Proceedings of the 2009 Winter Simulation Conference
M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, eds.

ECONOMETRIC SIMULATION OF THE INCOME TAX COMPLIANCE PROCESS FOR SMALL BUSINESSES

George Contos
Ardeshir Eftekharzadeh
John Guyton

Internal Revenue Service
Office of Research
500 North Capitol St
Washington, DC 20001

Brian Erard

B. Erard & Associates
2350 Swaps Court
Reston, VA 20191-2630

Scott Stilmar

IBM Corporation
12907 Federal Systems Park Dr.
Fairfax, VA 22033
and
Dept. of Economics
University of Virginia
P.O. Box 400182
Charlottesville, VA 22904

ABSTRACT

Econometric models can be very useful for estimating the marginal impacts of changes in policy. However, their broader application as a tool for micro-simulation analysis poses a number of challenges and limitations. This paper uses the context of modeling taxpayer compliance burden for small businesses to explore some extensions to standard econometric simulation techniques that provide more robust support of the distribution of the characteristics of interest. Key to the approach is explicitly simulating a random draw from the specified error distribution and a pair of calibration factors reflecting some of the technical limitations of a finite simulation. Further technical considerations regarding the retransformation of the dependent variable in a log-linear regression model are also discussed. Final comments include thoughts on potential refinements and implications for simulating the domain area of interest beyond the current scope of small business taxpayers.

1 INTRODUCTION

In 1998, the Internal Revenue Service (IRS) commenced a series of projects developing an improved methodology for measuring and modeling the compliance burdens imposed by the federal income tax system. These projects, and the resulting models, assist the IRS in its efforts to provide taxpayers with improved services and to help policymakers understand the full impact of changes in tax law.

IRS modeling objectives are to assess the impact of programs on taxpayer burden, to assess the role of burden in tax administration, and to fulfill IRS obligations to the Office of Management and Budget (OMB) for information required by the Paperwork Reduction Act. Official forecasts of total compliance burden are produced for each fiscal year. In addition, estimates of average compliance burden for each calendar year by tax form are published in the taxpayer instructions as a guide for the taxpayers.

The models also support tax policy making through “what-if” type analysis. Such analysis permits estimation of the impact of proposed legislation on taxpayer burden before it is enacted. To satisfy the third goal, in addition to estimating total and mean burden the model needs to perform satisfactorily for various subgroups of the total business population (e.g., subchapter C or S corporations) and across the population distribution.

978-1-4244-5771-7/09/$26.00 ©2009 IEEE

These three separate objectives require different estimates and provide the primary reason the IRS developed micro-simulation models for its studies of pre-filing and filing compliance burden (Arena et al. 2003, Connors et al. 2007). Such models are widely used to study the impact of public policies by examining the behavior of individual units at the micro-level (Gupta and O'Hare 2000).

The Individual Taxpayer Burden Model (ITBM) addressing the burden of Wage and Investment (W&I) and Self-Employed (SE) taxpayers was first deployed in January 2003. A subsequent model covering the Small Business taxpayer burden (SBBM) for both income tax and employment tax was first deployed in April 2006.

The ITBM and the SBBM models are driven by compliance burden estimates collected in three separate surveys. The W&I and SE surveys were conducted in 2000 and 2001 while the small business survey was conducted in 2004 and 2005. A new individual taxpayer survey for calendar year 2008 was conducted in May 2009, but has not yet been incorporated into the ITBM as of the time of the writing of this article. The models measure both the time and money that individuals and businesses spend on pre-filing and filing activities in response to the requirements of the U.S. federal tax system. Both models cover seven pre-filing and filing activities, such as recordkeeping, that would not have occurred without a federal tax system and excludes psychological costs and deadweight losses from changes in economic behavior (Guyton et al. 2003).

This paper discusses the development of the econometric equations for the SBBM. Particular focus is given to simulation methods used to preserve the distributional characteristics of the reported population when estimating the predicted burden in the population of the resulting econometric micro-simulation model.

2 MODELING APPROACH

As discussed earlier, the primary objective of the SBBM is to explain small business compliance burden. We developed a model reflecting the recent public and corporate finance literature and uses current statistical techniques. In addition, we wanted a model that could easily be adapted to changes in the tax system and the economy overall. Finally, we wanted to develop a model that had the potential to be adapted and generalized to model compliance burden for other taxpayer populations, such as large and medium-size businesses, individual taxpayers, and tax exempt entities.

2.1 Economic Model

To model compliance burden for small businesses we assume that business entities select the combination of capital and labor that allows them to fully respond to the requirements of the U.S. federal tax system while minimizing compliance costs. (Labor in this scenario is the time spent on pre-filing and filing activities by firm owners and employees as well as by paid professionals.) This assumption may not hold true for all firms all the time but we believe that for-profit entities tend to adopt a compliance process that reduces costs. For example, small and young entities have limited budgets so they tend to handle all pre-filing and filing activities in-house. The owners maintain the financial books, other business records, review the tax rules, prepare tax records, complete, and submit all tax forms. As firms grow they have more business transactions to account for and the business owners face higher opportunity costs on the time spent dealing with payroll, recordkeeping, and other paperwork. So they may invest in recordkeeping software and hire full-time recordkeeping staff or employ paid professionals for business activities, such as payroll. The improved infrastructure leads to less time needed for the tax-related activities. In addition, the firms' management becomes more familiar with the federal tax system and its requirements or hires paid tax preparers leading to further reduced compliance costs. Given this assumption, to model compliance burden for small businesses we tested the hypothesis that as business entities grow, their compliance costs increase at a decreasing rate.

