DEVELOPING AN ECONOMETRIC MODEL FOR MEASURING TAX NONCOMPLIANCE USING OPERATIONAL AUDIT DATA

Brian Erard, B. Erard and Associates, and Chih-Chin Ho, Internal Revenue Service
Presented at the 2002 American Statistical Association

Traditionally, the Internal Revenue Service (IRS) has relied on the Taxpayer Compliance Measurement Program (TCMP) as its primary source of data for estimating the tax gap--the difference between Federal income taxes owed and Federal income taxes voluntarily reported. Ignoring detection issues, the estimation of underreported taxes from TCMP data is straightforward. In particular, since the data represent a random sample from the overall return population, we can simply apply the sample weights to the detected levels of noncompliance associated with each return and aggregate.

The key advantages of operational audit data are the number of cases and the frequency of data collection. Whereas operational examinations proceed on essentially a continuous basis and involve over 700,000 returns per year, such special studies as the TCMP are undertaken only periodically and involve only about 50 thousand returns.

The main disadvantage of employing operational audit data for tax gap estimation is that returns targeted for operational examinations are not randomly selected. Rather, they are typically chosen specifically because the IRS believes they are likely to contain substantial errors. As a result, we cannot simply follow the TCMP methodology of directly projecting from the audited sample of returns the amount of noncompliance in the general return population.

As an illustration of how operational audit data can be employed to estimate the individual income tax gap, we develop a model for estimating an important element of the gap--improper claims for the Earned Income Tax Credit (EITC). Our econometric model is motivated by the specification developed by Erard and Feinstein (2001) for evaluating the level of non-compliance associated with understated self-employment income from operational audit data. Other related research efforts to measure noncompliance using operational audit data are listed among the references. We apply our model to data from one IRS district for a selected tax year.

Econometric Specification

For returns that have been audited, the aggregate level of noncompliance may be computed directly by summing the relevant adjustments on each return. However, to estimate the magnitude of under-reporting on returns that have not been audited, it is necessary to predict the magnitude of the adjustments that would have been made if the returns had been audited.

There are two reasons why the extent of noncompliance on returns subject to operational audits will tend to differ from the extent of noncompliance on non-examined returns. First, the two groups are likely to have important differences in their recorded return characteristics. In particular, the former group can be expected to contain a disproportionate share of returns with characteristics known or believed to be associated with substantial levels of noncomp liance.

Second, when deciding whether to proceed with an audit, IRS classifiers may employ information beyond what has been recorded from the return. For example, they may examine the taxpayer's prior audit and reporting history, or they may examine supporting information that has been attached to the return. As a result, examined and non-examined returns may differ in terms of unrecorded characteristics that are associated with non-compliance.

To control for these differences, both in recorded and unrecorded return characteristics, an econometric specification for the likelihood that a return will be audited is estimated jointly with a specification for noncompliance. A probit equation is used to describe the likelihood of an audit, while a tobit specification is used to describe the magnitude of noncompliance. The full model is as follows.

The first equation is a (reduced form) probit specification of the decision whether to audit a given return:

$$A^* = b_A' X_A + e_A \qquad (1)$$

The term A^* represents an index of the likelihood that a return with observed characteristics X_A will be audited. The term e_A represents a standard normal random disturbance, and b_A is a vector of coefficients

to be estimated. From the data, we can deduce whether A^* is greater than zero (indicated by whether an audit has been performed).

The second equation is a tobit specification for the magnitude of the EITC overclaim:

$$N^* = b_N' X_N + g_N \Phi\left(b_A' X_A\right) + e_N \qquad (2)$$

The term N^* represents an index of the propensity of a taxpayer with observed characteristics X_N to overstate his or her EITC claim. The term $\Phi\left(b_A' X_A\right)$ represents the probability of audit computed from Equation (1); the symbol $\Phi(z)$ refers to the value of the standard normal cumulative density function evaluated at z. The probability of audit is included as a regressor to account for the possibility that taxpayers who are at a higher risk of audit are relatively less likely to overstate their claims.

For returns that have been audited, we observe N--the actual amount of the overclaim. In particular, we observe an overclaim of the amount $N=N^*$ if N^* is greater than zero. Otherwise, we observe $N=0$, signifying no overstatement. The term e_N represents a standard normal random disturbance, while the parameter $?_N$ and the vector b_N represent coefficients to be estimated.

Although there is an upper bound on the amount by which the EITC can be overclaimed (the maximum permissible claim amount), very few overclaims achieve this bound, and we ignore it in our application. In addition, there are a small number of cases in our audit sample for which the amount of the credit was increased as a result of the audit. We set our measure of noncompliance (N) to 0 for these cases.

A key feature of our methodology is to allow the error term of the audit selection specification (e_A) in Equation (1) to be correlated with the error term of EITC overclaim specification (e_N) in Equation (2). This accounts for the possibility that unobserved factors which influence whether a return is audited may also be associated with the magnitude of the overclaim. By estimating this correlation term (r_{AN}), we are able to test explicitly the hypothesis of selection bias. We can correct for such bias if it is found to be pr esent by incorp orating the correlation term into our expression for predicting the magnitude of EITC overstatement on returns not subjected to audit.

