serie ec
WP-EC 2011-09

How different are the Spanish self-employed workers by underreporting their incomes?

Diego Martinez

Ivie

Edita / Published by: Instituto Valenciano de Investigaciones Económicas, S.A.

Depósito Legal / Legal Deposit no.: V-3641-2011

Impreso en España (octubre 2011) / Printed in Spain (October 2011)

WP-EC 2011-09

How different are the Spanish self-employed workers by underreporting their incomes?*

Diego Martínez**

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 (lower) concealment of income by salary workers. Our results show that the reported income by self-employed has to be increased by about 25-30 percent to obtain the level of income which would equal the level of underreporting by employees. Our estimates are robust to changes in specification, endogeneity and non-linearities.

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

JEL Classification: D12, H26, O17.

Resumen

Este artículo ofrece estimaciones de la ocultación de rentas por parte de los trabajadores autónomos españoles usando Encuestas de Presupuestos Familiares en el periodo 2006-2009. Para ello empleamos el conocido modelo de Pissarides y Weber (1989), cuya interpretación extendemos para admitir también la (menor) ocultación de renta por parte de los trabajadores asalariados. Nuestros resultados muestran que la renta reconocida por los trabajadores autónomos españoles debe incrementarse entre un 25 y 30 por ciento para obtener el nivel de renta que igualaría el grado de ocultación de los empleados. Dichas estimaciones son robustas a cambios en la especificación, endogeneidad y no linealidades.

Palabras clave: ocultación, encuestas familiares, consumo de comida, evasión fiscal.

Clasificación JEL: D12, H26, O17.

* The author would like to thank the hospitality of the Department of Economics at Uppsala University and the comments by Bertil Holmlund, Per Engstrom and a referee of the Ivie. All errors are my sole responsibility. I also acknowledge financial support from the Junta de Andalucia (Proyectos de Excelencia SEJ-02479 and SEJ-6882) and the Spanish Ministry of Science and Technology (ECO2010-15553 and ECO2010-21706).

** Department of Economics, University Pablo Olavide. 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 income 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.

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 provided 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 wages[1]. Moreover, as we will show later, ignoring underreporting by employees in the model when is present in the data results in empirical estimates with no correspondence with the theoretical framework used.

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 and testing non-linearities in the relationship between income and food expenditure, and potential problems of endogeneity.

In this context, we can summarize the main contributions of the paper as follows. First, 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. Second, 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.

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 procedure and shows the results. Finally, section 5 concludes.

2 The model

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

[1] National 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.

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:

$$Y_i = p_i Y_i^p, \tag{1}$$

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:

$$Y_i = k_i Y_i^{'}, \text{ with } k_i > 1. \tag{2}$$

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:

$$\ln Y_i^p = \ln Y_i^{'} - \ln p_i + \ln k_i, \tag{3}$$

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

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

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:

$$\ln p_i = \mu_i^p + u_i \tag{5}$$

$$\ln k_i = \mu_i^k + v_i, \tag{6}$$

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 $\sigma^2_{u_i}$ and $\sigma^2_{v_i}$.

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

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

The estimation of this equation requires further algebra manipulation using the properties of log-normal distributions. Particularly,

$$\ln \bar{p}_i = \mu_i^p + \frac{1}{2}\sigma^2_{u_i} \tag{8}$$

$$\ln \bar{k}_i = \mu_i^k + \frac{1}{2}\sigma^2_{v_i}, \tag{9}$$

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 ($\ln \bar{p}_{SE} = \ln \bar{p}_{SW}$), the third term in RHS of (7) can be written as

$$\beta(\mu_i^k - \mu_i^p) = \beta\left[\theta - \frac{1}{2}\left(\sigma^2_{v_{SE}} - \sigma^2_{v_{SW}}\right) + \frac{1}{2}\left(\sigma^2_{u_{SE}} - \sigma^2_{u_{SW}}\right)\right], \tag{10}$$