2.2 The Data Set

The compliance burden data used by the SBBM are 7,049 surveys collected by the IRS. The sample frame was small business taxpayers who filed a return during Processing Year 2003, that is all returns that were processed between January 1, 2003 and December 31, 2003. The population of small business taxpayers was defined as businesses (filers of Forms 1065, 1120, 1120-S, 1120-REIT, 1120-RIC, 1120-L, 1120-PC, 1120-F, 1120-FSC, 1120-SF, and 1120-H) with end-of-year assets of less than $10 million. The sample is a stratified sample design which, when weighted, represents the small business population.

The survey collected information on both the time and money that businesses spend on pre-filing and filing activities. Each survey was then linked to the matching administrative record to create the estimation data set. The administrative record includes selected items from the primary tax forms and various secondary forms and schedules. Both the survey and administrative records were extensively reviewed and cleaned for memory recall, administrative, or processing errors.

2.3 Econometric Model

To model the conditional distribution of taxpayer compliance burden, we employ a log-linear regression specification in which the natural log of burden is linearly related to a set of explanatory variables. This type of a model is supported by the results of the small business survey as well as the findings of a Large and Mid Size Business (LMSB) taxpayer survey conducted by Slemrod and Venkatesh (Slemrod and Venkatesh 2002). They found that the relationship between compliance burden and the size of the firm is best estimated using a log-linear specification.

Given that only one year of small business compliance burden data is available and burden estimates are produced on a yearly basis, all but one of the independent variables of the econometric model were based on administrative data. Running through the SBBM administrative files from subsequent years allows us to produce burden estimates for those years. The dependent variable, $\log(Burden)$, is of course based on survey data. As discussed earlier, the survey collected information on both the time and money that businesses spend on pre-filing and filing activities. In order to control for substitution of time and money and to aggregate across burden activities, we created a single measure of compliance burden. The key choice was whether to monetize the value of time and add it to the out-of-pocket costs or rather to chronotize the out-of-pocket costs and add it to time. We opted for the former for both technical and program management reasons. (To monetize the value of time the average labor cost for each entity was estimated. The average labor cost includes the wages paid to employees and all other overhead costs incurred by the firm such as health insurance premiums, pension contributions, etc. If the estimated average labor cost was below the minimum wage plus overhead costs or above the average fee charged by paid professionals for each particular activity, these limit values were used to monetize time for that taxpayer. Separate maximum limits were set for each particular activity, for example, the maximum hourly cost for recordkeeping time was set equal to the fees charged by professional bookkeepers.) Total monetized burden is equal to the sum of monetized time spent on pre-filing and filing activities and out-of-pocket costs.

Following the corporate finance literature, the model controls for two key firm characteristics; size of the entity and industrial classification. As a proxy of size we use the log of total receipts in the current period. (Total receipts are defined as the sum of gross receipts, rental real-estate income, interest income, dividend income, royalties income, and other income.) Total receipts or total assets are two of the most commonly used size proxies in corporate finance literature. We selected total receipts as our proxy since certain small businesses (total assets or receipts of $250,000 or less) are not required by the Internal Revenue Code (IRC) to file balance sheets. We tested log of total assets as a proxy of size for firms with total assets greater than $1 million and we got similar results as when using total receipts. Following the work of Slemrod and Venkatesh (Slemrod and Venkatesh 2002) we also include a dummy for zero total receipts. The dummy is set to one for firms with zero total receipts and zero for entities with total receipts greater than zero. The firm's industrial classification is defined at the two digit North America Industry Classification System (NAICS) level with further subclassification within the finance, insurance, and real estate industries. To better support "what-if" type analysis for various small business subgroups a number of dummies and interaction terms were included in the model. The dummies were based on the preparation method (equal to one for self preparers) and on the type of tax form each entity filed. The most unique aspect of modeling compliance burden is the need to account and control for the type and volume of activities performed by each individual taxpayer in response to their federal tax obligations. To do so we developed a proxy for the type of activities performed. Each tax item from the primary forms and schedules was organized into one of three complexity categories; low, medium, and high. The complexity categories are based on the notion that burden increases as a function of both the number and the type of tax-related activities. More specifically, if a business has to complete an additional tax item this year, keeping everything else the same, compliance burden will increase since the business will need to adjust its recordkeeping, familiarize itself with the relevant taxpayer instructions or pay higher preparation fees, etc. The increase in burden will also be a function of the extent to which the activity differs from the non-tax activities involved in managing a business (e.g., the business related recordkeeping and planning activities).

Separate complexity proxies were created for each one of the main Tax Forms. For example, a separate proxy was created for business entities that filed Form 1120, 1120-S, 1065, and groupings of similar special purpose corporate income tax forms (Forms 1120-REIT, 1120-RIC, 1120-L, 1120-PC, 1120-F, 1120-FSC, 1120-SF, and 1120-H). This was done for three reasons: first, the IRC has different provisions for the same line item across form types; second, different form types require reporting on different tax items; finally, it allowed the proxy to reflect differences in tax planning associated with the requirements and elective tax benefits of each form type.

