An Empirical Analysis of Factors That Distinguish Those Who Evade on Their Tax Return from Those Who Choose not to File a Return

Steven E. Crane
*Marquette University, steven.crane@marquette.edu*

Farrokh Nourzad
*Marquette University, farrokh.nourzad@marquette.edu*

An Empirical Analysis of Factors That Distinguish Those Who Evade on Their Tax Return from Those Who Choose not to File a Return

Steven E. Crane* and Farrokh Nourzad**

Abstract

This paper presents an empirical model distinguishing evaders who cheat by filing fraudulent income tax returns from those who do not file. Using a maximum-likelihood procedure that corrects for sample selection bias, and data from Michigan's amnesty program, we estimate a linear probability model which relates the probability of filing to various economic and demographic characteristics. The results indicate that higher true income and automatic withholding raise the probability of filing, while males and single individuals are less likely to file. The evidence regarding a grouping of occupations often thought to be associated with evasion is inconclusive.

I. Introduction

Income tax evasion is a widely recognized and growing problem that costs governments throughout the world a great deal of revenue. In the United States, the Internal Revenue Service (IRS) has estimated that the annual shortfall in income tax revenue due to evasion exceeds $100 billion. A portion of this is by individuals who file fraudulent returns that underreport their true tax obligation. Another portion of the tax gap is attributable to the estimated 9.8 million who do not file tax returns at all.

In recent years we have learned a great deal from both theoretical and empirical analyses of a taxpayer's decision to underreport his or her income (Cowell, 1990). In contrast, we know very little about nonfilers. This is partly because data availability

* Associate Professor of Economics, Department of Economics, Marquette University, Milwaukee, Wisconsin, USA.
** Associate Professor of Economics, Department of Economics, Marquette University, Milwaukee, Wisconsin, USA.

problems are more formidable; little information concerning nonfilers can be produced because no records exist. In addition, there is little theoretical guidance concerning this group. In conventional microtheoretic evasion models nonfilers represent corner solutions that are explicitly ruled out so that comparative statics analysis can be performed on interior solutions that reflect underreporting.

While formal models of nonfilers have only recently begun to appear in the literature (e.g., Yaniv, 1988), some progress can still be made at the empirical level. The growing popularity of income tax amnesty programs offers an opportunity to address the data availability problem. The experience in the U.S. indicates that amnesty participants include not only income underreporters but also nonfilers, many of whom were not otherwise known to the tax authorities. This means that, in principle, one can use data generated from amnesty programs to study tax evasion by underreporters and nonfilers.

In practice, however, one is faced with certain limitations and complications. Because many amnesty programs are hastily put together with little advanced planning for subsequent research, the data are often limited in terms of both quality and quantity. Moreover, special econometric procedures are needed to deal with the complications caused by the self-selected nature of amnesty data.

Elsewhere, we have studied the problem of income tax evasion at the individual level using data from tax amnesty programs offered by California and Michigan (Crane and Nourzad, 1990, 1992). Both studies focused on testing certain comparative statics propositions derived from microtheoretic models of income tax evasion. Because of this, we concentrated exclusively on participants who were amending their original returns under amnesty. In other words, we discarded the observations on individuals who had not filed an original return but came forth under amnesty and voluntarily provided information about their status for the year in question. Since nonfilers made up the majority of the samples, our previous work did not fully exploit the information available from these amnesty programs.

In the present paper we use a sample of both types of participants in the Michigan tax amnesty program to identify certain economic and demographic factors that distinguish nonfilers from underreporters. The results provide information regarding how these characteristics affect the probability that an evader chooses to be a nonfiler rather than an income underreporter. This information may be useful for developing a profile of those who opt to remain outside the tax system, and for designing policies to induce these individuals to enter the system.

Of course, this information still represents only an additional piece of the overall tax evasion puzzle. A truly complete analysis would require an extremely rich data set, one that would include not only information about the cheaters examined in this paper, but also completely honest taxpayers. With such data, one could examine the sequential process from the initial decision to cheat or not, through the decision regarding the form of cheating (i.e., not filing vs. filing a fraudulent return). Unfortunately, a data set of this
richness has never been assembled, and it is unlikely that it ever will be, since it would require a highly improbable level of cooperation (and honesty) from all groups.  

In the absence of this ideal data set, researchers must be content to focus on one or more pieces of the puzzle. Most research has focused on samples such as those from the IRS Tax Compliance Measurement Program, which allow one to distinguish between "honest" and dishonest filers, but contain no information on the nonfiling population. Our work uses a sample of tax amnesty data on dishonest individuals that allows us to examine the nature of the dishonesty. By combining the information that emerges from these separate approaches, analysts should be able to move closer to filling in the complete tax compliance picture.