where $\theta = \ln \bar{k}_{SE} - \ln \bar{k}_{SW} = \ln \frac{\bar{k}_{SE}}{\bar{k}_{SW}}$, that is, the degree of underreporting of income by self-employed relative to the extent of underreporting by salary worker households. Then, the food expenditure function (7) becomes:

$$\ln F_i = \boldsymbol{\alpha}\mathbf{X}_i^{'} + \beta \ln Y_i^{'} + \gamma DSE_i + \eta_i, \tag{11}$$

where $\gamma = \beta\left[\theta - \frac{1}{2}\left(\sigma^2_{v_{SE}} - \sigma^2_{v_{SW}}\right) + \frac{1}{2}\left(\sigma^2_{u_{SE}} - \sigma^2_{u_{SW}}\right)\right]$, DSE_i is dummy variable that takes the value 1 if the main holder of household 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.

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:

$$\ln Y_i^{'} = \boldsymbol{\Omega}\mathbf{X}_i^{'} + \boldsymbol{\Gamma}\mathbf{Z}_i^{'} + \xi_i, \tag{12}$$

where $\mathbf{Z}_i$ is a vector of variables used as instruments in IV-2SLS estimates of expression (11), given the potential endogeneity of $Y_i^{'}$. Again, the error term ξ_i has three components: unexplained variations in household food expenditures, 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 omiting 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

$$\sigma^2_{\xi_{SE}} - \sigma^2_{\xi_{SW}} = \sigma^2_{u_{SE}} + \sigma^2_{v_{SE}} - 2cov(uv)_{SE} - \sigma^2_{u_{SW}} - \sigma^2_{v_{SW}} + 2cov(uv)_{SW}. \tag{13}$$

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

$$\theta = \frac{\gamma}{\beta} + \frac{1}{2}(\sigma^2_{v_{SE}} - \sigma^2_{u_{SE}} + \sigma^2_{u_{SW}} - \sigma^2_{v_{SW}}). \tag{14}$$

Note that (14) 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 $\sigma^2_{v_{SW}}$ 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 obtained[2]. Taken the variances for salary workers as parameters, we see that the minimum value for θ is obtained when σ^2_{vSE} reaches its lowest value, that is, when it is equal to $\sigma^2_{v_{SW}}$. Under such a case,

$$\theta = \frac{\gamma}{\beta} - \frac{1}{2}(\sigma^2_{\xi_{SE}} - \sigma^2_{\xi_{SW}}), \tag{15}$$

where (13) has been used. By contrast, it is easy to see that (14) reaches its maximum value when $\sigma^2_{u_{SE}}$ is at its minimum feasible value, which in our model is like in Pissarides and Weber (1989): $\sigma^2_{u_{SE}} = \sigma^2_{u_{SW}}$.[3] This gives an upper bound for the extent of underreporting of income by self-employed households:

$$\theta = \frac{\gamma}{\beta} + \frac{1}{2}(\sigma^2_{\xi_{SE}} - \sigma^2_{\xi_{SW}}) \tag{16}$$

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 $\theta^{'}$ (to distinguish from the previous one) which is derived from their theoretical framework. In the case of the lower bound, they initially set up $\sigma^2_{v_{SE}}$ equal to 0; this leads to the following expression of $\theta^{'}$:

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

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

$$\theta' = \frac{\gamma}{\beta} - \frac{1}{2}(\sigma^2_{u_{SE}} - \sigma^2_{u_{SW}}). \tag{17}$$

After that, on the basis of this setting, Pissarides and Weber (1989) substitute the terms between parenthesis of (17) by $(\sigma^2_{\xi_{SE}} - \sigma^2_{\xi_{SW}})$ and that is what they really estimate. But comparing the parenthesis of (17) with (13), we see that both of them are not equal with employees concealing income ($\sigma^2_{v_{SW}} \neq 0$). Consequently, the expresion 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, $\theta = \theta' + \frac{1}{2}\sigma^2_{v_{SW}}$.