To develop the complexity categories we initially placed the various tax items into categories based on the recordkeeping intensity, tax planning activities, and overall complexity of extracting that information from the entity's financial books. More specifically, the low category includes items that are recorded and reported at an aggregate level. The medium category includes items that require additional recordkeeping and are reported to the IRS separately. Many of the items included in the medium category require attaching worksheets or otherwise documenting how the totals were determined. Finally, the high category includes items that may require a separate recordkeeping system or a process with potentially separate rules for each item. Tracking records across years is an additional component for most tax items in this category. To test the assignment criteria, the model was then run with each item as a separate right-hand side variable. The magnitude of the estimated coefficients was compared with the rest of the items in that complexity category. Items that had coefficients significantly different than their peers were moved to a more suitable category. (At the time of this writing, the complexity category assignments are under review by a number of stakeholders, government, academic, and industry experts.) As a proxy for the volume of activities we used the money amounts reported for each item. This is based on the notion that the larger the amount reported on a tax item the more transactions should typically be associated with the activities related to that line. The value of each complexity category is equal to the sum of the logs of one plus the amount reported for each item. By utilizing the properties of logarithms in the complexity categories the equation acquires a desirable property. Each tax item included in the categories acts as a separate regressor but with the coefficients of all items in the same category restricted to be the same. The equation estimated is:

$$\log(Burden_i) = b_0 + b_1\ X_i + b_2\ Low_i + b_3\ Medium_i + b_4\ High_i + \sum_{j=1}^{21} b_{5j}\ Industry\,dummy_{ij} + b_6\ Nopaid_i + b_7\ D_i + \varepsilon_i, \quad (1)$$

where the subscript i indexes the business entity: log (Burden) is the log of total monetized compliance burden; X includes the set of firm-level controls; Low, Medium, and High are the measures based on the volumes of activity associated with our three complexity categories; the industry dummy variables account for variations in burden across key industries; Nopaid is a dummy set to one for entities that prepared their tax returns in-house and zero for firms that used a paid preparer; and D is a set of dummies based on the preparation method and the type of tax form each entity filed and interaction terms.

3 ROBUST REGRESSION

The small business population is very diverse and covers businesses in a large range of asset classes; however, the majority of the entities are concentrated in the lower asset classes. Table 1 shows the average ratio of burden to total receipts by decile. It is clear that the small business survey data are skewed with a heavy tail.

Table 1: Average Burden As a Percentage of Total Receipts, By Decile

Decile	Upper Range of Decile(Total Receipts)	Average Burden As a Percentage of Total Receipts
10	\$402	822.14%
20	\$13,040	70.30%
30	\$39,595	13.20%
40	\$79,803	7.03%
50	\$134,662	4.32%
60	\$225,372	3.31%
70	\$397,542	2.38%
80	\$721,083	1.81%
90	\$1,746,796	0.98%
100	\$456,134,188	0.44%

Our log-linear regression specification addresses the inherent skewness in the compliance burden data (Manning and Mullahy 2001). Although there are a variety of alternative functional forms to address skewness, a Box-Cox test for the optimal transformation of the dependent variable confirmed a logarithmic transformation as the best option. (It is worth noting that following the model selection process described by Manning and Mullahy (Manning and Mullahy 2001) we researched whether a Generalized Linear Model (GLM) would perform better than OLS. First, the kurtosis of the log-scale residual was calculated from one of the consistent GLM estimators. Since the kurtosis was less than 3, the Park test was used to select the appropriate GLM. The estimated λ was equal to 1.58. If λ is equal to 1, this indicates that the raw-scale variance is proportional to the raw-scale prediction, and the Park test suggests considering a Poisson-like model. Alternatively, if λ is equal to 2, this means that the raw-scale variance is quadratic in the raw-scale prediction, and the Park test suggests considering either a gamma model or the homoskedastic log-linear OLS model. Since the estimated value for λ was between 1 and 2, all three specifications were tried. The results of the alternative specifications were all qualitatively similar, so the simplest log-linear OLS specification was selected. Although, both the survey and administrative data were cleaned and standardized early in the process, there was still concern that outliers could affect the robustness of the model. The detection of potential outliers was of particular interest to the IRS since the survey required the respondents to recall the intensity of the various activities performed and to separate and report only the part of the activity that would not have been undertaken in the absence of a federal tax system.

Given the complexity of the multivariate outlier detection process, robust regression was used to identify and adjust the weights of observations with reported values farthest away from the initial regression line. Robust regression is an iterative process that reduces the importance of observations with high residuals by lowering their weights (based on a weight function) and then re-runs the regression with the new weights repeating the process until it converges.

The process is particularly attractive since different weight functions are available so that it allows the user to tailor the process to the data. An advantage of robust regression is that it treats extreme responses while allowing the model to utilize the reduced weight observations. Returns reporting extreme values are kept in the sample but their sample weights are reduced for purposes of estimating the model. To maintain the integrity of the sample, for the observations with reduced weights a second record is inserted into the dataset with a weight equal to the original sample weight minus the reduced weight but with the burden data for this second record set to missing. Using a multiple imputation method, burden values are then estimated for the new records as well as for all other observations with missing values (Allison 2001).

Several weight functions were tested and their parameters were adjusted to meet the following criteria: the weights of observations within 1.5 standard deviations of the mean residual value of zero are not altered; observations at 3 standard deviations get a weight equal to 0.05% of original weight; observations outside this range are excluded from the sample; and there is a somewhat steady increase in the percent change in the weight of observations between 1.5 and 3.15 standard deviations. We concluded that the Hampel function best fit the model's goals since it removed a small number of influential outliers and decreased the cumulative weights by less than 5 percent. Figure 1 shows the graphical representation of the Hampel function on the record weights as a function of their proximity to the regression line.

