Journal of Public Economics 39 (1989) 17–32. North-Holland

AN EXPENDITURE-BASED ESTIMATE OF BRITAIN'S BLACK ECONOMY

Christopher A. PISSARIDES
London School of Economics, London WC2A 2AE, UK

Guglielmo WEBER*
University College, London WC1E 6BT, UK

Received March 1988, revised version received January 1989

We estimate the size of Britain's black economy (defined narrowly as unreported taxable income) by using income and expenditure data drawn from the 1982 Family Expenditure Survey. Our working assumptions are that all income groups report expenditure on food correctly; employees in employment report income correctly; and that the self-employed under-report their income. We estimate food exependiture equations for all groups and then invert them to arrive at the conclusion that on average true self-employment income is 1.55 times as much as reported self-employment income. This implies that the size of the black economy is about 5.5 percent of GDP.

1. Introduction

The black economy consists, by definition, of activities concealed from the law. Its measurement, therefore, suffers from the problem that if the source of income and expenditure data is the Inland Revenue, no black economy activities will be in it; and if it is not, some black economy activities may be in it and some not. Despite assurances about confidentiality, people may have no incentive to reveal the true extent of their activities to the data collector from fear that they may not be, after all, protected from the law. Thus, there is no single method which can give an accurate measure of the black economy. The many different approaches suggested and explored in the literature use mainly expenditure data, which may be reported more accurately than income data. Good surveys of this literature exist, and suggest that in the United Kingdom the size of the black economy is somewhere in the region of 5 percent of reported activity.[1]

*This study began as part of the Institute of Fiscal Studies project on the shadow economy, which produced the book by Stephen Smith (1986). We are grateful to the Institute for supporting our research and to Stephen Smith for his advice and comments at all stages of the research. We have also had helpful discussions with Jan Magnus, Costas Meghir and Manuel Arellano.

[1] See Smith (1986), from which we draw heavily. The theoretical literature on tax evasion was surveyed by Cowell (1985), who also lists some of the measurement literature.

Our definition of the black economy is the narrow one of activity that should normally be reported and taxed but is not. Smith's (1986) thorough review of research in this area concluded that black economy activity is concentrated amongst the self-employed, who run mainly small family businesses. There is a concentration of this kind of enterprise in agriculture, forestry and fishing (which is a very small sector), construction and distribution, and repairs: the plumber who offers a discount for payment in cash is the typical black-economy practitioner. Altogether, the self-employment sector employs about 10 percent of the civilian labour force and adds as much to GDP. The national accounts are adjusted upwards by something like 15 percent of reported self-employment income to account for tax evasion, so the size of the black economy implicit in this is not more than 1.5 percent of GDP. This figure, however, is probably a lower bound, as it is based on the assumption of accurate reporting of all expenditures on GDP.

The assumptions underlying the measurement of the black economy in this paper draw on the findings of Smith's study and are consistent with the assumptions underlying the adjustment of the national accounts.[2] They are two:

(1) the reporting of expenditure on some items by all groups in the population is accurate;

(2) the reporting of income by some groups in the population is accurate.

The data that we use for our analysis come from the Family Expenditure Survey (FES). The item of expenditure in the FES where we believe reporting is most accurate is food. Foor expenditure is recorded on a daily basis from a diary kept by a member of the household for a week. It is highly unlikely that the person filling in the diary (usually not the income earner in one-income families) will conceal some seemingly small food expenditure for tax reasons, or that he or she will call 'business expenses' part of the household food bill. In contrast, expenditure on other items may be concealed if it is conspicuous and believed to arouse curiosity, or some expenditure, e.g. on clothing or car maintenance, may come under 'business expenses'.[3]

With regard to our second assumption, we assume that only the self-employed under-report their income. We experimented with other pro-

[2] On the latter point, see also MacAfee (1980), and Dilnot and Morris (1981).

[3] We ran a few regressions with other items of expenditure to check the consistency of our estimates with food. The items we tried were clothing, services (repairs, car maintenance, personal services, etc.) and, 'pre-commitments' (mortgage, rent, television licence, etc.). The results were mixed with no clear-cut evidence of under- or over-reporting. Closer inspection of the data convinced us that the only expenditure item that could be relied upon to give accurate estimates of income under-reporting was food. Apart from the reasons given in the text, the other items of expenditure are less regular so there are several zero entries and several positive outliers in the reference week. Dealing with these types of data requires different techniques from the ones we employ here, so the estimates obtained from other items by application of our techniques are not strictly comparable with our food-based estimates.

fessional groups but we could not find evidence of under-reporting. All professional groups except for the self-employed exhibit a similar pattern of expenditure on food. The self-employed stand apart. We assume that one of the factors behind this discrepancy is under-reporting of income by the self-employed. Employees in employment are assumed to report their income correctly.