A similar argument can be managed for the case of the upper bound. The model by Pissarides and Weber (1989) sets $\sigma^2_{u_{SE}} = \sigma^2_{u_{SW}}$ (as in this paper), that applied on (14) gives

$$\theta' = \frac{\gamma}{\beta} + \frac{1}{2}\sigma^2_{v_{SE}}, \tag{18}$$

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 $(\sigma^2_{\xi_{SE}} - \sigma^2_{\xi_{SW}})$ in their estimation, that is, $\sigma^2_{v_{SE}} - \sigma^2_{v_{SW}}$ instead of $\sigma^2_{v_{SE}}$ in (18). Therefore, their definition of θ' must be corrected in order to have what is obtained from their empirical estimation: $\theta = \theta' - \frac{1}{2}\sigma^2_{v_{SW}}$.

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 an 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 lower) degree of underreporting of income by salary workers. Our approach also allows to deal with an inconsistency which can be found in the paper by Pissarides and Weber (1989) in the presence of such underreporting by employees; this discrepancy arises because what they empirically estimate is not what their model set up to be estimated.

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 year[4]. 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

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

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 items[5].

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 null[6]. We also show results below 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 income of household as a whole and the second is the net income of the main holder. 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 main holder, where all the data available here come from the answers of participants.

Salary worker household is defined as that in which the main holder 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 househods) but econometric estimates behave worse and a number of inconsistencies with other items of the survey were present[7]. As is usual in this type of papers, households in which the head holder works in agriculture, cattle farming or fishing have been removed from the sample; we aim at avoiding that the relationship between food consumption and income to be affected by the particular consumption pattern of these households.

Table 1 shows the main intuition behind this paper. Households whose main holder is a self-employed worker declare to spend in food the same or more than the households headed by a salary worker. But households with a self-employed main holder systematically report in the EPF less income than the corresponding salary worker households. In line with previous research, this is a

[5] The 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.

[6] For the sake of simplicity, we only report below econometric estimates with no imputations in food expenditures.

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

Table 1: Descriptive statistics

	Imputations in food		No imputations in food	
	SE	SW	SE	SW
Ln (food w/o meals out)	8.21 (0.69)	8.19 (0.73)	8.17 (0.70)	8.15 (0.74)
Ln (food)	8.76 (0.65)	8.76 (0.63)	8.73 (0.68)	8.73 (0.65)
Ln (net household income)	9.67 (0.67)	9.83 (0.61)	9.64 (0.68)	9.80 (0.61)
Ln (net main holder income)	9.33 (0.59)	9.53 (0.51)	9.30 (0.58)	9.52 (0.51)

Note: Standard deviations between parentheses.

clear indication that self-employed households underreport part of their income. 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 $\sigma^2_{\xi_{SE}} - \sigma^2_{\xi_{SW}}$; this is compatible with a more volatile pattern for self-employed income, as we set up in the theoretical framework.

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 main holder	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)
	Collaboration	0.015 (0.16)	0.007 (0.11)
	Dummy for male main holder	0.72 (0.44)	0.72 (0.44)
	Dummy for married main holder	0.11 (0.32)	0.18 (0.38)
	# of members recipient of income	1.60 (0.73)	1.56 (0.74)
Schooling	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)
	Dummy for University	0.15 (0.36)	0.24 (0.43)
Housing	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)
	Neighbourhood	3.69 (1.59)	3.56 (1.42)
Other consumptions	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 "collaboration" 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.

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 main holder 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 households[8].

[8]This is an interesting fact that also occurs in the Swedish case (Holmlund and Engstrom, 2009) but not in the US

Regarding housing charateristics, 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 "collaboration" 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 (11) states that the food consumption of both types of households depends on reported income, on a dummy distinguishing whether the main holder 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 (11), we have run a number of regressions under several specifications and methods[9]. In all of them 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 main holder of household (called MH income). Obviously, a dummy variable for the condition of being self-employed worker has been added among the regressors.

