Estimating the Income Reporting Function for the Self-Employed

Lindsay M. Tedds

School of Public Administration, University of Victoria

PO Box 1700 STN CSC, Victoria, BC, Canada, V8W 2Y2

Voice: 250-721-8068

Fax: 250-721-8849

Email: ltedds@uvic.ca

The author would like to gratefully acknowledge the financial support from the Ontario Graduate Scholarship Program and the Social Sciences and Humanities Research Council (Doctoral Fellowship #752-2004-1096 and INE grant #501-2002-0107).

Abstract

There is considerable interest in measuring tax noncompliance using microeconomic data. One such method estimates income under-reporting by self-employed households by assuming a known, parametric form of the Engel curve and making the further parametric assumption that self-employed households under-report their income by a constant fraction, independent of income. This paper proposes a nonparametric approach which avoids functional form restrictions and allows the form of the reporting function to be uncovered rather than being imposed a priori. The approach is illustrated by estimating the effect of the Canadian Goods and Services Tax on income under-reporting by a sample of self-employed households.

Keywords: *Underground Economy, Income Under-reporting, Nonparametric Estimation, Engel Curve*

JEL Classification: C14, D12, O17

1. Introduction

There has been a recent resurgence in interest in measuring the underground economy and this interest has been stimulated predominantly by the perception that the underground economy is sizeable and growing. In broad terms, the phrase "underground economy" refers to output that is produced and income that is generated by agents who hide this fact from authorities. Knowledge of the size and structure of the underground economy is important for a number of reasons. First, because underground activities are unmeasured, they are not taken into account in the information-set that is used to assist economic policy-makers. Second, the underground economy effectively re-distributes both income and wealth in ways that are not necessarily consistent with the re-distributional goals of the taxation system. Third, the shortfall in income-reporting that is associated with underground activities leads to an erosion in the tax base and tax revenue with subsequent implications for both public expenditure and taxation policies. Finally, enforcement activities are unlikely to be successful (and may have counterproductive consequences) without detailed knowledge of the characteristics and types of activities of underground economy participants.

To date, research that seeks to measure the underground economy has predominately employed macro-methods. These macroeconomic measures, however, have been criticized for not being consistent with modern economic models of consumer behaviour, employing flawed econometric techniques, producing unreliable estimates, and providing limited guidance to policy makers (Thomas 1999). In

particular, the macro-methods developed to date do not distinguish between hidden legal activity and criminal activity, differentiate between hidden activity by individuals and firms, or provide any information regarding the characteristics of those participating in the underground economy. In order to obtain this type of information, a method that uses microeconomic data is required.

One such approach, popularized by Pissarides and Weber (1989) (hereinafter P&W) and modified by Lyssiotou *et al.* (2004) (hereinafter LP&S), employs household income and expenditure data to estimate the degree of income under-reporting (i.e. the amount by which household income should be scaled upwards to obtain true, or actual, income as opposed to reported income) by self-employed households. The basic principle of this *Expenditure-based method* is that true household income can be imputed from reported household expenditures. The method is premised on variations of several key assumptions, namely: the reporting of expenditures on some items by all households is accurate; wage and other income is reported accurately while those who report non-zero self-employment income may be under-reported; and the marginal propensity to consume out of unreported income is equal to the marginal propensity to consume out of reported income. Actual, or true, income is than imputed by comparing the expenditure levels of households with positive self-employment income to the expenditure-income bundles of households with zero self-employment income and similar characteristics. In practice, the method is implemented by estimating reliable expenditure functions (i.e. Engel curves) for wage earners that are than inverted to estimate true income for the self-employed.

Previous studies have implemented the *Expenditure-based method* using parametric restrictions on the income reporting function and arbitrarily classifying households that have at least 25% of total income from self-employment as self-employed. The parametric restrictions imposes that those households that are classified as self-employed under-report their income by a constant fraction. There is, however, no empirical evidence that supports this restriction and little, if anything, is actually known about the functional form of the reporting function. The classification of self-employed households results in the total income, not just self-employment income, reported by these households being considered as potentially under-reported. As many self-employed households have wage and other income assumed to be accurately reported (e.g. government transfer payments) in addition to self-employment income, the method violates one of the critical assumptions necessary to estimate income under-reporting; namely, that some income is accurately reported.

This paper considers an alternative way of implementing the *Expenditure-based method*. In particular, the parametric restrictions are relaxed and a nonparametric approach to the measurement of income under-reporting is explored. Further, the approach separates self-employment income from other types of household income and only self-employment income is considered when estimating income under-reporting. Specifically, a two-step approach to estimating a variable-with-income reporting function is proposed, within the framework of the *Expenditure-based method*. The approach is essentially as follows. First, a nonparametric inverse food Engel curve is estimated for the sample of households that report positive wage income in conjunction with zero self-employment income,

to obtain an estimate of true income given (accurately) reported expenditures for every household in the sample (including those with self-employment income). Second, a nonparametric reporting function for self-employment income is estimated. This approach improves on the implementation of the *Expenditure-based method* by minimizing the number of assumptions required for estimation.

The approach is illustrated by estimating the effect of the Canadian Goods and Services Tax (GST) on income under-reporting by married households without children present in the household with self-employment income. It is often argued that the implementation of this broadly based consumption tax increased the incentives and opportunities for tax evasion (e.g. Spiro 1993, and Hill and Kabir 1996) though the Government of Canada maintained that it would reduce the scope the tax evasion. The empirical analysis uses the Canadian Family Expenditure Survey (FAMEX), which contains household level information about income and expenditures.

Overall, this refinement to the *Expenditure-based method* produces results that demonstrate that the assumption that the self-employed under-report their income by a constant fraction does not apply to the Canadian data considered in this paper. Rather, it is found that prior to the implementation of the GST the gap between true and reported self-employment income is larger for households at the lower end of the self-employment income distribution. Subsequent to the implementation of the GST, the reporting function adjusted such that the gap between true and reported self-employment income is similar throughout the self-employment income distribution. Overall, the aggregate results support the hypothesis that the GST increased tax

noncompliance by those with larger amounts of self-employment income and unaffected tax noncompliance by those with small amounts of self-employment income.

The remainder of this paper is organized as follows. Section 2 critiques the existing empirical applications of the *Expenditure-based method* to provide context for the approach proposed in this paper. An economic model is presented in section 3 and the resulting econometric model is outlined in section 4. Section 5 details the application, including a discussion of the GST, the data, and the results. The paper than ends with some concluding comments.