The method we use to calculate the extent of under-reporting consists of two parts. First, we estimate expenditure functions in terms of household characteristics and reported income. In a second stage we invert the expenditure function and forecast income from reported expenditure. Our two assumptions enable us to estimate a reliable food expenditure function for employees in employment and then to invert it to calculate the 'true' income of the self-employed.

Section 2 discusses the estimated expenditure function when the self-employed under-report their income. The actual estimates are not of immediate interest and they are discussed in the appendix. Section 3 reports the small number of estimates that is used in the calculation of income under-reporting and calculates a factor by which reported self-employment incomes have to be multiplied to arrive at true incomes. This factor is approximately 1.55. Given that reported self-employment income is about 10 percent of GDP, our estimates imply that the size of the black economy is about 5.5 percent of GDP. Thus, our estimate is remarkably similar to the average figure that can be obtained from the several estimates in the literature [see Smith (1986)] but it is higher than the implicit estimate of the size of the black economy in the National Accounts.

2. Income and expenditure

Households report to the FES consumption on individual items C_{ij} (for household i on item j) and after-tax income Y'_i. The FES records also a vector of household characteristics, Z_i. We assume that C_{ij} for food is correctly reported by all households interviewed, Y'_i is correctly reported by employees in employment and Z_i is correctly recorded for all households. Let Y_i be the 'true' income of household i, then for employees in employment $Y_i = Y'_i$, but for all the self-employed:

$$Y_i = k_i Y'_i, \quad k_i \geqq 1. \tag{1}$$

k_i is a random variable showing the extent of under-reporting of income by self-employed household i. A bigger k_i indicates more under-reporting by household i.

For each item of expenditure j there is an expenditure function,

$$\ln C_{ij}=Z_i\alpha_j+\beta_j \ln Y_i^p+\varepsilon_{ij}, \tag{2}$$

where α_j is a vector of parameters, β_j is a scalar 'marginal propensity to consume' good j, and ε_{ij} is white noise. Y_i^p is the measure of income that influences consumption decisions. For food expenditure this measure is likely to be less variable than observed income. We refer to it as permanent income, without necessarily requiring that the expenditure function conforms exactly to the permanent income hypothesis.

The importance of the distinction between permanent income and measured income for our purposes is that for given permanent income, the measured income of the self-employed may be more variable than the measured income of employees in employment. If this is correct, our measure of income under-reporting by the self-employed will have to be adjusted accordingly. In general, we assume that permanent and measured income are related by

$$Y_i=p_iY_i^p, \tag{3}$$

where p_i is a random variable. The expected value of p_i for each household depends on aggregate events: in a 'good' year p_i will have a mean above unity. We make the critical assumption that the mean of p_i is the same for the employees in employment and the self-employed. Its variance may be different for each group and in general we expect the variance of p_i to be bigger for the self-employed than for the employees in employment.

Now (1) and (3) imply that the log of permanent income is

$$\ln Y_i^p=\ln Y_i'-\ln p_i+\ln k_i. \tag{4}$$

The assumptions underlying (1) and (3) imply that if we use reported income in place of unobserved permanent income in (2), there are two additional random regressors, $-\ln p_i$ and $\ln k_i$, each entering with coefficient β_j.

Since we do not have data on either p_i or k_i we make some assumptions on their distribution over households, to make estimation tractable. Thus, we assume that both p_i and k_i are log-normal, and write them as deviations from their means:

$$\ln p_i=\mu_p+u_i, \tag{5}$$

$$\ln k_i=\mu_k+v_i. \tag{6}$$

The random variables u_i and v_i have zero means and constant variances, σ_u^2

and σ_v^2, within each occupational group. We do not need to make assumptions about any covariation between u_i, v_i and ε_{ij} at this stage.[4]

We have argued that:

(a) the employees in employment report their incomes correctly, so for them $k_i=1$ and $\sigma_v^2=0$, whereas for the self-employed we should generally expect to find $\mu_k>0$ and $\sigma_v^2>0$;

(b) the mean of p_i, denoted by $\bar{p}$, is the same for each occupational group, but the variance σ_u^2 is bigger for the self-employed than for the employees in employment.[5]

By the log-normality of p_i the mean of p_i and the mean of its log, μ_p, are related by

$$\ln \bar{p}=\mu_p+\tfrac{1}{2}\sigma_u^2, \tag{7}$$

and so, if we use subscript SE to denote the self-employed and subscript EE to denote the employees in employment:

$$\mu_{p\mathrm{SE}}-\mu_{p\mathrm{EE}}=-\tfrac{1}{2}(\sigma_{u\mathrm{SE}}^2-\sigma_{u\mathrm{EE}}^2)\leqq 0. \tag{8}$$

By the convexity of logs, equality between the means of p_i implies that the mean of the log of p_i of the self-employed is less than the mean of the log of p_i of the employees.