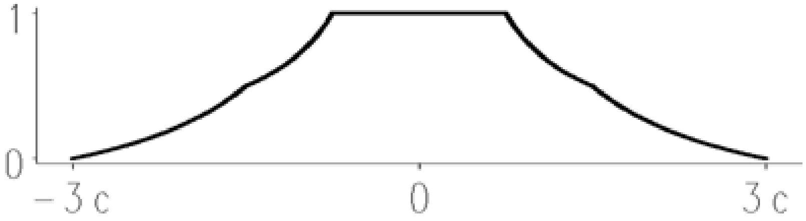

Figure 1: Hampel Function

4 SIMULATION ISSUES

Since total monetized compliance burden was transformed into logs for purposes of regression analysis, we had to retransform the estimates back to levels. This is an exercise that is not as trivial as it may seem. In a standard regression model, the error term (ε) has a mean of 0 and is thus ignored when predicting the values of the dependent variable. However, when one retransforms the dependent variable in a log-linear regression specification, the level of the dependent variable depends on the value of the anti-log of the error term ($exp\{\varepsilon\}$). In general, it is not safe to ignore the contribution of this non-linear function of the error term when predicting the level of the dependent variable. To see this, consider the log-linear specification:

$$\log(Y_i) = \beta' X_i + \varepsilon_i, \tag{2}$$

where i indexes observations, X_i is a column vector of explanatory variables, β is a column vector of coefficients, and (conditional on X_i) ε_i is a normally distributed error term with zero mean. In this specification, the natural log function has been used to transform the dependent variable Y_i. As in a standard regression, the mean of our transformed dependent variable is equal to $\beta' X_i$. However, when we retransform this specification to obtain the level of Y_i, we obtain:

$$Y_i = exp\{\beta' X_i\} exp\{\varepsilon_i\}. \tag{3}$$

Therefore, the conditional expectation of Y_i given X_i may be computed as:

$$E(Y_i|X_i) = exp\{\beta^{'} X_i\} E\left(exp\{\varepsilon_i|X_i\}\right). \tag{4}$$

Although $E(\varepsilon_i|X_i)$ is zero, the value of $E(exp\{\varepsilon_i|X_i\})$ turns out to be a nonlinear function of the error variance.

4.1 Simulation With Homoskedastic Errors

Under our assumption that the error term ε_i is homoskedastic (i.e., that it is normally distributed with constant variance σ^2), we have:

$$E(exp\{\varepsilon_i|X_i\}) = exp\left\{\frac{1}{2}\sigma^2\right\}. \tag{5}$$

Therefore, when predicting the level of the dependent variable based on the results of a log-linear regression analysis, one needs to take the average contribution of the error term into account. It is worth noting that if one chooses to ignore the average contribution of the error term (i.e., by setting the second factor in Equation (4) to (1), one obtains an estimate of the median value of Y_i rather than the mean value.) This point is illustrated in Figures 2 and 3, which relate to a hypothetical error term ε from a log-linear regression specification: the mean of the error term in this example is equal to zero, and the standard deviation is equal to 1. When one retransforms the natural log of the dependent variable in this specification back into levels, one obtains an expression like Equation (2), which involves the anti-log of the error term. Whereas the error term in Figure 2 is normally distributed, the anti-log of the error term in Figure 3 is log-normally distributed. Its mean takes the value $exp\{\sigma^2/2\} = exp\{1/2\} = 1.65$ In contrast, the median of $exp\{\varepsilon\}$ is equal to $exp\{0\} = 1$.

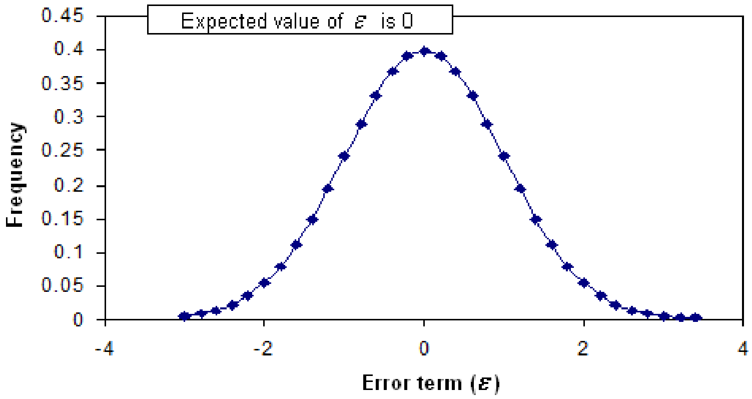

Figure 2: Log-linear Regression Error Term ε

4.2 Simulation with Heteroskedastic Errors

So far we have been considering the case in which the error term in our log-linear specification is homoskedastic. In a model where the regression error is heteroskedastic, the expectation of the anti-log of the regression error is no longer constant, which makes it more challenging to predict the level of the dependent variable. To determine if the variance of the error term in our model [Equation (1)] is homoskedastic or heteroskedastic the White test for heteroskedasticity was performed. It tests the null hypothesis that the variance of the residuals is homogenous. The estimated p-value for the test was less than 0.0001, which leads us to reject the null hypothesis and assume that heteroskedasticity is present. To account for the apparent heteroskedasticity in our model, we begin by considering the prediction formula for burden Y_i provided in Equation (4). Assuming that the conditional distribution of Y_i given the explanatory variables X_i is normal, Equation (4) simplifies to:

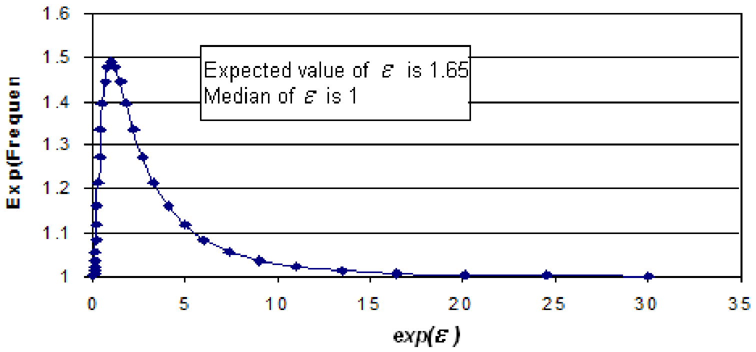

Figure 3: $E(exp\{\varepsilon\})$ is Lognormally Distributed with mean 1.65 and median 1

$$E(exp\{Y_i|X_i\}) = exp\{\beta^{'}X_i\}exp\left\{\frac{1}{2}\sigma_i^2\right\}. \tag{6}$$

The first term in this expression can be estimated by replacing β with its regression estimate. However, the presence of the second term requires us to estimate the variance of the error term (σ_i^2) for each observation in our sample. To address this problem, we have defined and estimated a parametric specification for the variance of the error term. We have followed the common practice of first applying OLS to our model [Equation (1)] and then regressing the squared residuals (e_i^2) from this analysis against a set of explanatory variables (Johnston and Dinardo 1984). The predicted values of the dependent variable from this auxiliary regression then serve as our estimates of the variance of the error term, which can be substituted in place of σ_i^2 in Equation (5). We go one step further in our analysis, by applying weighted least squares to our original burden regression equation to correct for the heteroskedasticity in this specification. Assuming that we have modeled the form of this heteroskedasticity correctly, our weighted least squares estimates of β will be (asymptotically) more efficient than our original OLS estimates.

The detailed steps of the procedure are:

A. Regress $ln(y_i)$ on X_i and obtain the estimated residuals e_i.

B. Define v_i equal to e_i^2. Regress v on x and then compute the predicted value ($\hat{v}_i$) of the dependent variable v_i for each observation.

C. Perform a weighted regression of $ln(y_i)$ on X_i using $1/\hat{v}_i$ as the weight variable. (In our case a new weight variable would be created using the sample $weights \times 1/\hat{v}_i$.)

D. Use the output from C to compute the predicted linear value of y as:

$$\hat{y}_i = exp(bX_i + \hat{v}_i/2), \tag{7}$$

where bX_i uses the estimated coefficients from step C and $\hat{v}_i$ is the estimated squared error from step B.

In applying the above procedure, we had a few cases where the modified weight in Step C was excessively large. In those cases, we substituted the original weight to avoid giving these few observations undue influence over the results. We plan to explore alternative ways of addressing this issue in future research.

4.3 Stochastic Micro-simulation

Given that one of the model's objectives is to support tax policy-making through "what-if" type analysis, the model needs to perform satisfactorily in estimating compliance burden for subgroups of the business population and across the overall population distribution. Up to this point, we have focused on using the regression results from our model to estimate the expected level of burden for a given taxpayer in our sample as a function of observed characteristics. It is common to use econometric predictions such as this as the basis for micro-simulation analysis of administrative or policy changes. Such an approach is non-stochastic in the sense that the simulated values of burden are a deterministic function of the explanatory variables in the model. An undesirable feature of assigning an estimate of the expected taxpayer burden to each taxpayer in the sample is that it causes the predicted burden values in the sample to be much less dispersed than the actual reported values. Given that taxpayer burden is highly skewed, this approach also causes the median of the predicted burden amounts in the sample to substantially exceed the median of the reported burden distribution. The failure of this approach to adequately simulate the distribution of reported burden among taxpayers is likely produce misleading inferences regarding the effects of administrative and policy changes on various aspects of that distribution. To better match the reported burden distribution, we have developed a stochastic micro-simulation methodology that simulates burden according to the distributional assumptions inherent in our model. The starting point for this approach is Equation (3), which shows the level of the dependent variable to be a function of the anti-log of the error term ε_i. Under the non-stochastic simulation approach, one replaces the unknown value of this function of the error term with its expectation, as in Equation (4), and then employs an estimate of the expected value to derive the predicted value of the dependent variable. Under our stochastic simulation approach, we instead draw random values from the normal distribution and use these random draws in place of the unobserved error terms. The mean of the normal distribution we draw from is set equal to zero and the variance is set equal to the estimated variance of the error term from our regression analysis; since we allow for heteroskedasticity in our analysis, the estimated variance varies across observations in our sample. We have elected to repeat this process 30 times for each observation, thereby yielding 30 simulated values of the dependent variable for each observation in our sample. The choice of 30 random draws as opposed to some other number was made at the authors' discretion. In future research, we plan to explore the sensitivity of our results to alternative values for the number of random draws.

4.4 Refinement of Stochastic Simulation Methodology

Since the support of the error term ε in Equation (4) is unbounded under the normal distribution, it is theoretically possible for compliance burden to be infinite. As a practical matter, however, businesses would cease to operate if the compliance burden was sufficiently onerous, so the actual burden distribution is in fact bounded. Consider, for example, Table 1. In this table, the reported compliance burden represents a fairly modest share of total receipts for all but the lowest decile of businesses, for which total receipts are less than \$402. To avoid implausibly large estimates of the compliance burden in our stochastic simulations, we censor our draws from the error distribution at plus or minus three standard deviations from the mean error of zero; in other words, any draw outside of this range is set equal to the threshold value. Henceforth, we will refer to this practice as "capping".