2. CRITIQUE OF THE EXISTING LITERATURE

In this section, attention is focused on two critical aspects of the existing empirical applications of the *Expenditure-based method* with the view of placing the empirical strategy proposed in this paper into context. These aspects concern: (1) functional form restrictions; and (2) the treatment of income.

2.1 Functional Form Restrictions

A critical aspect of the empirical work in this area is the specification of the expenditure and reporting functions. The pioneering work in the development of the *Expenditure-based method* was conducted by P&W.[1] To implement the method they

[1] The *Expenditure-based method* was developed following work conducted by Dilnot and Morris (1981) that calculated the difference between reported household income and expenditures and arbitrarily classify households as "black economy" households if expenditures exceeded income by at least 20 percent. Smith *et al.* (1986) propose a framework similar to that of P&W except that the slope

first categorize households as either being self-employed or wage earning. Second, they specify a constant elasticity Engel curve that is used to estimate the parameter θ in the reporting function for self-employed households, defined as

$$y_{t,h}^{TSE} = \theta y_{t,h}^{RSE} \quad 1$$

where $y_{t,h}^{TSE}$ represents true income for self-employed households, $y_{t,h}^{RSE}$ denotes reported income for self-employed households, and θ is assumed to be > 1. This method of estimating income under-reporting consists of two steps. First, an expenditure function is estimated for wage earners. Second, the expenditure function is inverted to calculate θ, the amount by which reported income for self-employed households must be scaled up by in order to obtain true income for self-employed households.

Figure 1 provides a graphical representation of the approach. Constant-elasticity Engel curves for wage (or employee) and self-employed households are shown. A self-employed household reports expenditures, E*, and income, Y, but the reported level of expenditures is actually consistent with true income, Y*. The amount by which reported income must be scaled up to obtain true income is calculated by taking the ratio of the distance 0Y*/0Y which is equivalent to the parameter θ in equation (1) above. As the reporting parameter is assumed to be constant, the Engel curve for self-employed households is assumed to be parallel to that of wage earners.

---

of the Engel curve was allowed to differ across wage earners and the self-employed and they make an assumption about the mean of true income.

LP&S propose a systems approach to the *Expenditure-based method*. They specify a system of Engel curves of quadratic-in-(log)income Working-Leser form. They base their demand system on nondurable goods only, namely: food, alcohol, fuel, clothing, personal goods/services, and leisure goods/services. The functional form for the Engel curve that is specified by LP&S raises two concerns. First, there is an implicit assumption of the *Expenditure-based method* that the Engel curve(s) employed in the estimation must be monotonic in income. In reference to Figure 1, if this critical assumption is violated, than a unique value of true income associated with a particular level of expenditures may not exist. The quadratic-in-(log)income Working-Leser form of the Engel curve specified by LP&S is not necessarily consistent with the monotonicity assumption, with particular goods, notably alcohol and clothing, known to violate this assumption (Banks *et al.* 1997). Second, the quadratic-in-(log)income Working-Leser form of the Engel curve is not invertible over all values due to the presence of asymptotes. While the presence of asymptotes is not a concern under the structure imposed by LP&S- the system of Engel curves is not (implicitly) inverted over all data points - it underscores the likelihood that the estimates are influenced, in whole or in part, by the parametric restrictions.

More generally, LP&S maintain the specification of the linear reporting function given in equation (1).[2] In fact, little is known about the form of the reporting

[2] LP&S also allow for what they call "preference heterogeneity". They suggest that income from self-employment may not be spent in the same way as income from other sources. In particular, they argue that households may spend wage income, which is predictable, on necessities and the self-employment income, which is subject to under-reporting and is unpredictable, on luxuries. Equally, they indicated that self-employed could just have different preferences. P&W assumed homogenous preferences

function and it is plausible that under-reporting will differ with income and household characteristics. In fact, Reinganum and Wilde (1985, 1986) conclude that taxpayers with greater true income under-report less than those with lower true income. This paper proposes a nonparametric approach which avoids functional form restrictions. The proposed method also works directly with an inverse Engel curve, avoiding problems associated with inversion, and continues with the tradition of the single equation approach. The single equation approach also allows the analysis to be restricted to a good for which the Engel curve is widely acknowledged to be monotonic in income.

2.2 The Treatment of Income

There are two significant concerns associated with income that need to be considered. These concerns relate to the treatment of income in each of the expenditure function and the reporting function.

2.2.1 Expenditure Function

There is a general belief that households base expenditures on permanent rather than transitory income. This implies that households save when they have positive transitory income and dissave when they have negative transitory income. If the *Expenditure-based method* is implemented using transitory income this may lead to biased estimates of income under-reporting. P&W acknowledge that permanent

---

among all households. LP&S allow for preference heterogeneity in their estimated system of budget shares through the inclusion of the self-employment proportion of reported income, which can enter the system nonlinearly. The preference heterogeneity term(s), however, are identified only by functional form and are not identified in the nonparametric framework proposed in this paper.

income is the measure of income that influences consumption decisions but stop short of requiring their expenditure function to conform exactly to the permanent income hypothesis, perhaps because the dataset used in their analysis (1982 British Family Expenditure Survey) did not contain information regarding household savings behaviour. They indicate 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." (P&W, pp. 20) Empirically, they implement this assumption by treating reported income as endogenous and than using instrumental estimation, which "…enables an independent estimate of the residual variance of reported income for each group which is exploited in the calculation of income under-reporting." (P&W, pp. 22)

Whether P&W Two-Stage Least Square (2SLS) approach is preferred to Ordinary Least Squares (OLS) depends on the quality of the instruments. Datasets that contain information on household expenditures and income may not contain relevant and suitable instrumental variables required for this analysis. Further, the approach requires the researcher to make additional and somewhat arbitrary assumptions which restrict the analysis. As a result, an alternative approach is employed in this paper. In particular, a measure of consumption income is used in the expenditure function rather than total income.

2.2.2 Reporting Function

As indicated in section 2.1 above, P&W define self-employed households as those households that report self-employment income as being at least 25 percent of total income. The reporting function is than estimated for these households using total reported after-tax income. However, many self-employed households also have wage income, as well as other accurately reported income such as government transfers, in addition to self-employment income. By including these sources of income in the reporting function, the authors violate one of the most critical assumptions necessary to estimate income under-reporting; namely, that some income is accurately reported.[3]

LP&S attempt to address this shortcoming by including wage and self-employment income independently in their system of Engle curves however it is not clear if the income variables are net of income taxes paid. Indeed, it is unlikely that the income variables are net of taxes since it is difficult to accurately allocate taxes paid to the various income components. Hence, the income used in the expenditure function may not accurately reflect the amount of income actually available to fund expenditures which may subsequently lead to errors when estimating income under-reporting. The approach proposed in this paper ensures that only self-employment income is included when estimating the income reporting function and that only income that is available to fund expenditures is included in the expenditure function.