The set of control variables includes in all specifications dummies for schooling of main holder, dummies for housing ownership (if mortgages, if rented), number of labour active members, a dummy for nacionality of main holder, a dummy for sex of main holder, a dummy for marital status, log of alcoholic drinks expenditures, log of durable goods for housing expenditures, log of durable goods for leisure expenditures, age of main holder, age squared of main holder, number of members in the households, 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, 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.

sample (Hurst et al, 2011).

[9] This first battery of results only provides point estimates of θ. Other estimates below do take into consideration the lower and upper bounds of this value.

Table 3 shows OLS estimates of expression (11). While the coefficient of income appears to be quite low when only food purchases are considered as dependent variable, results are different (and close to the mainstream of literature) if total food is used. One of the relevant findings in this table is the extent in which the degree of underreporting of income $\hat{\theta}$ increases (from around 1.225 to over 1.370) when the income of main holder is taken into consideration. This is line with previous papers. Engstrom and Holmlund (2009), by focussing the difference between underreporting in self-employed households and underreporting 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.

In this sense, it is reasonable to think that higher levels of underreporting will be found when only self-employment income is considered. By contrast, as long as many households have different income sources, although the main holder to be self-employed, the concealment of earnings will be lower -ceteris paribus-, at least in an inverse proportion to the share of salary incomes over the total family income.

Regarding the impact of control variables on the dependent variable, the results exhibit reasonable patterns and similar to previous studies[10]. For instance, food expenditures are negatively affected by the age squared of main holder and dummies for rented accommodation, by the single marital status of main holder and by year 2009, when the economic crisis was specially hard. By contrast, the effect of age, Spanish nacionality, number of members and labour active persons in the household, level of schoolling for the main holder and dummy for housing owner without mortgage is positive. Moreover, household expenditures in durable goods for leisure and housing and alcoholic drinks have a complementary relationship with food expenditures.

[10] Log files from Stata with detailed estimates are available upon request.

Table 3: OLS estimates of underreporting

	$\ln(food)$	$\ln(food)$	$\ln(total\ food)$	$\ln(total\ food)$
$\ln(total\ income)$	0.188(0.016)		0.330(0.014)	
$\ln(MH\ income)$		0.120(0.016)		0.263(0.015)
DSE	0.038(0.023)	0.046(0.024)	0.060(0.022)	0.082(0.023)
$\hat{\theta}$	1.225	1.466	1.202	1.369
$\bar{R}^2$	0.31	0.30	0.38	0.36
$Obs.$	6965	6812	6507	6360

Notes: Robust standard errors in parenthesis. Total food is the sum of food and meals away home. MH income is the income of main holder and total income the income of household

The likely non-linear relationship between food consumption and income has been a concern in previous papers on underreporting. Pissarides and Weber (1989) and Hurst et al (2011) do not find strong evidence in favour of this hypothesis. However, using nonparametric techniques, Lyssiotou et al (2004) and Tedds (2010) show that the analysis must not be contrained to linear functional forms. Our table 4 reports OLS estimates 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). At least in the case of the two first columns, some doubts arise as the coefficient of square log of income appears to be statistically significant.

Table 4: OLS estimates of underreporting with non-linearities

	$\ln(food)$	$\ln(food)$	$\ln(total\ food)$	$\ln(total\ food)$
$\ln(total\ income)$	1.422(0.367)		0.750(0.307)	
$\ln(MH\ income)$		0.799(0.366)		−0.015(0.369)
$\ln(total\ income)^2$	−0.062(0.018)		−0.021(0.015)	
$\ln(MH\ income)^2$		−0.035(0.019)		0.014(0.019)
DSE	0.045(0.023)	0.050(0.024)	0.063(0.022)	0.081(0.023)
$\bar{R}^2$	0.32	0.31	0.38	0.36
$Obs.$	6965	6812	6507	6360

Notes: Robust standard errors in parenthesis. Total food is the sum of food and meals away home. MH income is the income of main holder and total income the income of household

An additional check has been implemented to make sure that the relationship between food expenditure and income does not follow a non-linear pattern[11]. This test is based on the augmented partial residuals (APR) test by Mallows (1986). It consists of drawing a graph of log of income against the residuals from an estimation including the log of income and its square, with other regressors as well. Figures (1)-(4) in the appendix A show these graphs, where a fitted line through the residuals (in yellow) and a Lowess nonparametric regression (in red) have been included for reference. As can be seen, there are only small deviations from the linear pattern and particularly due to outliers concentrated in the tails of income distribution[12].