Substituting now from (4), (5) and (6) into (2) we get:

$$\ln C_{ij}=Z_i\alpha_j+\beta_j\ln Y_i'-\beta_j(\mu_p-\mu_k)-\beta_j(u_i-v_i)+\varepsilon_{ij}. \tag{9}$$

Suppose we estimate this equation separately for the self-employed and for employees in employment, but impose the restrictions that the α_j and β_j are common. Then the intercepts of the equations should differ because $\mu_p-\mu_k$ is not the same in each group. The variance of the errors of each equation should also differ, with the self-employed generally having bigger variance. These differences in the estimates can be used to obtain an estimate of income under-reporting for the self-employed.

[4] The assumption of log-normal k_i obviously violates the restriction $k_i\geqq 1$, i.e. that no one over-reports income. Estimation with truncated errors is not possible for this problem because the regression error is a composite of the three errors, u_i, v_i, and ε_{ij}, which cannot be individually identified. See the regression equation (9) below.

[5] Since p_i shows differences between normal and actual incomes due to time effects and chance events, its distribution may be taken to be independent of the distribution of Y_i^p. Then from (3) mean actual income is the product of $\bar{p}$ and mean permanent income. Our assumption of equal $\bar{p}$ for the self-employed and other employees implies that transitory aggregate events affect all occupational groups equi-proportionally.

3. An estimate of the black economy

We estimated eq. (9) for food expenditure by running regressions of the form:

$$\ln C_{ij} = Z_i \alpha_j + \beta_j \ln Y_i' + \gamma_j SE_i + \eta_i, \tag{10}$$

where SE_i is a dummy variable taking the value 1 if household i is headed by a self-employed individual and 0 if it is headed by an employee in employment. All other households were excluded from the estimation. We estimated the regression separately for two broad occupational groups, 'white-collar' and 'blue-collar'. The appendix gives the definitions of the statistical series used in the estimation of (10) and the adjustments made to them (where appropriate). A few remarks are in order here.

Self-employment is defined by our data source according to independent criteria, but we prefer to define it as consisting of all households with income from self-employment of at least 25 percent of total income. With this definition we capture some households who may classify themselves as employees, yet may have a sizeable part of their income from self-employment and so have ample opportunity to engage in black economy activities. Our definition and the FES one are, however, highly correlated and the results with each are virtually identical. We inevitably miss any under-reporting of small amounts of self-employment income earned by full-time employees, who have occasional opportunities to work on their own account. These employees cannot be identified from the FES, since most of them would report zero self-employment income.

Table 1 gives some descriptive statistics from our sample. As in the national averages, about 12 percent of white-collar workers and 8 percent of blue-collar workers in our sample are classified as self-employed. The income and food expenditure of the self-employed are higher than the corresponding entries for the employees in employment. But the most striking difference in the data is in the standard deviations of incomes, where for the self-employed it is twice as high as for the employees. A Kolmogorov–Smirnov test for log-normality of income shows that log-normality is not rejected at the 95 percent level for any group, at the conventional approximation used to derive critical values for the test. Thus, the data give some support for the assumptions made in the preceding section about income and its distribution.

The error η_i in the regression eq. (10) is, by assumption, heteroscedastic. We estimated (10) under the assumption that its variance takes only two values: one for the self-employed and one for the employees. The estimates obtained are consistent with our assumptions.

Reported income Y_i' was treated as endogenous and instrumented. This enables an independent estimate of the residual variance of reported income

Table 1
Some descriptive sample statistics.