The organization of the paper is as follows: In the next section we describe some econometric issues concerning the use of amnesty data for analyzing tax evasion. In section III, we describe our empirical model. In section IV, we present and discuss our estimation results. The paper ends with a section summarizing this work and discussing some policy implications.

II. The Econometrics of Analyzing Tax Amnesty Data

Analyzing evasion using amnesty data calls for special econometric procedures to deal with the potential bias caused by the self-selected nature of the sample. This requires modeling not only the initial decision to evade, but also the subsequent decision to participate in the amnesty program.

The Filing Decision

The evader’s decision regarding the form his or her evasion will take can be modeled as follows. Let $Y_i = 1$ if the $i$th individual evades by filing a tax return that underreports his or her true taxable income and $Y_i = 0$ if he or she is a nonfiler. Denote by $X_i$ the vector of the factors that affect the individual’s decision to file or not and let $\beta$ be a vector of unknown parameters. Suppose the probability that $Y_i = 1$ is $F(X_i, \beta)$, in which case $\text{Prob} (Y_i = 0) = 1 - F(X_i, \beta)$. For estimation purposes, the main question is how to specify $F(X_i, \beta)$. If we assume $F(X_i, \beta) = X_i \beta$, we have the linear probability model (LPM),

$$Y_i = X_i \beta + u_i$$  \hspace{1cm} (1)$$

where $u_i$ is a random error term with mean 0 and variance $\sigma^2$. On the other hand, if we let $F(X_i, \beta)$ be the cumulative standard normal (logistic) distribution function, we have the probit (logit) model.
Our task would be quite easy if we had a truly random sample. Most standard econometrics programs have built-in routines for estimating these discrete choice models. A sample of amnesty participants, however, is not entirely random because amnesty filers themselves decided to be in the program. As a result, the parameter estimates from any of the above models are likely to suffer from bias unless the participation decision is incorporated into the model.

The Participation Decision

In many applications, correcting for self-selection bias can be accomplished by using the two-step procedure developed by Heckman (1979). In this approach the participation decision equation is expressed as,

\[ I_i^* = Z_i \gamma - \varepsilon_i \]

where \( I_i^* \) is a latent variable which represents the propensity to participate in the program, \( Z_i \) is the vector of factors affecting the participation decision, \( \gamma \) is a vector of unknown parameters, and \( \varepsilon_i \) is a random error term with mean zero and variance 1. Since \( I_i^* \) is unobservable, the standard procedure is to define \( I_i = 1 \) if and only if \( I_i^* > 0 \) and \( I_i = 0 \) otherwise, and to use it as the left-hand-side variable in equation (2). In the first step of the Heckman procedure, equation (2) is estimated by probit. Using the resulting estimated parameters, a correction factor (the Inverse Mills Ratio) is constructed and used in the second step as an additional regressor in the base model, equation (1).

The type of sample selection that is associated with amnesty data is a special case that cannot be treated by this method. The problem is that amnesty samples are truncated in that we only observe those who participate in the programs and have no information on those who do not; that is, we observe \( Y_i, X_i, \) and \( Z_i \) only if \( I_i = 1 \). We cannot use the Heckman two-step procedure since in the first step the parameter vector of the participation function, \( \gamma \), cannot be estimated. An alternative sample-selection correction procedure for incorporating the participation decision is needed for amnesty samples.

Correcting for Sample-Selection Bias in Amnesty Data

Bloom and Killingsworth (1985), Maddala (1983), and Muthén and Joreskög (1983) have developed procedures that can be used for correcting self-selection bias in truncated samples. In contrast to Heckman's two-step approach, theirs is a one-step estimation procedure which involves maximizing a likelihood function that integrates information about the base decision of interest and the related decision to participate.
This procedure, which assumes the base decision is modeled as a conventional linear regression, can be used here provided the filing decision is specified as the LPM shown in equation (1). On the other hand, if the base decision is specified as a probit (logit) model, then the issue is more complicated.

In this paper we report results using a linear probability model for the filing decision while correcting for self-selection bias using the one-step approach mentioned above. This involves maximizing the likelihood function,

\[ \Pi_t [\Phi (Z_t \gamma)]^{-1} (1/\sigma) \exp \left[ - \left( \frac{1}{2} \sigma^2 \right) (Y_t - X_t \beta)^2 \right] \]

where \( \Phi(\cdot) \) is the cumulative distribution function and \( \rho \) is the correlation coefficient between \( u_i \) and \( e_i \), the error terms in the filing decision equation (1) and the participation equation (2), respectively. All other notations are as defined previously.