One of the most relevant technical issue by estimating expression (11) is the potential endogeneity of some regressors, particularly income. Table 5 gives IV-2SLS estimates of the extent of underreporting treating income as endogenous. The choice of reported specifications is based on i) the requirement of having good statistics for overidentification restrictions (Sargan test for validity of selected instruments) and ii) the confirmation that the variables treated as endogenous are really endogenous (Wooldridge, 1995)[13]; thus, the H_0 of this test is "variables are exogenous". Although there is not a great rejection of H_0 (except in the third column), it is also true that the acceptance of such hypothesis is not strong[14]. The sets of instruments are listed in the Appendix B.

The results are in line with the previous OLS estimates but in the lower range, that is, the self-employed households underreport about 25-30 percent of their income. These figures are very close (and even below) to those obtained with samples for other countries where tax morale is stronger than in Spain. A couple of reasons can be pointed out to explain this. The first is that we are measuring here the level of underreporting of income by self-employed with respect to salary workers; thus, we are not offering absolute values of hidden incomes but relative estimates to a given extent of concealment by employees. The second reason is a more technical point and non-exclusive of our study. As Schuetze (2002) says, when the dummy for self-employed workers is not treated as endogenous, a number of high income self-employed households, who conceal a great part of their income, can

[11] Alternatively, we tried to apply the approach by Hurst et al (2011) of constructing a measure of underreporting consistent with a log quadratic Engel curve (see their expression R3 in their appendix). But the results were not satisfactory: when reasonable estimates of underreporting were obtained, econometrics did not perform well, and vice versa.

[12] A preliminary check of non-linearities using the APR test consists of comparing the plots of augmented partial residuals (with the square of log of income in the regression) with the plots of partial residuals (without the square of log of income). These pairs of plots are almost identical, what implies an additional indication that non-linearities are not strong enough to be taken into consideration.

[13] This is the reason why the dummy for self-employed households, treated by other papers as endogenous (for instance, Pissarides and Weber, 1989), is not instrumented here. Schuetze (2002) claims a similar argument to ours.

[14] When non-robust standard errors are used, Hausman-Wu tests for endogeneity confirms these findings.

be misclassified as salary workers, and this would likely lead to a downward bias in the estimated coefficient for self-employed dummy.

In the case of IV-2SLS estimates, a non-linear relationship between food consumption and income does not seem to be a problem as the statistical significance of the coefficient of square of log of income is far from usual levels (see columns 2 and 4); anyway, additional checks -similar to those performed above for OLS estimations- have been run, confirming our first impression[15]. Also in this case, the coefficients and statistical significance of control variables are reasonable.

Table 5: IV estimates of underreporting with non-linearities

	$\ln(food)$	$\ln(food)$	$\ln(total\ food)$	$\ln(total\ food)$
$\ln(MH\ income)$	0.251(0.063)	−0.716(2.863)	0.337(0.030)	3.683(3.279)
$\ln(MH\ income)^2$		0.050(0.150)		−0.172(0.168)
DSE	0.056(0.025)	0.050(0.032)	0.088(0.023)	0.116(0.036)
$\hat{\theta}$	1.251		1.301	
$\bar{R}^2$	0.30	0.29	0.36	0.33
Sargan test	3.32(0.34)	3.20(0.20)	3.38(0.18)	2.24(0.13)
Wooldridge test	4.07(0.04)	4.80(0.09)	4.78(0.02)	5.25(0.07)
Obs.	6812	6812	6360	6360

Notes: Robust standard errors in parenthesis, except with statistics where p-values. Total food is the sum of food and meals away home. MH income is the income of main holder. Set of instruments: See appendix B.