	White-collar		Blue-collar	
	Self-employed	Employees	Self-employed	Employees
Number	101	824	95	1,188
Age	43.6 (11.9)	39.5 (10.9)	38.7 (10.5)	38.1 (11.5)
Number of children	1.3 (1.1)	1.2 (1.1)	1.5 (1.2)	1.3 (1.1)
House mortgaged	0.6 (0.5)	0.8 (0.4)	0.6 (0.5)	0.5 (0.5)
Ln (income)	4.67 (0.6)	4.61 (0.3)	4.35 (0.5)	4.31 (0.3)
Ln (food expenditure)	2.93 (0.4)	2.80 (0.4)	2.89 (0.4)	2.73 (0.3)
Kolmogorov–Smirnov test of normality of ln (income)	0.046 (0.136)	0.032 (0.047)	0.084 (0.139)	0.023 (0.039)

Note: The numbers in parentheses are standard deviations, except for the Kolmogorov–Smirnov statistic, where they are the approximate critical values at the 95 percent level (normality is rejected if the statistic exceeds the critical value). See Kendall and Stuart (1979, p. 481).

Table 2
Heteroscedasticity-corrected IV estimates.

	β_j	γ_j	$\sigma^2_{\eta\mathrm{SE}}$	$\sigma^2_{\eta\mathrm{EE}}$	σ^2_{YSE}	σ^2_{YEE}
White-collar ($N=925$)	0.270 (0.097)	0.092 (0.048)	0.185	0.138	0.250	0.065
Blue-collar ($N=1{,}283$)	0.235 (0.081)	0.107 (0.042)	0.157	0.083	0.146	0.060

Notes: $N=$number of observations. Asymptotic standard errors in parentheses. Refer to eq. (10) for the meaning of the first four symbols. σ^2_{Y} denotes the residual variance from a reduced-form income regression.

for each group, which we exploit in the calculation of income under-reporting. We discuss our estimation procedure and report the estimates of the coefficients of (10) in the appendix. Here we give a table of the coefficient estimates that we use in the calculation of income under-reporting. These are the marginal propensity to consume, β_j, the coefficient on the self-employment dummy, γ_j, and the residual variance of reported income, σ^2_{Y}. The variance of η_i for each group, $\sigma^2_{\eta\mathrm{SE}}$ and $\sigma^2_{\eta\mathrm{EE}}$, is also of interest and is given in the table (see table 2).

The marginal propensity to consume food is estimated to be 0.270 for households in white-collar occupations and 0.235 for households in blue-collar occupations. The ranking of the coefficients may be a little surprising, given that average incomes in white-collar occupations are higher than average incomes in blue-collar occupations but the estimates are very close to each other.

The marginal propensities to consume are the same for employees and for the self-employed, as our working assumptions require. We tested for differences in the coefficients of the self-employed and employees by introducing the dummy variable SE_i interactively with income but rejected the null. The t-statistic for the interactive term in the white-collar regression was -1.0 and in the blue-collar regression -0.9. We also tested for non-linearities in the effect of income on expenditure by introducing the square of the log of income in each regression as an additional regressor. The t-statistic in the white-collar regression was 0.9 and in the blue-collar regression 0.3. Thus, we also rejected the null of non-linearity.

The coefficient on the shift dummy, SE_i, is positive and similar in value for both classes of occupations. Thus, the self-employed appear to consume more than the employees in employment after controlling for income and household characteristics. We come to the interpretation of this estimate and its implications for under-reporting below. We tried several other shift dummies, for other occupational groups, but they were all insignificant.

The residual consumption variance for the self-employed is always higher than the residual consumption variance for the employees, as we hypothesized. This variance is a composite of the variances and covariances of three types of errors, as (9) makes clear. In order to get an independent estimate of the variance of errors in income, we calculated the residual variance from reduced-form regressions for income, of the form:

$$\ln Y_i' = Z_i\delta_1 + X_i\delta_2 + \zeta_i, \tag{11}$$

where X_i is a set of identifying instruments. The residual ζ_i is again a composite of three errors: unexplained variations in permanent income, deviations of actual from permanent income, u_i, and deviations of actual from reported income, v_i. By our previous argument concerning the properties of u_i and v_i for the self-employed and the employees, the residual income variance of the self-employed should exceed the residual income variance of the employees. Table 2 confirms this for both occupational classes.[6]

From the discussion of eqs. (5), (6) and (8) we conclude that γ_j estimates:

[6] We cannot use reduced forms like (11) to obtain a direct estimate of under-reporting because we do not know if the earnings function of the self-employed is the same as the earnings function of the employees. In general they are not, because of self-selection, e.g. the self-employed may be more risk-tolerant and their incomes may reflect it. Our working hypothesis is that given incomes and other household characteristics, the food consumption functions of the two groups are the same.

$$\gamma_j = \beta_j[\mu_k + \tfrac{1}{2}(\sigma^2_{u\mathrm{SE}} - \sigma^2_{u\mathrm{EE}})]. \tag{12}$$

The estimate we are interested in is of the mean value of k_i, the number by which average reported self-employment income has to be multiplied to give average true income. Let this number be $\bar{k}$. By the assumed log-normality of k_i we have:

$$\ln \bar{k} = \mu_k + \tfrac{1}{2}\sigma^2_{v\mathrm{SE}}. \tag{13}$$

Thus, to go from the estimated coefficient on the self-employment dummy to an estimate of mean under-reporting we need to know the variances of u_i and v_i. In general these are not known, but some inferences about them can be made from the estimated residual income variance.

