Measuring Tax Reporting Compliance: A Trichotomous Choice Model

Chih-Chin Ho and Alex Turk, Internal Revenue Service

This paper develops a system to predict aggregate reporting accuracy based on selected characteristics reported on individual income tax returns. The system includes two parts: a rule-based scheme to classify Federal individual income tax returns into five clusters; and for each cluster, a two-stage regression model to predict tax change as a function of return characteristics. Both clustering and modeling processes were developed using the 1988 Taxpayer Compliance Measurement Program (TCMP) individual filer post-audit survey data.

The clustering scheme is an economic theory-driven approach in which we select certain income sources and filing attributes to classify tax returns into five clusters that are generally homogenous in reporting compliance and yet mutually exclusive and exhaustive to one another. The clustering criteria pertain to filing complexity, income instability, and detection mechanism.

We develop a two-stage regression model to predict tax liability misreporting as a function of return characteristics as pre-audit observed influences for choosing a compliance state and the associated misreported amount. We expand the Turk, Ho, and Steuer (1997) binary choice model to a trichotomous choice model (TCM) that incorporates three observed states of reporting compliance: underreporting, full reporting, and overreporting of tax liability.

These extensions require two econometric refinements to ensure unbiased and consistent estimators of the model parameters. First, we estimate a multinomial logistic model to accommodate the trichotomous choice among the observed states of compliance. Second, we estimate two ordinary least square (OLS) regressions: one for overreporting and one for underreporting. In addition, we use a more generalized procedure to correct the selection bias inherent in the subsequent OLS regressions. Specifically, we follow the procedure outlined in Lee (1983) to rectify the selection bias resulting from multinomial logistic modeling of the trichotomous choice, as opposed to the Heckman (1976) procedure used in probit modeling of the binary choice.

We used two types of simulation to conduct an empirical evaluation of our trichotomous choice model. First, we used a tenfold cross-validation to generate "out of sample" prediction errors. Second, we used a deterministic simulation of the counter-factual state of full reporting compliance.

Theoretical Framework

Consider the following trichotomous choice model with mixed continuous and discrete dependent variables.

(1) $Y_s^* = X I_s + \varepsilon_s$
$Y_s = Z_s \beta_s + \omega_s \qquad \text{for } s = 1,2,3$

Note that X, Z_s are the vectors of return characteristics as pre-audit observed influences for choosing a compliance state (Y_s^*) and the associated misreported amount of tax liability (Y_s), respectively. ε_s and ω_s are the error terms representing the unobservable characteristics that influence Y_s^* and Y_s, respectively.

The continuous dependent variable Y_s is observed if and only if the state s is chosen.

(2) State s is chosen if and only if
$Y_s^* > \text{Max}\,(Y_j^*) \qquad \text{for } j = 1,2,3 \text{ and } j \neq s$

The discrete dependent variable Y_s^* is unobservable but has a dichotomous realization I*. Let I* be a trichotomous variable with values 1 to 3 where I* = s if state s is chosen. Equivalently,

(3) $I_s^* = s$ if and only if
$X\ I_s > O_s \quad \text{for } s = 1,2,3$
where $O_s = \text{Max}\ (I_j^* - \varepsilon_s) \text{ for } j = 1,2,3\ ; j \neq s$

Domencich and McFadden (1975) show that if ε_s is independent and identically Gumbel-distributed, the probability of choosing compliance state s is defined as:

$$\eta_s = \text{Prob}\,(I_s^* = s) = \frac{\exp(XI_s)}{\sum_{s=1}^{3} \exp(XI_s)} \qquad \text{for } s = 1,2,3$$

We observe Y_s if and only if I^*=s. We assume the error term, ω_s, represents unobserved characteristics that influence the amount of misreported tax liability.

The joint role of the unobserved characteristics, which influence the probability (η_s) and the amount of misreported tax liability (Y_s), means that these two error terms (ε_s and ω_s) may be correlated.