As usual in literature and in line with the theoretical framework of Section 2, we have computed the lower and upper bounds of the extent of underreporting of income using the expressions (15) and (16), respectively. Recall that on the basis of standard assumptions, the values for $\sigma^2_{\xi_{SE}}$ and $\sigma^2_{\xi_{SW}}$ can be obtained as the residual variances of (12) when a separate estimation for each type of household is done. Table 6 shows the results for the two linear IV-2SLS estimates of table 5. Appendix C reports the residual variances estimated for each type of household in the regression for income equation. It is clear that the limits of interval are not far away from the central values obtained previously. We add then more evidence supporting our finding that the extent of underreporting of income by Spanish self-employed households in relation to the underreporting of salary workers is about 25-30 per cent of their income.

[15] Available upon request.

Table 6: IV-2SLS estimates of underreporting with bounds

	$\ln(food)$	$\ln(total\ food)$
Central value	1.251	1.301
Lower bound	1.213	1.260
Upper bound	1.287	1.337

Notes: Residual variances used in Appendix C.

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, however, shows evidence contrary to this hypothesis. Our estimates range the magnitude of underreporting by between 25 and 30 per cent of the reported income recognized by the households headed by self-employed workers. As was said in the Introduction, 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.

A partial explanation of this unexpected result can be placed in 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 range of 25-30 per cent of underreporting by self-employed main holders is not relative to the true income of such as self-employed workers, but 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. 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 sense, previous references suffer a problem of consistency between the degree of underreporting which is derived from the theoretical framework and the empirical estimates which are actually shown. Despite the fact that the latter are the same than those obtained using our approach, they cannot be interpreted in line with the underlying model in the presence of underreporting by salary workers. By contrast, the approach proposed here not only allows to deal with the fact of salary workers hiding part of their earnings but also to adjust the theoretical framework to the empirical estimates.

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 in our case) 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-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.

References

[1] 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.

[2] 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.

[3] European Commission (2007). Undeclared Work in the European Union, Special Eurobarometer 284.

[4] 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.

[5] 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.

[6] Johansson, E. (2005). An estimate of self employment income underreporting in Finland. Nordic Journal of Political Economy, 31(1), 99-109.

[7] 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.

[8] 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.

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

[10] Mallows, C. L. (1986). Augmented partial residuals. Technometrics, 28 (4), 313-319.

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

[12] 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.

[13] Tedds, L. (2010). Estimating the income reporting function for the self-employed. Empirical Economics, 38(3), 669-687.

[14] Torrini, R. (2005). Cross-country differences in self-employment rates: the role of institutions. Labour Economics, 12(5), 661-683.

[15] Wooldridge, J. (1995). Selection Corrections for Panel Data Models under Conditional Mean Independence Assumptions. Journal of Econometrics, 68, pp. 115-132.