A consequence of imposing the caps is that the mean of the simulated values for compliance burden will tend to be somewhat lower than the mean of the reported values of burden in our sample. Below, we derive a correction factor that can be applied to address this issue.

Let $\hat{Y}_{ij}$ represent the j^{th} $(j = 1,\ldots,30)$ uncapped simulated value for the i^{th} observation in our sample, and let $\hat{Y}^c_{ij}$ represent the corresponding capped simulated value. Define the threshold values for the random draws for the error term as $\pm c_i$. Then $\hat{Y}^c_{ij}$ is related to $\hat{Y}_{ij}$ as follows:

$$\hat{Y}^c_{ij} \quad = \quad \begin{cases} e^{b'X_i+c_i} & \hat{Y}_{ij} > e^{b'X_i+c_i}; \\ \hat{Y}_{ij} & e^{b'X_i-c_i} \le \hat{Y}_{ij} \le e^{b'X_i+c_i}; \\ e^{b'X_i-c_i} & \hat{Y}_{ij} < e^{b'X_i-c_i}, \end{cases} \tag{8}$$

where b represents the estimated value of β in our log-linear regression specification provided in Equation (2). Given that we draw from the normal distribution to simulate the error term in Equation (2), $\hat{Y}_{ij}$ approximately follows a log-normal distribution with parameters $\beta'X_i$ and σ_i^2. Dropping subscripts to simplify the notation, the approximate corresponding probability density function (p.d.f.) of $\hat{Y}^c$ is defined as:

$$f(\hat{Y}^c) \quad = \quad \begin{cases} \frac{1}{\hat{Y}\,\sigma\sqrt{2\pi}} e^{-\frac{\left(ln\hat{Y}-\beta' X\right)^2}{2\sigma^2}} & e^{b' X - c_i} < \hat{Y} < e^{b' X + c_i}; \\ \frac{1}{2}\left[1 - erf\left(\frac{c}{\sigma\sqrt{2}}\right)\right] & \hat{Y}^c = e^{b' X \pm c}, \end{cases} \tag{9}$$

where $erf(w)$ defined as, (Abromowiz and Stegun 1972)

$$erf\,(w) = \frac{2}{\sqrt{\pi}} \int_0^w e^{-t^2} dt. \tag{10}$$

Based on the p.d.f. presented in Eq. (9), the (conditional) expected value of the capped simulated value $\hat{Y}^c$ may be expressed as the sum of three components:

$$E(\hat{Y}^c|X) \quad = \quad e^{b' X+c}\left[1 - erf\left(\frac{c}{\sqrt{2}\sigma}\right)\right] \; + \; \int_{e^{b' X-c}}^{e^{b' X+c}} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\left(ln\hat{Y}-b' X\right)^2}{2\sigma^2}} d\hat{Y} \; + \; e^{b' X-c}\left[1 - erf\left(\frac{c}{\sqrt{2}\sigma}\right)\right]. \tag{11}$$

After simplification, this may be expressed as:

$$E(\hat{Y}^c|X) \quad = \quad \frac{1}{2}\, e^{b' X \,+\, \frac{\sigma^2}{2}} \left[erf\left(\frac{1}{\sigma\sqrt{2}}\left(c - \sigma^2\right)\right) + erf\left(\frac{1}{\sigma\sqrt{2}}\left(c + \sigma^2\right)\right)\right] \; + \; e^{b' X} \cosh(c) \left[1 - erf(\frac{c}{\sqrt{2}\sigma})\right], \tag{12}$$

where $\cosh(c)$ represents the value of the hyperbolic cosine function evaluated at c (i.e., $(e^{+c} + e^{-c})/2$).

Notice that since $erf\,(\infty) = 1$, if the threshold values of $\pm c$ approach infinity in absolute value, the mean of the capped simulated value $\hat{Y}^c$ will approach the mean of the uncapped simulated value $\hat{Y}$; namely, $e^{\mu + \frac{\sigma^2}{2}}$. To account for the difference in the means of these two variables when the threshold values are finite, we define the following "correction factor":

$$\text{Correction Factor} \equiv \frac{E(\hat{Y})}{E(\hat{Y^c})} =$$

$$\frac{1}{\frac{1}{2}\left[erf\left(\frac{1}{\sigma\sqrt{2}}\left(c - \sigma^2\right)\right) - erf\left(\frac{1}{\sigma\sqrt{2}}\left(-c - \sigma^2\right)\right)\right] + e^{-\frac{\sigma^2}{2}} \cosh(c)\left[1 - erf(\frac{c}{\sqrt{2}\sigma})\right]}. \tag{13}$$

By applying this correction factor to adjust each of our capped simulated values for taxpayer burden, the mean of the simulated values in our sample will tend to be close to the mean that would be achieved from the uncapped distribution.

5 SIMULATED BURDEN ESTIMATES

Column 2 of Table 2 shows the distribution of the reported burden and column 3 shows the distribution of the predicted burden before any additional adjustments to account for the average contribution of the error term to the level of burden. As noted previously, such an approach effectively produces estimates of the median burden for each observation in the sample rather than the mean (expected) burden. Not surprisingly, the median of the predictions in column 3 is rather similar to the median of the reported burden distribution in column 2. On the other hand, the mean of the predictions in column 3 is well below the mean reported burden, which reflects the fact that the median of a highly right-skewed distribution falls well below the mean of the distribution. Column 4 shows the distribution of the predicted burden after the parametric approach is used to account for the average contribution of the error term to the level of taxpayer burden. The estimated mean (6,682) is much closer to the reported mean (6,644) than that reported in column 3 (4,080). Although our non-stochastic micro-simulation approach based on an econometric estimate of the expected level of taxpayer burden for each observation in the sample does a rather good job of estimating the mean burden, observe that it fails to adequately represent the percentiles of the reported burden distribution.