This situation is a bit more complicated than the standard selectivity bias correction procedure outlined in Heckman (1976) and was applied to the binary choice model in Turk, Ho, and Steuer (1997), because the error terms in the first stage (ε_s) are not Normally-distributed.

Lee (1983) develops a generalization of the Heckman procedure that can accommodate our trichotomous choice model. Define a selectivity bias correction measure as:

$$(5) \quad \lambda_s = \phi\,(\Phi^{-1}(\eta_s))\,/\,\eta_s \qquad \text{for } s = 1,2,3$$

Note that ϕ is a standard normal density function, and Φ^{-1} is an inverse of standard normal cumulative distribution function.

With λ_s included in the OLS regression as a selectivity bias correction measure, the estimator of expected values of the misreported amount, conditional on the realization of that particular compliance state, is **unbiased and consistent.**

When the marginal distributions of ω_s are Normal, N $(0,\sigma_s^2)$, we have the following equation:

$$(6) \quad E\,(Y_s \mid I_s^*{=}s) = Z_s\beta_s + \rho_s\sigma_s\lambda_s$$

for s = 1,2,3

Note that ρ_s is the correlation coefficient between the two error terms (ε_s and ω_s), and σ_s is the standard deviation, representing the scaling factor of estimating the amount of misreported tax liability.

In essence, the expected value of tax liability misreporting is an average of expected values of misreporting in each compliance state weighted by the probability of being in each state.

$$(7) \quad E\,(Y) = \sum_{s=1}^{3} \left[\, E\,(Y_s \mid I_s^*{=}s) * \eta_s \right]$$

for s = 1,2,3

Statistical Procedure

Agresti (1990) develops a multinomial logit model to estimate the probabilities of mutually exclusive events:

$$(8) \quad \text{Log}\,(\eta_s / \eta_{s*}) = X\, I_s \qquad \text{for } s = 1,2,3$$

where X is the vector of k explanatory variables of reported return characteristics as pre-audit observed influences for choosing a compliance state, I_s is the vector of k-associated parameters, and s* designates the baseline state.

Our trichotomous choice model has three mutually exclusive and exhaustive compliance states:

underreporting	(s=1,Y>0)
overreporting	(s=2,Y<0)
full reporting	(s=3,Y=0)

To remove the indeterminacy, we normalize the model by selecting full reporting as the baseline state, and thus estimate two log-odds ratios for underreporting and overreporting, respectively, versus full reporting as the following:

$$(8.1) \quad \text{Log}\,(\eta_1 / \eta_3) = X\, I_1$$

$$(8.2) \quad \text{Log}\,(\eta_2 / \eta_3) = X\, I_2$$

Note that I_s represents the change in the log-odds ratio of being in the compliance state s versus full compliance by a one-unit increase in X.

Consequently, we estimate two **conditional** probabilities upon full reporting **simultaneously**, one for underreporting and one for overreporting:

$$(9.1)\quad \eta_1 = \exp(X\,I_1) / [1+\exp(X\,I_1) + \exp(X\,I_2)]$$

$$(9.2)\quad \eta_2 = \exp(X\,I_2) / [1+\exp(X\,I_1) + \exp(X\,I_2)]$$

Note that $\exp(X\,I_s)$ represents the change in the odds ratio of being in the compliance state s versus full compliance by a one-unit increase in X.

Since full reporting is the baseline state, we estimate two misreported amounts of tax liability separately, one for underpayment and one for overpayment:

$$(10.1)\quad E(Y_1 \mid I_s^*=1) = Z_1\beta_1 + \rho_1\sigma_1\lambda_1$$

$$(10.2)\quad E(Y_2 \mid I_s^*=2) = Z_2\beta_2 + \rho_2\sigma_2\lambda_2$$

The predicted value for tax liability misreporting for a particular tax return can be calculated by summing over all three compliance states of the expected value of the misreported tax liability in a state weighted by the probability of being in that state. Since the expected value of full reporting is zero tax misstatement, that term is implicit.

$$(11)\quad E(Y) = \eta_1 * E(Y_1 \mid I_s^*=1) + \eta_2 * E(Y_2 \mid I_s^*=2)$$

■ Clustering Scheme

We classify tax returns into five homogenous clusters, primarily based on the source of income rather than the level of income. We aggregate all income items on individual income tax returns into four types of income.