---

[3] This classification of households also means that households who derive less than 25 percent of this income from self-employment were included in the base group that are assumed to accurately report their income.

3. Economic Model

As outlined above, to date, the *Expenditure-based method* has been implemented by estimating Engel curves which are implicitly or explicitly inverted to obtain an average estimate of income under-reporting. A more direct approach to estimating income under-reporting is to use an inverse Engel curve (i.e. with expenditures taking on the role of the dependent variable). The proposed approach responds directly to the concerns raised in section 2.

As is standard in the empirical literature, households are considered to have weakly inter-temporally separable preferences in consumption, leisure, and monetary assets. Following a standard two-stage budgeting process, at each time t, each household, h, allocates a portion of their wealth to consumption as follows:

$$E_{t,h} = y_{t,h}^{Reg} + y_{t,h}^{TSE} - T_{t,h} - \Delta S_{t,h} \qquad 2$$

where $y_{t,h}^{Reg}$ is all (assumed) accurately reported income, $y_{t,h}^{TSE}$ is true self-employment income, $T_{t,h}$ is taxes paid on all reported income, and $\Delta S_{t,h}$ is net financial savings. $\Delta S_{t,h}$ represents the amount that a household needs to allocate to or from savings in order to smooth consumption over time periods. If $\Delta S_{t,h}$ is negative than the household is drawing down savings to boost consumption in this time period whereas if $\Delta S_{t,h}$ is positive than the household is contributing to savings in this time period in order to boost consumption in the future. It is assumed that $y_{t,h}^{TSE}$ is only accurately observed for those households that report zero self-employment income.

Households then allocate expenditures at time t, $E_{t,h}$ across goods and the general expenditure function for any given good is given as:

$$F_{t,h} = f(E_t, \varepsilon_t) \qquad 3$$

where ε_t reflects a vector of taste components that will usually be related to E_t due to the inter-temporal optimization process assumed and $F_{t,h}$ represents expenditures on a particular good. Assuming that the expenditures noted by $F_{t,h}$ is accurate and the function f is monotonic and continuous, than it is possible to invert the function in (3) to obtain a truthful mapping from expenditures on the good under consideration to total consumption income in period t:

$$E_{t,h} = f^{-1}(F_{t,h}, \varepsilon_t) \qquad 4$$

The mapping process and the relationship noted in equation (2) can then be used to obtain a measure of $y_{t,h}^{TSE}$ for households that report positive self-employment income and this knowledge can be used to derive the form of the reporting function.

4. Econometric Model

The model proposed in section 3 is estimated using nonparametric methods. The use of nonparametric methods in this case has three advantages. First, it allows the form of the reporting function to be uncovered rather than being imposed a priori. Second, it provides the opportunity to test the null hypothesis that the reporting function takes the form specified in equation (1) and assumed in the previous

literature. Third, it avoids functional form restrictions on the Engel curve which may influence the estimates of under-reporting.

Estimating Engel curves using nonparametric techniques is quite common and demographic household characteristics, used to account for observable heterogeneity, are often included by specifying a partially linear additive semi-parametric specification. However, "restrictions from consumer theory are not innocuous both on the form of the Engel curve relationship and on the way in which observable heterogeneity (demographics in our case) can enter." (Blundell *et al.* 1998, pp. 436) In particular, Blundell *et al.* (1998) have demonstrated that "…the additive structure between demographic composition and income that underlies the partially linear semiparametric model implies strong and unreasonable restrictions on behaviour." (Blundell *et al.* 1998, pp. 459) Rather, to be consistent with consumer theory, the demographics that enter the Engel curve specification must also scale the income term. The nonparametric estimation strategy proposed here cannot be implemented if income and demographic terms enter non-additively, hence, semiparametric estimation was not pursued. Instead, estimation is conducted separately on an identified homogenous sub-population (i.e. married couples without dependent children present in household).[4]

To achieve estimation, some initial assumptions are required. The three fundamental assumptions of P&W are maintained but classifying households as either self-employed or not is avoided, following LP&S. The three fundamental

[4] This is not to say that a semi-parametric approach cannot be pursued within the framework proposed but is beyond the scope of this paper.

assumptions are as follows. First, food expenditures are used in the analysis and it is assumed that the reporting of food expenditures by all households is accurate. The arguments for using food as opposed to any other commodity or group of commodities are that: there is no social stigma associated with food consumption which could cause expenditures to be reported inaccurately (counter examples would include tobacco and alcohol); food expenditures are more likely to be reported accurately by households engaging is tax evasion since individual expenditures on food are small and are unlikely to rouse suspicion; tastes for food are more likely to be uniform across employment groups and over time; it is very difficult for a household to postpone food consumption; most food purchases cannot be included as a business expense; and, the food Engel curve is widely acknowledged to be continuous and monotonic. Second, only self-employment income can be under-reported.[5] Third, the marginal propensity to consume out of unreported income is constrained to be equal to the marginal propensity to consume out of reported income.

[5] Self-employment income is reported and taxed at year end (though many self-employed are required to make estimated tax payments during the year in order to ensure that they meet their tax obligation in a timely manner) by the individual who earned the income and there is no third party who also reports this income. That is, there is no check and balance within the system to ensure that the individual is accurately reporting their self-employment income, unlike with almost all other income sources. As a result, there is an opportunity for some self-employment income to be under-reported. That said, the assumption that only self-employment income is under-reported is likely not entirely accurate. For example, employers can pay their employees in whole or partially in cash as a way to evade income and payroll taxes. The extent that this assumption is not valid will lead to the resulting estimate of the degree of under-reporting to be biased toward zero.

The estimation strategy is as follows. The first stage of the procedure is to nonparametrically estimate an inverse food Engel curve to obtain true consumption income given (accurately) reported food expenditures, given by

$$E_{t,h} = g(F_{t,h}) + \varepsilon_t \qquad 5$$

where $E_{t,h}$ is consumption income as defined in equation (2) and $F_{t,h}$ is household food expenditures. By assumption, $F_{t,h}$ is accurately observed for all households but $E_{t,h}$ is only accurately observed for those households that have zero self-employment income. This implies that $g(F_{t,h})$ can be consistently estimated for households that report zero self-employment income. Estimation of equation (5) yields the conditional expected value $E[E_{t,h} \mid F_{t,h}]$ of total consumption income for households that report zero self-employment income which is used to obtain an accurate estimate of total consumption income for households with positive self-employment income based on food expenditures. As a result, consistent estimates $\hat{g}(F_{t,h})$ of total household consumption income are obtained for every household.