Columns 6-8 present the results based on our stochastic micro-simulation methodology under which we randomly draw values from the distribution of the error term in our regression model and employ these random draws in our prediction formula for the taxpayer burden distribution. Column 6 represents our original capped simulation of the distribution. In column 7, we apply the correction factor defined by Equation (13). The inverse of this correction factor is the distortion factor associated with capping the error at 3 σ. Column 5 reflects column 4 with the further application of the distortion factor for better compatibility with column 6. Lastly, in column 8, we adjust the simulated values from column 7 for downward bias associated with the finite number of random draws that were performed. As the results indicate, our stochastic micro-simulation approach does a much better job of representing the overall distribution of reported burden than the non-stochastic micro-simulation methodology.

Table 2: Reported and Predicted Income Tax Compliance Burden

Quantile	Reported Burden	Predicted without Transformation adjustment	Predicted with Transformation adjustment	Transformation Adjustment Distorted	Simulated Draw with Capping	Simulated Draw Capping Correction	Simulated Draw Cap and Draw Correction
1	2	3	4	5	6	7	8
100% Max	212,555	31,919	41,216	40,752	293,581	299,573	215,875
99%	50,043	17,214	24,309	24,001	51,204	52,398	48,291
95%	24,545	10,510	15,584	15,290	23,951	24,508	23,947
90%	15,995	7,968	12,528	12,251	15,782	16,144	16,149
75% Q3	7,745	5,280	8,548	8,354	7,608	7,791	7,913
50% Median	3,166	3,138	5,587	5,430	3,219	3,301	3,383
25% Q1	1,265	1,934	3,603	3,489	1,277	1,315	1,339
10%	507	1,060	2,105	2,032	498	514	518
5%	268	642	1,311	1,268	265	274	273
1%	70	141	289	279	64	67	65
0% Min	0	28	64	62	0	0	0
Mean	6,644	4,080	6,682	6,524	6,632	6,794	6,682

6 ESTIMATED COEFFICIENTS

Table 3 shows the results of the robust OLS regression of the complete small business econometric model. The estimated coefficient for log (total receipts) is as expected positive, 0.236, and significant at the 1% level. The same is true for the No Receipts coefficient, 2,625. Both coefficients are qualitatively similar to the corresponding coefficients estimated by Slemrod and Venkatesh (Slemrod and Venkatesh 2002), 0.4639 and 8.6283 respectively. All three coefficients for the complexity categories are statistically significant at the 1% level; equal to 0.003 for Low, 0.005 for Medium, and 0.009 for High. Since the coefficients are positive additional increases in the volume of an activity will increase total burden. In addition, the magnitudes of these coefficients confirm the make up of the complexity categories. An additional dollar increase in a medium complexity item, all else held constant, will increase burden more than an additional dollar increase in a low complexity item. The industry, tax form, and interaction terms will not be discussed in detail in this paper but are generally in line with our expectations. (After robust regression was implemented: the weight of 84.83 percent of observations was unaffected; 14.05 percent of observations had reduced weight; and 1.13 percent of observations were removed. Overall, the population weights were decreased by 4.41 percent with the burden data for this population effectively set to missing and imputed.)

7 CONCLUSION AND OUTLOOK

This paper presents a stochastic micro-simulation approach for modeling the distribution of small business taxpayer compliance burden. As discussed, additional issues come into play when using a log-linear regression specification to model the dependent variable. Future work on this approach is expected to focus on technical refinements to the error simulation process. It may prove useful to attempt to estimate a population max error cap from the sample max error cap and to develop error caps for sub-populations. It may also prove useful to perform sensitivity analysis on the impact of the number of random error draws on the simulation results and to use such an analysis to inform the determination of the desired practical number of error draws.

From a subject matter domain perspective, these results show promise in extending the type of micro-simulation analysis of compliance burden that can be performed. The improved fit of the stochastic micro-simulation approach over the non-stochastic approach is expected to provide a better foundation for analyzing how administrative and policy changes would impact the overall distribution of taxpayer burden. Future work in this area is expected to explicitly model the choice of preparation methods using a methodology that accounts for its endogeneity with expected compliance burden.

ACKNOWLEDGMENTS

There have been a lot of hands involved in various parts of this project. The authors would like to thank Michael Brick, Paul Chen, Joe Cordes, Jose Garcia, Patrick Langetieg, Wu-Lang Lee, Allen Lerman, Karen Masken, Mark Mazur, Susan Nelson, John O'Hare, Justin Rinker, Joel Slemrod, David Teale, Melissa Vigil, and Karen Yeager for a variety of contributions ranging from helpful domain input and technical comments to extensive research support.

A Table 3: Regression Results

Here are the results for regression analysis.