Suppose that the unexplained variations in permanent income in (11) have the same variance for both the employees and the self-employed. This is not an unreasonable assumption, given that these variations are due to omitted variables and that the self-employment dummy is one of the regressors in (11).[7] Then by the assumption that the employees do not under-report their income we have:

$$\operatorname{var} \zeta_{\mathrm{SE}} - \operatorname{var} \zeta_{\mathrm{EE}} = \operatorname{var}(u-v)_{\mathrm{SE}} - \operatorname{var} u_{\mathrm{EE}}. \tag{14}$$

Expanding the variance of $u-v$ and translating to the symbols of table 2 we write (14) as:

$$\sigma^2_{Y\mathrm{SE}} - \sigma^2_{Y\mathrm{EE}} = \sigma^2_{u\mathrm{SE}} + \sigma^2_{v\mathrm{SE}} - 2\operatorname{cov}(uv)_{\mathrm{SE}} - \sigma^2_{u\mathrm{EE}}. \tag{15}$$

From (12) and (13) we obtain,

$$\mu_k + \tfrac{1}{2}\sigma^2_{v\mathrm{SE}} = \gamma_j/\beta_j + \tfrac{1}{2}(\sigma^2_{v\mathrm{SE}} - \sigma^2_{u\mathrm{SE}} + \sigma^2_{u\mathrm{EE}}). \tag{16}$$

Comparison of (15) with (16) shows that our estimates of the residual income variances of the two occupational groups do not give us enough information to calculate the degree of under-reporting. But it is possible to calculate a range for mean under-reporting, by reasoning as follows.

Consider variations in the variances $\sigma^2_{v\mathrm{SE}}$ and $\sigma^2_{u\mathrm{SE}}$ that satisfy condition (15), for given values of the other variances (which we treat as parametric)

[7] We have no way of checking this assumption from our data (or from any other readily available data source). We make it as an identifying assumption, since without a restriction on the variance of permanent income we cannot derive the variances of u and v that are needed in our estimates. If self-employed permanent incomes have more variance than other permanent incomes, in contrast to our hypothesis, the estimated variance of u and v for the self-employed in (15) would be less, with ambiguous effect on our estimates of average under-reporting. However, the difference would not be big, as the discussion that follows will make clear.

and given the partial correlation coefficient ρ between u_{SE} and v_{SE}. The question we pose is whether our estimate of mean under-reporting in (16) varies within a small range when $\sigma^2_{v\mathrm{SE}}$ and $\sigma^2_{u\mathrm{SE}}$ vary over their feasible range. It is easy to check that if $\rho=0$, there is a small well-determined range for mean under-reporting; but if $\rho\neq 0$ some difficulties arise, which we discuss below.

If $\rho=0$ (15) implies that $\sigma^2_{v\mathrm{SE}}$ and $\sigma^2_{u\mathrm{SE}}$ are negatively related, so (16) gives a lower bound for mean under-reporting when $\sigma^2_{v\mathrm{SE}}$ takes its lowest value and an upper bound when $\sigma^2_{u\mathrm{SE}}$ takes its lowest value. The lowest feasible value for $\sigma^2_{v\mathrm{SE}}$ is 0, so (15) and (16) imply that the lower bound satisfies,

$$\mu_k+\tfrac{1}{2}\sigma^2_{v\mathrm{SE}}=\gamma_j/\beta_j-\tfrac{1}{2}(\sigma^2_{Y\mathrm{SE}}-\sigma^2_{Y\mathrm{EE}}). \tag{17}$$

With the estimates given in table 2 this lower bound is

$$\begin{aligned}\mu_k+\tfrac{1}{2}\sigma^2_{v\mathrm{SE}}&=0.248 \text{ for white-collar}\\ &=0.412 \text{ for blue-collar.}\end{aligned}$$

Translating this into the mean value of k_i, the number by which mean declared incomes have to be multiplied to arrive at true incomes, by taking antilogs we find

$$\begin{aligned}\bar{k}_l&=1.28 \text{ for white-collar}\\ &=1.51 \text{ for blue-collar.}\end{aligned}$$

We qualify $\bar{k}$ by subscript l to indicate that this estimate of mean under-reporting is a lower bound.