To identify the model in a non-parametric sense, the vector X_A in Equation (1) must include at least one exogenous regressor that is not contained in vector X_N from Equation (2). Otherwise, the parameters of the noncompliance equation [Equation (2)] would only be identified on the basis of the assumed functional form of our specification.

In our analysis, we include as a regressor in X_A a measure of the district audit coverage rate, which varies according to the examination class to which a return has been assigned. The district audit coverage rate is exogenous because it cannot be influenced by the amount reported on a single return.

This coverage rate serves as the starting point for assigning an audit probability to a return. Factors beyond the return's examination class that influence the risk of audit are accounted for in the remaining regressors in X_A.

We exclude the district audit coverage rate from Equation (2), because we believe that the propensity of a taxpayer to overclaim the EITC should depend on his or her individual-specific audit risk rather than simply the district average for the taxpayer's examination class.

Likelihood Function

We estimate our model using the method of maximum likelihood. The observations in our data can be constructively divided into three categories, according to whether an audit took place and the outcome of the audit. We specify the likelihood expressions associated with each case below.

Case 1: No Audit

The first category contains those returns that were not subjected to an audit. For returns in this category, only the audit equation applies, and the likelihood expression (L_1) simply represents the probability that the return would not be audited:

$$L_1 = 1 - \Phi\left(b_A' X_A\right),$$

where $\Phi(z)$ represents the standard normal cumulative distribution function evaluated at z.

Case 2: Audit, No EITC Overclaim

The second category contains audited returns that were found to have reported the EITC properly. For a return in this category, the likelihood expression

(L_2) represents the joint probability of the return being audited and no adjustment being made to the EITC claim amount:

$$L_2 = BN\left(-\frac{b'_N X_N + \Phi\left(b'_A X_A\right)}{s_N}, b'_A X_A, r_{AN} \right),$$

where $BN(z_1, z_2, r)$ represents the standard bivariate normal cumulative distribution function evaluated at z_1 and z_2, for correlation ?.

Case 3: Audit, EITC Overclaim

The third category contains audited returns that were found to have overstated the amount of EITC to which the taxpayer was entitled. For a return in this category, the likelihood expression (L_3) represents the probability density function for the observed EITC overclaim amount times the conditional probability of the return being audited, given the observed overclaim amount:

$$L_3 = \frac{1}{s_N} f\left(\frac{N - b'_N X_N - g_N \Phi\left(b'_A X_A\right)}{s_N} \right) * \Phi\left(\frac{b'_A X_A + r_{AN}\left[\frac{N - b'_N X_N - g_N \Phi\left(b'_A X_A\right)}{s_N} \right]}{\sqrt{1 - r_{AN}^2}} \right),$$

where $f(z)$ represents the standard normal probability density function evaluated at z.

Data Sources

We rely on two IRS data sources for estimation of our model: the Examination Operational Automation Database (EOAD) and the Individual Returns Transaction File (IRTF). The EOAD contains detailed audit results from operational audit cases that have been closed, including the values of adjustments made by the examiner to specific line items on the tax return. The IRTF contains detailed line item tax information from returns filed for a given tax period, including information from supplemental forms and schedules.

To illustrate our methodology, we combine the information from these two data sources to derive a sample containing detailed information from audited and unaudited individual income tax returns filed in the Chicago district for Tax Year 1996. Our data base for analysis is a choice-based sample containing all audited, timely Tax Year 1996 EITC claimants from the Chicago district identified in the EOAD and a 1-percent random sample of all unaudited Tax Year 1996 EITC claimants from the Chicago district. As displayed below in Table 1, our data base includes 7,300 randomly selected unaudited returns that claimed the EITC, representing 730,000 returns, and 728 audited returns that claimed the EITC.

Table 1: Choice-Based Sample Design

Type of return claiming EITC	Unweighted sample size	Weighted sample size
Audited	728	728
Unaudited	7,300	730,000

Results and Discussion

Table 2 summarizes the adjustments made to EITC claims in our audit sample. Overall, 82.7 percent of the returns selected for examination were found to have overstated the amount of EITC to which the taxpayer was entitled. In 66.5 percent of all examinations, the EITC amount claimed was entirely disallowed. Over all returns, claims were reduced by an average of $1,202. The high rate and dollar value of adjustments are at least partly attributable to the effectiveness of IRS audit selection criteria, which target returns that are deemed likely to require a substantial adjustment.

Table 2: Audit Sample Statistics

EITC adjustment rate	82.7%
% EITC claims entirely disallowed	66.5%
Mean EITC adjustment	$1,202
Median EITC adjustment	$1,000

As discussed earlier, we attempt to account for the role of audit selection in our model by jointly estimating an equation describing the likelihood that a return will be audited with a specification describing the likelihood and magnitude of noncompliance. To protect the confidentiality of IRS audit selection criteria, we restrict our presentation of estimation results to the portion of the model that pertains to noncompliance; more specifically, the parameters associated with Equation (2) and the

correlation term r_{AN} between the disturbances of equations (1) and (2).