Labor Income (LI) includes wage and salaries, Social Security benefits, unemployment compensation, pension income, State tax refunds, alimony, and IRA distributions.

Capital Income (KI) includes interest and dividends, capital gains, supplementary income, and income from sales of business equipment (Form 4797).

Business Income (BI) includes non-farm sole proprietor income (Schedule C). Risk Income (RI) includes farm income (Schedule F) and other income.

These four income types differ considerably in three main areas: detection mechanism, income instability, and filing complexity.

Table 1 summarizes characteristics of these four types of income.

Table 1
Characteristics of Four Major Income Types

	Detection Mechanism	Income Instability	Filing Complexity
Labor Income	High	Low	Low
Capital Income	Moderate	Moderate	Moderate
Business Income	Low	High	High
Risk Income	Low	High	High

The clustering approach is a rule-based and economic theory-driven approach in which we use the four main income sources and selected filing attributes to classify individual tax returns into five mutually exclusive and exhaustive clusters. Table 2 summarizes the rules of classification.

■ Empirical Evaluation

The empirical evaluation of our trichotomous choice model was accomplished with two different types of simulations. First, we used a tenfold cross-validation to generate "out of sample" prediction errors. Second, we used a deterministic simulation of the counter-factual state of all individuals correctly reporting their tax liabilities.

A tenfold cross-validation was used to evaluate the model predictions and sensitivity to sample selection. The procedure consisted of dividing the observations in each cluster into 10 mutually exclusive groups. For each cluster, the estimation procedure was replicated 10 times. In each replication, one group was held back for prediction purposes. The results are summarized in Table 3.

As a benchmark, the same cross-validation proce-

Table 2
Definitions of Five Mutually Exclusive
And Exhaustive Clusters

Rule 1: Assign Returns to Group A and B	Group A	LT >=0.80 **AND** LI >=(KI* + BI* + RI*)
	Group B	LT< 0.8 **OR** LI< (KI* + BI* + RI*)
Rule 2: Assign Group A to Cluster 1 and 2	Cluster 1	{Either Claiming Standard Deduction OR Filing as Other Than Married Filing Jointly} **AND** (Filing No Schedules)
	Cluster 2	**ELSE in GROUP A**
Rule 3: Assign Group B to Cluster 3, 4, and 5	Cluster 3	KI > Max (BI, RI) **AND** KI* > Max (BI*, RI*)
	Cluster 4	{KI<= Max(BI,RI) OR KI* <= Max (BI*,RI*) } **AND** GR_C > (GR_F + ABS_O)
	Cluster 5	{KI<= Max(BI,RI) OR KI* <= Max (BI*,RI*) } **AND** GR_C <= (GR_F + ABS_O)

NOTE:

LT = Proportion of Labor Income to Total Income

KI*= Sum of Absolute Values of Capital Income Components
BI* = Absolute Value of Business Income
RI* = Sum of Absolute Values of Risk Income Components

GR_C = Gross Receipts of Schedule C
GR_F = Gross Receipts of Schedule F
ABS_O = Absolute Value of Other Income

dure was applied to an OLS regression of tax misreporting, Y, on the set of unique elements in X, Z_1, and Z_2. Columns I and II contain the average prediction error for the TCM and the OLS models, respectively, as a proportion of the average observed tax change. Column III summarizes the mean square errors (MSE) of the residuals for predicting the amount of tax misreporting. The residuals are defined as actual tax misreporting minus predicted tax misreporting. The TCM MSE is reported relative to the OLS MSE.