We have argued earlier that self-employed incomes have at least as much variance as salaries and wages, so the minimum feasible value for $\sigma^2_{u\mathrm{SE}}$ is $\sigma^2_{u\mathrm{EE}}$. Thus, the upper bound on the degree of under-reporting is obtained when $\sigma^2_{u\mathrm{SE}}=\sigma^2_{u\mathrm{SE}}$. In this case (15) and (16) imply,

$$\mu_k+\tfrac{1}{2}\sigma^2_{v\mathrm{SE}}=\gamma_j/\beta_j+\tfrac{1}{2}(\sigma^2_{Y\mathrm{SE}}-\sigma^2_{Y\mathrm{EE}}). \tag{18}$$

Comparison with (17) shows that our range for mean under-reporting is symmetric. With the estimates in table 2 we get in this case

$$\begin{aligned}\bar{k}_u&=1.54 \text{ for white-collar}\\ &=1.64 \text{ for blue-collar.}\end{aligned}$$

A comparison of the upper and lower bounds for under-reporting shows a small range in the estimate, especially for blue-collar workers. The reason is that for blue-collar workers in particular, the estimated randomness in the reported income of the self-employed is not much higher than the estimated randomness in the income of the employees in employment. In the case where the covariance between u and v is zero, the difference in the estimated randomness of incomes imposes an upper bound on the sum of the variances of u and v.

But if the covariance between u and v is not zero, the sum of the variances of u and v can be higher. For this reason, it is worth considering what the covariance indicates and whether it is reasonable to assume that it is equal to zero.

Whether this covariance is zero or not depends on what assumptions we make about the propensity of some people to under-report their incomes by more than others. It is plausible to argue that if a self-employed person landed an exceptionally good job in one year, he or she might be less inclined to declare the income from it than if the source of income was regular. If the income is exceptional, a tax return that does not show it is not likely to arouse the taxman's curiosity. Similarly, if in one year a self-employed person happens to have below normal income, he or she may be more inclined to declare it. The tax on it may be proportionally less and not declaring part of it may give rise to a declared income that is small enough to arouse curiosity. Behaviour of this kind gives rise to a positive covariance between u_i and v_i. Zero covariance requires that whatever the income of the self-employed person turned out to be in any particular year, his or her tendency would be to report the same percentage of it to the tax authorities as in previous years.

Unfortunately, if the covariance between u and v is positive it is not possible to calculate a range for under-reporting without further information on the distribution of u and v. But we can show by example that even high correlations between u and v do not affect our previously estimated range by very much.

If the correlation between u and v is small, it is possible to show that our estimate of under-reporting is an under-estimate (this is likely to be the case also for high correlations). We consider, therefore, what happens to our estimated upper bound, $\bar{k}_u$, as the correlation coefficient ρ becomes positive. In this case (15) and (16) imply

$$\mu_k + \tfrac{1}{2}\sigma^2_{v\mathrm{SE}} = \gamma_j/\beta_j + \tfrac{1}{2}(\sigma^2_{Y\mathrm{SE}} - \sigma^2_{Y\mathrm{EE}}) + \mathrm{cov}(uv)_{\mathrm{SE}}. \qquad (19)$$

Calculation of the covariance requires knowledge of the partial correlation coefficient ρ and the standard deviations $\sigma_{u\mathrm{SE}}$ and $\sigma_{v\mathrm{SE}}$. An estimate of an upper bound on $\sigma_{u\mathrm{SE}}$ is available. The variance of u_i cannot exceed the residual income variance of the employees in employment, since the latter is made up of the sum of the variance of u_i and the residual variance of permanent income. Taking the estimate of $\sigma^2_{Y\mathrm{EE}}$ in table 2 as an upper limit for $\sigma^2_{u\mathrm{EE}}$, and noting that $\sigma^2_{u\mathrm{SE}}=\sigma^2_{u\mathrm{EE}}$ in the case we are now investigating, we obtain from (15),

$$\sigma^2_{Y\mathrm{SE}}-\sigma^2_{Y\mathrm{EE}}=\sigma^2_{v\mathrm{SE}}-2\rho\sigma_{Y\mathrm{EE}}\sigma_{v\mathrm{SE}}. \tag{20}$$