Maximizing (3) addresses the self-selection issue as follows. The fact that nonparticipants are not represented in the sample means that a portion of the overall distribution is missing. As a result, standard estimation procedures produce coefficients whose means are not centered on the true parameter value. Further, the missing portion of the distribution means that the area under the density function does not sum to one, thereby invalidating the usual hypothesis tests. Equation (3) "corrects" for these problems by incorporating information about the decision to participate, and using it to re-scale the distribution for participants. This re-scaling re-centers the estimators, and assures that the area under the distribution curve sums to one.

In addition to correcting for the sample-selection bias, this maximum-likelihood procedure has the advantage that the resulting inverse Hessian matrix yields correct asymptotic standard errors. Thus, we avoid the problem caused by heteroscedasticity when the model implied by (3) is estimated by nonlinear least squares.

III. The Empirical Model

In this section we specify an empirical model for analyzing the effect of various socio-economic factors on an evader's choice between filing a fraudulent tax return and not filing. We begin by describing the data used in the analysis. Next, we present our empirical counterpart of the filing decision equation (1) followed by that of the participation decision equation (2).

The Sample

The data used in this study are taken from the tax amnesty database constructed by the Michigan Treasury Department. The original data set contained information taken
from 4,203 returns, 2,985 of which pertained to individual income taxes. Of these, 588 were filed by individuals who were amending a return and 2,397 were by individuals who had not filed previously for the year in question.

The sample used in the present study is a subset of the individual income tax amnesty returns. It consists of 1,748 returns filed under amnesty. Of these, 213 amended a return and 1,535 were "new" returns by individuals who had not filed previously for the year in question. This subsample is the result of an extensive data evaluation/verification process in which data were checked for internal consistency, missing observations, and the like.

Factors Affecting the Filing Decision

Given the limited theoretical guidance concerning the file/no-file decision, our specification of the filing decision is based largely on intuition and data availability. We use a number of variables as potential discriminants between underreporters and nonfilers. One such variable is the individual's true income. For this we use adjusted gross income as reported on the amended return.\(^4\) We also control for differences in the opportunity to evade. First, we account for the impact of automatic tax withholding. Second, we identify individuals with certain occupations that are suspected to have higher opportunities to evade. We also distinguish between the two groups of evaders based on two demographic characteristics, gender and marital status.

We specify our empirical version of equation (1) as follows,

\[
Y_i = \beta_0 + \beta_1 AGI_i + \beta_2 WITHHELD_i + \beta_3 OCCUPATN_i + \\
\beta_4 MALE_i + \beta_5 MARRIED_i + u_i
\]

(4)

where \(Y_i\) is a dummy variable that equals 1 if the individual is an income underreporter and 0 if he or she is a nonfiler; \(AGI\) is the individual's true adjusted gross income; \(WITHHELD\) is a dummy variable that is 1 if any income had been withheld and 0 otherwise; \(OCCUPATN\) is a dummy variable which equals 1 if the individual's occupation is one or more of the following: self-employed, sales, farming, foods and beverages, construction, and personal services; \(MALE\) is a dummy variable that takes the value 1 if the individual is male and 0 otherwise; and \(MARRIED\) is a dummy variable identifying married individuals.

Factors Affecting the Participation Decision

In specifying our empirical version of the participation decision equation (2), we draw on the recent work by Alm and Beck (1991) and by Fisher, Goddeeris, and Young (1989). They emphasize the importance of perceptions about increases in penalty and
the detection probability that occur after amnesty. We postulate that the perceived increase in the detection probability post-amnesty is related to expectations regarding the subsequent enforcement regime.

We control for this factor in three ways. First, we identify certain types of income that are more likely to attract attention. Second, we make use of the fact that the Michigan Department of Treasury planned to target the returns of professionals who were licensed by the state. Third, we identify the returns of evaders who might have feared they had been uncovered either through IRS audits or the Michigan Treasury Department document matching programs.

We complete our specification of the participation function by controlling for a nonpecuniary influence (Fisher, Goddeeris and Young, 1989). We identify the amnesty returns that were accompanied by a special letter of explanation from the individual concerning his or her filing for amnesty. Given that the amnesty provisions required no such explanation, this letter might indicate that the individual felt some degree of guilt or remorse.