As indicated in equation (2) above, total household consumption income is comprised of four elements, namely the household's: true self-employment income ($y_{t,h}^{TSE}$), reported other income ($y_{t,h}^{\mathrm{Re}g}$), income taxes paid ($T_{t,h}$) and net financial savings ($\Delta S_{t,h}$). If $y_{t,h}^{\mathrm{Re}g}$ is subtracted from and $T_{t,h}$ and $\Delta S_{t,h}$ are added to the conditional expected value of total consumption income, one obtains an estimate of

expected true self-employment income for those households that report positive self-employment income. That is, a measure of $y_{t,h}^{TSE}$ can be obtained as follows:

$$E[y_{t,h}^{TSE} \mid F_{t,h}, y_{t,h}^{Reg}, T_{t,h}, \Delta S_{t,h}, y_{t,h}^{RSE} > 0] = E[E_{t,h} \mid F_{t,h}] - y_{t,h}^{Reg} + T_{t,h} + \Delta S_{t,h} \quad 6$$

where $y_{t,h}^{RSE}$ represents reported self-employment income. The expected amount of self-employment income that is unreported by households, $y_{t,h}^{USE}$, is obtained by subtracting reported self-employment income from expected true self-employment income

$$E[y_{t,h}^{USE} \mid F_{t,h}, y_{t,h}^{Reg}, T_{t,h}, \Delta S_{t,h}, y_{t,h}^{RSE} > 0] = E[y_{t,h}^{TSE} \mid F_{t,h}, y_{t,h}^{Reg}, T_{t,h}, \Delta S_{t,h}] - y_{t,h}^{RSE} \quad 7$$

Total unreported income is found by summing over every household with positive reported self-employment income.

The second step estimates the nonparametric form of the reporting function, the parametric form of which is given by equation (1), for those households that report positive self-employment income. The nonparametric form of the reporting function is given by:

$$E[y_{t,h}^{TSE} \mid F_{t,h}, y_{t,h}^{Reg}, T_{t,h}, \Delta S_{t,h}, y_{t,h}^{RSE} > 0] = h(y_{t,h}^{RSE}) \quad 8$$

where $y_{t,h}^{RSE}$ represents reported self-employment income.

Nonparametric estimation of equations (5) and (8) is achieved by employing the locally-weighted least-squares procedure, using the Gaussian kernel, and adaptive bandwidth. Equation (5), the inverse Engel curve, is estimated at every point in the data but assigns a weight of zero to households with positive self-employment income in the estimation process.

5. Application

The nonparametric application of the *Expenditure-based method* is illustrated here by estimating the effect of the Canadian Goods and Services Tax (GST) on income under-reporting by self-employed households. The implementation of the GST in 1991 represents an interesting opportunity to explore changes in income under-reporting by the self-employed in Canada. The GST is a federal value-added tax that applies at a rate of 7% to the supply of most goods and services, including services offered by the self-employed, in Canada and replaced a less comprehensive manufacturers' sales tax (MST).[6] Most businesses, including the self-employed, are required to register for the GST and registrants are obligated to collect and remit the GST. As a GST registrant, businesses can reclaim the GST they have paid on business purchases. However, "small suppliers" are not *required* to register for and collect the GST.[7] The Canada Revenue Agency (CRA) defines a GST Small Supplier as a sole proprietor, partnership, or corporation whose total taxable revenues before expenses are $30,000 or less annually. Additionally, even if a business does qualify as a GST Small Supplier, the business can still register for the GST. Unfortunately, the data used in this analysis contains no information regarding the GST registrant status of the self-employed contained in the data sample.

---

[6] The data used for this analysis contains no information regarding the GST registrant status of the self-employed.

[7] Small suppliers that do not register for the GST are also unable to reclaim the GST they have aid on business purchases.

Prior to introducing the GST, the federal government argued that the GST would reduce the scope for tax evasion because it is applied successively at different stages of processing. That is, businesses, including the self-employed, are required to pay the GST on all its inputs but this is credited against the GST it collects from its own customers. In order to obtain the credit, however, the business is required to produce receipts showing that it paid the GST on its inputs. For this reason, the tax is said to apply only to the value added by a business. Another promoted virtue of the GST was that, as a consumption tax, it is a tax that even the hard-to-tax (e.g. those earning their full income in the underground economy) would have to pay since they must purchase at least some of their goods and services in the observed economy. On the other hand, it is often argued that the implementation of the GST increased the incentives and opportunities for tax evasion. First, the business can choose not to report some fraction of their sales, avoiding both their income and GST tax liability, while still claiming their whole input tax credit. Second, the business and customer can collude and avoid collecting and paying the GST, respectively.

5.1 Data

The data come from the public use Canadian Family Expenditure Surveys (FAMEX).[8] The FAMEX is a cross-sectional household expenditure survey that collects detailed information on household expenditures and characteristics, including income, savings, and tax payments. The FAMEX is unusual in that it does not have a

[8] In 1997, the Survey of Household spending (SHS) replaced the FAMEX and has been conducted annually since. The SHS, however, does not provide detailed information regarding the sources of household income so this data cannot be used for this analysis.

diary component, like many national expenditure surveys. Instead, it is a recall survey where information is gathered from face-to-face interviews and pertains to the most recently completed calendar year. A considerable amount of effort is made to ensure the quality of the data and respondents are asked to consult bills and receipts and, if necessary, multiple visits are made to a household. Unfortunately for the purposes of this study, records where expenditures exceed all sources of income by 20% or more are rejected and it is unknown how many records are rejected using this rule. As a result, it is reasonable to assume that the estimates obtained will be a lower bound estimates.

Total gross income in the FAMEX is separated into five income sources, four of which are assumed to be accurately reported. Income from wages and salaries, investments, government transfers, and miscellaneous income (including pension payments) are all assumed to be accurately reported since most of the components of these four sources of income are subject to third party reporting in Canada. Self-employment income is the component of income that is assumed to be subject to underreporting and includes income from farm and non-farm activities and is net of allowable business expenses. The advantage of using self-employment income defined in this way, rather than income before expenses, is that if the business expenses are legitimate than it accurately reflects the amount of income available to fund consumption. If, however, expenses are overstated than the effect will be the same as when an individual under-reports their income. While the method cannot discern between income that is underreporting and expenses that are overstated, the net effect is the same as both activities represent tax noncompliance. Food

expenditures are converted to real 1996 dollars using the food price index developed by Browning and Thomas (1999). Income components, income taxes, and the change in asset term[9] are converted to real 1996 dollars using a general price index.