In general, the models' performances are very similar. The relative aggregate prediction errors are practically identical in all but Cluster 3. Since the predictions are for the 10 cross-validation holdout samples, the OLS residuals do not sum to zero. In four of the five clusters, there was a small reduction in the MSE of prediction.

Table 3
Cross-Validation Prediction Errors
for Average Tax Misstatement

	Relative Average Prediction Error		Relative MSE of Prediction
	TCM	OLS	TCM/OLS
Cluster 1	0.093	0.099	0.999
Cluster 2	-0.001	-0.002	1.026
Cluster 3	0.029	-0.002	1.000
Cluster 4	0.009	-0.002	0.984
Cluster 5	0.105	0.099	0.988

The second step was to evaluate the model predictions in a simulated or counterfactual state of no tax misreporting. While the first evaluation focused on the performance of the models with different samples, this evaluation holds the model parameter estimates constant and instead focuses on the performance of the model in a different paradigm, one of compliance. Presumably, an estimation procedure that produces unbiased estimates of the true parameters should provide better predictions in simulated scenarios. This paradigm is the special case in the continuum between compliance and non-compliance.

This evaluation was accomplished by essentially assuming in our counter-factual state that all of the information reported on each return was identical to what the auditor determined to be the correct value in the TCMP data. Under the assumptions of the counter-factual state, the observations are assigned to the appropriate clusters, and the predicted misstatements were calculated using the TCM and the OLS models. In the simulated scenario, misreporting declines are zero for each individual.

Table 4 is a summary of the predictions of the TCM and the OLS model in the counter-factual state. The values in the table represent the percentage change in the predicted misreporting for the actual and the counter-factual states. The results for each model are reported for the five clusters. It appears that the TCM model does a better job predicting compliance in the simulated scenario. The change in TCM-predicted misreporting is closer to the simulated scenario (a decline of 100 percent) in four of the five clusters. It is important to remember that both the OLS model and the TCM parameters were estimated with the same clustering scheme and have essentially the same predictor variables. The difference is that the TCM model explicitly considers the censoring process and controls for the potential bias in second-stage regression. The results reported in Table 4 suggest that it is important to consider this bias when using the estimates to predict individual behavior when policy or other exogenous factors change. In short, it reinforces the theoretical motivation for the TCM model.

Table 4
Model Prediction in the Simulated State of No Tax Misreporting

	Percentage Change Predicted Misreporting TCM	Percentage Change Predicted Misreporting OLS
Cluster 1	-31%	-27%
Cluster 2	-5%	0%
Cluster 3	3%	6%
Cluster 4	-3%	-9%
Cluster 5	-42%	-11%

Conclusions

The model developed here provides a framework that is intuitively and theoretically appealing for modeling reporting compliance. The clustering scheme is primarily based on selected income sources and filing attributes. The clustering criteria pertain to detection mechanism, income instability, and filing complexity.

For each of the five mutually exclusive and exhaustive clusters, we develop two sequential components. The first model is a multinomial logistic model for estimating respective probabilities of being in one of the three mutually exclusive compliance states; and the second model is a regression model with selectivity bias correction for estimating the misreported tax liability conditional on a respective compliance state.

The empirical evaluations suggest that the model will provide aggregate predictions that appear to be at least as good as OLS estimates that do not adequately deal with the censoring observed in audit data. The evaluations in the simulated scenario of perfect compliance suggest that the approach of explicitly considering the selection process has enhanced our reporting compliance model.

References

Agresti, A. (1990), Categorical Data Analysis, John Wiley and Sons.

Domencich K. and McFadden D. (1975), Urban Travel Demand, North Holland Publishing Company, Amsterdam.

Heckman, J. (1976), "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Variables and a Simple Estimator of Such Models," Annals of Economic and Social Measurement.

Lee L. (1983), "Generalized Econometric Models with Selectivity," Econometrica.

Turk, A.; Ho, C.; and Steuer, A. (1997), "Alternative Methods for Modeling Income Tax Reporting Compliance," American Statistical Association 1997 Proceedings of the Business and Economics Statistics Section.