We thus specify the following participation decision equation,

\[ \hat{r}_i = \gamma_0 + \gamma_1 RENTROYL_i + \gamma_2 LICENSED_i + \gamma_3 IRSAUDIT_i + \]

\[ + \gamma_4 LETTER_i + \gamma_5 GUILT_i - \epsilon_i \]

where \( \hat{r}_i \) is the latent variable representing the propensity to participate in the amnesty program; \( RENTROYL \) is a dummy variable that equals 1 if the individual's income includes rents, royalties, or business income; \( LICENSED \) is a dummy variable that indicates whether the individual is licensed by the state for practice in one or more of the following fields: architecture, medicine/health, law, personal services, and transportation; \( IRSAUDIT \) is a dummy variable that is equal to 1 if the participant was under audit by the IRS prior to or during the amnesty program; \( LETTER \) is a dummy variable that takes the value 1 if the individual was sent a letter by the Michigan Department of Treasury inquiring why he or she had filed a federal return but had failed to file a state return; and \( GUILT \) is a dummy variable indicating returns filed under amnesty that were accompanied by a letter of explanation.

Recall that the filing decision equation (4) and the participation equation (5) are not estimated separately; they are integrated into the likelihood function (3) which is then maximized. This yields unbiased estimates for \( \beta \)s in equation (4), but the estimates of \( \gamma \)s in equation (5) are unreliable (Maddala, 1983). In light of this, we focus exclusively on the parameters of the filing equation.

While the sign of the income coefficient ultimately depends upon risk attitudes, intuition suggests one might expect a positive coefficient. As one's income rises, it becomes increasingly difficult to avoid leaving trails for the tax authorities to follow. Moreover, opportunities to engage in legal tax avoidance increase with income. Therefore, as income rises those who decide to evade are more likely to do so on the return rather than by not filing. Along the same lines, we expect the withholding variable to
have a positive coefficient. This captures the fact that the individual was known to the authorities as having earned income that might be taxable.

Turning to the occupation variable, it is difficult to form a clear-cut sign expectation. Although it is generally agreed that people in these occupations have more opportunities to evade, there is no consensus as to what form their evasion takes. Some of these occupations may provide the opportunity not to file at all, whereas others offer the individual various ways to misstate different items on his or her return. This uncertainty is compounded by the fact that the model already contains the withholding variable which partially controls for the opportunity not to file. Thus OCCUPATN captures only the effect of the incremental opportunity not to file, above and beyond that available in the absence of automatic withholding.

As for the marital status variable, we expect it to have a positive coefficient because it is more likely that two people leave traceable impact in the economy. Finally, we have no sign expectation on the gender variable.

### IV. Estimation Results

The maximum-likelihood (ML) results are reported in Table 1. For comparison purposes, we have also included results from estimating the filing decision model by ordinary least squares (OLS) which does not correct for sample selection.

<table>
<thead>
<tr>
<th></th>
<th>ML Estimates*</th>
<th>OLS Estimates**</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coeff</td>
<td>t-Rat</td>
</tr>
<tr>
<td>CONSTANT1</td>
<td>0.843</td>
<td>2.526</td>
</tr>
<tr>
<td>AGI</td>
<td>0.013</td>
<td>1.726</td>
</tr>
<tr>
<td>WITHHELD</td>
<td>0.116</td>
<td>5.495</td>
</tr>
<tr>
<td>OCCUPATN</td>
<td>-0.021</td>
<td>-1.014</td>
</tr>
<tr>
<td>MALE</td>
<td>-0.055</td>
<td>-2.832</td>
</tr>
<tr>
<td>MARRIED</td>
<td>0.059</td>
<td>3.625</td>
</tr>
<tr>
<td>CONSTANT2</td>
<td>-4.426</td>
<td>-1.528</td>
</tr>
<tr>
<td>RENTROYL</td>
<td>5.273</td>
<td>0.438</td>
</tr>
<tr>
<td>LICENSED</td>
<td>-0.003</td>
<td>-0.028</td>
</tr>
<tr>
<td>IRSAUDIT</td>
<td>5.043</td>
<td>0.852</td>
</tr>
<tr>
<td>LETTER</td>
<td>-0.424</td>
<td>-1.084</td>
</tr>
<tr>
<td>GUILT</td>
<td>1.235</td>
<td>1.276</td>
</tr>
<tr>
<td>(\sigma)</td>
<td>0.317</td>
<td>4.116</td>
</tr>
<tr>
<td>(\rho)</td>
<td>-0.566</td>
<td>-1.807</td>
</tr>
</tbody>
</table>

\* With sample-selection correction.
Consider first the ML results. The estimated coefficient of the income variable is positive and significant at the 8% level. This indicates that evaders are more likely to file as true income increases, which suggests that nonfilers are more likely to be low income individuals. The estimated coefficient of the withholding dummy variable is positive and highly significant, implying that the probability that an evader is a nonfiler decreases if the individual is subject to withholding.