Food expenditures used in this analysis were limited to food and other groceries (excluding nonfood items) purchased from stores, farmer stalls, and home delivery. There are three arguments for excluding restaurant expenditures from the analysis. First, in many countries, including Canada, the self-employed are able to write off for tax purposes food that is consumed in restaurants much more easily than wage employees and, as a consequence, self-employed households may have different a different expenditure function than wage employees. Second, as the measure of self-employment income in the FAMEX is net of expenses, including expenditures that are subsequently netted out of the income used in the first stage regressions will likely produce biased results. Third, most grocery items are zero rated under the GST while restaurant expenditures are taxable expenditures. Despite these concerns, similar results to those reported in the next section were achieved using total food expenditures.

For reasons discussed in section 4, the sample for this analysis is limited to married couples without children present in the household and it is assumed that the household unit acts as a single decision maker regarding expenditure and income

[9] Net change in assets and liabilities includes total net change in assets (including cash held in banks, money owed to households, money deposited against future purchases, net contributions less withdrawals to Registered Retirement Savings Plans, financial assets, sales of personal property, real estate, and investments in unincorporated business or farms) less net change in debts (including loans with regular payments and other money owed).

reporting.[10] The sample is further restricted to households: which constitute one economic family; where the head and spouse are of working age (25-64 years of age); where the head is employed and is not working in the primary occupation category[11]; and that have positive food expenditures. Finally, households with negative reported self-employment income were also excluded from the analysis.[12] To ensure that there are sufficient observations included in each stage of the analysis the FAMEX data was pooled. Results from using FAMEX data for the years 1982 and 1986 are compared to those obtained using data for the years 1992 and 1996. The implicit restriction made by pooling the data in this way is that the marginal propensity to consume food is the same for each of the two years contained in each of the pooled samples.

Six additional households with reported self-employment income above $120,000 in the pooled 1982/1986 sample were excluded from the analysis as well as eight households with reported self-employment income above $150,000 in the pooled 1992/1996 sample. These households did not have sufficient observations within their vicinity to obtain valid nonparametric estimates in the second stage of the estimation process. Pooling, along with the restrictions noted here, left a total of

---

[10] The unit of analysis, ideally, would be individuals, as it would avoid assuming households act as single decision makers and because in Canada, taxes are assessed on the individual rather than the household. In the FAMEX, however, expenditures are only surveyed at the household level and there are insufficient observations to conduct the analysis on single adult households.

[11] This last restriction will exclude farm households, which are likely to have much different expenditure patterns on food than those in other occupations

[12] Only a small number of households report negative self-employment income and the results are robust to their inclusion in the model.

1,976 households in the 1982 and 1986 pooled sample, of which 323 were self-employed and a total of 1904 households in the 1992 and 1996 pooled sample, of which 387 are self-employed. The increase in the ratio of self-employed households to non self-employed households between the two samples is not unexpected, given that the Canadian self-employment rate rose from 13% in 1979 to 18% by 1997 (Picot *et al.* 1998).

The sample includes households that are living in both rural and urban areas. Limiting the analysis to households living only in urban areas, resulted in insufficient observations. It is possible that households in urban and rural environments have different levels of food expenditures at similar income levels for reasons that are unassociated with income under-reporting. For example, households in rural environments may be more likely to: grow food for consumption in a household garden; face reduced food prices due to the presence of local producers and suppliers; and engage in the trade of goods and services for food products. To the extent that this is true, food expenditures for rural households with no self-employment income will act as a poor counterfactual for urban households with positive self-employment income and vice versa.

Table I provides some summary statistics of the data. The top half of the table presents statistics for households with zero self-employment income, while the bottom half of the table presents statistics for households with positive self-employment income. The left column shows statistics for the 1982/1986 pooled sample and the right column for 1992/1996. The two household groups report comparable average total incomes, changes in assets, and expenditures on food in

each of the two samples, but self-employed households have greater variability in their financial assets and income.

5.2 Results

Table II reports household population estimates of income under-reporting by married couples without children present in the household and that report positive self-employment income for 1982/1986, presented in the column on the left, and 1992/1996 in the column on the right. The total amount of income under-reporting is found by subtracting reported self-employment income from expected true self - employment income and summing up over households. The first row of table II shows the population estimates for total income under-reporting. Total income under-reporting more than doubled between the 1980's and the 1990's, amounting to just over $2.6 billion in the 1982/1986 pooled sample and increasing to approximately $5.8 billion in the 1992/1986 pooled sample. The associated 90% bootstrapped confidence intervals are noted in the parenthesis. the confidence intervals indicate that average income under-reporting in each sample is statistically significant. Further, the results are economically significant, indicating that income underreporting increased by over $3 billion after the implementation of the GST. That is, the results support the notion that the GST increased, rather than decreased, income under-reporting by married households with positive reported self-employment income.

As the number of self-employed households increased between the two pooled samples, as shown in the second row of table II, it could be that the increase in total

income under-reporting was simply due to the increase in self-employed households over the sample period, rather than due to the implementation of the GST. In order to determine if there was a change in the amount of income under-reporting per household, the average per household income under-reporting is calculated. Despite the fact that the number of self-employed households increased between these two pooled samples, there was an increase in the average amount of self-employment income that went unreported. Income under-reporting per married household, presented in the third row, amounted to $9,803 in the 1982/1986 pooled sample and $13,929 in the 1992/1996 pooled sample. The 90% bootstrapped confidence intervals for these per household amounts are presented in the final row of the table. Again, for both samples, the confidence intervals indicate that average income under-reporting is statistically significant in both samples. The results indicate the income under-reporting per household in the sample increased by $4,126.

Figure 2 presents graphs of the nonparametrically estimated reporting function that were obtained using equation (8). Again, the graph on the left is for the 1982/1986 pooled sample while the graph on the right is for 1992/1996. Estimated true self-employment income is plotted on the vertical axis and reported self-employment income is plotted on the horizontal axis. Also shown are 90% bootstrapped confidence intervals obtained using the "wild" bootstrap procedure (Wu 1986) which allows for heteroskedastic errors. The forty-five degree line in the figures shows reported self-employment income. When the plot of estimated true self-employment income is above the forty-five degree line, a household is under-reporting their self-employment income.

