectos de Excelencia SEJ-02479 and SEJ-6882) and the Spanish Ministry of Science and Technology (ECO2010-15553 and ECO2010-21706).

References

- [1] Ahn, N. and De la Rica, S. (1997). The underground economy in Spain: an alternative to unemployment. *Applied Economics*, 29 (6), 733-743.
- [2] Alm, J. and Torgler, B. (2006). Culture differences and tax morale in the United States and in Europe. *Journal of Economic Psychology*, 27 (2), 224-246.
- [3] Arrazola, M., de Hevia, J., Mauleón, I. and Sánchez, R. (2011). La economía sumergida en España. Discussion Paper, University Rey Juan Carlos de Madrid.
- [4] Bentolila, S., Dolado, J. J., Franz, W. and Pissarides, C. (1994). Labour flexibility and wages: lessons from Spain. *Economy Policy*, 4 (1), 55-99.
- [5] Engstrom, P. and Holmlund, B. (2009). Tax evasion and self-employment in a high-tax country: evidence from Sweden. *Applied Economics*, 41(19), 2419-2430.
- [6] European Commission (2007). Undeclared Work in the European Union, Special Eurobarometer 284.
- [7] Freire-Seren, M. J. and Panades, J. (2008). Does tax evasion modify the redistributive effect of tax progressivity?. *The Economic Record*, 84 (267), 486-495.
- [8] Hurst, E., Li, G. and Pugsley, B. (2011). Are household surveys like tax forms: Evidence from income underreporting of the self-employed. Finance and Economics Discussion Series #2011-06, Federal Reserve Board Also published as NBER Working Papers with number 16527.
- [9] Johansson, E. (2005). An estimate of self employment income underreporting in Finland. *Nordic Journal of Political Economy*, 31(1), 99-109.

- [10] Kleven, H. J., Knudsen, M. B., Kreiner, C. T., Pedersen, S. and Saez, E. (2011). Unwilling or unable to cheat? Evidence from a randomized tax audit experiment in Denmark. *Econometrica*, 79(3), 651–692.
- [11] Kolm, A.-S. and Nielsen, Soren B. (2008). Under-reporting of Income and Labor Market Performance, *Journal of Public Economic Theory*, 10(2), 195-217.
- [12] Lyssiotou, P., Pashardes, P. and Stengos, T. (2004). Estimates of the black economy based on consumer demand approaches. *Economic Journal*, 114 (497), 622-640.
- [13] Pissarides, C. and Weber, G. (1989). An expenditure-based estimate of Britain’s black economy. *Journal of Public Economics*, 39(1), 17-32.
- [14] Schuetze, H. J. (2002). Profiles of tax non-compliance among the self-employed in Canada: 1969 to 1992. *Canadian Public Policy*, 28(2), 219-237.
- [15] Schneider, F. (2012). Size and development of the shadow economy of 31 European and 5 other OECD countries from 2003 to 2012. Working Paper, University of Linz.
- [16] Tedds, L. (2010). Estimating the income reporting function for the self-employed. *Empirical Economics*, 38(3), 669-687.
- [17] Torrini, R. (2005). Cross-country differences in self-employment rates: the role of institutions. *Labour Economics*, 12(5), 661-683..

Appendix A: Instruments used in Table 3

- Column (1) of table 3: dummy variable for nationality of household head, two dummy variables indicating the education of household head (primary and post-compulsory secondary school), and the product of self-employed dummy with age, age squared, and the two previous schooling dummy variables.
- Column (2) of table 3: The same than in column (1).
- Column (3) of table 3: two dummy variables indicating the education of household head (primary and post-compulsory secondary school), and the product of self-employed dummy with the two previous schooling dummy and compulsory secondary school.
- Column (4) of table 3: two dummy variables indicating the education of household head (primary and post-compulsory secondary school), and the product of self-employed dummy with age, age squared, and the two previous schooling dummy variables.

Appendix B: Residual variances in income equations

Table B1: Residual variances in income equations

	$\sigma_{\xi_{SE}}^2$	$\sigma_{\xi_{SW}}^2$
Model [1]	0.326	0.289
Model [2]	0.250	0.209
Model [3]	0.250	0.212
Model [4]	0.326	0.293