Our measure of EITC noncompliance is the amount by which the EITC claim has been reduced as a result of the audit. For the purpose of estimation, we divide our measure by $1,000 as a normalization. Table 3 defines the explanatory variables used in our specification of EITC noncompliance described by Equation (2).

Table 3: Definitions of Explanatory Variables for Specification of EITC Overclaim Amount

Variable	Definition
TPI	**Amount of total positive income divided by $1,000.**
SELF-EMPLOYED	**Dummy variable for the presence of Schedule C (self-employment) income or loss.**
HOMEOWNER	**Dummy variable for the presence of a deduction for home mortgage interest.**
SCHED. E INCOME	**Dummy variable for the presence of any rental, royalty, or partnership income on Schedule E.**
AUDIT RISK	**Probability of audit computed based on the estimated parameters of Equation (1).**

Each of the explanatory variables defined in the above table is constructed from the information originally reported on the income tax return. Our estimation results are summarized in Table 4. The results indicate that, all else being equal, self-employed EITC claimants are relatively less likely to overstate the amount of the credit to which they are entitled, while homeowners are relatively more likely to do so. The presence of rental, royalty, or partnership income has a negative, but statistically insignificant, association with overclaiming the credit. The level of income appears to play no role in EITC noncompliance.

As discussed earlier, we expected the coefficient for the audit probability to be negative, signifying that, all else being equal, a taxpayer is relatively less likely to overstate the EITC if the risk of audit is high. The actual estimate is negative, but it is not statistically significant. Similarly, the estimated value of the correlation term r_{AN} is positive as expected, signifying that returns selected for audit tend to have larger EITC overclaims than returns with similar recorded characteristics that are not selected. Again, however, the parameter estimate is not statistically significant.

Table 4: Estimation Results

Parameter	Estimate	t-statistic
CONSTANT TERM	0.5554	0.233
TPI	-0.0100	-1.284
SELF-EMPLOYED	-0.4340	-2.649
HOMEOWNER	0.7438	2.229
SCHED. E INCOME	-0.3204	-1.626
AUDIT RISK	-0.3572	-0.397
s_e	1.1867	23.351
r_{AN}	0.0556	0.101

Table 5 summarizes the fit of our model in terms of its ability to predict the aggregate rate and dollar value of EITC overclaim adjustments within our audit sample. In terms of the aggregate adjustment rate, the model fits very well, predicting an 82.76-percent adjustment rate compared to an actual rate of 82.69 percent. The model also fits reasonably well in terms of the aggregate dollar adjustment amount, predicting an aggregate adjustment of $924,459 compared to an actual adjustment of $875,023. This represents an over-prediction error of 5.65 percent.

In Table 6, we compare the mean predicted EITC overclaim amounts for audited and unaudited returns. As expected, the predicted overclaim amount is, on average, higher for returns that have been subjected to audit than those that were not audited. However, the magnitude of the difference (about 5 percent) is lower than we anticipated. These results are based on a very preliminary variable specification for both the likelihood of an audit and the magnitude of

Table 5: Aggregate Rate and Dollar Value of EITC Overclaims--Actual vs. Predicted

Audit Population Parameter	Actual	Predicted
Aggregate Adjustment Rate	82.69%	82.76%
Aggregate Adjustment Amount	$875,023	$924,459

noncompliance, in both cases based solely on our judgment of what variables were likely to be important. While this suffices for the purposes of our illustration, we suspect that a more rigorous variable selection methodology would lead to an improved specification that would better capture the selection bias associated with the audit selection process.

Table 6: Mean Predicted EITC Overclaim Amount: Audited vs. Unaudited Returns

Type of Return	Mean Predicted EITC Overclaim Amount
Audited	$1,270
Unaudited	$1,213

References

Alm, James; Erard, Brian; and Feinstein, Jonathan S. (1996), "The Relationship Between State and Federal Tax Audits," in *Empirical Foundations of Household Taxation*, Martin Feldstein and James Poterba, editors, University of Chicago Press, Chicago, pp. 235–273.

Erard, Brian (1999), "*Estate Tax Underreporting Study*," report prepared by B. Erard and Associates for the IRS Economic Analysis and Modeling Group, Order Number TIRNO-98-P-00406, March 4.

Erard, Brian and Feinstein, Jonathan S. (2001), "*Estimating the Federal Income Tax Gap Using Operational Audit Data*," report prepared by B. Erard and Associates for the IRS Economic Analysis and Modeling Group, Order Number TIRNO-00-P-01128, November 6.

Eller, Martha Britton; Erard, Brian; and Ho, Chih-Chin (2000), "The Magnitude and Determinants of Federal Estate Tax Noncompliance," in *Rethinking Estate and Gift Taxation*, William G. Gale, James R. Hines, Jr., and Joel Slemrod, editors, The Brookings Institution, Washington, DC, pp. 375–410.