The graphs in Figure 2 show that the reporting function appears to be somewhat linear but rather than a ray through the origin, as implied by equation (1), it appears as though the reporting function includes a constant. It is possible, in this framework, to test whether or not the reporting function takes the form specified in equation (1), as assumed previously in the literature, or takes the form

$$y_{t,h}^{TSE} = \alpha + \theta y_{t,h}^{RSE} \tag{9}$$

A description of the testing method, which is well known, is found in Yatchew (1998). The two different null hypotheses are each tested against the nonparametric alternative. Table III summarizes the results of these tests. The results for the test of the null that the reporting function takes the form of equation (1) against the alternative that the reporting function takes the nonparametric specification of equation (8) are shown in the top half of the table. The results for the test of the null that the reporting function takes the form of equation (9) against the alternative that the reporting function takes the nonparametric specification of equation (8) are shown in the bottom half of the table. The value of the test statistic is noted along with the associated p-value. Unsurprisingly, the null hypothesis for the first test is rejected in both pooled samples at all of the usual significance levels while we fail to reject the null hypothesis at all the usual significant levels in both pooled samples for the second test. The figure and the test results call into question the assumption that self-employment income is underreported by a constant fraction, however, some caution should be exercised in interpreting the test results since this test statistic is known to suffer from severe size and power distortions. (Li and Wang 1988)

Returning to Figure 2, for the 1982/1986 pooled sample the results suggest that prior to the implementation of the GST, the gap between true and reported self-employment income is larger for households at the lower end of the reported self-employment income distribution. The results are contrary to that previously assumed but does provide some support for the fundamental finding of Reinganum and Wilde (1985, 1986) that "…taxpayers with greater true income under-report less than those with lower true income.…" (Reinganum and Wilde 1986, p. 741) Subsequent to the implementation of the GST, the reporting function adjusted such that the gap between true and reported self-employment income is similar throughout the self-employment income distribution. Overall, the aggregate results support the hypothesis that the GST increased tax noncompliance by those with larger amounts of self-employment income and unaffected tax noncompliance by those with small amount of self-employment income.

The results should not be interpreted as suggesting that those with low reported household income under-report to a greater extent since the analysis only considers self-employment income. Fewer than 5 percent of households in the 1982/1996 sample and fewer than 2 percent of households in the 1992/1996 sample report earning no other income than self-employment income and reported self-employment income is only weakly positively correlated with gross household income. This indicates that self-employment income is a poor measure of total household income. In addition, the self-employment income term is net of expenses and those households that report low self-employment income net of expenses may in fact be able to deduct a large amount of expenses. That is, the results include both

income under-reporting and expense over-stating behavior and it is impossible to disentangle these activities without more information. However, the change in reporting behavior after the implementation of the GST is noteworthy as it suggests that, *on average*, self-employment individual's under-report/overstate expenses by the same dollar amount, regardless of reported self-employment income net of expenses. It could be that, based on Canada Revenue Agency's auditing rules, that there is a certain amount of income that can be understated before arousing suspicions of the tax authority. In addition, once engaged in underreporting behavior, it is likely unwise to deviate much from this behavior since significant changes in reported income and/or expenses deductions are likely to increase the probability of an audit. This might explain why the unaffected tax noncompliance by those with small amount of self-employment income; they were already underreporting an optimal amount.

6. Conclusion

This paper proposes a nonparametric approach for estimating income under-reporting by households with self-employment income. The use of nonparametric methods is shown to have several advantages over previous parametric approaches. First, it allows the form of the reporting function to be uncovered rather than being imposed a priori. Second, it provided the ability to test, and find evidence against, the previously held hypothesis that the reporting function takes a specific linear form. Third, the framework allows addresses series concerns with the empirical strategies

followed in the previous literature. A further advantage of this method is the ease in which population estimates can be generated. In particular, the total amount of unreported income in the population could be obtained directly, whereas previous studies could only extrapolate this information by using national accounts data. Overall, the approach outlined in this paper calls into question many of the assumptions made in the parametric applications of the *Expenditure-based method.*

The approach outlined in this paper is illustrated by estimating the effect of the Canadian Goods and Services Tax on income under-reporting by married households with self-employment income. The results demonstrate that the assumption that the self-employed under-report their income by a constant fraction does not apply to the Canadian data considered in this paper. Rather, it is found that prior to the implementation of the GST the gap between true and reported self-employment income is larger for households at the lower end of the reported self-employment income distribution (net of expenses). Subsequent to the implementation of the GST, the results support the hypothesis that the GST increased tax noncompliance by those with larger amounts of self-employment income and unaffected tax noncompliance by those with small amounts of self-employment income.

Caution needs to be exercised in interpreting these specific results, as the reliability of the estimate depends on the quality of the data and on the various assumptions made. In addition, by using survey data, only those households that elected to take part in the survey can be studied. Households that are heavily involved in underground activity, particularly those households that are involved in

illegal activity (for example, drug trafficking, human smuggling, and prostitution), are unlikely to participate in the survey or may elect to modify their reported amount of expenditures to ensure they are not perceived to be living beyond their means. Further, caution must also be exercised in interpreting and comparing the results presented here to those obtained by alternate methods. In particular, the results presented here should not be interpreted as representing a measure of the total underground economy. Households with self-employment income but with different demographic characteristics (e.g. households with children, single person households etc.) may engage in income under-reporting at different rates than the households studied here. Additionally, income under-reporting by the self-employed, represents only a portion of underground activity. Finally, the method presented in this paper, estimates income that is not reported to tax authorities, which is quite distinct from measuring production or income that is missed by the statistical offices when they calculate the value of the national product. Many methods employed in estimating underground activity use the latter calculation.

There are several avenues for further exploration in terms of refining the *Expenditure-based method* to improve consistency with available data and knowledge concerning participation in the underground economy. In particular, redefining the base group is warranted, as it will include those who are engaging in income under-reporting behavior themselves. It is also worthwhile to consider the unstated assumptions about the possibility of self-selection into self-employment being unrelated to consumption tastes, an assumption made when estimating the inverse Engel curve for those who report zero self-employment income and applying the

identical function to those who report positive self-employment income. This is an important issue that has, to date, been ignored in the literature.