Turning to the occupation variable, we observe that its estimated coefficient is negative, although not statistically significant at conventional levels. It may be that the occupational grouping used to construct this variable is too broad. Or it may be that the opposing effects mentioned earlier cancel each other out. On the other hand, after controlling for the effect of withholding, there may be little incremental opportunity not to file associated with these occupations. In fact, when the model was re-estimated without WITHHELD, the estimated coefficient of OCCUPATN remained negative but became highly statistically significant.

Finally, consider the two demographic variables. Both have statistically significant effects on the probability of filing. Males are more likely to be nonfilers, whereas married individuals are more likely to evade by filing fraudulent returns.

We conclude our discussion of the ML results by noting that the estimated standard error of the filing equation, σ, and the estimated correlation coefficient between the error terms, ρ, are both statistically significant at reasonable levels. The latter result is of particular interest because it suggests that our attempt to model the participation decision has met with some success. We can get a sense of the impact of the selection bias on the estimated parameters of the filing decision equation by comparing the ML estimates with the OLS results.

This comparison indicates that the two sets of results are similar in terms of signs and significance. As far as magnitudes of these estimates are concerned, the most notable difference is the coefficient of income. The ML estimate of this parameter is considerably smaller than its OLS counterpart, both in absolute terms and relative to the other statistically significant parameters. This may be important for developing an accurate profile of nonfilers.

V. Concluding Remarks

In this paper we demonstrated how amnesty data can be used to study tax evasion. Using data from the amnesty program by the state of Michigan, we specified and estimated a model that examined the effect of economic and demographic factors on probability of filing a fraudulent income tax return versus not filing at all. In doing so, we employed a maximum likelihood procedure that corrects for the sample selection bias in the data.

We find that there is a positive correlation between the level of income and the probability of filing. An implication of this is that, despite the large number of nonfilers
in the population at large, the potential return to compliance efforts directed toward these individuals may be limited. However, if for reasons other than revenue generation it is desirable to bring nonfilers into the tax system, then automatic withholding appears to be an effective enforcement tool. This is suggested by the very strong positive correlation that we find between withholding and the probability of filing.

In contrast, we are unable to establish a link between nonfiling and a grouping of occupations often thought to be associated with evasion. Thus targeting these occupations as a way of reducing nonfiling behavior may not be an effective compliance policy. Perhaps being subject to automatic tax withholding constrains the opportunity to be a nonfiler which typically provides the rationale for targeting these occupations. In fact, gender or marital status appears to be a better discriminant of nonfiling than occupation since we find that males and single individuals are more likely to be nonfilers.

The analysis reported in this paper can be extended and improved in several ways. First, given the weak performance of the occupation variable, it would be wise to re-examine this issue using more detailed occupational groupings. Second, in order to gain insight into possible differences between the behavior of repeated and single-year evaders, control should be made for the returns of multiple-year amnesty filers. A third, yet related, issue is that there are two types of nonfilers: those for whom there is some previous record and those for whom no such record exists. Distinguishing between these two groups may be relevant for both the base equation and the participation equation. Fourth, we have paid no attention to characteristics of the returns themselves. It may be that the degree of complexity of the individual’s return has some impact on the file/nonfile decision. Finally, experimentation with alternative econometric approaches to estimating the base equation should be undertaken. Estimation using probit, logit, or perhaps even a tobit specification should provide an indication of the robustness of the estimates.

Notes

1. Data pertaining to income underreporters can be generated from a number of sources including special audit programs such as the IRS Tax Compliance Measurement Program, TCMP.
2. Without complete honesty, it would be very difficult to distinguish honest filers from dishonest ones. At best, "honest" would be defined as those who successfully withstood a comprehensive audit, which still misses the most sophisticated evaders. Without full honesty, it would be very difficult to identify evaders who are nonfilers. At best, some of the nonfilers could be identified through a massive effort to follow any "traces" these individuals might have left in the tax system (e.g., a previous return, withholding, etc.), while missing the nonfilers who had left no traces to follow.
3. The program, which lasted from May 12 through June 30 of 1986, covered all forms of state taxes. For more on the Michigan amnesty program see Bowman and Martin (1987, 1988) and Fisher, Goddeeris, and Young (1989).
4. In using amnesty data, our assumption is that the return filed under amnesty is filled out truthfully.
5. Prior to estimation the income variable was standardized in order to improve the convergence properties of the model.
In fact, in our sample the mean value of AGI for the nonfilers is $20,842, whereas that of the underreporters is $52,931.

We also estimated the filing decision equation by probit and obtained results which were virtually identical qualitatively to the OLS results.

References