A Appendix: Figures for augmented partial residuals tests

Figure 1: Column (1) of Table 3

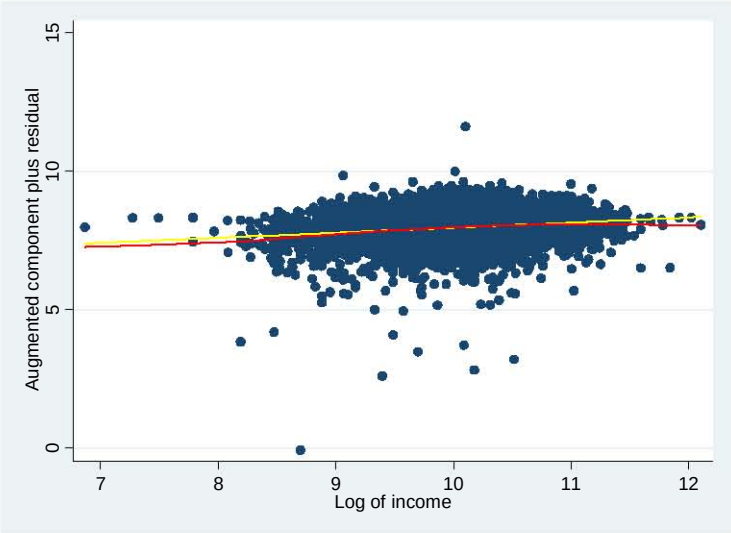

Figure 2: Column (2) of Table 3

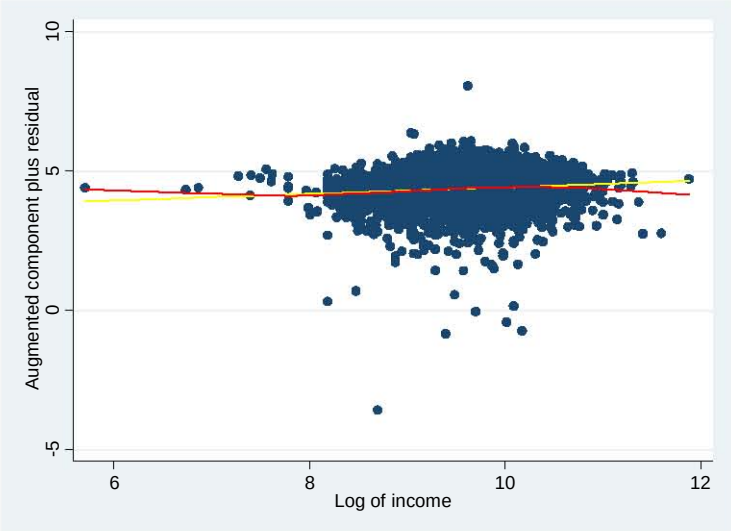

Figure 3: Column (3) of Table 3

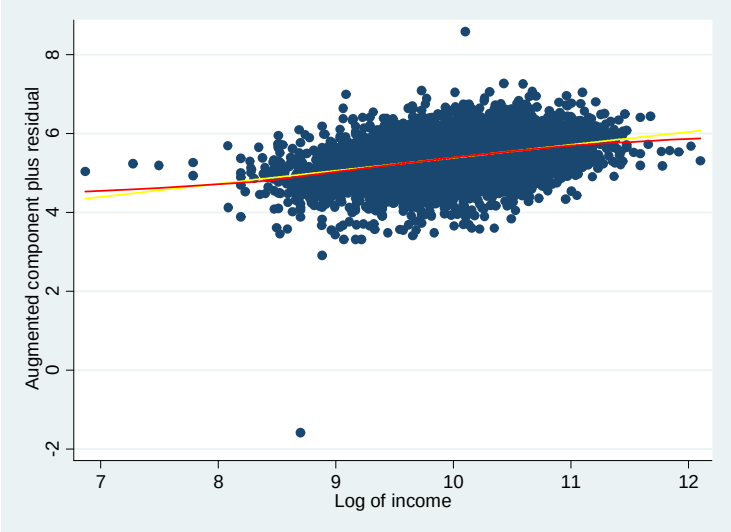

Figure 4: Column (4) of Table 3

B Appendix: Instruments in IV specifications

- Column (1) of table 4: Dummy variable for housing ownership (1 if financed by mortgage, 0 otherwise), dummy variable for sex of main holder of household (1 if male, 0 otherwise), dummy variable for nacionality of main holder (1 if Spanish, 0 otherwise), and number of other housing available for household in addition to the main house.
- Column (2) of table 4: The same than in column (1).
- Column (3) of table 4: Dummy variables for schooling of main holder of household.
- Column (4) of table 4: The same than in column (3).

C Appendix: Residual variances in income equations

Table 7: Residual variances in income equations

	$\sigma^2_{\xi_{SE}}$	$\sigma^2_{\xi_{SW}}$
IV estimates of column (1)	0.692	0.633
IV estimates of column (3)	0.689	0.629

Ivie

Guardia Civil, 22 - Esc. 2, 1º
46020 Valencia - Spain
Phone: +34 963 190 050
Fax: +34 963 190 055

Website: http://www.ivie.es
E-mail: publicaciones@ivie.es