In addition, there are several shortcomings related to the use of expenditure data that are shared by comparable data sets for other countries to consider. The most important of these is that income reported in the FAMEX may not be the same as income reported to the tax authority by households because households are not required to produce any proof of income. Tax filer data, on the other hand, would have exact information regarding income and, hence, would provide more accurate estimates of income under-reporting. Unfortunately, tax filer data does not contain detailed information regarding expenditures. It does, however, contain information regarding expenditures on goods and services that are subject to tax credits and deductions and it may be possible to use this information and the method outlined in this paper to obtain more accurate estimates of income under-reporting. In addition, it is well known that measurement error is a universal feature of micro data and is likely an area of concern with the expenditure and income data used to implement the *Expenditure-based method*. Using mismeasured variables in parametric methods usually results in biased estimates and the problem becomes even more complex in nonlinear methods, such as those employed in this paper. Methods have been proposed to correct the data before analysis to eliminate these biases and these should be explored in future work.

Acknowledgements

I would like to thank Tom Crossley and Mike Veall and David Bjerk, all of McMaster University, David Giles, University of Victoria, and participants at seminars at McMaster University, University of Victoria, and the Canadian Economics Association (2004) for their helpful comments and invaluable guidance. I would like to thank the editor, Bernd Fitzenberger, associate editor, and two anonymous referees for their useful suggestions that resulted in a much improved version of the paper.

REFERENCES

Banks J, Blundell R, and Lewbel A (1997) Quadratic engel curves and consumer demand. Rev. Econ. Stat. 4: 527-538.

Blundell R, Duncan A, and Pendakur K (1998) Semiparametric estimation and consumer demand. J. Appl. Econ. 13: 435-461.

Browning M, Thomas I (1999) Prices for the FAMEX: methods and sources. Working Paper, Department of Economics, McMaster University.

Dilnot AW and Morris CN (1981) What do we know about the back economy? Fisc. Stud. 2: 58-73.

Hill R, Kabir M (1996) Tax rRates, the tax mix, and the growth of the underground economy in Canada: What Can We Infer. Can. Tax J. 44: 1552-1583.

Li Q, Wang S (1998) A simple consistent bootstrap test for a parametric regression function. J. Econ. 87: 145–165.

Lyssiotou P, Pashardes P, Stengos T (2004) Estimates of the black economy based on consumer demand approaches. Econ. J. 114: 622-639.

Picot G, Manswer M, Lin Z (1998) The role of self-employment in job creation in Canada and the United States. Can. Econ. Obs. 12: 3-17.

Pissarides CA, Weber G (1989) An expenditure-based estimate of britain's black economy. J. Pub. Econ. 39: 17-32.

Reinganum JF, Louis LW (1985) Income tax compliance in a principal-agent framework. J. Pub. Econ. 26: 1-18.

_____ (1986) Equilibrium verification and reporting policies in models of tax compliance. Int. Econ. Rev. 27: 739-760.

Smith S, Pissarides CA, Weber G (1986) evidence from survey discrepancies. In: Smith S (ed) Britain's shadow economy, Clarendon Press, Oxford, pp. 137-153.

Spiro PS (1993) Evidence of a post-GST increase in the underground economy. Can. Tax J. 41: 247-258.

Thomas J (1999) Quantifying the black economy: measurement without theory yet again. Econ. J. 109: F381-F337.

Wu CFJ (1986) Jackknife, bootstrap, and other resampling methods in regression analysis. Ann. of Stat., 14: 1261-1350.

Yatchew A (1998) Nonparametric regression techniques in economics. J. Econ. Lit. 36: 669-721.

TABLES

Table I: Data Summary[a]

	1982 & 1986 Pooled FAMEX				1992 & 1996 Pooled FAMEX			
	Households with positive wage income and zero self-employment income							
	Unweighted Sample Size=1,653 Weighted Sample Size=1,272,127				**Unweighted Sample Size=1,517 Weighted Sample Size=1,685,505**			
	Mean	Std dev	Min	Max	Mean	Std dev	Min	Max
Gross Regular Income	61,487	27,928	6,296	269,133	66,204	32,543	9,700	283,000
Gross Self-Employment Income (less Expenses)	0	0	0	0	0	0	0	0
Total Taxes Paid	12,975	9,561	-12,080	86,001	15,994	12,152	-3,232	97,375
Change in Financial Position	6,182	13,5841	-89,709	99,692	6,558	16,486	-124,008	156,621
Food Expenditures	4,381	1,670	227	14,773	4,100	1,758	379	17,600
	Households with positive self-employment income							
	Unweighted Sample Size=323 Weighted Sample Size=266,437				**Unweighted Sample Size=387 Weighted Sample Size=416,133**			
	Mean	Std dev	Min	Max	Mean	Std dev	Min	Max
Gross Regular Income	38,070	33,006	0	185,230	39,888	32,866	0	243,200
Gross Self-Employment Income (less Expenses)	19,324	19,957	56	95,962	22,377	25,119	106	126,356
Total Taxes Paid	9,789	9,036	-2,090	49,154	13,263	13,469	-89,00	95,000

Change in Financial Position	5,686	16,548	-59,796	80,641	5,7180	18,404	-98,402	180,335
Food Expenditures	4,285	1,876	621	13,083	4,161	1,992	556	15,758

Notes: [a]Amounts are calculated for married households without dependent children present in the household, are in real (1996) Canadian dollars and are rounded to the nearest dollar, and obtained using the survey weights provided in the FAMEX by Statistics Canada to obtain population amounts.

Table II: Estimates of Income Under-Reporting[a]

	1982 & 1986 Pooled FAMEX	1992 & 1996 Pooled FAMEX
Total amount of Income Under-Reporting (90% Bootstrapped Confidence Interval)	$2.612 Billion ($1.936B; $3.335)	$5.796 Billion ($4.544B; $6.613B)
Population Size	266,437	416,133
Average Income Under-Reporting (90% Bootstrapped Confidence Interval)	$9,803 ($7,414; $12,193)	$13,929 ($11,655 $15,642)