Table 3: Regression Results

Variable	Estimate	T-stat
Intercept	4.710	30.44
Log Total Receipts	0.236	20.39
No Receipts Indicator	2.625	18.33
Low Complexity	0.003	3.16
Medium Complexity	0.005	2.87
High Complexity	0.009	4.43
Mining, Quarrying, and Oil and Gas Extraction	-0.186	-0.73
Utilities	0.363	1.24
Construction	0.291	2.98
Manufacturing	0.256	2.23
Wholesale Trade	0.028	0.27
Retail Trade	0.113	1.09
Transportation and Warehousing	0.294	2.36
Information	0.301	1.90
Insurance Companies	0.156	1.19
Funds, Trusts, and Other Financial Vehicles	0.549	2.64
Finance, except Insurance and Funds	0.120	0.99
Real Estate and Rental and Leasing	0.179	2.00
Professional, Scientific, and Technical Services	0.288	3.13
Management of Companies	0.854	3.10
Administrative and Support and Waste Management and Remediation Services	-0.007	-0.06
Educational Services	-0.195	-0.93
Health Care and Social Assistance	0.099	0.93
Arts, Entertainment, and Recreation	-0.148	-1.11
Accommodation and Food Services	0.192	1.57
Other Services (except Public Administration)	0.203	1.85
Unknown NAICS	0.450	3.70
No Paid	-0.248	-3.56

Continued on next page

Table 3 – continued from previous page

Variable	Estimate	T-stat
Partnership	0.110	2.34
Partnership and No Paid	-0.112	-1.06
Partership, Nopaid, and No Receipts Indicator	-1.603	-5.72
Partnership, No Paid, and Insurance	-1.031	-0.6
Partnership, No Paid, and Funds	-1.174	-4.22
Partnership, No Paid, and Finance	-0.266	-1.27
C Corp	0.000	0.01
C Corp and No Paid	-0.295	-2.73
Form 1120A indicator	-0.488	-4.87
Form 1120 Special for Profit	0.327	1.05
Form 1120 Special for Profit and Nopaid	-0.693	-0.93
Form 1120 Special Non-Profit	0.301	1.79
Form 1120 Special Non-Profit and Nopaid	-1.764	-7.50
Form 1120 Special for Profit and No Receipts Indicator	0.746	1.68
Form 1120 Special Non-Profit and No Receipts Indicator	-1.622	-5.90
No Paid and No Receipts Indicator	-0.474	-2.43

REFERENCES

Abromowiz, M., and I. A. Stegun. 1972. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. New York: Dover.

Allison, P. D. 2001. *Missing data*. 1st ed. California: Sage Publications, Inc.

Arena, P., J. F. O'Hare, and M. P. Stavrianos. 2003. Measuring taxpayer compliance burden: A microsimulation approach. In 95th *Annual Conference on Taxation*, 333–341.

Connors, D., A. Greenland, J. L. Guyton, E. L. Morrison, and M. Sebastiani. 2007. IRS post-filing processes simulation modeling: A comparison of DES with econometric microsimulation in tax administartion. In *Proceedings of the 2007 Winter Simulation Conference*, ed. S. G. Henderson, B. Biller, M.-H. Hsieh, J. Shortle, J. D. Tew, and R. R. Barton, 1268–1274. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Gupta, A., and J. O'Hare. 2000. Practical microsimulation models. In *Economic Analysis: Microsimulation Modeling in Government*. North-Holland.

Guyton, J. L., J. F. O'Hare, M. P. Stavrianos, and E. J. Toder. 2003. Estimating the compliance cost of the US individual income tax. *National Tax Journal* 56:673–688.

Johnston, J., and J. Dinardo. 1984. *Econometric methods*. 3rd ed. New York: McGraw-Hill.

Manning, W. G., and J. Mullahy. 2001. Estimating log models: to transform or not to transform? *Journal of Health Economics* 20:461–494.

Slemrod, J., and V. Venkatesh. 2002. The income tax compliance cost of large and mid-size businesses. Technical report, IRS, LMSB Division.

AUTHOR BIOGRAPHIES

GEORGE CONTOS is a senior economist with the IRS Office of Research. He has been the lead technical analyst for the Taxpayer Compliance Burden Studies for two years. Before joining the Office of Research he was a senior economist with the IRS Statistics of Income for ten years. He received a Ph.D. in Economics from American University. His email address for these proceedings is <George.Contos@irs.gov>.

ARDESHIR EFTEKHARZADEH is an operation research analyst with the IRS Office of Research. His Ph.D. is from University of Maryland at College Park, in 2007 in theoretical physics. His research interests are micro-simulation and agent based simulation. He can be reached at <Ardeshir.Eftekharzadeh@irs.gov>.

BRIAN ERARD operates a consulting practice B. Erard & Associates in the Washington, DC area. Dr. Erard specializes in developing and implementing innovative econometric applications in the areas of compliance, enforcement, and administration. He has published extensively in academic journals and scholarly conference proceedings, and he has consulted on a wide range of issues for federal, provincial, and state government agencies in the U.S., Canada, and abroad. Prior to establishing his consulting practice, Dr. Erard spent a decade in academia on the economics faculties of the University of Toronto and Carleton University, and as a visiting research fellow at the University of Michigan Office of Tax Policy Research. <BEandAssoc@Aol.com>.

JOHN GUYTON is branch chief of the Forecasting and Service Analysis group in the IRS Office of Research. John has 13 years experience developing, using, maintaining, and managing tax micro-simulation models in the tax practice of a Big 4 accounting firm, public sector consulting, and most recently at the IRS Office of Research. For the past nine years, a major portion of this work has involved modeling the compliance burden faced by taxpayers. John has a bachelors degree from the University of Oklahoma and a PhD from the University of Maryland. His email address is <John.Guyton@irs.gov>.

SCOTT STILMAR is a Managing Consultant for IBM Business Consulting Services. He has over five years experience serving as an analyst for the Taxpayer Compliance Burden Studies. Currently on a leave of absence, Scott is enrolled at the University of Virginia seeking a graduate degree in Economics. Prior to working at IBM, he received a B.S in Mathematical Economics from Wake Forest University. His email address is <sstilmar@gmail.com>.