Equation (20) and the estimates in table 2 are used to solve for $\sigma^2_{v\mathrm{SE}}$, which when combined with our assumed upper bound for $\sigma^2_{u\mathrm{SE}}$, give an estimate of the upper bound on the covariance of $(uv)_{\mathrm{SE}}$ for given ρ. If $\rho=1$ this estimate is,

$$\begin{aligned}\operatorname{cov}(uv)_{\mathrm{SE}}&=0.192 \text{ for white collar}\\&=0.156 \text{ for blue collar.}\end{aligned}$$

Hence, from (19) we obtain the new estimates for our former upper bounds on mean under-reporting,

$$\begin{aligned}\bar{K}'_u&=1.87 \text{ for white collar}\\&=1.92 \text{ for blue collar.}\end{aligned}$$

Thus, even $\rho=1$ does not raise our previously obtained estimates by very much. Smaller correlation coefficients reduce the effect further. If $\rho=0.5$, the estimates fall to 1.66 for white-collar and 1.74 for blue-collar workers respectively. We may conclude that although a positive correlation between chance variations in income and chance variations in under-reporting may raise our estimate of average under-reporting, the effect it is likely to have on the estimate is not very big. A reasonable mid-point estimate for mean under-reporting, even with positive covariance, is a little below 1.5 for white-collar workers and about 1.6 for blue-collar workers, averaging at about 1.55 for all self-employed individuals.

4. Conclusions

Our estimate of the size of the black economy in Britain, defined in the narrow sense of income-earning activities that are concealed from the taxman, relies on two assumptions. First, expenditure on food by all groups in the population is correctly reported in the Family Expenditure Survey. Second, employees in employment report their income correctly but the self-employed may conceal part of it. On the basis of these assumptions and some others on the stochastic distribution of actual and reported income, we estimated consistent food expenditure equations for individual households. By inverting these equations we were then able to arrive at the conclusion that on average reported self-employment incomes have to be multiplied by a factor of 1.55 to give true incomes. Since income from self-employment amounts to about 10 percent of GDP, our estimate implies that the size of the black economy is about 5.5 percent of GDP. This estimate is higher than the Inland Revenue's implicit estimate of 1.5 percent of GDP but it is consistent with most of the other microeconomic evidence.

We found some evidence that 'blue-collar' workers conceal on average a little more of their income than 'white-collar' workers, 60 percent as compared to 50 percent or less. But our estimate of under-reporting by white-collar workers has a bigger range and may be as high as 80%.

Appendix: Data and estimation

The data we used come from the 1982 Family Expenditure Survey and relate to individual households. We selected our sample according to the following criteria:

(a) each household has two adults;

(b) the head is either employed or self-employed;

(c) the second adult is the wife;

(d) the household lives in Great Britain;

(e) no obvious inconsistencies can be detected in the data used;

(f) the quality code (A268) for self-employment income is less than 2 (the interviewer actually saw a written record of income receipts from self-employment).

After selection, we had 2,208 households, of which 925 were classified as white-collar (the occupation code for the head taking values 1 to 5), and the remaining 1,283 as blue-collar (occupation 6 to 8).

The data on incomes given in the survey had to be adjusted in one important respect: while earnings from employment recorded in the survey were received in the few weeks or months before the interview, self-employment income related to the last available record, usually for the previous year. If we were to use the data without correcting for the different timing, we would risk finding spurious income 'under-reporting' for the self-employed due to inflation. Fortunately, information is provided on the precise date that the reported self-employment income was received (codes A226 and A227), and rescaling is possible.

We proceeded as follows: we took annual figures for income from self-employment (CSO, Blue Book, 1985), net of tax-evasion corrections, and divided these by the number of self-employed individuals (Department of Employment Gazette). An annual per capita series was thus obtained, and monthly observations were calculated by linear interpolation. We thus obtained a monthly rate of inflation of self-employment incomes which we used to update the recorded incomes from self-employment. Self-employment income was computed as the sum of codes 326 and 328 (thus excluding self-supplied goods). With knowledge of gross incomes, we obtained net current income by subtracting recorded tax payments.