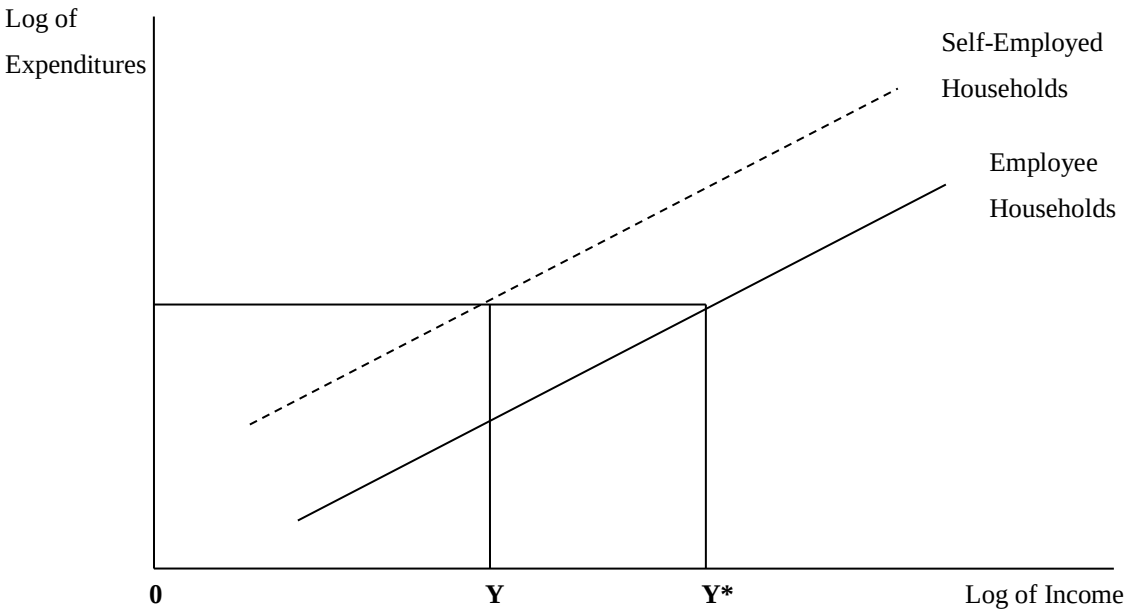
Notes [a] Amounts are calculated for married households without dependent children present in the household, are in real (1996) Canadian dollars, and obtained using the survey weights provided in the FAMEX by Statistics Canada to obtain population amounts.

Table III: Testing Linearity of the Reporting Function

	Test: H₀: $y_{t,h}^{TSE} = \theta y_{t,h}^{RSE}$ **vs. Hₐ:** $y_{t,h}^{TSE} = h(y_{t,h}^{RSE})\, H_a : y_{SE,h}^{*} = f(y_{SE,h})$	
	1982 & 1986 Pooled FAMEX	**1992 & 1996 Pooled FAMEX**
Test Statistic	3.590	6.680
p-value	0.000	0.000
	Test: H₀: $y_{t,h}^{TSE} = \alpha + \theta y_{t,h}^{RSE}$ **vs. Hₐ:** $y_{t,h}^{TSE} = h(y_{t,h}^{RSE})$	
	1982 & 1986 Pooled FAMEX	**1992 & 1996 Pooled FAMEX**
Test Statistic	-0.817	3.683
p-value	0.207	0.247

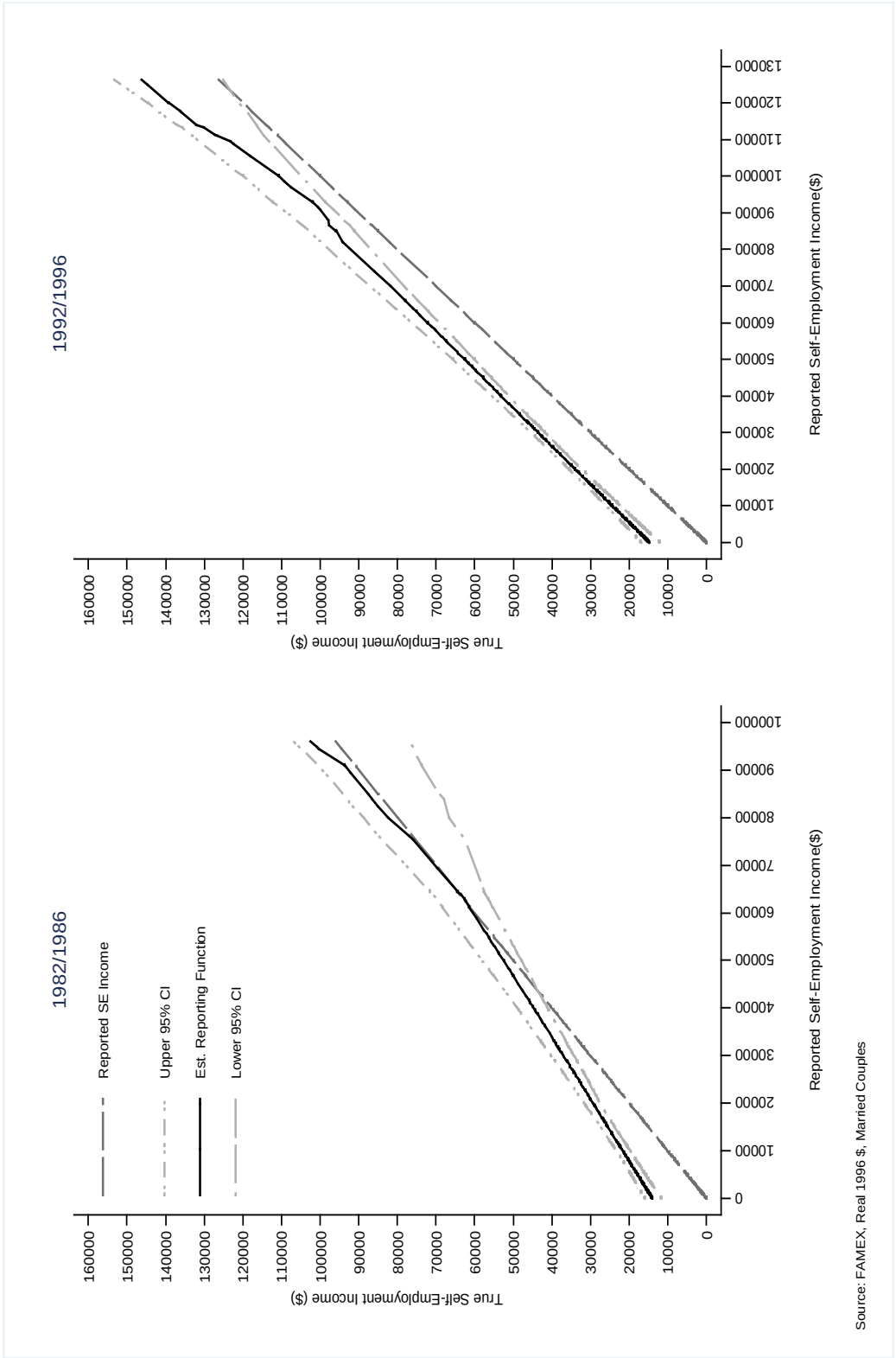
FIGURES

Fig. 1 Income Under-reporting in the Single Equation Expenditure-based method



Created in Word

Fig. 2 Estimated Reporting Functions



Source: FAMEX, Real 1996 $, Married Couples

Created in Stata