The following variables were constructed:

LFOOD = logarithm of expenditure on food
WCW = wife in white-collar job (occ. code 1 to 5)
BCW = wife in blue-collar job (occ. code 6 to 8)
AGE = age of the head of household minus 40
AGESQ = $(\text{age} - 40)^2$
NCH = number of children
NCHSQ = $(\text{number of children})^2$
LAT = dummy for local authority tenants
REN = dummy for other rented accommodation
OOM = dummy for owner-occupiers with mortgage
GLC = dummy for Greater London
NOR = dummy for Northern regions (Scotland, Yorkshire and Humberside, North, Northwestern)
CH = dummy for centrally-heated accommodation
WM = dummy for ownership of a washing machine
TV = dummy for ownership of a TV set
ROOM = number of rooms in the house
KIDS1 = number of children aged 0 to 4
KIDS2 = number of children aged 5 to 16
S1 = dummy variable for interviews that took place in the first quarter
S2 = dummy variable for interviews that took place in the second quarter
S3 = dummy variable for interviews that took place in the third quarter
SED1 = dummy variable for self-employed household: it takes value 1 if at least 25 percent of household income is from self-employment
SED = dummy variable for self-employed households: it takes value 1 if economic position of the head is given as self-employed (code 1)
LNCY = logarithm of net current income (adjusted self-employement income where appropriate)
CAR = number of cars owned by the household
RAT = rateable value of the house
NCD = dummy variable: it takes value 1 if current and 'normal' income differ
SEDW = dummy variable for self-employed wife (economic position code = 1)
FTW = dummy variable for full-time employed wife
PTW = dummy variable for part-time employed wife.

The coefficient estimates given in table 2 of the text are the income

The following food equations were estimated by generalised 2SLS:

	White-collar	Blue-collar
Dependent variable: logarithm of food expenditure		
Constant	1.2234 (0.046)	1.5953 (0.320)
WCW	−0.0024 (0.033)	0.0385 (0.032)
BCW	−0.0597 (0.041)	0.0554 (0.025)
AGE	0.0047 (0.002)	0.0030 (0.001)
AGESQ/10	−0.0029 (0.001)	−0.0016 (0.001)
NCH	0.1449 (0.057)	0.2778 (0.045)
NCHSQ	−0.0083 (0.010)	−0.0172 (0.006)
LAT	0.0970 (0.067)	−0.0243 (0.034)
REN	0.0772 (0.073)	−0.0159 (0.043)
OOM	−0.0362 (0.048)	−0.0706 (0.034)
GLC	0.0755 (0.042)	0.0703 (0.034)
NOR	0.0143 (0.027)	−0.0093 (0.017)
CH	−0.0048 (0.040)	0.0131 (0.019)
WM	−0.0194 (0.068)	−0.0427 (0.034)
TV	0.1081 (0.035)	0.0189 (0.023)
ROOM	0.0120 (0.012)	0.0053 (0.009)
KIDS1	−0.0634 (0.054)	−0.1135 (0.041)
KIDS2	−0.0069 (0.052)	−0.0741 (0.041)
S1	−0.0192 (0.035)	−0.0095 (0.024)
S2	0.0567 (0.036)	−0.0860 (0.023)
S3	0.0030 (0.035)	−0.0700 (0.024)
SED1	0.0919 (0.048)	0.1065 (0.042)
LNCY	0.2695 (0.096)	0.2354 (0.081)
$\bar{R}^2$	0.1811	0.2966
N	925	1,283
χ^2_{22} (SC)	23.75	25.58
χ^2_1 (Het)	11.93	36.61

Notes: *SED1* and *LNCY* were treated as endogenous. SC is the test for overidentifying restrictions. Het is a heteroscedasticity test run on the original 2SLS equations.

Additional instruments: *NCD*, *CAR*, *RAT*, *SEDW*, *FTW*, *PTW*, and the product of *SED* with *RAT*, *TV*, *CAR*, *AGE*, *AGESQ*, *NCH*, *NCHSQ*, *KIDS1*, *KIDS2*, *CH*, *WM*, *ROOM*, *LAT*, *REN*, *SEDW*, *FWO*, *PTW*.

coefficients (β_j) and the coefficients estimated for *SED1* (γ_j) in the regressions overleaf. The variances of η are the residual variances of these regressions, for the observations pertaining to the self-employed and to the employees in employment separately. The other variances given in table 2 are the residual variances from regressions of log income (*LNCY*) on all the other independent variables in the food regressions [the Z variables of eq. (11)] and the 'additional instruments' listed below (the X variables). The reduced form income regressions are of no particular interest so they are not reported.

References

Cowell, F.A., 1985, The economics of tax evasion, Bulletin of Economic Research 37, 163–193.

Dilnot, A.W. and C.N. Morris, 1981, What do we know about the black economy?, Fiscal Studies 2, 58–73.

Kendall, M. and A. Stuart, 1979, The advanced theory of statistics, 2 (Charles Griffin & Co., London), fourth edition.

MacAfee, K., 1980, A glimpse of the hidden economy in the national accounts, Economic Trends 316, 81–87, Feb.

Smith, S., 1986, Britain's shadow economy (Oxford University Press, Oxford).